Durham University

Durham E-Theses

Controllable Generation with Diffusion Models and Applications in Medical Imaging

Junjie Shentu

How to cite:

Shentu, Junjie (2026) Controllable Generation with Diffusion Models and Applications in Medical Imaging. Doctoral thesis, Durham University.

Use policy

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a https://etheses.durham.ac.uk/id/eprint/16656/ is made to the metadata record in Durham E-Theses
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Please consult the full Durham E-Theses policy for further details.

Academic Support Office, The Palatine Centre, Durham University, Stockton Road, Durham, DH1 3LE
e-mail: e-theses.admin@durham.ac.uk Tel: +44 0191 334 6107
https://etheses.durham.ac.uk

Controllable Generation with Diffusion Models and Applications in Medical Imaging

Junjie Shentu

A Thesis presented for the degree of
Doctor of Philosophy

Department of Computer Science
Durham University
United Kingdom
May 2026

Abstract

Deep generative modeling has undergone rapid advancement in recent years, with diffusion models (DMs) emerging as a dominant framework for high-fidelity and diverse image synthesis. A core strength of DMs is their inherent controllability, offering the ability to steer generation through various conditions. Despite these advantages, achieving reliable controllability remains challenging, as issues such as concept entanglement, multimodal misalignment, and limited adaptability to specialized domains continue to restrict the practical deployment of DMs. These challenges become even more pronounced in medical imaging, where controllability is essential for tasks such as targeted data augmentation, but research on controllable generation using DMs within medical imaging remains limited. In this thesis, I investigate controllable generation with DMs from both theoretical and application-oriented perspectives. First, I advance the understanding of controllability in general-purpose text-to-image DMs by introducing an attention-driven framework for disentangling and customizing multiple visual concepts from a single image. This framework effectively mitigates issues of feature fusion and asynchronous learning across concepts that degrade customization quality, improving fidelity and control in customized generation. Building on the insights into attention-based guidance, I then explore controllable generation in medical imaging, focusing primarily on dermoscopic and chest X-ray modalities. To address the scarcity and imbalance of dermoscopic data, I propose a text-guided diffusion-based synthesis framework, incorporating dynamic prompt construction and region-aware fine-tuning to strengthen visual-textual alignment and enable controllable generation of lesion-mask pairs. Furthermore, a dual-branch DM is developed to tackle low-contrast bias in skin lesion segmentation by jointly controlling lesion layout and style, enabling the creation of targeted synthetic data that substantially improves segmentation performance on challenging cases while preserving overall accuracy. Finally, I extend controllable generation to multimodal medical content by proposing an integrated vision-language model capable of synthesizing clinically coherent chest X-ray images and their accompanying radiology reports. Through a novel prompt formulation and a self-supervised

report generation module, the model enhances both the visual realism and clinical validity of synthetic image-report pairs. These contributions demonstrate how DMs can be endowed with fine-grained, reliable controllability, and how such controllability can be leveraged to address domain-specific challenges in medical imaging. The presented methods provide effective frameworks for the development of controllable DMs with both general and clinical utility.

Declaration

The work in this thesis is based on research carried out at the Department of Computer Science, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Copyright © 2026 by Junjie Shentu.

"The copyright of this thesis rests with the author. No quotations from it should be published without the author's prior written consent and information derived from it should be acknowledged".

Acknowledgements

Standing at the final stage of my PhD career and reflecting on this remarkable journey, I'm deeply aware that this thesis could never have been completed without the guidance, support, and encouragement of many individuals to whom I owe my sincere gratitude.

First and foremost, I would like to express my deepest thanks to my PhD supervisor, my mentor, and my friend—Prof Noura Al Moubayed. You gave me the invaluable opportunity to pursue a PhD in a field I had never explored before, and you continually encouraged me throughout this challenging yet fascinating journey. Whenever I found myself trapped in self-doubt or hesitating to take the next step, your words always come back to me: "Why not?"

I'm also grateful to everyone in Noura's research group: Danial, James, Sean, Nour, Dean, Matthew, Jamie, Tom H., Tom W., Patrick, Chenghao, Strahinja, and Sid. You made me never feel alone, even when I was on the other side of the world from home. Group meetings, lunchtime debates, after-work drinks, hot pot gatherings, Christmas parties, and many other shared moments have become memories I truly cherish and will always carry with me.

My sincere thanks also go to my Chinese friends in Durham for the sense of belonging and warmth you offered. I treasure the trips we took together and the long conversations we shared about the present and the future, especially during times when we felt uncertain about our lives.

To my Mum and Dad, thank you for your unwavering support, patience, and understanding, even during moments when I was overwhelmed by negativity. To my Grandma, I hope I have grown into someone you would be proud of, and your memory continues to inspire me to move forward with courage.

Finally, a passing thank you to myself—for making the decision to embark on this PhD, and for the perseverance, effort, and time devoted over the past four years.

Contents

Abstract **ii**

Declaration **iv**

Acknowledgements **v**

List of Figures **xi**

List of Tables **xx**

1 Introduction **1**

1.1 Motivation 4

1.2 Publications 6

1.3 Thesis Structure and Contributions 7

2 Literature Review **9**

2.1 Visual Deep Generative Model 10

2.1.1 Variational Autoencoders 10

2.1.2 Generative Adversarial Networks 13

2.1.3 Denoising Diffusion Probabilistic Models 15

2.1.4 Latent Diffusion Models 19

2.2 Attention Mechanism in Diffusion Models 20

2.2.1 Attention in the Transformer . . . 21
2.2.2 Attention in Diffusion models . . . 23
2.2.3 Interpretation of Attention Maps . . . 26
2.3 Controllability of Visual Generation . . . 30
2.3.1 Controllability in Early Deep Generative Models . . . 31
2.3.2 Controllability in Diffusion Models . . . 36
2.4 Applications of Diffusion Models in Medical Imaging . . . 51
2.4.1 Chest X-ray . . . 52
2.4.2 Magnetic Resonance Imaging . . . 60
2.4.3 Computed Tomography . . . 65
2.4.4 Dermoscopic Imaging . . . 73

3 Attention-based Disentanglement of Multiple Concepts for Text-to-Image Customization 80

3.1 Introduction . . . 81
3.2 Related Work . . . 83
3.2.1 Diffusion models and T2I customization . . . 83
3.2.2 Application of attention in diffusion models . . . 84
3.2.3 Disentangling multiple concepts from a single image . . . 84
3.3 Proposed Method . . . 85
3.3.1 Attention-guided mask creation . . . 85
3.3.2 Adaptive sampling ratio estimation based on attention scores . 88
3.3.3 Feature-retaining training framework . . . 91
3.4 Experiments . . . 92
3.4.1 Experimental settings . . . 92
3.4.2 Qualitative comparisons . . . 94
3.4.3 Quantitative comparisons . . . 97
3.4.4 Ablation studies . . . 97
3.4.5 Generalizing to more concepts . . . 100
3.5 Conclusion . . . 101
3.6 Appendix . . . 102
3.6.1 Datasets . . . 102

3.6.2 Additional details for preliminary experiment 103
3.6.3 Additional details for main experiment 106

4 Controllable Synthesis of Dermoscopic Images for Enhanced Computer Aided Diagnosis and Detection 110

4.1 Introduction 111
4.2 Related Works 115
4.2.1 CAD systems for skin lesion diagnosis 115
4.2.2 Dataset augmentation for skin lesion images 116
4.2.3 Controllable generation with diffusion models 117
4.3 Proposed Method 119
4.3.1 Preliminary 120
4.3.2 Attribute-aware DermPrompt 122
4.3.3 Region-aware Finetuning of SD 123
4.3.4 Training-free Pipeline for Dermoscopic Lesion-mask Pair Generation 125
4.4 Experiments and Results 128
4.4.1 Dataset 128
4.4.2 General generation quality 129
4.4.3 Dataset Augmentation for Multi-class Classification using Controllable Generation 131
4.4.4 Dataset Augmentation for Segmentation using Controllable Generation 135
4.4.5 Ablation study 139
4.4.6 Further applications in controllable generation 142
4.5 Discussion 145
4.6 Conclusion 150
4.7 Appendix 150
4.7.1 Evaluation Metrics 150
4.7.2 Implementation details 152

5 Mitigating Low-Contrast Bias in Skin Lesion Segmentation using a Dual-Branch Controllable Diffusion Model 154

5.1 Introduction . . . 155

5.2 Related Works . . . 157

5.2.1 Bias in segmentation of dermoscopic images . . . 157

5.2.2 Skin tone analysis and annotation . . . 158

5.2.3 Generative models for dermoscopic images . . . 158

5.3 Methods . . . 159

5.3.1 Workflow overview . . . 159

5.3.2 Attribute annotation of dermoscopic images . . . 160

5.3.3 Dual-branch controllable dermoscopic generation model . . . 161

5.4 Bias analysis of skin lesion segmentation . . . 168

5.4.1 Skin tone annotation of dermoscopic images . . . 168

5.4.2 Identifying bias in skin lesion segmentation . . . 170

5.5 Synthetic dataset for segmentation bias mitigation . . . 173

5.5.1 Experimental setup . . . 173

5.5.2 General quality of generated images . . . 175

5.5.3 Mitigating segmentation bias with generated data . . . 176

5.5.4 External validation on HAM10000 . . . 181

5.5.5 Ablation studies . . . 185

5.6 Conclusion . . . 188

6 Controllable Generation of Clinically Accurate Chest X-Ray Image-Report Pairs using an Integrated Vision-Language Model 189

6.1 Introduction . . . 190

6.2 Related Work . . . 192

6.2.1 Generative models for CXR image generation . . . 192

6.2.2 Generation of CXR reports . . . 193

6.3 Method . . . 194

6.3.1 Text-to-image generation and optimization with the diffusion model . . . 194

6.3.2 CXR report generation with self-supervised learning . . . 197

6.4 Experiments . . . 199
6.4.1 Dataset . . . 199
6.4.2 Baselines and evaluation metrics . . . 199
6.4.3 Evaluation of CXR images . . . 200
6.4.4 Evaluation of CXR reports . . . 202
6.5 Ablation . . . 204
6.5.1 Extracting representative text embedding . . . 204
6.5.2 Image vs. Image embedding . . . 207
6.5.3 Choice of L_{prior} . . . 207
6.6 Conclusion . . . 208

7 Concluding Remarks **209**
7.1 Contributions . . . 209
7.2 Limitations and Future Work . . . 213
7.3 Epilogue . . . 215

List of Figures

2.1 Computation graph of VAEs (Adapted from Lai et al. (2025)) 11
2.2 Computation graph of a GAN (Adapted from Lai et al. (2025)) . . . 14
2.3 Illustration of DDPMs (Adapted from Ho et al. (2020)) 16
2.4 Illustration of LDMs (Adapted from Rombach et al. (2022)) 20
2.5 Architecture of transformer (Adapted from Vaswani et al. (2017)) . . 22
2.6 Illustration of (a) Scaled Dot-Product Attention and (b) Multi-Head Attention (Adapted from Vaswani et al. (2017)) 24
2.7 Illustration of cross-attention and self-attention operations in SD (Adapted from Liu et al. (2024a)) 25
2.8 (a) Method overview and (b) editing capabilities of P2P. Cross-attention layers produce attention maps of visual latents and textual embeddings. During generation, the spatial layout and shape of cross-attention maps can be modified in different ways (e.g., swapping, refining, and re-weighting) to affect the generated images, achieving various editing operations. (Adapted from Hertz et al. (2022)) 27
2.9 Visualization of self-attention maps extracted from different U-Net layers using PCA (Adapted from Tumanyan et al. (2023)) 28
2.10 Style-aligned generation using shared attention layers (Adapted from Hertz et al. (2024)) 29

2.11 Enhancing the cross-attention maps using the self-attention maps with $\tau = 1, 2, 4$, given a text prompt "A bike is parked in a room; bicycle" for generating the first image (Adapted from Nguyen et al. (2023)) 30
2.12 Example results of Pix2Pix (Adapted from Isola et al. (2017)) 33
2.13 Example results of style mixing between Source A and Source B using StyleGAN. Source A is overridden with styles from B at specific resolutions. Coarse styles (4^2-8^2) transfer high-level semantics like pose, face shape, and glasses. Middle styles (16^2-32^2) inherit smaller facial features and hair details. Fine styles (64^2-1024^2) transfer the color scheme and microstructure. (Adapted from Karras et al. (2019)) . . . 35
2.14 Example results of ControlNet that applies both textual and spatial control over generated images. (Adapted from Zhang et al. (2023a)) . 38
2.15 Model architecture of ControlNet. (Adapted from Zhang et al. (2023a)) 39
2.16 Model architecture of T2I-Adapter. (Adapted from Mou et al. (2024)) 40
2.17 Model architecture of IP-Adapter. (Adapted from Ye et al. (2023)) . 41
2.18 Examples of generated samples with image prompt and additional structural conditions using IP-Adapter. (Adapted from Ye et al. (2023)) 42
2.19 Illustration of Progressive Guidance. The vanilla guidance fails to condition the generated sample as a leopard, whereas Progressive Guidance (Ours) uses the tiger, panther, and leopard to influence the image content in the initial state to form the critical features of the leopard. The right part shows more failure cases corrected using Progressive Guidance. The darkness of the gradients denotes the associated information degree values.(Adapted from Dinh et al. (2023)) 45
2.20 Adaptation of self-attention for style transfer.(Adapted from Chung et al. (2024)) 47
2.21 Generated images and cross-attention maps for each subject token with and without Attend-and-Excite over vanilla SD.(Adapted from Chefer et al. (2023)) 48

2.22 Illustration of the results by (a) BoxDiff and (b) ZestGuide (Adapted from Xie et al. (2023); Couairon et al. (2023)) 49
2.23 Illustration of the results of Blend Latent Diffusion. (Adapted from Avrahami et al. (2023a)) 50
2.24 Illustration of samples in MIMIC-CXR dataset. (Adapted from Johnson et al. (2019a)) 54
2.25 Examples of synthetic images from RoentGen. (Adapted from Chambon et al. (2022a)) 55
2.26 Comparison of anomaly detection using Diff3M and DDPM. (Adapted from Kim et al. (2025)) 58
2.27 Example of axial brain MRI images from fastMRI dataset with different contrasts: (a) FLAIR; (b) T1 weighted; (c) T1 weighted with contrast agent (T1 POST); (d) T2 weighted (Adapted from Zbontar et al. (2018)) 61
2.28 Examples of ground truth and synthetic MRIs for cross-modality synthesis task (Adapted from Hu et al. (2025)) 64
2.29 Examples from RSNA PE CT dataset: (a) central PE; (b) right-sided PE and left-sided PE; (c) chronic PE; (d) true filling defect not PE; (e) flow artifact; (f) RV/LV ratio; (g) RV/LV ratio; (h) QA-motion; (i) QA-contrast. LV = left ventricle, QA = quality assurance RV = right ventricle (Adapted from Colak et al. (2021)) 67
2.30 Example of qualitative results of a 25% does CT image (a) ground truth; (b) FBP; (c) PWLS; (d) Noise2Noise; (e) Noise2Sim; (f) SS-DDNe; (g) DR2; (h) GDP; (i) Dn-Dp; (j) SPDiff; (k) NEED (Adapted from Gao et al. (2025)) 69
2.31 Reconstruction results from simulated projections for different methods with a scanning angular range of 90° (Adapted from Han et al. (2024)) 71
2.32 Examples of volumetric generation on MRI and CT (Adapted from He et al. (2024)) 72
2.33 Examples of dermoscopic images from the ISIC datasets 74

2.34 Examples of newly generated benign and malignant skin mole data using Derm-T2IM (Adapted from Farooq et al. (2024)) 76
2.35 Illustration of adding or removing artifacts from dermoscopic images using MaskMedPaint (Adapted from Jin et al. (2024)) 78
2.36 Qualitative comparisons of different methods on dermoscopic images (Adapted from Bozorgpour et al. (2023)) 79

3.1 I propose *AttenCraft*, an optimized method for disentangling multiple concepts in a single image. Baseline models present two key issues: (a) feature fusion; (b) asynchronous learning. My method significantly mitigates these issues and realizes robust concept disentanglement and feature learning. 83
3.2 Method overview. Given an image with multiple concepts, within a few steps in the pre-processing stage, I create accurate masks for each concept and adaptively estimate the sampling ratio for multiple concepts to enhance learning synchronicity. I also propose an optimized training framework by introducing different loss functions for sampled subsets of varying sizes to prevent feature fusion. 86
3.3 Process of attention-guided mask creation. By applying the cross-attention and self-attention maps, precise masks can be created without specialized models or human inputs. 88
3.4 Results of the token initialization experiment. (a) Variation of single-concept CLIP-I scores with training step; (b) The highest cross-attention score of [V] concerning different initialization patterns. . . . 90
3.5 Qualitative results for concept disentanglement and feature fusion. *CusDiff* cannot disentangle multiple concepts, and both *DisenDiff* and *BAS* present feature fusion. My method not only disentangles the target concepts, but also mitigates the feature fusion problems . . 95
3.6 Qualitative results for learning synchronicity. *DisenDiff* and *BAS* show asynchronous learning in different forms, while my method achieves a more synchronous feature learning. 96

3.7 Visualization of cross-attention maps. My method presents proper attention activation for multiple conceptions. 96
3.8 Qualitative results for ablating attention-guided mask creation. All three techniques are vital for mask creation, and disabling them will cause failure in certain datasets. 98
3.9 Qualitative results of ablation studies on feature-retaining training framework. My proposed framework can effectively prevent feature fusion during training. 99
3.10 Qualitative results of ablation studies on adaptive estimation of sampling ratio. The numbers on images denote the sampling ratio determined by the method. 101
3.11 Qualitative results for *AttenCraft* applied on input images containing more than two concepts. My proposed method can be seamlessly applied to input images containing more than two concepts. 102
3.12 Illustration of datasets 103
3.13 Variation of single concept CLIP-I scores with training step 105
3.14 Initial masks for each dataset created by my method 109

4.1 Overview of my proposed method DiDGen including three novel technical contributions: (a) DermPrompt for producing attribute-rich text prompts; (b) region-aware finetuning that facilitate the establishment of the semantic visual-textual alignments between text prompts and visual representations, and (c) training-free lesion-mask generation pipeline for synthesizing high-quality images of lesion-mask pairs 119
4.2 Average volunteer rankings of captions generated by different VLMs . 122
4.3 Samples of synthetic dermoscopic images generated by different methods 130
4.4 Qualitative comparison of synthetic dermoscopic images for different diagnostic categories generated by different methods 132
4.5 Limitation of DermPrompt in layout-guided generation. The intrinsic limitation of the text guidance cannot provide fine-grained layout guidance, meaning that the layout of the generated image does not resemble the real image. 136

4.6 Qualitative comparison of different methods in the generation of dermoscopic lesion-mask pairs 137
4.7 Qualitative results for ablation studies: (a) visualization of generated images and cross-attention maps of the P-Tokens from models finetuned with/without region-aware finetuning; (b) visualization of generated images and processed attention maps A_{CS} of P-Token <*lesion*> from models with/without self-attention guidance. 143
4.8 Text-guided attribute customization. My proposed DermPrompt can achieve high-level controllable generation including but not limited to skin color, lesion color, and other marks. 144
4.9 Bounding-box guided generation. My finetuned SD can be combined with the training-free pipeline BoxDiff (Xie et al., 2023) to generate dermoscopic images wherein the lesions comply with the layout of bounding boxes. 145
4.10 Domain-specific image editing. My finetuned SD can be combined with attention-based image editing methods, such as Null-text Inversion (Mokady et al., 2023), and edit real domain-specific dermoscopic images. For example, changing the color and texture of benign lesions to resemble the characteristics of melanoma lesions. 146
4.11 Distribution of per-model sample-wise IoU gains versus original IoU, with sample counts and mean IoU gains by interval, collected from the results of (a) U-Net; (b) AttenU-Net; (c) DCSAU-Net; (d) XBound-Former, highlighting that significant improvements occur on the poorly-segmented cases. 147

5.1 The proposed workflow for identifying and mitigating low-contrast bias in skin lesion segmentation. Dermoscopic images are first annotated with skin tone, ITA-based color attributes, and structural attributes; segmentation models are then trained and their errors analyzed with respect to these attributes to identify the main bias source; finally, targeted training samples are selected as input to synthesize image-mask pairs with the proposed dual-branch diffusion model, and these are used to finetune the segmentation models to mitigate bias. . 160

5.2 Illustration of the dual-branch controllable diffusion model, detailing the (left) region-aware finetuning strategy and (right) dual-branch generation pipeline. In the region-aware finetuning stage, special *<lesion>* and *<skin>* tokens are aligned with lesion and skin regions via an attention loss that matches cross-attention maps to the corresponding masks. In the dual-branch generation pipeline, a style reference image is first inverted to obtain its latent and attention statistics through a DDIM inversion, the final noisy latent is used as the initial latent for generation. Then the latent is jointly optimized for layout and style before splitting into a layout branch (driving spatial structure and mask extraction) and a style branch (injecting style via attention), yielding aligned lesion images and masks in a single denoising pass. 162

5.3 Attention-based mechanism for layout and style control within the dual-branch framework. Queries from the layout branch encode spatial structure, while keys and values from the style reference control appearance. By reusing layout queries when combining them with style keys and values, the model enforces that the stylized output follows the lesion layout of the layout branch while adopting the color and texture statistics of the style image. 167

5.4 Distributions of ISIC 2018 test set and poorly segmentation samples on various image attributes. The distributions over skin tone and geometric attributes remain broadly similar, whereas poorly segmented cases are strongly concentrated at small or negative ITA Difference values, highlighting low lesion-skin color contrast as the dominant and consistent source of segmentation bias. 171
5.5 Examples of dermoscopic images, ground-truth masks, and predictions from DCSAU-Net for cases with different ITA Differences. The top example, with a large positive ITA Difference (high-contrast), is segmented accurately, while the bottom example, with a small ITA Difference (low-contrast), is severely under-segmented, visually demonstrating how insufficient color contrast between lesion and surrounding skin leads to segmentation failure. 173
5.6 Qualitative comparison of synthetic images generated by my proposed method versus baseline models, and mask annotation from my method. My dual-branch framework generates dermoscopic images that simultaneously preserve realistic lesion structure and faithfully inherit stylistic features from the reference, while producing masks that tightly follow lesion boundaries, whereas baselines either lack style control, introduce artifacts, or misalign structure and style. . . 177
5.7 Qualitative results showing ground-truth and predicted masks from different segmentation models before and after finetuning on various data sources for challenging low-contrast samples (first three rows) and an originally well-segmented sample (last row). They illustrate that finetuning on my synthetic low-contrast data consistently improves prediction accuracy on difficult cases, while preserving good performance on easy cases. 179

5.8 Validation of bias source on the HAM10000 dataset, showing the distributions of ITA Skin, ITA Lesion, and ITA Difference for the full HAM10000 dataset and for poorly segmented subsets at three IoU thresholds ($S_{0.4}$, $S_{0.6}$, $S_{0.8}$) across all segmentation models. As in ISIC 2018, the poorly segmented samples cluster around ITA Difference values near zero, confirming on an external dataset that low lesion-skin contrast is the main factor driving segmentation errors. 183

5.9 Qualitative comparison of the visual fidelity of the proposed dual-branch architecture and a single-branch variant under the same style reference and layout conditions. My dual-branch model preserves both lesion geometry and rich stylistic properties such as color saturation and fine texture, whereas the single-branch model produces more washed-out, structurally less faithful results, visually demonstrating the benefit of separating layout and style into two coordinated branches. 186

5.10 Comparison of model performance for DCSAU-Net finetuned with different synthetic generation ratios ($n = 1, 2, 3, 4$) per low-contrast real sample. It shows that increasing the number of synthetic images yields only modest gains on challenging samples but progressively harms performance on the full test set, supporting the choice of a 1:1 generation ratio as a good trade-off between targeted debiasing and overall generalization under the consideration of computational cost. 187

6.1 CXR image with radiology report 191

6.2 An overview of the inference process of *CXR-IRGen* 194

6.3 Illustration of the training process of the vision module (* denotes the frozen part) 196

6.4 Illustration of the training process of the language module during (a) first stage and (b) second stage (* denotes the frozen part) 198

6.5 Comparison of different representative text embedding 206

List of Tables

2.1 Summary of image numbers in the ISIC 2016-2020 datasets 74

3.1 Results of quantitative comparisons 97
3.2 Ablation results of feature-retaining training framework 99
3.3 Ablation results of adaptive sampling ratio estimation 100
3.4 Patterns for initialization of identifier tokens 104
3.5 Highest cross-attention scores of [V] using different initialization patterns 104
3.6 Performance of adopted methods under a learning rate value of 1×10^{-4} 108

4.1 General generation quality of different methods 129
4.2 Performance of classifiers trained on original dataset and datasets augmented by different methods on the ISIC 2018 dataset 134
4.3 Class-wise F1 score of various classifiers trained on the original dataset and dataset augmented by my method 134
4.4 Averaged performance of classifiers trained on original dataset and datasets augmented by different methods on the Derm7pt dataset . . 135
4.5 Performance of segmentation models trained on the original and augmented datasets on the ISIC 2018 dataset 138
4.6 Averaged performance of segmentation models on the PH2 dataset . . 139

4.7 Ablation results of the effects of cross-attention guidance in the lesion-mask generation pipeline . . . 141

5.1 Performance comparison of skin tone annotation methods on the DDI dataset . . . 169
5.2 Quantitative evaluation of general synthetic image quality for the proposed method and baselines . . . 178
5.3 Bias mitigation results on ISIC 2018, comparing segmentation performance after finetuning on different data sources . . . 180
5.4 External validation of bias mitigation on the HAM10000 dataset . . . 184
5.5 Ablation study on the impact of layout guidance on mask diversity and downstream segmentation performance . . . 184

6.1 General metrics of CXR images generated by different models (RIE: reference image embedding) . . . 201
6.2 AUROC values of the binary classification task on original CXR images and CXR images generated by different models (RIE: reference image embedding) . . . 203
6.3 Comparison of *CXR-IRGen* and baselines models on original CXR images (Results with * are taken from the original paper (Endo et al., 2021)) . . . 203
6.4 F1 scores of *CXR-IRGen* and baseline models on CXR images generated by the vision module *CXR-IRGen* (RIE: reference image embedding) . . . 205
6.5 Comparison of different variants of *CXR-IRGen* . . . 205

CHAPTER 1

Introduction

Over the last decade, deep learning (DL) has emerged as a state-of-the-art technique in numerous fields, with computer vision being one of the most prominent cases (Voulodimos et al., 2018). With the resurgence of convolutional neural networks (CNNs), AlexNet (Krizhevsky et al., 2012) has marked a breakthrough by demonstrating the effectiveness of large-scale CNN architecture for image classification. Subsequent models, such as VGGNet (Simonyan and Zisserman, 2014), GoogLeNet (Szegedy et al., 2015), and ResNet (He et al., 2016), were inspired by and continued to improve performance. Parallel advancements in fully connected networks (FCNs) and encoder-decoder architectures, such as U-Net (Ronneberger et al., 2015), extended from recognition to dense prediction problems, enabling accurate segmentation. More recently, the introduction of Transformer-based models (Vaswani et al., 2017), represented by Vision Transformers (ViT) (Dosovitskiy, 2020), reshaped representation learning by replacing convolutions with self-attention, pushing scalability and cross-domain performance to the next stage.

The rapid evolution of DL theories and architectures has been further translated into widespread real-world applications. Large-scale image classification models are playing a pivotal role in visual recognition systems, and have been embedded in com-

mercial organization platforms such as Google Cloud Vision and Amazon Rekognition (Hosseini and Hasan, 2023). Object detection frameworks based on DL models have been widely adopted in human face recognition (Balaban, 2015), intelligence surveillance (Sreenu and Durai, 2019), and autonomous driving(Grigorescu et al., 2020), enabling reliable detection of humans, vehicles, products, and anomalies in real time. Moreover, DL models have advanced dense pixel-level interpretation tasks, including semantic segmentation and scene understanding (Hao et al., 2020; Yang et al., 2018), supporting autonomous navigation, monitoring, and planning. Meanwhile, the introduction of ViT has further facilitated multimodal perception systems and remote sensing by improving robustness under distribution shifts (Zhou et al., 2024a; Bi et al., 2022).

Apart from deterministic models, deep generative models are gaining prominence and demonstrating substantial practical impact in visual applications. Early approaches based on variational autoencoders (VAEs) (Kingma and Welling, 2013) can compact images into latent representations and realize image reconstruction, enabling multiple tasks such as anomaly detection and image compression. Moreover, generative adversarial networks (GANs) (Goodfellow et al., 2020) introduced adversarial learning for image synthesis, leading to photorealistic image generation in applications ranging from image super-resolution to domain translation (Cherian and Sullivan, 2019; Zhu et al., 2020). Diffusion models (DMs) have emerged as a new paradigm for image generation with unprecedented realism in image generation, surpassing GANs in generation quality and avoiding mode collapse (Ho et al., 2020; Dhariwal and Nichol, 2021). More importantly, DMs have presented huge potential for controllable generation with their abilities to incorporate additional conditional signals in various modalities, including text, images, and audio (Rombach et al., 2022; Zhang et al., 2023a; Tang et al., 2023a). The controllability of DMs has enabled user-driven content creation, powering design prototyping, visual effects production, and creative media generation (Geng and Yang, 2025).

In the context of visual generative models, controllability refers to the ability to guide the generation process toward user-specified content while preserving realism and coherence. Such control can be manifested at different levels, including

high-level semantic attributes (e.g., object category or medical condition), spatial structure (e.g., shape, layout, or lesion location), and visual appearance (e.g., style, color, or texture). In Stable Diffusion (SD), which is built on latent diffusion models, this controllability is realized by conditioning the denoising process in the latent space, most commonly through text embeddings injected via cross-attention (Rombach et al., 2022). This design not only supports prompt-based image synthesis, but also provides a flexible foundation for incorporating richer control signals, making SD a representative and highly influential framework for studying controllable image generation.

The realm of medical imaging is a critical subset of computer vision, as well as a hot area for the application of DL models. With the continual development of digital imaging techniques in medicine, the fidelity, clarity, and accuracy of medical images are improving, creating an increasing potential for the deployment of DL models (Luo et al., 2025). Classification and segmentation are the two representative tasks in DL-assisted medical imaging analysis, with wide application cases across various modalities, such as dermoscopy, X-ray, computed tomography (CT), and magnetic resonance imaging (MRI) (Thomas et al., 2021a; Baltruschat et al., 2019; Gao et al., 2017; Liu et al., 2018; Dalmaz et al., 2022). Moreover, some studies aim to learn medical visual presentations with enhanced robustness to enhance the performance of downstream DL models (Zhang et al., 2022; Liu et al., 2023a). Furthermore, deep generative models have introduced new applications and challenges to the medical imaging community. GAN-based approaches are extensively applied for data augmentation, cross-modality translation, image reconstruction (Armanious et al., 2020; Bamoriya et al., 2022; Tian et al., 2022). DMs improve the results in these tasks with their advantages in synthesis quality and training stability, and play their unique role in other tasks like anatomy-preserving edits and uncertainty-aware sampling (Maksudov et al., 2025; Jeong et al., 2024a).

Despite the remarkable advances of DMs and their rapid adoption in visual generation tasks, achieving reliable and diverse forms of controllability remains a nontrivial challenge. As one of the most critical properties of DMs, controllability steers countless possibilities for generation by altering structural and stylistic attributes.

The exploration of controllability in DMs is still a trending research topic as issues like entangled representation and weak multimodal alignment remain unsolved (Cao et al., 2025). In parallel, work on controllable deep generative modeling in medical imaging, particularly using DMs, is still insufficient, with studies primarily focusing on unconditional or lightly conditional synthesis (Hung et al., 2023). This gap is especially evident in medical applications where controllability is essential, such as targeted data augmentation, disease progression simulation, and bias-aware generation. This thesis represents my contributions in exploring the two problems across four different areas under the unified theme of *controllable image generation and its application in medical imaging*.

1.1 Motivation

Although modern text-to-image (T2I) DMs have demonstrated impressive generative capabilities, their controllability remains insufficient for many real-world applications. The widely used SD model (Rombach et al., 2022) can respond to high-level textual descriptions, yet text alone is often an imprecise medium for encoding complex spatial, semantic, or stylistic constraints. Many tasks in human creative and technical activities require finer-grained forms of control that go beyond what natural language can adequately express. For example, specifying a detailed object layout, geometric arrangement, or scene structure is difficult to describe textually, resulting in outputs that deviate from user intent. In addition, a particularly demanding case arises in customized generation, where users seek to recreate an object or identity from a small set of images and place it into new contexts (Gal et al., 2022; Ruiz et al., 2023a). These scenarios require controllability that is both more precise and more flexible than what standard text conditioning provides. Consequently, there is a clear need to investigate the theoretical foundations and practical mechanisms of controllability in general-purpose DMs, motivating the first research direction of this thesis.

Beyond general-purpose image synthesis, the need for more powerful and fine-grained controllability becomes even more pronounced in medical imaging, where

data scarcity and imbalance remain significant challenges (Huynh et al., 2022; Upadhyay and Bhandari, 2024). Dermoscopic imaging is a representative example of this issue. In the widely used dermoscopic dataset, International Skin Imaging Collaboration (ISIC) 2018 (Codella et al., 2019), certain pathology categories only contain a limited number of samples, constraining the performance and generalizability of classification models trained on such imbalanced data. Moreover, high-quality pixel-level annotations are available for only a subset of images, limiting the effectiveness of segmentation models that rely heavily on accurate ground-truth masks. As a result, there is an urgent need for controllable generation methods capable of producing targeted synthetic samples to support robust training of downstream models, resulting in improved overall performance and mitigated class imbalance.

In addition, imbalance within medical imaging datasets can also lead to systematic biases in downstream models. When certain patient groups, imaging styles, or anatomical characteristics are underrepresented, models trained on such skewed data may exhibit degraded performance on minority subsets. A representative example is dermoscopic image segmentation, where recent studies have shown that models tend to underperform on images from patients with darker skin tones (Benčević et al., 2024). To address this issue, generation frameworks with flexible and fine-grained controllability of the image color, texture, and appearance are required to augment underrepresented minority subsets, thereby mitigating model bias.

A further limitation in current medical generative modeling lies in the narrow focus on unimodal data. For example, most existing approaches in the chest X-ray (CXR) field concentrate on generating either CXR images or radiology reports exclusively. However, clinical workflows inherently rely on multimodal information, where the image and its accompanying report form a complementary pair that conveys diagnostic reasoning and clinical context. Such multimodal pairs are essential for training models in report generation, visual–textual representation learning, and downstream tasks that require cross-modal alignment or interpretability. Nevertheless, multimodal datasets are particularly expensive to construct, as they require both high-quality imaging and expert-authored textual descriptions. This motivates the development of a controllable generation framework capable of producing coher-

ent multimodal medical content, such as CXR image–report pairs, to expand the richness of available training data.

The three challenges related to applications in medical imaging motivate the second research direction of this thesis: the development of controllable DMs that support richer forms of guidance for medical image generation.

1.2 Publications

The work presented in this thesis is the result of papers that have been submitted for publication or published in peer-reviewed publications or conference proceedings throughout my PhD. The publication and corresponding chapters are listed below:

- Chapter 3 contains work presented in Junjie Shentu, Matthew Watson and Noura Al Moubayed **"AttenCraft: Attention-based Disentanglement of Multiple Concepts for Text-to-Image Customization."** *IEEE Transactions on Multimedia*, 2025.

- Chapter 4 contains work presented in Junjie Shentu, Matthew Watson and Noura Al Moubayed **"DiDGen: Diffusion-based Dual-task Synthesis for Dermoscopic Data Generation"** *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*, 2025 & Junjie Shentu, Matthew Watson and Noura Al Moubayed **"Controllable Synthesis of Dermoscopic Images Using Diffusion Models for Enhanced Computer Aided Diagnosis and Detection"** *Medical Image Analysis*, 2026.

- Chapter 5 contains work presented in Junjie Shentu, and Noura Al Moubayed **"Beyond Skin Tone: Mitigating Low-Contrast Bias in Skin Lesion Segmentation using a Dual-Branch Controllable Diffusion Model"** *Medical Image Analysis*, 2026, (Under Revision Review).

- Chapter 6 contains work presented in Junjie Shentu and Noura Al Moubayed **"CXR-IRGen: An Integrated Vision and Language Model for the Generation of Clinically Accurate Chest X-Ray Image-Report Pairs"**

Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024.

1.3 Thesis Structure and Contributions

Building on the motivations outlined in the previous section, this thesis is organized around two research directions identified in Chapter 1.1. These directions constitute a unified exploration of both the theoretical mechanisms and practical applications of controllable image generation in medical imaging. This thesis starts with Chapter 2. In Chapter 2, I present a comprehensive literature review about the theoretical and practical background of this thesis. The literature review starts with the introduction of visual deep generative models, and demonstrates the mathematical foundation for each model. Then, the attention mechanism in the DMs is analyzed, with their rules in data generation being deeply discussed. Furthermore, I discuss the controllability in visual generation, including various means of controllability adopted in different deep visual generative models. Finally, I introduce the development of DMs in the medical image field, including but not limited to representative datasets of various imaging modalities and typical tasks for the application of DMs in these modalities.

Chapter 3 to Chapter 6 are structured into four core research components in sequence, each corresponding to a major contribution. Chapter 3 advances controllability in general-purpose T2I DMs by proposing an attention-based framework for multi-concept disentanglement. By leveraging self-attention maps to autonomously derive concept-specific masks and introducing an adaptive sampling mechanism for balanced concept learning, the proposed method alleviates the issues of feature entanglement and asynchronous optimization in T2I customization. This contribution deepens the understanding of internal mechanisms governing concept representation and demonstrates reliable multi-concept control from a single example. It also lays the groundwork for the further investigation of controllable generation of dermoscopic images in Chapter 4 and Chapter 5.

Chapter 4 introduces a diffusion-based dermoscopic image synthesis framework

designed to enhance computer-aided diagnosis and detection. This framework incorporates dynamic text prompting and region-aware finetuning to improve visual-textual alignment, while a training-free mechanism enables simultaneous generation of lesion-mask pairs. The method improves both the fidelity and diversity of synthetic dermoscopic images and demonstrates downstream benefits for classification and segmentation models, illustrating the utility of controllable generation in addressing data scarcity and variability. The proposed method for controlled simultaneous generation of lesion–mask pairs is further applied in Chapter 5 and combined with finer-grained control over generated content.

Chapter 5 tackles a clinically significant bias source in skin lesion segmentation: low color contrast between the lesion and the surrounding skin. A dual-branch controllable DM is developed to disentangle lesion structure and style, enabling precise manipulation of low-contrast characteristics. Combined with the method for simultaneous image-mask pair generation introduced in Chapter 4, controlled generation of paired image-mask samples with styles similar to the targeted image subset. Targeted finetuning using these synthetic samples markedly improves segmentation robustness on challenging subsets while maintaining overall performance, showcasing how controllable generation can be leveraged as a principled tool for bias mitigation in medical imaging.

Chapter 6 extends controllable diffusion modeling to multimodal medical data generation by developing an integrated vision-language model for synthesizing chest X-ray image-report pairs. Through a hybrid prompt design that combines textual and visual embeddings, and a self-supervised report generation module tailored for radiology, the proposed approach enhances both the perceptual quality and clinical validity of the generated data. The resulting synthetic image–report pairs provide a unified resource for multimodal model training, addressing the scarcity of large-scale clinically annotated datasets.

Finally, Chapter 7 summarizes the contributions of the thesis by chapter, concludes the limitations, and proposes potential topics for future work.

CHAPTER 2

Literature Review

This chapter reviews the literature that forms the foundation of this thesis. It first introduces representative visual deep generative models, including variational autoencoders, generative adversarial networks, and diffusion models, in order to position diffusion models within the broader development of generative learning. It then discusses attention mechanisms in diffusion models, with a particular focus on their architectural role and interpretability, before reviewing the main research directions in controllability for visual generation. Finally, the chapter surveys recent applications of diffusion models in medical imaging across several representative modalities, highlighting both current progress and remaining challenges.

This review also serves to connect the background literature with the rest of the thesis. The discussion of latent diffusion models, attention mechanisms, and controllability provides the basis for Chapter 3, which investigates attention-driven controllability in general text-to-image generation. The review of medical imaging applications further motivates Chapter 4 to Chapter 6, where controllable diffusion models are studied in dermoscopic image synthesis, bias-aware generation for segmentation improvement, and multimodal chest X-ray image-report generation. In this way, Chapter 2 establishes the theoretical and application context for the

methods and problems addressed in the following chapters.

Compared with the existing literature, the novelty of this thesis lies in studying controllable diffusion models from both methodological and application-oriented perspectives under a unified framework. While prior work has often treated controllability in general visual generation and medical imaging as separate topics, this thesis connects them through a coherent line of research. Specifically, it explores attention-driven controllability and disentanglement in diffusion models, extends these ideas to dermoscopic image synthesis, develops bias-aware controllable generation to improve downstream segmentation under low-contrast conditions, and further investigates multimodal generation for chest X-ray images and reports. Together, these contributions distinguish the thesis from prior studies by linking fine-grained controllability mechanisms with practical medical imaging applications.

2.1 Visual Deep Generative Model

Deep generative models are neural networks that learn the probabilistic distribution of high-dimensional and complex data in an end-to-end learning and generate new data falling within the learned distributions (Suzuki and Matsuo, 2022). The target data modality can be various, including text, image, 3D point cloud, and audio. Among which, visual deep generative models designed for image generation have drawn a great deal of attention, as images are one of the most common and perceptible modalities. In this section, several prominent visual deep generative models proposed in the last decade are introduced to establish a theoretical background for this thesis.

2.1.1 Variational Autoencoders

A natural and intuitive form of a neural network for generating new data is the Autoencoders (AEs) (Baldi, 2012). AEs adopt an encoder-decoder architecture, where the encoder compresses the high-dimensional input data into low-dimensional latents, while the decoder reconstructs the input from the latents. The encoder and decoder are usually parameterized by CNNs in image generation tasks (Van den

Oord et al., 2016). Although the end-to-end training of AEs enables learning of compact representations and low-loss reconstruction of input, the lack of distributional constraint on the latent space usually results in meaningless output during generation, limiting the generation quality and flexibility (Shrivastava et al., 2024).

To address this limitation, the Variational Autoencoders (VAEs) (Kingma and Welling, 2013) utilize a probabilistic structure on the latent space to transform the AEs from a simple reconstruction model into a model capable of generating realistic and new data within the learned distribution in the target data. An illustration of the computation graph of VAEs is presented in Fig. 2.1.

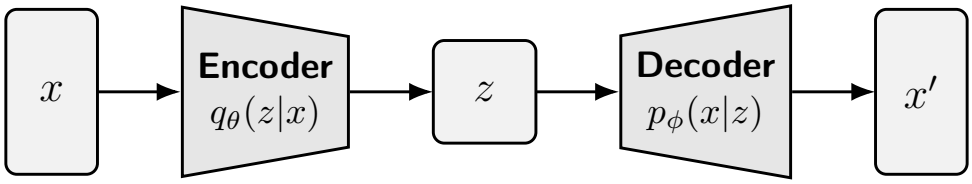


Figure 2.1: Computation graph of VAEs (Adapted from Lai et al. (2025))

The architecture of the VAEs resembles the AEs, but VAEs assume that each input data x is generated from a latent variable z that captures the hidden factors of x, such as the layout, color, or texture. Moreover, VAEs assume that z is sampled from a simple prior distribution (e.g., standard Gaussian distribution $z \sim p_{prior} := \mathcal{N}(\mathbf{0}, \mathbf{I})$).

During generation, z is projected back to the data space through the decoder that defines a conditional distribution $p_\phi(x|z)$. The decoder distribution is often chosen to be a Gaussian distribution with fixed variance:

$$p_\phi(x|z) := \mathcal{N}\left(x; \mu_\phi(z), \sigma^2\mathbf{I}\right) \tag{2.1}$$

where μ_ϕ is a neural network, $\mu_\phi(z)$ is the mean vector predicted by the decoder, and $\sigma > 0$ is a small constant controlling the variance. Such a simple distribution forces the learning to extract reliable latent features to produce realistic outputs instead of simply memorizing input data. When sampling new data, VAEs first sample a random $z \sim p_{prior}$, and decode into x via $x \sim p_\phi(x|z)$. The genration process can be thereby described by a marginal likelihood:

$$p_\phi(x) = \int p_\phi(x|z)\, p(z) dz \tag{2.2}$$

In the reverse process that projects x into z, the distribution $p_\phi(z|x)$ can be described by Bayes' theorem:

$$p_\phi(z|x) = \frac{p_\phi(x|z)\, p(z)}{p_\phi(x)} \tag{2.3}$$

However, the computation of $p_\phi(x)$ integrates over the entire latent space, and is computationally impossible. Therefore, VAEs approximate it with a parametric, easy-to-sample distribution as the encoder:

$$q_\theta(z|x) \approx p_\phi(z|x) \tag{2.4}$$

The encoder maps each data x to a distribution over latents, and z is then sampled from it. Similarly, the encoder $q_\theta(z|x)$ is commonly modeled as a Gaussian distribution in a form of:

$$q_\theta(z|x) = \mathcal{N}\left(z; \mu_\theta(x), diag\left(\sigma_\theta^2(x)\right)\right) \tag{2.5}$$

where $\sigma_\phi^2(z)$ is the vector of variances of x, and $diag(\cdot)$ converts the vector of variances into a diagonal covariance matrix.

Since the real data distribution $p(x)$ is never known, the generative models can be optimized by maximizing the likelihood $logp_\phi(x)$ over the dataset. Although $logp_\phi(x)$ is intractable either, it can be maximized by maximizing a lower bound, which is named the Evidence Lower Bound (ELBO):

$$\text{ELBO}_\theta(x) = \mathbb{E}_{q(z|x)}\left[\log p_\phi(x|z)\right] - \mathcal{D}_{KL}\left(q_\theta(z|x) \parallel p(z)\right) \tag{2.6}$$

where $\mathcal{D}_{KL}$ is KL-divergence. The first term is the reconstruction term that evaluates the performance of the decoder, while the second term is the latent regularization term that assesses the performance of the encoder. The optimization objective of VAEs is therefore maximizing the ELBO over the training set $\mathcal{X}$:

$$\underset{\phi,\theta}{argmax} \sum_{x \in \mathcal{X}} \text{ELBO}(x) \tag{2.7}$$

Nevertheless, VAEs suffer from the problem of blurry outputs because the decoders are typically modeled as a simple noise distribution, such as a Gaussian. It encourages the neural network to output the mean of all possible results when uncertainty exists, averaging different possible textures and edges and producing smooth, blurry results (Kingma and Welling, 2013). Therefore, new paradigms for high-fidelity image synthesis are required.

2.1.2 Generative Adversarial Networks

Generative adversarial networks (GANs) are a class of deep generative models that use an alternative training methodology for new data generation to alleviate the issues of blurry outputs and slow computation associated with VAEs (Goodfellow et al., 2014). Specifically, GANs adopt the philosophy of the minimax optimization problem in adversarial machine learning (Huang et al., 2011). In the context of a classification task, adversarial machine learning builds a defender and an attacker, which are essentially neural networks, where the defender searches over the model's parameter space to find the parameters that minimize the classification loss, while the attacker searches over possible input perturbations to produce samples that maximize the loss.

In the generation task that GANs focus on, the defender and attacker are renamed by the generator G and discriminator D. Given a dataset containing n data samples with d dimensions (i.e., $\left\{x_i \in \mathbb{R}^d\right\}_{i=1}^n$). The generator G takes a random noise $z \in \mathbb{R}^p$ with p dimensions as input, and produces a new data sample $x' \in \mathbb{R}^d$. The generated data x' should fall within the distribution of the real data x, thus the discriminator D is required to judge the quality of x' and compare its similarity with real data. D is a binary classifier that classifies the generated data as real or fake (generated) data:

$$D(x) := \begin{cases} 1, & \text{if } x \text{ is real} \\ 0, & \text{if } x \text{ is fake} \end{cases} \tag{2.8}$$

During the training of GANs, G and D play the minimax game where G is aiming to produce new data similar to real data, while D tries to discriminate the data generated by G and the data sampled from the real data distribution. G and D become stronger in a simultaneous training process in which they compete with each other. Furthermore, when one of them gets stronger abilities in training, the other would also be stronger to compete with it. The typical structure of GANs is elaborated in Fig. 2.2.

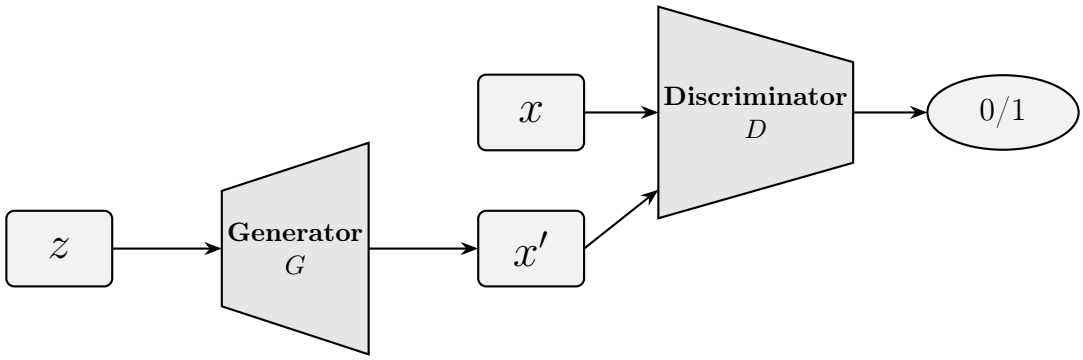


Figure 2.2: Computation graph of a GAN (Adapted from Lai et al. (2025))

In practice, G and D are initialized with neural networks. The optimization object for them is:

$$\min_G \max_D \; V(D,G) = \mathbb{E}_{x\sim p_{data}(x)} \left[\log D(x)\right] + \mathbb{E}_{z\sim p_z} \left[\log\left(1 - D(G(z))\right)\right] \tag{2.9}$$

where p_{data} and p_z are distributions of the real dataset and random noise, respectively. $\mathbb{E}[\cdot]$ denotes the expectation calculation. $V(D,G)$ is a binary entropy function.

In Eq. (2.9), the first term is an expectation over the real data distribution. It only affects D as it does not depend on G and therefore remains constant from G's perspective. According to Eq. (2.8), D outputs the value one for real data samples, so maximizing this term corresponds to encouraging D to correctly assign the "real"

label to real data. The second term in Eq. (2.9) is an expectation over the random noise z. Since a competent D should classify generated data as "fake", it seeks to minimize $D(G(z))$, which is equivalent to maximizing $log(1 - D(G(z)))$. On the other hand, G only affects the second term, and optimizes the opposite direction of this term, with its goal being to fool D into misclassifying generated data as real ones, and maximizing $D(G(z))$. During training, Eq. (2.9) is solved by alternating the following two gradient updates:

$$\begin{aligned} \theta_D^{t+1} &= \theta_D^t + \lambda^t \bigtriangledown_{\theta_D} V(D^t, G^t) \\ \theta_G^{t+1} &= \theta_G^t + \lambda^t \bigtriangledown_{\theta_G} V(D^{t+1}, G^t) \end{aligned} \tag{2.10}$$

where θ_D and θ_G are parameters for D and G, respectively. $\lambda$is the learning rate, and t denotes the iteration numvber of optimization.

Despite the superior generative capability compared to VAEs, GANs still suffer from several limitations. The minimax optimization often faces convergence problems, leading to instabilized training (Arjovsky and Bottou, 2017). Moreover, GANs frequently exhibit mode collapse, where G produces new data from only a subset of the distribution of real data, neglecting other modes (Arjovsky et al., 2017).

2.1.3 Denoising Diffusion Probabilistic Models

Diffusion models (DMs) are a novel branch of deep generative models that is designed to learn the complex distribution $p(x)$ of a set of observed samples x (Luo, 2022). DMs start from the introduction of diffusion probabilistic models, which model the data distribution by learning to reverse a gradual and multi-step noising process (Sohl-Dickstein et al., 2015). Denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020), inspired by considerations from nonequilibrium thermodynamics, represent a cornerstone of diffusion modeling and achieve huge success in high quality image synthesis. Compared to GANs, DDPM does not require the specific and sophisticated design of training and optimization, and has a lower risk of mode collapse.

As presented in Fig. 2.3 DDPMs can be interpreted from a variational perspective. DDPMs describe two distinct stochastic processes for image generation: the

forward process and the reverse denoising process.

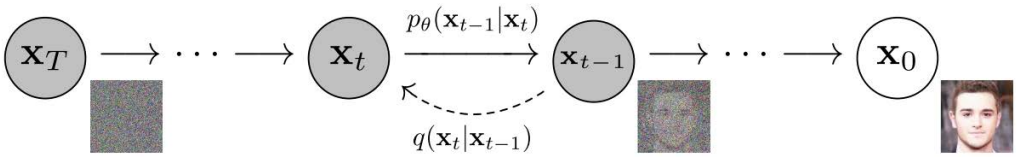


Figure 2.3: Illustration of DDPMs (Adapted from Ho et al. (2020))

The forward process corrupts the original data x_0 by gradually adding Gaussian noise via a transition $q(x_t|x_{t-1})$, as described in Eq. (2.11). The forward process takes multiple steps (typically 1000 steps in DDPMs), and transforms data into pure Gaussian noise $p_{prior} := \mathcal{N}(\mathbf{0}, \mathbf{I})$ over T steps, where the final corrupted sample is $X_T \sim p_{prior}$. The transition $q(x_t|x_{t-1})$, given in Eq. (2.12), can also be viewed as a prefixed encoder that does not require learning.

$$q\left(x_{1:T} \mid x_0\right) := \prod_{t=1}^{T} q\left(x_t \mid x_{t-1}\right) \tag{2.11}$$

$$q\left(x_t \mid x_{t-1}\right) := \mathcal{N}\left(x_t; \sqrt{\alpha}_t x_{t-1}, (1-\alpha_t)\mathbf{I}\right) \tag{2.12}$$

where $\sqrt{\alpha}_t x_{t-1} T$ and $(1-\alpha_t)\mathbf{I}$ are mean and variance, respectively. $\sqrt{\alpha}_t$ and $(1-\alpha_t)$ are scalars.

The reverse denoising process introduces a learnable decoder (usually a neural network) to iteratively denoise the corrupted data via a parameterized distribution $p_\theta(x_{t-1}|x_t)$:

$$p_\theta\left(x_{1:T} \mid x_0\right) := \prod_{t=1}^{T} p_\theta\left(x_{t-1} \mid x_t\right) \tag{2.13}$$

The ELBO of DDPMs can be obtained with respect to three stages in the forward/reverse process (Chan et al., 2024):

$$
\begin{aligned}
\mathrm{ELBO}_\theta(x) = \mathbb{E}_{q(x_1|x_0)} & \left[\log \underbrace{p_\theta(x_0|x_1)}_{\text{how good the initial block is}} \right] \\
& - \mathbb{E}_{q(x_{T-1}|x_0)} \left[\underbrace{\mathbb{D}_{\mathrm{KL}}\left(q(x_T|x_{T-1}) \| p(x_T)\right)}_{\text{how good the final block is}} \right] \\
& - \sum_{t=1}^{T-1} \mathbb{E}_{q(x_{t-1}, x_{t+1}|x_0)} \left[\underbrace{\mathbb{D}_{\mathrm{KL}}\left(q(x_t|x_{t-1}) \| p_\theta(x_t|x_{t+1})\right)}_{\text{how good the transition blocks are}} \right]
\end{aligned}
\tag{2.14}
$$

where the initial block only focuses on the denoising from x_1 to x_0, the final block only implements the noising from x_{T-1} to x_T, while transition blocks implement both the denoising and noising between x_{t-1} and x_t ($2 \leq t \leq T-1$). Eq. (2.14) is intractable since the samples (x_{t-1}, x_{t+1}) should be drawn from a joint distribution $q_\theta(x_{t-1}, x_{t+1}|x_0)$, which is unknown. By applying Bayes' theorem and further deduction, nevertheless, Eq. (2.14) can be simplified to:

$$
\begin{aligned}
\mathrm{ELBO}_\theta(x) = \mathbb{E}_{q(x_1|x_0)} \left[\log p_\theta(x_0|x_1)\right] & - \underbrace{\mathbb{D}_{\mathrm{KL}}\left(q(x_T|x_0) \| p(x_T)\right)}_{\text{nothing to train}} \\
& - \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)} \left[\frac{1}{2\sigma_q^2(t)} \left\| \mu_q(x_t, x_0) - \mu_\theta(x_t) \right\|^2 \right]
\end{aligned}
\tag{2.15}
$$

$$
\mu_q(x_t, x_0) = \frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\bar{\alpha}_t} x_t + \frac{(1-\alpha_t)\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t} x_0 \tag{2.16}
$$

Since the KL-divergence term in Eq. (2.15) does not contain parameters, there is nothing to learn during training, and it can be dropped. Equation (2.15) suggests that a network μ_θ should be found to minimize the loss $\frac{1}{2\sigma_q^2(t)} \left\| \mu_q(x_t, x_0) - \mu_\theta(x_t) \right\|^2$, where $\sigma_q(t) = \sqrt{1-\alpha_t}$ so μ_θ can be parameterized in the form in Eq. (2.17). In this conetxt, Eq. (2.15) can be further transformed into Eq. (2.18):

$$
\underbrace{\boldsymbol{\mu}_\theta(x_t)}_{\text{a network}} \stackrel{\text{def}}{=} \frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\bar{\alpha}_t} x_t + \frac{(1-\alpha_t)\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t} \underbrace{\hat{z}_\theta(x_t)}_{\text{another network}} \tag{2.17}
$$

$$
\begin{aligned}
\text{ELBO}_\theta(x) &= \mathbb{E}_{q(x_1|x_0)}\left[\log p_\theta(x_0|x_1)\right] \\
&\quad - \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}\left[\frac{1}{2\sigma_q^2(t)}\left\|\mu_q(x_t,x_0)-\mu_\theta(x_t)\right\|^2\right] \\
&= -\frac{1}{2\sigma_q^2(1)}\left\|\tilde{x}_\theta(x_1)-x_0\right\|^2 \\
&\quad - \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}\left[\frac{1}{2\sigma_q^2(t)}\left\|\mu_q(x_t,x_0)-\mu_\theta(x_t)\right\|^2\right] \\
&= -\sum_{t=1}^{T} \frac{1}{2\sigma_q^2(t)}\frac{(1-\alpha_t)^2\bar{\alpha}_{t-1}}{(1-\bar{\alpha}_t)^2}\mathbb{E}_{q(x_t|x_0)}\left[\left\|\hat{x}_\theta(x_t)-x_0\right\|^2\right]
\end{aligned}
\tag{2.18}
$$

where $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon_0 \quad \epsilon_0 \sim \mathcal{N}(0,I)$. Given a training set $\mathcal{X}$ containing m samples, the expectation in Eq. (2.18) can be approximated using Monte Carlo:

$$
\begin{aligned}
&\underset{\theta}{argmax} \sum_{x_0\in\mathcal{X}} \text{ELBO}(x_0) \\
&= \underset{\theta}{argmin} \sum_{x_0\in\mathcal{X}}\sum_{t=1}^{T} \mathbb{E}_{q(x_t|x_0)}\left[\frac{1}{2\sigma_q^2(t)}\frac{(1-\alpha_t)^2\bar{\alpha}_{t-1}}{(1-\bar{\alpha}_t)^2}\left\|\hat{x}_0\left(\sqrt{\bar{\alpha}_t}x_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0\right)-x_0\right\|^2\right] \\
&= \underset{\theta}{argmin} \sum_{x_0\in\mathcal{X}}\sum_{t=1}^{T}\frac{1}{M}\sum_{m=1}^{M} \frac{1}{2\sigma_q^2(t)}\frac{(1-\alpha_t)^2\bar{\alpha}_{t-1}}{(1-\bar{\alpha}_t)^2}\left\|\hat{x}_0\left(\sqrt{\bar{\alpha}_t}x_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0^{(m)}\right)-x_0\right\|^2
\end{aligned}
\tag{2.19}
$$

Therefore, optimizing Eq. (2.19) involves training a denoiser $\hat{x}_\theta(x_t)$ with step-dependent importance weights $w_t = \frac{1}{2\sigma_q^2(t)} \cdot \frac{(1-\alpha_t)^2\bar{\alpha}_{t-1}}{(1-\bar{\alpha}_t)^2}$. Furthermore, the signal prediction $\hat{x}_\theta \rightarrow x_0$ can be reformulated as a prediction of the noise. Given that $x_0 = \frac{x_t-\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0}{\sqrt{\bar{\alpha}_t}}$, Eq. (2.16) can be reformulated as Eq. (2.20) by substituting x_0 with it. Thus, the approximate denoising transition mean μ_θ can be set in a similar form, as presented in Eq. (2.21). Accordingly, the ELBO can be reformulated as Eq. (2.22).

$$
\mu_q(x_t,x_0) = \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\epsilon_0 \tag{2.20}
$$

$$\mu_\theta(x_t) = \frac{1}{\sqrt{\alpha_t}} x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}} \hat{\epsilon}_\theta(x_t) \tag{2.21}$$

$$\begin{aligned} \text{ELBO}_\theta(x_0, \boldsymbol{\epsilon}_0) &= -\sum_{t=1}^{T} \mathbb{E}_{q(x_t|x_0)} \left[\frac{1}{2\sigma_q^2(t)} \frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t} \left\| \hat{\boldsymbol{\epsilon}}_\theta(x_t) - \boldsymbol{\epsilon}_0 \right\|^2 \right] \\ &= -\sum_{t=1}^{T} \mathbb{E}_{q(x_t|x_0)} \left[\frac{1}{2\sigma_q^2(t)} \frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t} \left\| \hat{\boldsymbol{\epsilon}}_\theta \left(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t} \boldsymbol{\epsilon}_0 \right) - \boldsymbol{\epsilon}_0 \right\|^2 \right] \end{aligned} \tag{2.22}$$

where $\hat{\boldsymbol{\epsilon}}(x_t)$ is a neural network that learns to predict the noise $\boldsymbol{\epsilon}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ that determines x_t from x_0. Finally, the loss function of DDPMs from the perspective of noise prediction can be simply expressed as:

$$L_{DDPM} = \mathbb{E}_{t, x_0, \boldsymbol{\epsilon}} \left[|| \ \boldsymbol{\epsilon}_\theta(x_t) - \boldsymbol{\epsilon} ||^{\ 2} \right] \tag{2.23}$$

Using the trained $\hat{\boldsymbol{\epsilon}}(x_t)$, the inference process starts from a white noise $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. For each step $t = T, T-1, \cdots, 1$, the input noise is updated according to:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \hat{\boldsymbol{\epsilon}}_\theta(x_t) \right) + \sigma_q(t) \boldsymbol{z} \quad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2.24}$$

2.1.4 Latent Diffusion Models

In the early practice, the training and sampling of DDPMs cost excessive amounts of computational resources in the high-dimensional space of RGB images, motivating the exploration of image generation in lower-dimensional spaces. The latent diffusion models (LDMs) (Rombach et al., 2022) adapt DDPMs into a latent space from the pixel space. The schematic diagram of LDMs is presented in Fig. 2.4.

LDMs can be divided into three stages. The images $x \in \mathbb{R}^{H \times W \times 3}$ in RGB space are first compressed into latent representations z by the encoder $\mathcal{E}$ of VAEs as $z = \mathcal{E}(x)$. Then DMs are applied to the latent z by gradually adding Gaussian noise until z becomes pure noise z_T, and then learn a neural network (usually a convolutional-based U-Net) to predict the residual noise in each timestep during the denoising process. After denoising, the images are reconstructed from z by using the decoder

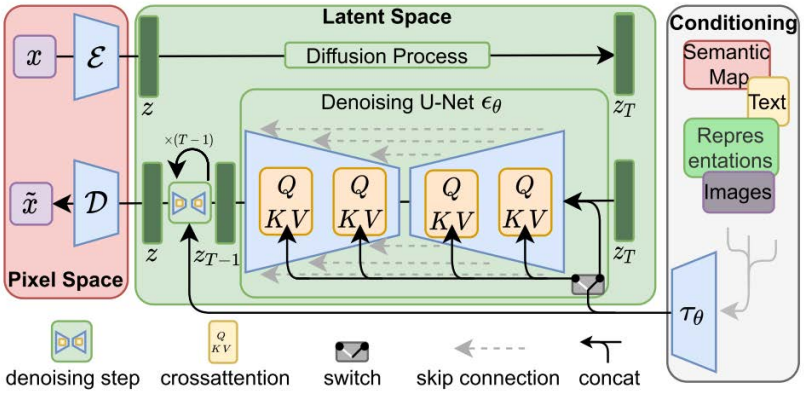


Figure 2.4: Illustration of LDMs (Adapted from Rombach et al. (2022))

of VAEs $\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(z))$. The dimension of latent representations is remarkably smaller than the original images since the encoder $\mathcal{E}$ downsamples the images by a factor of $f = 2^m$ with $m \in \mathbb{N}$. Therefore, the number of model parameters can be limited, allowing more space for model scaling to achieve high-resolution generation. Note that the VAE has been proven to be capable of preserving image information in reconstruction (Chambon et al., 2022b). More importantly, the cross-attention layer is introduced into the U-Net and LDMs can be conditioned on y in any modality, such as text and images. The optimizing object of conditional LDMs becomes:

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x),y,\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),t}\left[|| \boldsymbol{\epsilon}_\theta(z_t, t, \tau_\theta(y)) - \boldsymbol{\epsilon}||_2^2\right] \tag{2.25}$$

2.2 Attention Mechanism in Diffusion Models

Having established the probabilistic theories involved in DMs, I then turn the focus to the architectural design that plays a pivotal role in the DM process: the attention mechanism. While VAEs and GANs rely heavily on the inductive bias of CNNs to enforce local spatial correlations, LDMs necessitate architectures capable of modeling long-range dependencies and integrating complex, multimodal conditioning information. Attention mechanism, adapted from the Transformer architectures (Vaswani et al., 2017) originally developed for natural language processing, has fulfilled the requirement by providing a differentiable means of routing information globally across the latent image space. In this section, the architectural integration

of the attention mechanism within the U-Net backbone in LDMs is elucidated, and the interaction between latent image features and conditioning signals is discussed.

2.2.1 Attention in the Transformer

Before the introduction of the transformer, sequence modeling that learns temporal and structural patterns in sequences of ordered data was dominated by recurrent neural networks (RNNs) that processed data sequentially (Hochreiter and Schmidhuber, 1997). Such a paradigm does not support parallelization across sequence positions, and introduces a hidden state bottleneck, where the hidden state of the final token is expected to compress the entire semantic history of the sequence (Li et al., 2018). Moreover, CNNs face a similar limitation as a pixel's value of a generated image is determined by the receptive field, whose growth is slow with the depth of the network, resulting in weak encoding of high-level global structure (Seif and Androutsos, 2018). In contrast, the attention mechanism fundamentally addresses this problem by allowing every position from a single sequence to directly attend to all positions in the same sequence (self-attention) or all positions from another sequence (cross-attention) to compute a representation of the sequence. The attention operation enables highly expressive, content-dependent modeling of dependencies without regard to their distance in the sequences.

The basic architectural components of the transformer are presented in Fig. 2.5. Each transformer block contains N encoder-decoder pairs. In every encoder, a self-attention layer captures dependencies among all tokens in the input sequence. In contrast, each decoder contains a masked self-attention layer that only attends to previous tokens, and a cross-attention layer that takes the encoder's output information as conditions for the decoder. Such a combined-attention design enables the transformer to capture both intra-sequence and inter-sequence dependencies.

The core computation within the attention layer is the Scaled Dot-Product Attention (SDPA), which is illustrated in Fig. 2.6(a). Formally, given an input sequence, SDPA first projects the input data into three distinct values, which are queries (Q), keys (K), and values (V). These projections are learned via linear transformations:

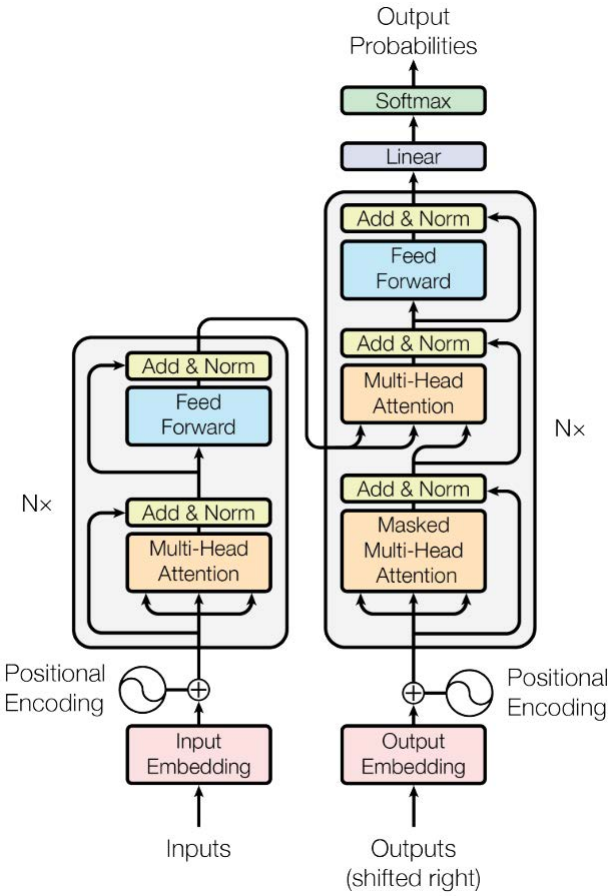


Figure 2.5: Architecture of transformer (Adapted from Vaswani et al. (2017))

$$Q = XW^Q, \quad K = YW^K, \quad V = YW^V \tag{2.26}$$

where $W^Q \in \mathbb{R}^{d_x \times d_k}$, $W^K \in \mathbb{R}^{d_y \times d_k}$, and $W^V \in \mathbb{R}^{d_y \times d_v}$ are learnable weight matrices. The source inputs $X \in \mathbb{R}^{N_x \times d_x}$ and $Y \in \mathbb{R}^{N_y \times d_y}$ are from the same input sequence in self-attention, enabling each position to aggregate information from all positions within the same sequence. Self-attention is bidirectional in the encoder, and is masked to preserve autoregressivity in the decoder. Cross-attention uses queries from the target sequence but keys and values from a separate source sequence from the encoder output, thereby fusing representations across modalities or sequences while preserving the core computational structure. The output of SDPA is a weighted sum of these values:

$$Attention(Q, K, V) = softmax\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2.27}$$

The term $softmax\left(\frac{QK^\top}{\sqrt{d_k}}\right)$ results in an attention map of size $N_x \times N_y$. An element (i, j) of the attention map represents the weight of the i-th query token in on the j-th key token. The scaling factor $1/\sqrt{d_k}$ prevents the softmax gradients from diminishing as the dimension grows.

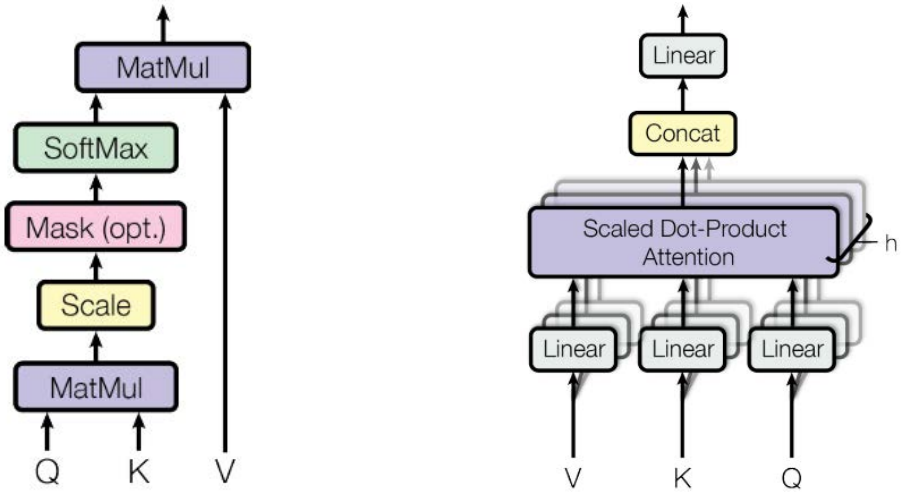
Furthermore, since the model needs to focus on different positions simultaneously, a single attention head would limit its ability to develop such an ability. To address this limitation, Multi-Head Attention (MHA), containing h attention heads that run independent attention functions Eq. (2.27) in parallel, is applied in the transformer, as shown in Fig. 2.6(b). The outputs of all attention heads are concatenated and projected via an output matrix W^O:

$$MultiHead(Q, K, V) = Concat\left(head_1, \ldots, head_h\right) W^O \tag{2.28}$$

In the vision community, the advent of the vision transformer (ViT) has demonstrated that the transformer architecture with attention mechanism applies to visual tasks, and outperforms CNNs (Dosovitskiy, 2020). By dividing an input image into 16×16 non-overlapping patches, ViT transforms computer vision problems into sequence-to-sequence problems that can be processed with the attention mechanism. ViT also provides the foundation for multimodal models like Contrastive Language-Image Pre-training (CLIP (Radford et al., 2021)), which generate aligned embeddings for paired images and texts.

2.2.2 Attention in Diffusion models

The standard backbone for DDPMs is a time-conditional U-Net, whose structure was originally developed for medical image segmentation. It has a symmetric structure consisting of an encoder and a decoder. The encoder progressively downsamples the spatial resolution of the feature maps while increasing the channel dimension, compressing the input image into a high-level latent representation. The decoder upsamples the latent features back to the original resolution, with skip connections concatenating features from the encoder to the decoder to preserve high-frequency spatial information lost during downsampling. However, self-attention blocks are



(a) (b)

Figure 2.6: Illustration of (a) Scaled Dot-Product Attention and (b) Multi-Head Attention (Adapted from Vaswani et al. (2017))

only inserted at the 16×16 resolution level between convolutional blocks due to the heavy computation required at higher-resolution levels.

On the other hand, LDMs facilitate heavier usage of the attention mechanism by moving the diffusion process from the pixel space to a latent space compressed by VAEs (e.g., a 512×512 image is compressed into a 64×64 latent). Such a low-resolution latent space makes it possible to insert attention blocks at multiple scales without high computational costs. Specifically, attention blocks are applied at four resolution levels in SD's U-Net [1], including 64×64, 32×32, 16×16, and 8×8. At low-resolution levels, the attention mechanism processes coarse semantic structures and global layout. In contrast, at high resolutions, the attention mechanism refines the details, ensuring that the textures and styles of target objects are consistent (Nguyen et al., 2024a).

A major architectural breakthrough in LDMs is the explicit formalization of conditioning through cross-attention. Earlier conditional GANs primarily injected conditioning signals by concatenating class labels or global projection vectors, offering only coarse and inflexible control (Mirza and Osindero, 2014). In contrast, LDMs inject conditioning information directly into the U-Net via cross-attention.

[1] https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5

Concretely, in the T2I generation, cross-attention maps textual embeddings $\tau_\theta(y)$ encoded by a domain-specific encoder τ_θ (e.g., BERT (Devlin et al., 2019), CLIP (Radford et al., 2021), or T5 (Raffel et al., 2020)) onto the flattened latent features z_t via Eq. (2.27), where Q is calculated from z_t, K and V are calculated from $\tau_\theta(y)$. The cross-attention operation allows each spatial location in the latent features to dynamically query the entire text sequence, establishing semantic correspondence, as detailed in Chapter 2.2.3.

Moreover, self-attention layers operate on the spatial dimension of U-Net features, which simultaneously serve as queries, keys, and values. By allowing every spatial patch to attend to all others, self-attention layers explicitly model long-range dependencies, such as texture consistency across distant regions or maintaining object symmetry Dhariwal and Nichol (2021). An illustration of cross-attention and self-attention is showcased in Fig. 2.7.

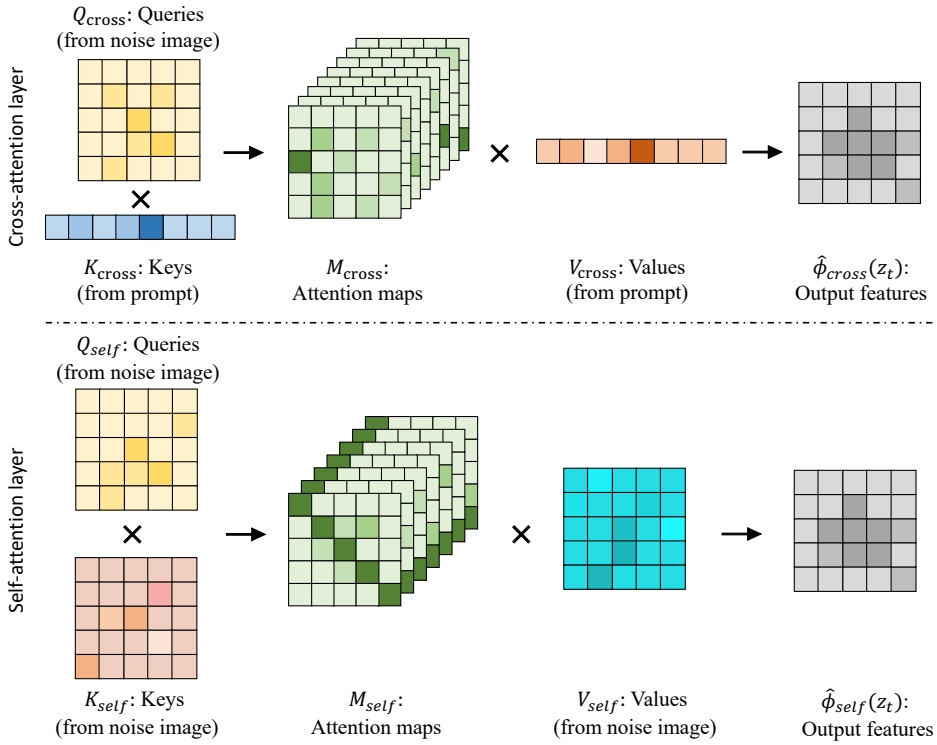


Figure 2.7: Illustration of cross-attention and self-attention operations in SD (Adapted from Liu et al. (2024a))

2.2.3 Interpretation of Attention Maps

Due to their nature as deep neural networks, deep generative models present limited interpretability (Moran and Aragam, 2025). Understanding the internal mechanics of DMs is a prerequisite for safety, controllability, and optimization. The primary locus of this understanding lies in the attention mechanism, which offers a unique window into the black box. This section explores the research that utilizes the attention mechanism to understand the internal reasoning of DMs.

Starting with the exploration of cross-attention in DMs, the most direct question associated with the interpretability is "Which parts of the image correspond to which words?" Diffusion Attentive Attribution Maps (DAAM) is the first work to analyze the semantic connections between textual tokens and visual latent from a visuolingustic perspective(Tang et al., 2023b). Specifically, per-token attribution maps based on cross-attention over generated images are established by aggregating the cross-attention scores across time steps, layers and heads. The attribute maps act as 2D heat maps per input text token, indicating which image regions that word most influences. DAAM reveals a solid semantic relationship between generated images and nouns in the prompts, with each noun's attribute map clearly presenting the position and shape of the corresponding objects in generated images. Furthermore, the authors revealed that the syntactic dependencies in the text prompt influence the spatial layout of generated image regions, and adjectives tend to exert diffuse, global influence rather than localized modification. Collectively, these findings demonstrate the promise of attention-based attribution for diffusion interpretability in T2I models.

A closely related line of work is the Prompt-to-Prompt (P2P), which demonstrates that cross-attention maps are not only interpretable but can also be manipulated to achieve fine-grained image editing (Hertz et al., 2022). P2P allows users to modify an image by altering the text prompt, while largely preserving the global spatial layout and semantic structure of the original image without the requirement for retraining. It introduces the concept of cross-attention injection, where attention maps associated with specific tokens or layers are smoothly steered, enabling the model to change specific concepts, attributes, or styles while maintaining com-

position, pose, and background coherence, as presented in Fig. 2.8. Importantly, P2P can be combined with DDIM inversion (Song et al., 2020a), which reverses a generated or real image into its latent diffusion trajectory, thereby enabling editing of existing content rather than only generating new samples from scratch. In addition, optimized inversion methods (Mokady et al., 2023) were subsequently proposed to enhance the reconstruction quality of the given real image, thereby facilitating meaningful and intuitive editing.

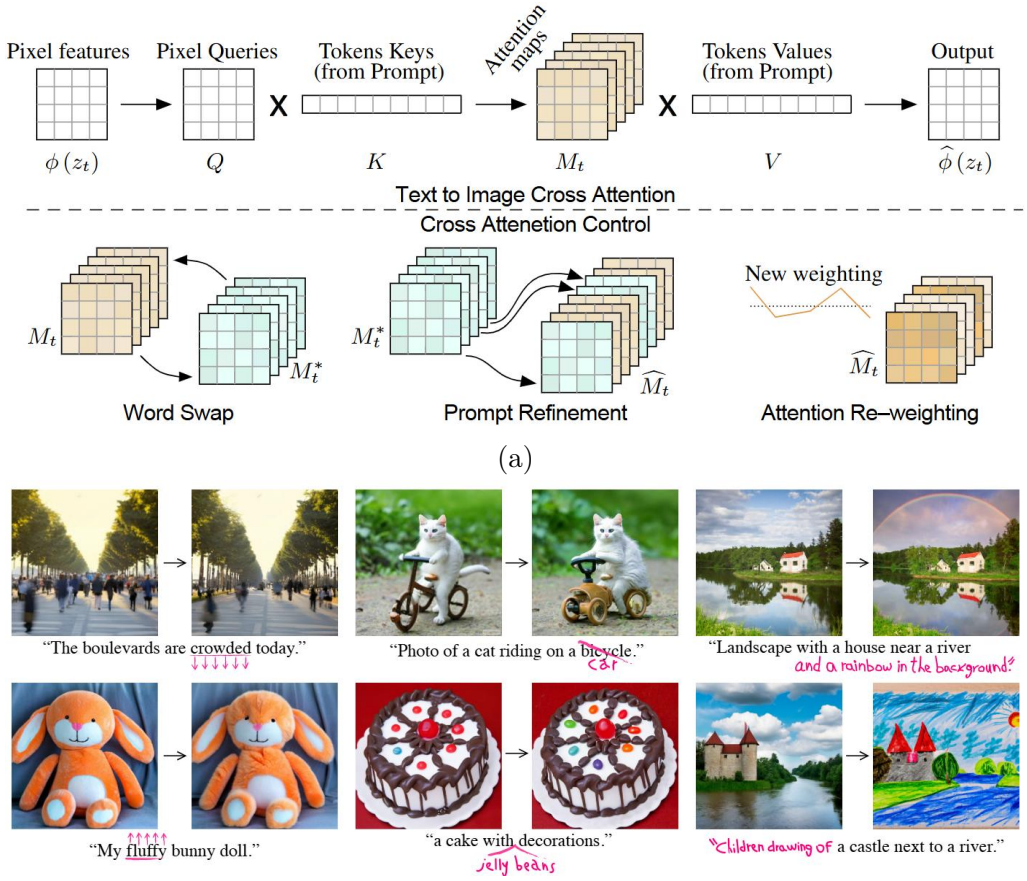


Figure 2.8: (a) Method overview and (b) editing capabilities of P2P. Cross-attention layers produce attention maps of visual latents and textual embeddings. During generation, the spatial layout and shape of cross-attention maps can be modified in different ways (e.g., swapping, refining, and re-weighting) to affect the generated images, achieving various editing operations. (Adapted from Hertz et al. (2022))

In addition, apart from textual tokens from prompts, special tokens, the end-of-sentence token (EOT) and the start-of-sentence token (SOT) play a unique role on cross-attention maps. As the EOT often acts as a global aggregator of the

sentence's meaning in the CLIP, its cross-attention map for the token attends to the entire image or salient foreground object (Wu et al., 2024a; Mun et al., 2024). The SOT, on the other hand, is reported to carry rich background information, reflected by its cross-attention map (Chen et al., 2024a).

As cross-attention specifies "what does where", self-attention answers "how does it hold together". In the context of image-to-image translation, Plug-and-Play (PnP) finds that the structure of the generated image can be controlled through self-attention layers inside the U-Net of DMs during the generation process (Tumanyan et al., 2023). The self-attention computes pixel-to-pixel (or patch-to-patch in the latent space) affinity, which demonstrates the shape of an object independent of its semantic class. As presented in Fig. 2.9, a visualization of self-attention maps processed by principal component analysis (PCA) indicates that self-attention maps can accurately capture the layout of the image, and self-attention maps with higher resolutions (i.e., extracted from high-level layers in U-Net) tend to present finer structural information. Therefore, the topology from the source image can be maintained by preserving these affinities, and then injected into the newly generated image containing distinct content.

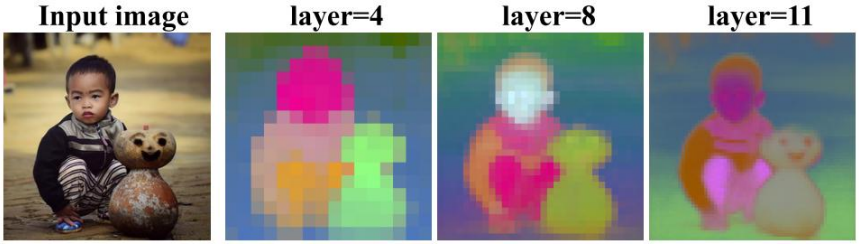


Figure 2.9: Visualization of self-attention maps extracted from different U-Net layers using PCA (Adapted from Tumanyan et al. (2023))

Apart from the structural and geometric information, self-attention layers also encode rich stylistic information, such as color, texture, and general appearance style. MasaCtrl (Cao et al., 2023) introduces a mutual self-attention that can query correlated local contents and textures from source images for consistency. Hertz et al. (2024) propose a shared self-attention method that applies AdaIN (Huang and Belongie, 2017) over the self-attention queries and keys of the target image using the reference queries and keys, generating style-aligned images with varying contents

as shown in Fig. 2.10. Furthermore, this characteristic of self-attention is directly applied to the style transfer of real images by swapping the self-attention keys and values (Chung et al., 2024). Therefore, self-attention in DMs offers a dual role that supports both semantic consistency and stylistic control.

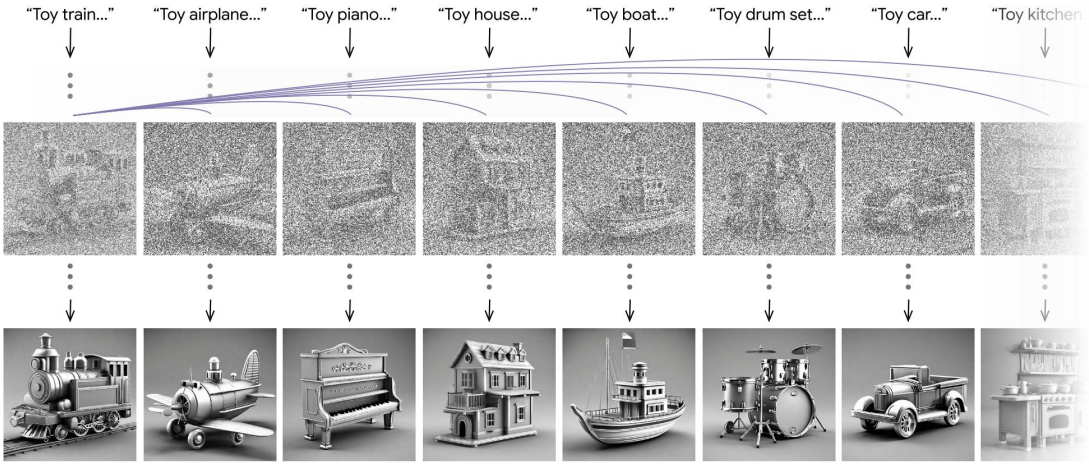


Figure 2.10: Style-aligned generation using shared attention layers (Adapted from Hertz et al. (2024))

In light of these findings about the characteristics of cross-attention and self-attention, their roles in image segmentation are further exploited. Cross-attention maps between nouns in text prompts and their corresponding objects in generated images are often processed to be pseudo-masks as auxiliary tools for object locating in various tasks, such as image generation (Cao et al., 2023), image editing (Couairon et al., 2022), and image personalization (Ham et al., 2024). Moreover, the rich semantic and structural information from these attention maps can directly assist image segmentation or the augmentation of segmentation datasets. Yoshihashi et al. (2023) directly utilizes pseudo-masks post-processed from cross-attention maps as segmentation labels for synthesized images, and the synthesized image-mask pairs are used for training downstream segmentation models. To improve the accuracy of such pseudo-masks, Wu et al. (2023a) proposes to sequentially use a fixed threshold value γ, a refining tool, DenseCRF (Krähenbühl and Koltun, 2011), which excels in finding local relationships defined by color and distance of pixels, and finally an AffinityNet (Ahn and Kwak, 2018) to give an estimation for pixels with a middle confidence score. This generation paradigm "cross-attention maps + labeling de-

coder" for puedo-labels was then extended to more modalities, including semantic mask, instance mask, depth map, and pose estimation (Wu et al., 2023b). Furthermore, self-attention maps are capable of refining spatial details of cross-attention maps (Nguyen et al., 2023; Ma et al., 2023a). By simply multiplying τ-th powere of self-attention map A_S to the cross-attention map A_C, the structural layout of A_C can be enhanced by the pixel-to-pixel affinity matrices contained in A_S:

$$A_C^* = (A_S)^\tau \cdot A_C \tag{2.29}$$

An illustration of this attention enhancement with different values of τ is shown in Fig. 2.11. Based on these insights, SLIME (Khani et al., 2023) proposes a novel weighted accumulated self-attention map that helps show the boundaries of target objects in more detail, and can be directly used for image segmentation.

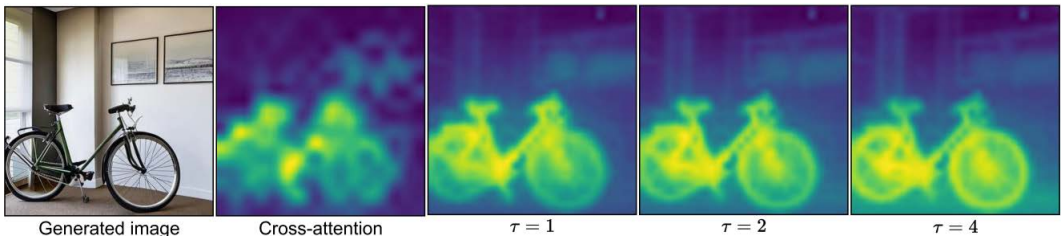


Figure 2.11: Enhancing the cross-attention maps using the self-attention maps with $\tau = 1, 2, 4$, given a text prompt "A bike is parked in a room; bicycle" for generating the first image (Adapted from Nguyen et al. (2023))

2.3 Controllability of Visual Generation

While early deep generative models demonstrate the capacity to synthesize plausible data samples from the learned distributions, they offer users minimal control over the semantic or structural attributes of the output (Chen et al., 2016). In this section, I'll systematically analyze the evolution of controllable visual generation, tracing the methodological development from the latent space disentanglement efforts in VAEs and GANs to the gradient-guided mechanisms in the current era of DMs. The controllability of generation discussed in this section is not only class conditioning but also the capacity to render high-dimensional attributes, including spatial layout,

semantic composition, stylistic texture, and subject identity, without degrading the generative prior of the foundation model (Cao et al., 2025).

2.3.1 Controllability in Early Deep Generative Models

The primary theoretical framework for controllability in early deep generative models is grounded in the concept of disentanglement. The central hypothesis is whether a generative model could learn a representation where the distinct dimensions of the latent code z corresponded linearly to independent semantic factors of variation in the data space x (e.g., rotation, scale, color), then control could be achieved via simple vector arithmetic or traversal along these latent axes.

While standard VAEs can successfully compress data samples into the latent space and restore them, the latent spaces are often entangled, with multiple semantic attributes encoded in a single encoded latent dimension, causing difficulties in controllable generation. β-VAE (Higgins et al., 2017) introduces an enforced disentanglement by introducing a hyperparameter $\beta > 1$ to the KL-divergence term of the ELBO, and the objective becomes:

$$L_{\beta\text{-VAE}} = \mathbb{E}_{q(z|x)}[\log p_\theta(x|z)] - \beta \mathcal{D}_{KL}(q_\theta(z|x)||p(z)) \tag{2.30}$$

By imposing a heavier penalty on $\mathcal{D}_{KL}$, the new objective forces the posterior to align more closely with the prior. Since the prior has independent components, the posterior is encouraged to learn statistically independent factors of variation in the data, thereby achieving disentanglement. However, such a formulation also introduces a trade-off between reconstruction and disentanglement. Increasing β effectively constrains the information capacity of the latent bottleneck. Although this encourages statistical independence among latent dimensions and thus improves disentanglement, it simultaneously restricts how much information about the input x can be preserved in z. As a consequence, the decoder is forced to reconstruct from a poor latent representation, often resulting in blurry or low-fidelity outputs (Eddahmani et al., 2023).

To achieve disentanglement without uniformly increasing the KL weight as in

β-VAE, FactorVAE (Kim and Mnih, 2018) decomposes the KL-divergence term to isolate the penalty responsible for disentanglement, and explicitly penalizes the total correlation of the marginal distribution of the latent $q(z) = \mathbb{E}_{p(x)}[q(z \mid x)]$. The total correlation (TC) is defined as the KL-divergence between a joint distribution and the product of its marginals:

$$TC(z) = \mathcal{D}_{KL}(q(z)||\prod_{j=1}^{d} q(z_j)) \tag{2.31}$$

where z denotes a d-dimensional latent random vector produced by the encoder of VAEs, and z_j is the j-th scalar latent variable. $q(z_j)$ is the marginal distribution of the j-th latent dimension, obtained by integrating out all other latent variables. $q(z)$ is the joint distribution over all latent dimensions.

Minimizing $TC(z)$ encourages the distribution of the latent to be independent across dimensions without forcing the posterior $q(z|x)$ to collapse to the prior $p(z)$ for each data sample. FactorVAE optimizes $TC(z)$ using a discriminator network that distinguishes between samples from $q(z)$ and samples from the factorized distribution $\bar{q}(z) = \prod_j q(z_j)$.

Similarly, vanilla GANs employ an adversarial loss to produce visually realistic images, but they provide little control over the semantic or structural attributes of the generated samples. In the absence of explicit constraints or supervision, variations in the latent code lead to entangled and unpredictable changes in the output, making precise manipulation of individual factors difficult. The earliest form of explicit control in GANs is the Conditional GAN (cGAN) (Mirza and Osindero, 2014), which conditions both the generator and discriminator with an auxiliary variable y, such as a class label. While the cGAN can effectively realize categorical control (e.g., "cat" or "dog"), it lacks the ability to manipulate continuous and fine-grained attributes within a class (e.g., "a white cat with a short tail" or "a standing dog").

Building upon this idea, conditional adversarial learning is rapidly extended to image-to-image translation, where the conditioning variable y becomes a structured visual input. The goal of image-to-image translation is to learn a mapping between

two image domains while preserving the underlying spatial structure of the input image. A representative model for this task is Pix2Pix (Isola et al., 2017), which formulates paired image-to-image translation as a conditional GAN problem, where the generator learns a mapping from an input image (e.g., a semantic map) to a target image (e.g., a photograph), as shown in Fig. 2.12. By leveraging paired training data and an additional reconstruction loss, Pix2Pix encourages structural alignment between the input and output, enabling explicit control over the generated image through structural conditions. This paradigm demonstrates that cGANs can effectively translate high-level spatial layouts into realistic visual images, and subsequent study has extended it for high-resolution image translation from semantic maps (Wang et al., 2018). However, it heavily depends on the availability of paired datasets, which are often expensive or infeasible to collect.

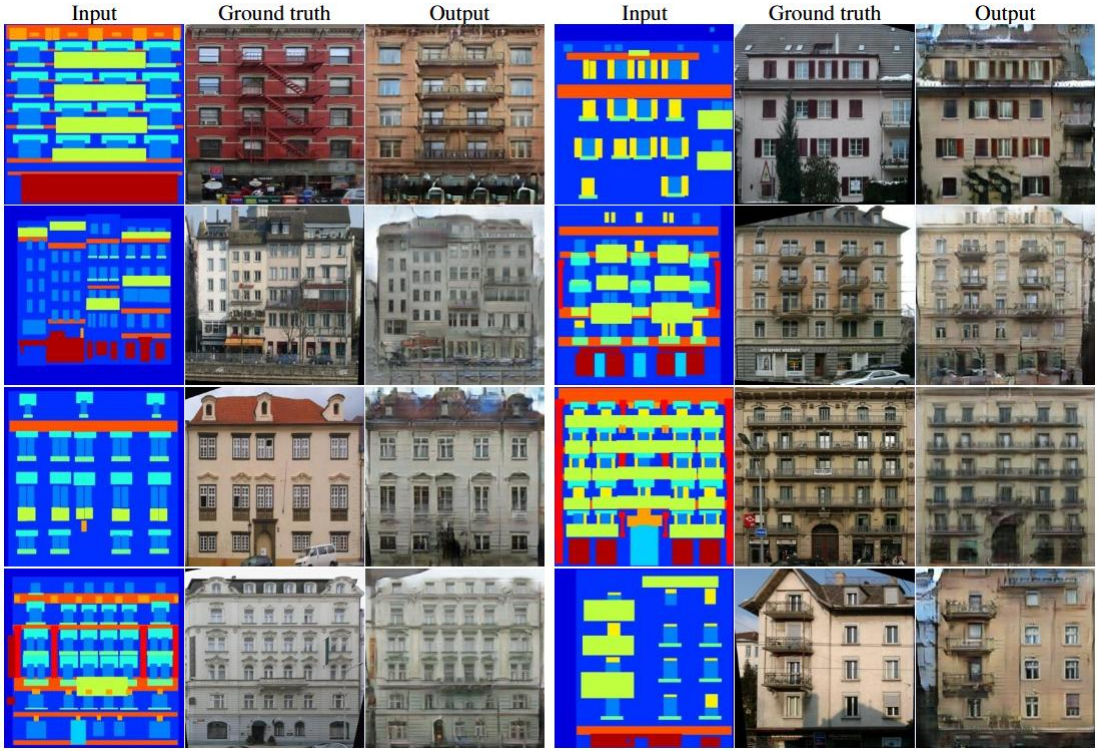


Figure 2.12: Example results of Pix2Pix (Adapted from Isola et al. (2017))

To address the data limitation, CycleGAN introduced a framework for unpaired image-to-image translation (Zhu et al., 2017). It employs two generators and a cycle-consistency constraint, enforcing that a sample translated from one domain to another and back should reconstruct the original input (i.e., $G(F(x)) \approx x$). This

design enables learning translations between domains without paired supervision. Moreover, it inspires the exploration of style-level controllability in unpaired image-to-image translation (Kim et al., 2017). Furthermore, SPADE (Park et al., 2019) resolves the issue of semantic information washing away by injecting semantic information into multiple network layers, allowing users to specify a detailed semantic layout.

Additionally, StyleGAN has significantly developed unsupervised controllability for image generation. In StyleGAN, the input latent vector z, sampled from a simple prior distribution, is first mapped through a learned mapping network into an intermediate latent space $\mathcal{W}$, which is then used to control the generator via layer-wise modulation. StyleGAN modulates the activations of each convolutional layer using Adaptive Instance Normalization (AdaIN) parameters derived from $\mathcal{W}$. As a result, different layers of the generator become responsible for different levels of visual abstraction: earlier layers primarily control coarse, high-level attributes such as pose, shape, and global structure, while later layers influence fine-grained details such as texture, color, and microstructure (Karras et al., 2019). This hierarchical organization leads to an emergent form of disentanglement, where semantic attributes can often be manipulated by modifying specific components of the latent representation. An example of the StyleGAN results is presented in Fig. 2.13.

Prior to the emergence of DMs, GANs had attempted T2I generation, aiming to generate images conditioned on natural language descriptions. StackGAN (Zhang et al., 2017) separates T2I generation into a two-stage process, where a Stage-I generator produces low-resolution images that capture coarse shape and basic color information conditioned on a text embedding, and a Stage-II generator subsequently refines these images to higher resolution by adding fine-grained details. Furthermore, AttnGAN (Xu et al., 2018) introduces an attention mechanism to improve fine-grained text-image correspondence. Instead of conditioning solely on a global sentence embedding, AttnGAN aligns individual word embeddings with spatial subregions of the generated image through a word–region attention module, enabling the model to generate images that better reflect detailed textual descriptions.

Despite the successes in controllable image generation, GAN-based approaches

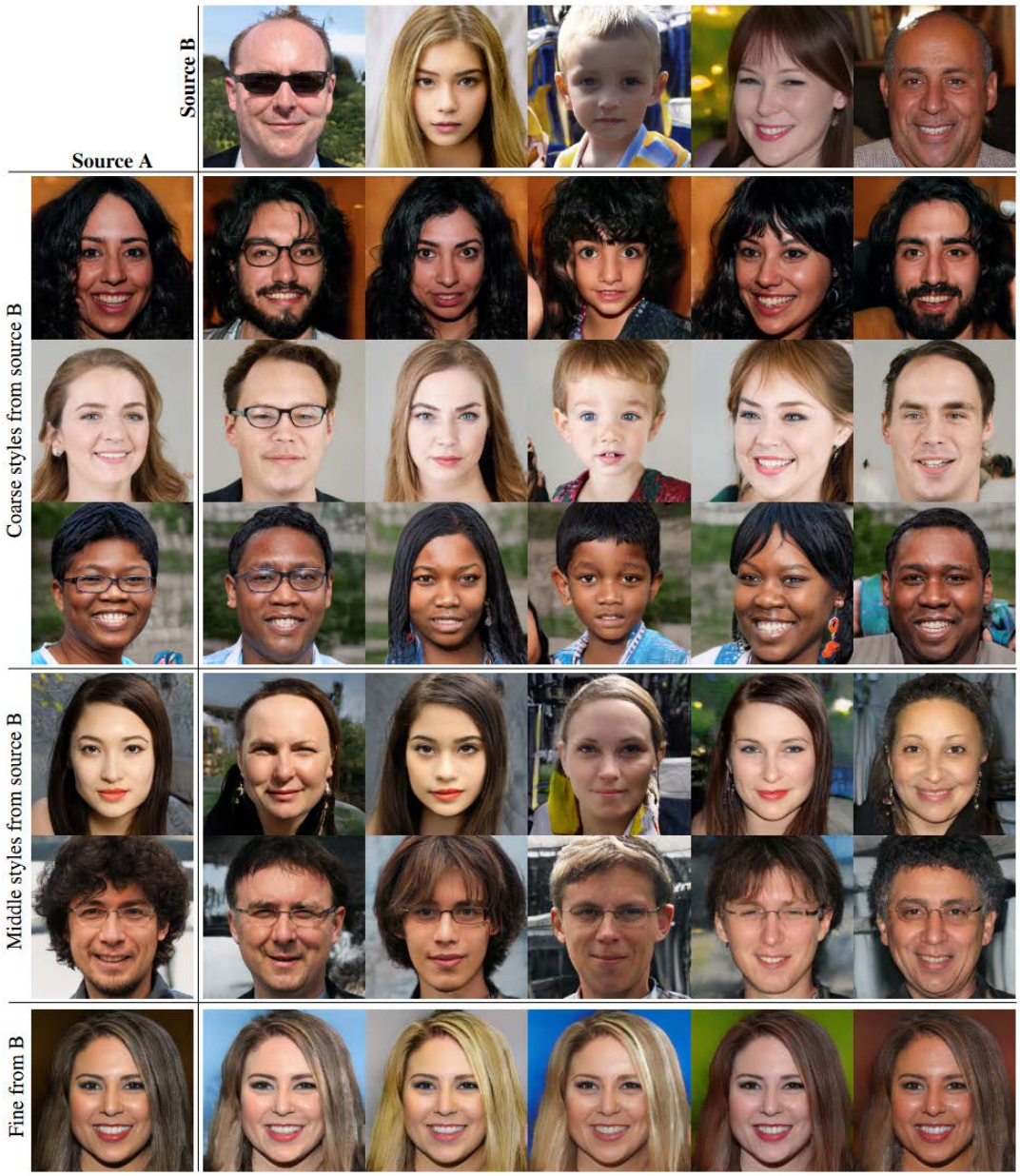


Figure 2.13: Example results of style mixing between Source A and Source B using StyleGAN. Source A is overridden with styles from B at specific resolutions. Coarse styles (4^2-8^2) transfer high-level semantics like pose, face shape, and glasses. Middle styles (16^2-32^2) inherit smaller facial features and hair details. Fine styles (64^2-1024^2) transfer the color scheme and microstructure. (Adapted from Karras et al. (2019))

still suffer from several intrinsic limitations that restrict their flexibility, robustness, and scalability. First, editing real images typically requires GAN inversion that projects the real samples into the latent space, but this inversion process is ill-posed and often leads to reconstruction artifacts, identity drift, or loss of fine de-

tails, particularly for out-of-domain inputs (Anirudh et al., 2018). Second, although architectures based on StyleGAN introduce a degree of latent disentanglement, their control remains largely global and entangled due to the lack of spatially localized or object-aware conditioning, resulting in unintended changes in other regions or attributes when manipulating a single semantic attribute. Last but not least, the adversarial training objective inherent to GANs prioritizes sample realism over data likelihood, making these models susceptible to mode collapse and limited distribution coverage, which in turn constrains output diversity (Sajjadi et al., 2018).

2.3.2 Controllability in Diffusion Models

The transition from VAEs/GANs to DMs not only boosts the image quality, but also brings a fundamental rethinking of controllability in generative modeling. By formulating image synthesis as a stochastic denoising process that progressively refines samples from noise, DMs provide multiple natural intervention points, during training, conditioning, and inference, through which control signals can be injected, modulated, or edited. To systematize the rapidly growing literature, I organize controllable diffusion methods into three evolutionary paradigms:

- Implicit controllability via pretraining
- Auxiliary training of modular components
- Inference-time controllability

Implicit controllability via pretraining

Classifier-free guidance (CFG) (Ho and Salimans, 2022) leverages pretrained conditional knowledge through a modified pretraining strategy to bias sampling trajectories toward desired conditions. Different to explicit controllability learned end-to-end (discussed in the following sections), this conditional generation arises from the recombination of pretrained probabilistic components in an implicit pathway during sampling.

CFG trains a single diffusion model using conditioning dropout, where the condition embedding c is randomly replaced with a null token $\emptyset$ with probability p_{drop} (typically 10-20%) during pretraining. As a result, the model implicitly learns two score fields: a conditional Score: $\epsilon_\theta(x_t, c, t)$ and an unconditional Score: $\epsilon_\theta(x_t, \emptyset, t)$. Recalling the score of the conditional distribution $\nabla \log p(c|x) \propto \nabla \log p(x|c) - \nabla \log p(x)$, CFG approximates the guidance term by the difference between the conditional and unconditional scores, and the modified score for sampling is defined as:

$$\tilde{\epsilon}_\theta(x_t, c, t) = \epsilon_\theta(x_t, \emptyset, t) + w \cdot (\epsilon_\theta(x_t, c, t) - \epsilon_\theta(x_t, \emptyset, t)) \tag{2.32}$$

where w is the guidance scale. CFG does not introduce explicit control variables into the diffusion process. Instead, controllability emerges implicitly from how the model was pretrained and how its outputs are combined during inference.

Despite its success, CFG still has inherent limitations, most notably the increased inference cost resulting from dual forward passes. Furthermore, the method is prone to extrapolation artifacts at high guidance scales, where the guided score pushes samples beyond the training distribution, leading to oversaturation and contrast amplification. Although techniques such as Dynamic Thresholding partially mitigate these issues, they do not fundamentally alter the implicit nature of the control signal (Saharia et al., 2022).

Auxiliary training of modular components

The CFG discussed in the previous section realizes controllability implicitly through pretraining objectives. Although it is highly effective for semantic control that specifies what content should appear in an image, it remains limited in its ability to impose fine-grained spatial constraints, such as precise object layout, geometry, or pose (where/how things should appear). This limitation stems from the fact that conditioning signals act globally on the score field and do not explicitly interface with the internal spatial representations of the DMs. At the same time, finetuning the entire large-scale diffusion backbone (e.g., SD) for each new spatial conditioning task is computationally unacceptable. Moreover, there is a risk of catastrophic forgetting, which causes the model to lose previously acquired general visual knowl-

edge (Kirkpatrick et al., 2017). These challenges motivated the emergence of modular controllability, a novel paradigm in which the pretrained diffusion backbone is frozen and lightweight auxiliary modules are trained to inject control signals in a structured and task-specific manner.

Within this paradigm, controllability becomes explicitly encoded in dedicated modules, rather than implicitly emerging from pretraining. ControlNet (Zhang et al., 2023a) is a landmark in modular controllability for DMs. It provides a general-purpose framework for incorporating diverse spatial conditions, including but not limited to canny edges and human pose, as shown in Fig. 2.14, and human pose skeletons, into a frozen pretrained DM, while preserving its generative capacity. The core design of ControlNet is the trainable copy strategy. Starting from a pretrained SD, ControlNet freezes the original U-Net parameters, thereby preserving semantic knowledge learned from large-scale pretraining. Then, it creates a trainable copy of the encoder with an identical architecture. The trainable copy is connected to the locked model via 1×1 convolution layers initialized with both weights and biases set strictly to zero, termed zero convolution layers. The architectural design of ControlNet is presented in Fig. 2.15.

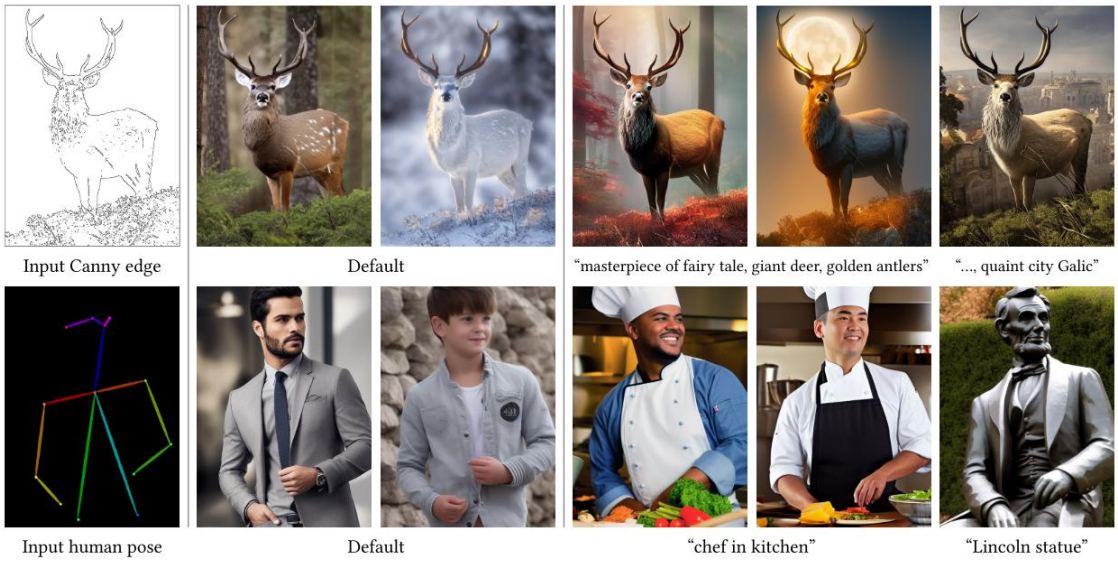


Figure 2.14: Example results of ControlNet that applies both textual and spatial control over generated images. (Adapted from Zhang et al. (2023a))

The external condition image c is fed into the trainable copy of the encoder. The features extracted by the trainable encoder are then injected into the frozen U-Net

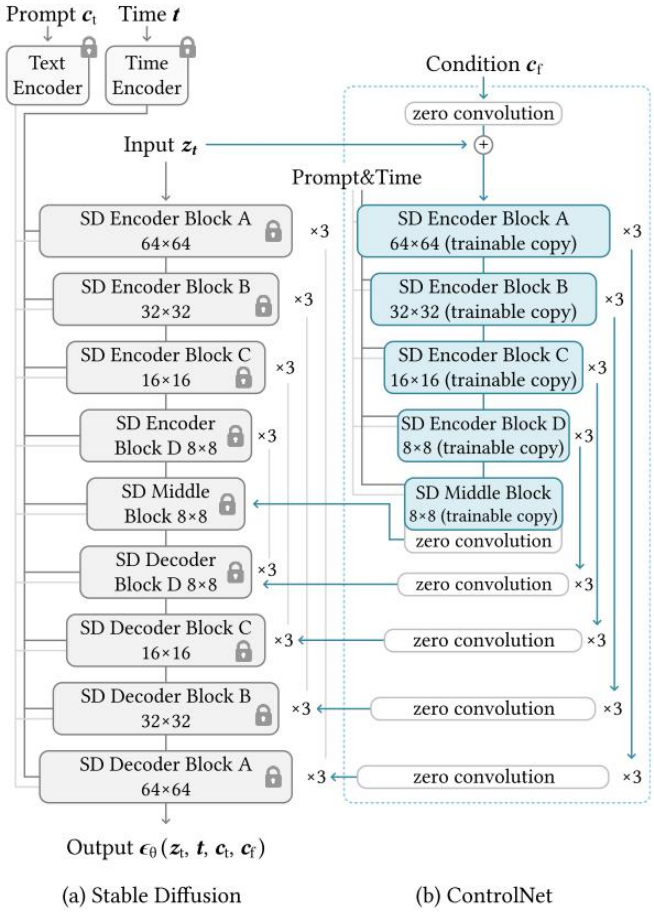


Figure 2.15: Model architecture of ControlNet. (Adapted from Zhang et al. (2023a))

through zero convolution layers. The output y of a ControlNet block is given by:

$$y = F(x; \Theta_{locked}) + Z(F(x + Z(c; \Theta_{z1}); \Theta_{trainable}); \Theta_{z2}) \tag{2.33}$$

where $F(\cdot)$ denotes U-Net feature transformations and $Z(\cdot)$ represents the zero convolution operation. Θ_{z1} and Θ_{z2} are parameters of two zero zero convolution layers.

This design ensures that structural information is introduced in a controlled and stable manner, without disrupting the pretrained backbone. The zero convolution plays a beneficial role in the paradigm of adapter-based finetuning, as it applies identity initialization, which assures the model behaves identically to the original pretrained SD at the start of the training, and prevents unstable gradients from randomly initialized layers from propagating into the frozen backbone, enabling smooth and stable adaptation. Furthermore, the outputs of the trainable encoder are added to the corresponding skip features of the frozen U-Net at multiple spatial

resolutions (e.g., 64×64, 32×32, 16×16), enabling pixel-level alignment with structural conditions such as edges or pose skeletons.

Concurrent to ControlNet, T2I-Adapter (Mou et al., 2024) offers a more lightweight solution by introducing a standalone lightweight convolutional network (70–80M parameters) that processes the condition image. As illustrated in the model architecture design (Fig. 2.16), the adapter consists of a sequence of downsampling and residual blocks designed to produce feature maps that are spatially and channel-wise aligned with the internal representations of the U-Net. The extracted features are injected directly into the encoder blocks of the frozen U-Net via additive fusion. The adapter outputs are pre-aligned to the corresponding resolutions and channel dimensions, ensuring seamless integration without modifying the backbone architecture.

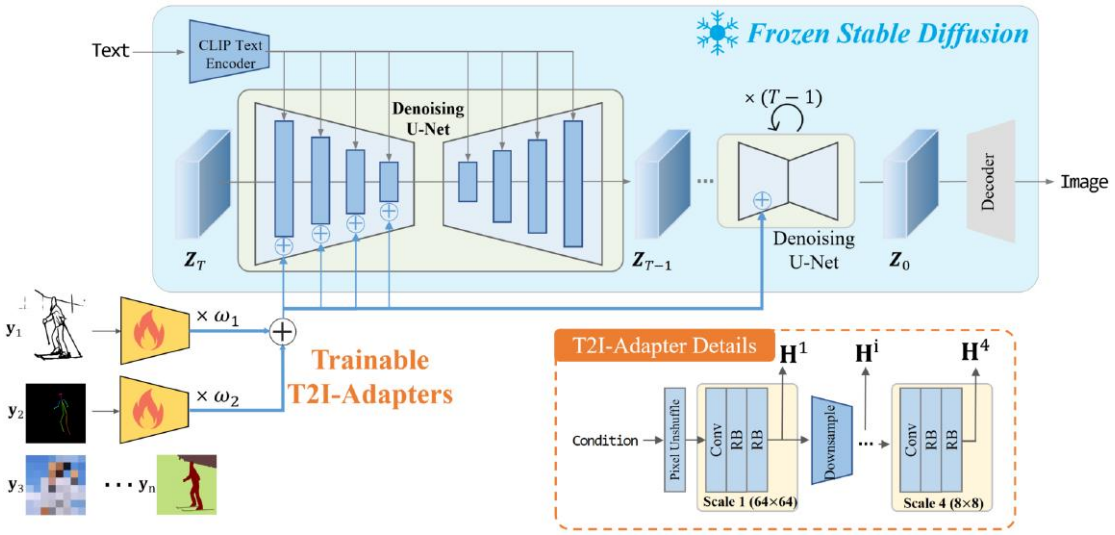


Figure 2.16: Model architecture of T2I-Adapter. (Adapted from Mou et al. (2024))

While ControlNet and T2I-Adapter excel at injecting explicit spatial structure, they are less suited for image-based semantic or stylistic conditioning, such as generating an image in the style of a reference image. IP-Adapter (Ye et al., 2023) addresses this limitation by introducing modular control at the attention level, rather than through convolutional feature injection. An illustration of model design is presented in Fig. 2.17.

IP-Adapter decouples text and image conditioning within the cross-attention mechanism. The original U-Net and text cross-attention layers are frozen, and a

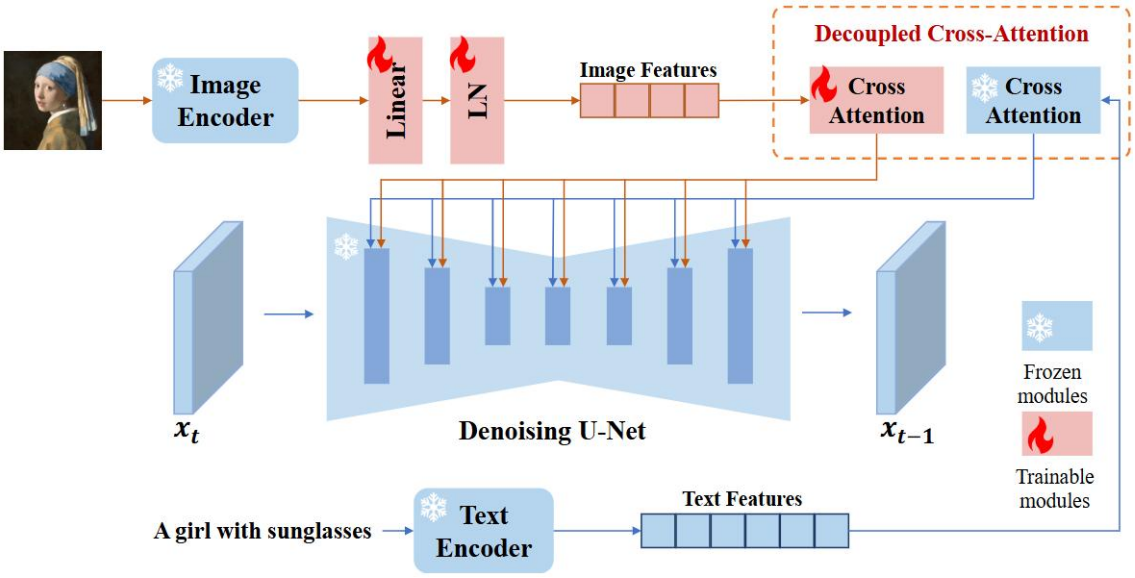


Figure 2.17: Model architecture of IP-Adapter. (Adapted from Ye et al. (2023))

parallel attention branch is introduced specifically for image features. It augments the standard cross-attention operation with a second, image-specific branch:

$$\text{Output} = \text{Softmax}\left(\frac{QK_{text}^{\top}}{\sqrt{d}}\right)V_{text} + \text{Softmax}\left(\frac{QK_{image}^{\top}}{\sqrt{d}}\right)V_{image} \tag{2.34}$$

where K_{image} and V_{image} are projected from CLIP image features via newly introduced trainable layers. By sharing the query Q while separating key-value spaces, IP-Adapter enables additive multimodal conditioning, allowing users to combine text prompts with image references in a controllable and interpretable manner. Examples of application are presented in Fig. 2.18.

Inference-time controllability

Score-based guidance DMs frame data generation as the reverse of a predefined stochastic process that progressively injects noise into observed data. This formulation provides not only a principled probabilistic foundation, but also a natural interface for incorporating external control signals through conditional distributions. While the previous discussion introduced DMs from a variational inference perspective, it is often more illuminating to analyze controllability through the perspective

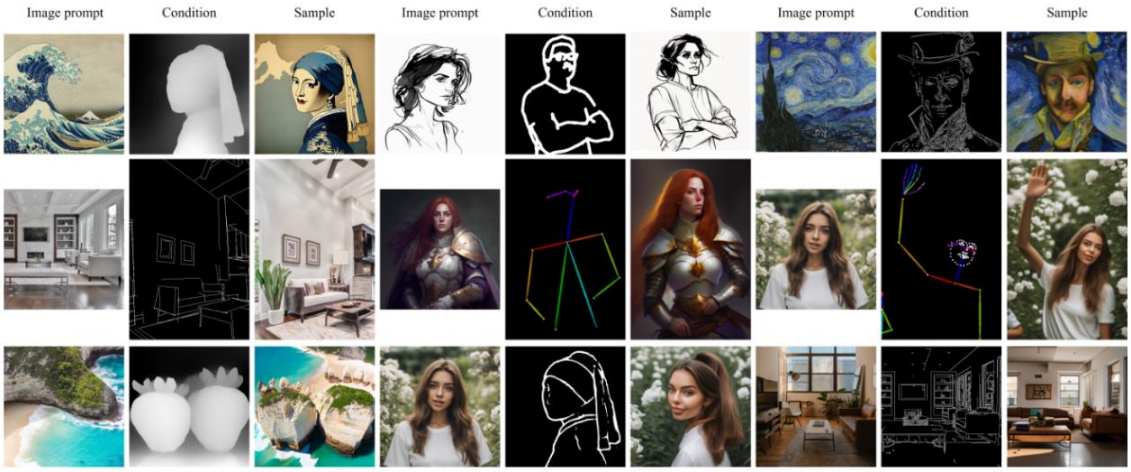


Figure 2.18: Examples of generated samples with image prompt and additional structural conditions using IP-Adapter. (Adapted from Ye et al. (2023))

of score-based generative modeling, which provides a continuous-time formulation and a more direct interpretation of conditional guidance.

Recall that DMs are defined as a Markov chain that gradually perturbs data samples with Gaussian noise, as described in Eq. (2.12). In continuous time, the discrete forward process can be described by a Stochastic Differential Equation (SDE) of the form:

$$dx = f(x,t)dt + g(t)dw \tag{2.35}$$

where $f(x,t)$ is the drift coefficient, $g(t)$ is the diffusion coefficient, and w denotes standard Brownian motion (Song et al., 2020b). The generative capability of DMs arises from simulating the corresponding reverse-time SDE, which gradually transforms noise back into structured data:

$$dx = [f(x,t) - g(t)^2 \nabla_x \log p_t(x)]dt + g(t)d\bar{w} \tag{2.36}$$

where $d\bar{w}$ denotes Brownian motion evolving backward in time, and $\nabla_x \log p_t(x)$ is the score function, which is the gradient of the log-density of the data distribution at time step t. However, the reverse process depends on the score function of the marginal distribution $p_t(x)$, which is generally intractable and must be approximated. The central objective of diffusion model training is therefore to learn a

neural network $s_\theta(x, t)$ to approximate the true score:

$$s_\theta(x, t) \approx \nabla_x \log p_t(x) \tag{2.37}$$

Controllable generation of DMs requires sampling from a conditional distribution $p(x|c)$, where c denotes an external control signal. Within the diffusion framework, controllability can be formally analyzed through the decomposition of the conditional score function. By applying Bayes' rule $p(x|c) = p(c|x)p(x)/p(c)$, the score of the conditional distribution can be decomposed as:

$$\nabla_x \log p_t(x|c) = \nabla_x \log p_t(x) + \nabla_x \log p_t(c|x) \tag{2.38}$$

The decomposition reveals that the conditional score consists of two components. The first term $\nabla_x \log p_t(x)$ is the unconditional score that captures the intrinsic geometry of the data manifold learned from unlabeled data. It encodes prior knowledge about what constitutes a plausible sample, thereby ensuring realism and sample fidelity. The second term $\nabla_x \log p_t(c|x)$ is the guidance term, which represents the gradient of the log-likelihood of the condition given the noisy sample. Functionally, it acts as a steering force that biases the reverse diffusion trajectory toward regions of the latent space consistent with the desired condition.

This formulation implies that controllability does not necessarily require retraining the generative model itself. As long as the guidance term can be estimated, it can be injected into the reverse process at sampling time. Thus, guidance-based controllability emerges from the manipulation of the reverse diffusion process. The most representative methods of guidance-based controllability are Classifier Guidance (Dhariwal and Nichol, 2021). Rather than modifying the DMs' architecture or retraining them conditionally, Classifier Guidance relies on a noise-aware classifier $p_\phi(c|x_t, t)$ pretrained on noisy diffusion states to inject conditional information. During reverse diffusion, controllability is introduced by injecting the classifier gradient into the sampling process. In the noise-prediction formulation, the guided update is:

$$\hat{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) - s \cdot \sqrt{1-\alpha_t} \nabla_{x_t} \log p_\phi(c|x_t) \tag{2.39}$$

where s is the guidance scale controlling the relative influence of the classifier.

The diffusion model itself remains unconditional and unchanged, whereas the condition affects the generation process only through the pretrained classifier, making controllability an emergent property of the interaction between pretrained models, rather than an explicitly learned control mechanism. Nevertheless, Classifier Guidance requires training specialized classifiers for every new condition (e.g., text, depth maps, and semantic maps) under various noise levels, raising high demand for computational resources.

The drawbacks of classier guidance necessitate universal approaches that can bridge the gap between noisy latents and clean-image discriminators. Bansal et al. (2023) propose Universal Guidance to eliminate the need for noise-aware training. They demonstrate that standard, off-the-shelf vision models trained only on clean images can guide the diffusion process by leveraging the DMs' own single-step denoising approximation $\hat{x}_0$ as a proxy input for the guidance function. The proposed algorithm introduces two mechanisms: Forward Universal Guidance and Backward Universal Guidance. Forward Universal Guidance adjusts the score estimate based on the gradient of the guidance loss with respect to $\hat{x}_0$, while Backward Universal Guidance optimizes the clean estimate directly to satisfy constraints and projects these changes back to the noisy state. This approach enables the use of diverse guidance signals, including segmentation maps, object detection, and face recognition, without finetuning the diffusion backbone or the guidance networks.

However, the reliance on the single-step denoising approximation to compute gradients has limitations, as Wallace et al. (2023) argue that gradients derived from this one-step approximation are often misaligned with the final generation, leading to suboptimal control, particularly when complex guidance signals are involved. To resolve this, they introduce Direct Optimization of Diffusion Latents (DOODL), which shifts the optimization target from intermediate steps to the initial noise vector x_T. By utilizing an invertible diffusion process (EDICT), which allows for memory-efficient backpropagation through the entire diffusion chain. By maintain-

ing two coupled latent vectors, EDICT allows the reconstruction of intermediate activations during the backward pass with constant memory complexity. This enables full-chain guidance, which significantly improves performance on finegrained vocabulary expansion and aesthetic alignment compared to methods relying on single-step approximations.

While Bansal et al. (2023) and Wallace et al. (2023) focus on the mechanics of applying off-the-shelf guidance, Dinh et al. (2023) identify that standard Classifier Guidance often leads to diversity suppression and adversarial effects, where the model generates high-confidence samples that lack robust semantic features or collapse into a single mode, such as over-exploiting specific background textures. To mitigate this, they propose a generalized approach named Progressive Guidance to modify the guidance objective. As shown in Fig. 2.19, rather than aggressively optimizing for a single target class, Progressive Guidance incorporates gradients from relevant, semantically similar classes during the early, noisy stages of sampling. By progressively refining the guidance signal from broad semantic concepts to specific class details along the temporal dimension, this method enhances sample diversity and feature robustness while maintaining inference-time controllability.

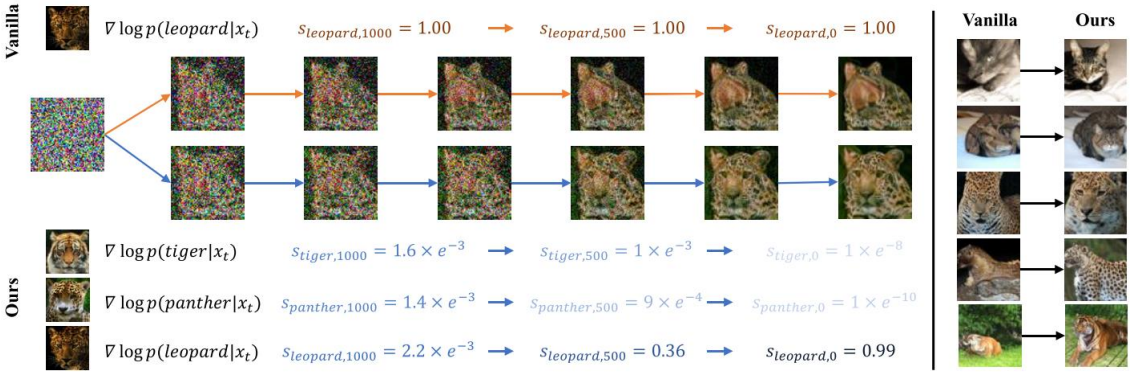


Figure 2.19: Illustration of Progressive Guidance. The vanilla guidance fails to condition the generated sample as a leopard, whereas Progressive Guidance (Ours) uses the tiger, panther, and leopard to influence the image content in the initial state to form the critical features of the leopard. The right part shows more failure cases corrected using Progressive Guidance. The darkness of the gradients denotes the associated information degree values.(Adapted from Dinh et al. (2023))

Attention-based control The second major category of inference-time control targets the attention mechanism of DMs. As discussed in Chapter 2.2.3, the cross-attention demonstrates semantic connections between textual and visual features, and self-attention contains rich spatial relationship within the latents. By leveraging these characteristics of attention mechanism, direct manipulation of the internal representations at inference time can be achieved.

Attention-based control are commonly applied with inversion techniques to realize image editing, as demonstrated by P2P (Hertz et al., 2022). It performs two diffusion processes in parallel: one driven by the original prompt and another by an edited prompt. During sampling, selected cross-attention maps from the source generation are injected into the target generation at corresponding layers enforcing spatial alignment between the two generations, and ensuring that objects remain in the same locations even when their semantic identities change. While P2P excels at text-driven manipulation, it remains limited when the desired control signal is another image rather than a textual description, raising the requirement for self-attention–based image guidance (Cao et al., 2023; Alaluf et al., 2024). Methods such as MasaCtrl (Cao et al., 2023) extend the attention mechanism by allowing the target image to attend not to its own features, but to those of a reference image. Specifically, the queries are computed from the target latent, while the keys and values are drawn from the source image latent, effectively transferring appearance-related information from the source to the target, while preserving the target's spatial structure.

Furthermore, the self-attention is adopted for style-transfer (Chung et al., 2024; Zhang et al., 2023b; Hu et al., 2024). StyleID (Chung et al., 2024) demonstrates a training-free method to adapt a large pretrained diffusion model for style transfer directly through self-attention manipulation. As shown in Fig. 2.20, it reveals that self-attention layers encode rich appearance and texture information apart from spatial features, making them suitable for style modulation. During sampling, it replaces the key-value pairs of self-attention layers in the target generation with those extracted from a reference style image, while preserving the query from the content image. This enables the model to retain structural layout while inheriting

stylistic patterns from the reference.

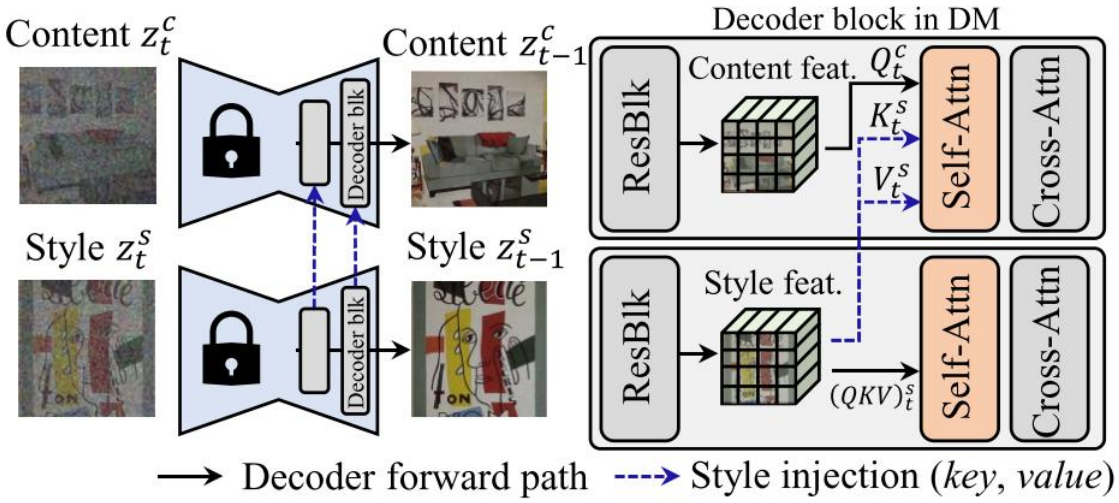


Figure 2.20: Adaptation of self-attention for style transfer.(Adapted from Chung et al. (2024))

In addition, attention-based control can be combined with score-based guidance to realize more diverse, robust, and interpretable inference-time control and optimization. As demonstrated by DAAM (Tang et al., 2023b), the manipulation of textual tokens on generated visual feature works through cross-attention, but one of the most significant limitations identified through cross-attention analysis is the semantic neglect phenomenon. When a prompt contains two semantically similar words such as "a cat and a dog", the attention maps for "cat" and "dog" frequently overlap or bleed into one another because the text encoder (often CLIP) projects semantically similar words into approximate regions of the embedding space, leading to generated images where the "cat" might have dog-like features or where attributes intended for one object (e.g., "spotted dog") are omitted. Attend-and-Excite (Chefer et al., 2023) addresses this limitation by explicitly optimizing attention activations during inference. At each denoising step, the method evaluates the maximum activation of each target token in the cross-attention map. If a token's activation falls below a threshold, a gradient is computed to increase its contribution:

$$z_t \leftarrow z_t + \eta \nabla_{x_t} max(A_{token}) \tag{2.40}$$

where η is a coefficient adjusting the scale of gradient, A_{token} denotes the cross-

attention map of a certain token. Attend-and-Excite effectively transforms attention into a controllable optimization objective, ensuring semantic coverage without altering the model or training process. An illustration of the optimization of it on image synthesis is shown in Fig. 2.21. Based on Attend-and-Excite, subsequent studies keep improving the semantic consistency between text prompt and generated images via additional loss functions during inference-time (Li et al., 2023a; Rassin et al., 2023).

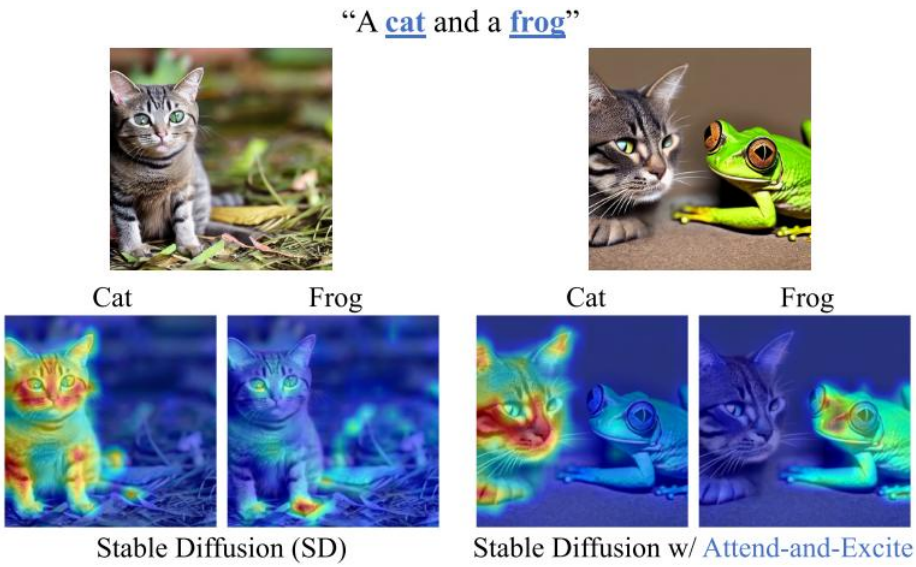


Figure 2.21: Generated images and cross-attention maps for each subject token with and without Attend-and-Excite over vanilla SD.(Adapted from Chefer et al. (2023))

Beyond semantic control, attention manipulation is extended explicit spatial constraints for direct image synthesis without auxiliary training. Several works design specific spatial constraints to align generated objects with bounding boxes. BoxDiff (Xie et al., 2023) proposes to apply a set of training-free constraints, including inner-box, outer-box, and corner constraints, to cross-attention maps to restrict object synthesis to user-defined regions. Similarly, Attention Refocusing (Phung et al., 2024) introduces cross-attention refocusing (CAR) and self-attention refocusing (SAR) losses, optimizing the latent to maximize attention within target regions while minimizing it elsewhere. Moreover, BACON (Chen et al., 2024b) employs boundary-constrained attention losses (region, boundary, and regularization) to refine object boundaries and counts. In a similar manner, R&B (Xiao et al., 2023) gradually modulates the attention maps through a boundary-aware loss to

strengthen object discriminability within the corresponding regions. Furthermore, approaches leverage internal representations of attention maps for guidance. Self-Guidance (Epstein et al., 2023) extracts properties such as size and location from internal activations to steer the sampling process, similar to Classifier Guidance but using the model's own signals. For segmentation-level control, ZestGuide (Couairon et al., 2023) extracts implicit segmentation maps from cross-attention layers and optimizes the noise estimation to align with a provided zero-shot segmentation mask. To offer a better understanding, illustrations of the results by BoxDiff and ZestGuide are presented in Fig. 2.22.

(a) (b)

Figure 2.22: Illustration of the results by (a) BoxDiff and (b) ZestGuide (Adapted from Xie et al. (2023); Couairon et al. (2023))

Latent and feature-space optimization Beyond modifying attention mechanisms or applying score-based optimization, a third influential paradigm for controllable diffusion focuses on manipulation of latent variables z_t and intermediate feature representations. This paradigm enables finegrained spatial control, flexible semantic editing, and high-fidelity preservation of structure, without retraining the underlying model.

One of the most intuitive approaches to latent manipulation is Blended Latent Diffusion (Avrahami et al., 2023a), which addresses the challenge of seamless local editing and compositional generation, as illustrated in Fig. 2.23. At each denoising step t, the latent variable is constructed as a spatial blend of a background latent

obtained by inverting the original image and a foreground latent generated from a new prompt. The blending is governed by a binary or soft spatial mask M, given by:

$$z_t^{blend} = M \odot z_t^{foreground} + (1 - M) \odot z_t^{background} \tag{2.41}$$

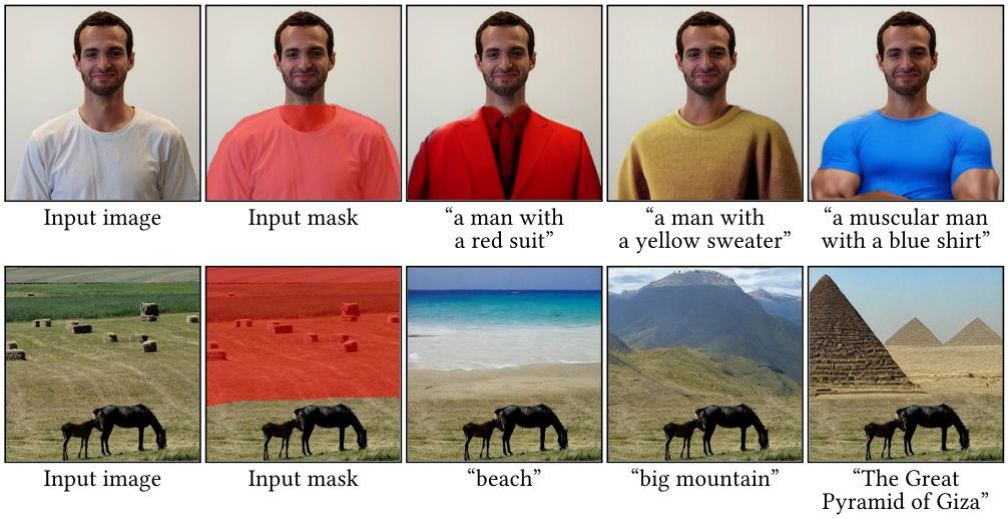


Figure 2.23: Illustration of the results of Blend Latent Diffusion. (Adapted from Avrahami et al. (2023a))

Building on this principle, MultiDiffusion (Bar-Tal et al., 2023) generalizes blending to support large-scale and multi-region generation. It runs multiple diffusion processes in parallel over overlapping spatial regions or tiles, each potentially guided by a different text prompt. Moreover, PnP (Tumanyan et al., 2023) extracts intermediate U-Net features from a reference image during inversion and reinjects them during generation for tasks such as structure-preserving translation. FreeControl (Mo et al., 2024) extends this idea by observing that principal components of feature maps capture stable semantic structures such as edges and depth cues. By enforcing alignment between the principal subspaces of a reference image and the generated output, FreeControl enables structure-aware guidance without relying on explicit spatial annotations or additional networks.

Furthermore, DragonDiffusion (Mou et al., 2023) realizes geometric manipulation, such as moving an object or reshaping a semantic part, by utilizing feature-level correspondence within the U-Net. Since the intermediate feature maps encode dense

semantic descriptions that are spatially aligned with the image content, DragonDiffusion constructs an energy function based on the cosine similarity between feature vectors at these locations, and then encourages the semantic content at the handle to move toward the target position by minimizing the energy function.

Additionally, several methods reinterpret controllability as a problem of optimizing the entire diffusion trajectory rather than individual states. Specifically, Null-Text Inversion (NTI) (Mokady et al., 2023) addresses the mismatch between conditional and unconditional branches in CFG. While standard DDIM inversion recovers a latent trajectory corresponding to a given image, it does not ensure consistency between the conditional and unconditional score functions, often leading to reconstruction errors and semantic drift during editing. NTI optimizes the unconditional text embedding (the null token) rather than the latent code itself, such that the diffusion process can faithfully reconstruct the input image under CFG, enabling stable and high-fidelity edits without modifying model weights or introducing additional guidance losses. ReNoise (Garibi et al., 2024) further refines inversion quality by repeatedly applying forward diffusion and denoising steps, averaging predictions to converge toward a stable latent fixed point. This iterative refinement reduces the semantic gap in DDIM-based inversion, without incurring the heavy optimization cost of techniques like NTI. These novel inversion approaches highlight the importance of trajectory-level consistency in achieving robust and controllable diffusion generation and editing.

2.4 Applications of Diffusion Models in Medical Imaging

The integration of deep generative models into medical image analysis has witnessed a paradigm shift with the emergence of DMs (Luo et al., 2025; Kazerouni et al., 2023). Although earlier generative models, particularly VAEs and GANs, have been applied in this domain, their technical characteristics impose important limitations in medical contexts. VAEs commonly trade visual fidelity for latent regularization, often yielding overly smooth reconstructions in which subtle anatomical boundaries

or small lesions may be blurred. GANs, in contrast, can generate sharper images, but their adversarial training is inherently unstable and susceptible to mode collapse, which is problematic when rare pathological patterns and population diversity must be preserved. Moreover, GAN-based approaches generally lack an explicit probabilistic formulation, making uncertainty estimation and likelihood-based reasoning less straightforward, despite the fact that these properties are highly relevant in clinically oriented applications. By comparison, a central advantage of DMs lies in their formulation as iterative denoising processes that progressively transform noise into structured samples. This perspective naturally aligns with many medical imaging problems, which can be interpreted as inverse or ill-posed reconstruction tasks (Webber and Reader, 2024). Furthermore, the probabilistic nature of DMs has proven particularly advantageous for addressing the scarcity of annotated medical data, enabling robust data augmentation strategies for rare pathologies and unsupervised anomaly detection in screening workflows.

From a practical perspective, diffusion-based approaches have demonstrated consistent advantages over earlier generative paradigms such as GANs and VAEs, particularly in terms of training stability, sample diversity, and robustness to mode collapse. These advantages are especially important in clinical contexts, where reliability, interpretability, and anatomical plausibility are critical. Consequently, DMs are now being actively explored across a wide range of medical imaging applications, spanning both 2D and 3D data, multiple imaging physics, and diverse clinical tasks.

In the following section, I illustrate these developments through four representative medical imaging modalities, which are chest X-ray (CXR), magnetic resonance imaging (MRI), computed tomography (CT), and dermoscopic imaging, highlighting the typical datasets, problem formulations, and diffusion-based solutions that have emerged in recent literature.

2.4.1 Chest X-ray

CXR represents the most ubiquitous, cost-effective, and historically significant imaging tool, accounting for over 40% of all diagnostic imaging procedures globally. However, despite its prevalence, the automated interpretation of CXR still faces

challenges: the scarcity of large-scale, expertly annotated datasets; the inherent class imbalance of rare pathologies; and the "black box" nature of DL models that impair clinical trust (Ahmad et al., 2023; Álvarez-Rodríguez et al., 2022). DMs have emerged not merely as an incremental improvement, but as a foundational architectural shift for addressing the current limitations.

Datasets

MIMIC-CXR and CheXpert have emerged as the two most influential benchmarks for CXR data. MIMIC-CXR (Johnson et al., 2019a) comprises over 377,000 chest radiographs paired with free-text radiology reports, making it the largest publicly available CXR dataset with rich linguistic annotation. Its scale and multimodal nature have made it s foundation dataset for T2I generation in the medical domain, supporting recent diffusion-based models. Unlike structured label datasets, MIMIC-CXR enables models to learn fine-grained correspondences between textual descriptions and radiographic patterns, thereby facilitating more expressive conditional generation and enabling prompt-based synthesis paradigms analogous to those in natural image generation. An example of the data sample in MIMIC-CXR is presented in Fig. 2.24.

On the other hand, CheXpert (Irvin et al., 2019) provides a more structured but semantically constrained alternative. It contains over 224,000 chest radiographs annotated with 14 predefined observation labels, automatically extracted using a rule-based natural language processing system. Owing to its standardized label space, CheXpert has been widely adopted for class-conditional diffusion modeling and for evaluating the utility of synthetic images via downstream classification performance. From a generative modeling perspective, this structured supervision simplifies conditional control and stabilizes training, making CheXpert particularly suitable for controlled synthesis and ablation studies.

Text-to-image and report-to-image generation

Text-to-image and report-to-image synthesis are the most mature and intuitive directions of DMs' application in CXR analysis. Although healthcare systems generate

EXAMINATION: CHEST (PA AND LAT)

INDICATION: ___ year old woman with ?pleural effusion // ?pleural effusion

TECHNIQUE: Chest PA and lateral

COMPARISON: ___

FINDINGS:

Cardiac size cannot be evaluated. Large left pleural effusion is new. Small right effusion is new. The upper lungs are clear. Right lower lobe opacities are better seen in prior CT. There is no pneumothorax. There are mild degenerative changes in the thoracic spine

IMPRESSION:

Large left pleural effusion

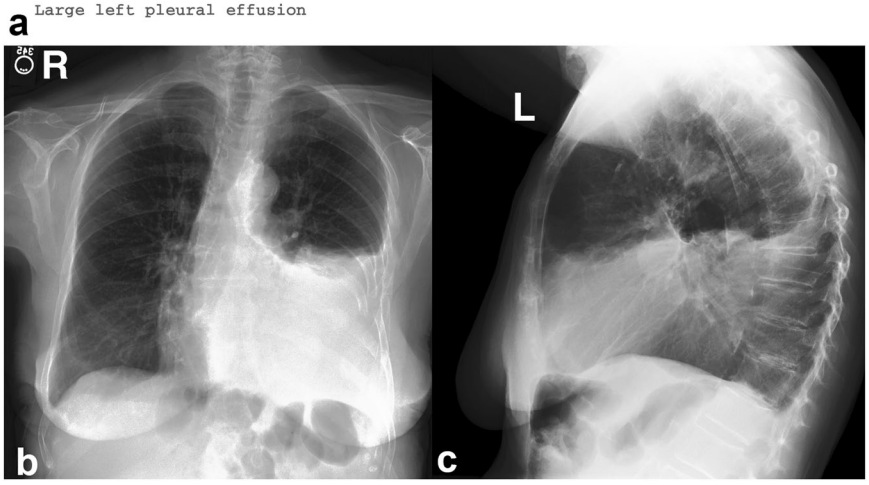


Figure 2.24: Illustration of samples in MIMIC-CXR dataset. (Adapted from Johnson et al. (2019a))

millions of radiographs annually, access to large-scale, well-annotated datasets for specific diseases, especially rare pathologies or pediatric conditions, remains severely limited due to privacy, cost, and annotation constraints. DMs offer a principled mechanism for alleviating this imbalance by synthesizing realistic and semantically controllable radiographs conditioned on clinical text, thereby augmenting data availability while preserving patient privacy.

Unlike earlier GAN-based approaches, DMs exhibit greater stability, higher fidelity, and improved mode coverage, making them particularly well-suited for high-dimensional and structurally constrained CXR data. RoentGen (Chambon et al., 2022a) studies the systematic adaptation of the SD framework to the domain of chest radiography. Built upon LDMs, RoentGen enables efficient training while preserving high-frequency anatomical detail. The model is initialized from a SD checkpoint pretrained on the LAION-5B dataset and subsequently finetuned using paired radiograph–report data from MIMIC-CXR dataset. A critical methodological

contribution of RoentGen lies in its domain adaptation strategy. Rather than training from scratch, RoentGen proposes to use a general-purpose T2I model and adapt its text encoder to radiological language through continued pretraining. This process effectively aligns the semantic embedding space with domain-specific terminology, enabling the model to associate complex phrases with their corresponding visual manifestations. Thereby, RoentGen implicitly constructs a form of visual–semantic lexicon, wherein recurring radiological descriptors become grounded in consistent spatial and textural patterns. As presented in Fig. 2.25, RoentGen demonstrates a strong capacity to generate visually plausible chest radiographs conditioned on free-text prompts, marking a significant step beyond earlier GAN-based synthesis methods.

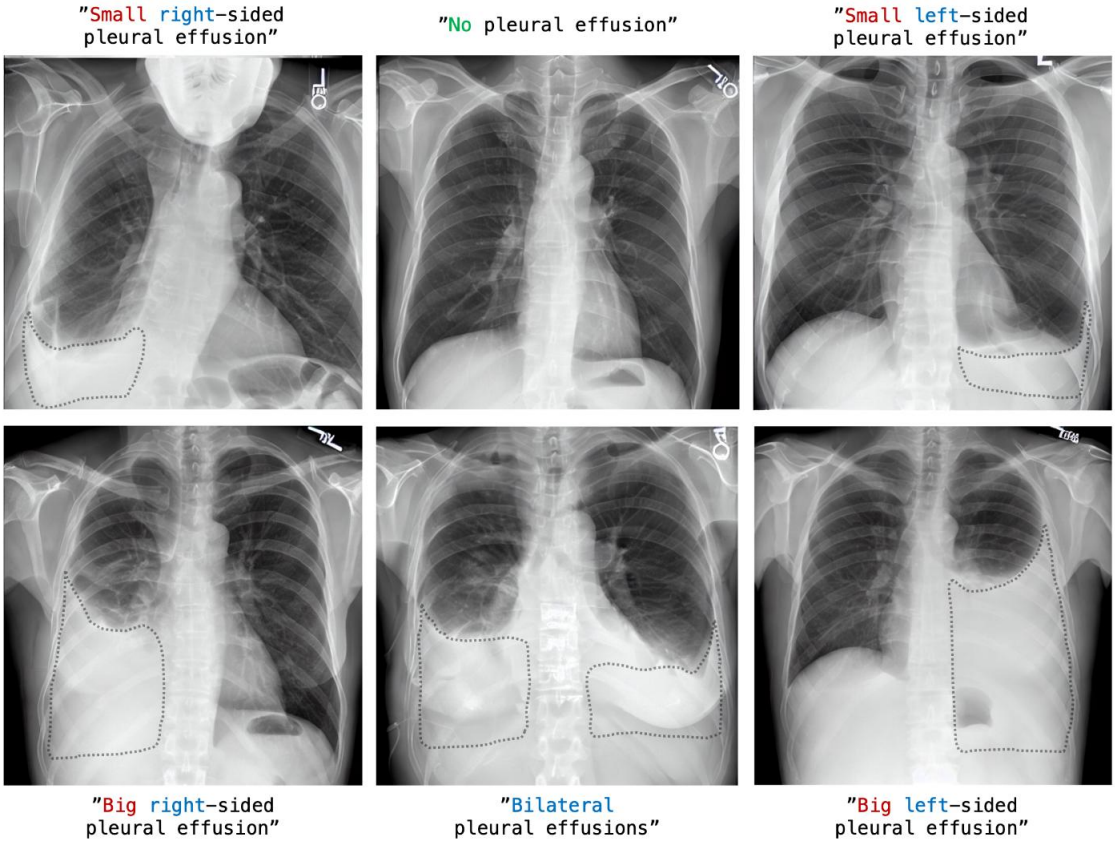


Figure 2.25: Examples of synthetic images from RoentGen. (Adapted from Chambon et al. (2022a))

CheXGen replaces the convolutional U-Net with a Diffusion Transformer (DiT) (Peebles and Xie, 2023). In the context of chest radiography, global reasoning is essential, as clinical interpretations often depend on bilateral symmetry, relative lung

volumes, mediastinal alignment, and cross-regional comparisons that are difficult to capture through localized convolutional operations alone. CheXGen leverages large-scale pretraining to learn a rich joint distribution between visual and textual modalities. The transformer architecture allows every spatial token to attend to all others, enabling coherent global structure from early layers onward. This design choice directly addresses the anatomical inconsistencies observed in earlier CNN-based diffusion models and results in more stable global layouts, particularly for features such as lung symmetry, heart size, and mediastinal shift.

Despite these advances, the evaluation of generative models in CXR imaging remains a contentious issue. The widely adopted Fréchet Inception Distance (FID) has been routinely applied to CXR generation. However, FID relies on feature embeddings extracted from an ImageNet-pretrained Inception-v3 network, which encodes semantics irrelevant to radiographic interpretation. As a result, models may achieve favorable FID scores by producing visually sharp images that nonetheless violate anatomical or physiological plausibility. Recent work has advocated for task-aware and clinically grounded evaluation metrics. One such alternative is the classification accuracy score, which measures the performance of classifiers trained on synthetic data and evaluated on real-world datasets. Improvements in downstream AUROC provide indirect but meaningful evidence that generated images preserve diagnostically relevant features. Human evaluation also remains a resource-intensive but important benchmark. Radiologists' attempts to distinguish real from synthetic images reveal that modern DMs have significantly narrowed the realism gap (Schuit et al., 2025).

Unsupervised anomaly detection

Unsupervised anomaly detection (UAD) has emerged as one of the most compelling applications of DMs in CXR analysis. While large volumes of "normal" imaging data are readily available, abnormal findings are both underrepresented and highly heterogeneous. Diffusion-based UAD methods address this imbalance by learning a generative model exclusively from healthy data, thereby implicitly modeling the distribution of normal anatomy. When presented with an anomalous input during

inference, the model is unable to faithfully reconstruct pathological regions, and the resulting discrepancy between the input and the reconstruction serves as a spatially localized anomaly signal. Unlike discriminative classifiers, which require curated labels for each pathology, diffusion-based UAD reframes anomaly detection as a density estimation problem.

Early diffusion-based UAD methods adopt the reconstruction-based paradigm in which a model trained solely on healthy CXRs attempts to denoise a corrupted version of an input image. The underlying assumption is that pathological patterns, being out-of-distribution, will not be faithfully reconstructed and will therefore appear in the residual between the input and the reconstruction. In practice, however, this assumption is fragile due to the identity preservation problem. When the diffusion process is conditioned directly on the input image, the model may reconstruct a visually plausible but patient-agnostic version of the anatomy, effectively erasing not only pathology but also legitimate inter-subject anatomical variation. This leads to elevated false-positive rates, particularly for patients with uncommon but healthy anatomical traits.

One notable method addressing this issue is the SSDM (Syed et al., 2025), which is built upon diffusion probabilistic modeling. The U-Net-based Gaussian diffusion process is trained exclusively on normal CXR images to learn the distribution of healthy lung anatomy. During training, the model progressively corrupts input images with Gaussian noise and learns to reconstruct them by reversing this process; at inference time, when presented with an image containing pathology, the model's reconstruction tends to remove abnormal structures that do not conform to learned normal priors. Anomaly scores are consequently derived from reconstruction error between the original input and its reconstructed counterpart. SSDM has shown promising performance for lung anomaly detection in CXR, illustrating the feasibility of diffusion models for unsupervised CXR anomaly detection.

While noise modeling improves reconstruction behavior, purely image-based DMs remain fundamentally limited in their ability to account for clinical context. Anatomical normality is not absolute, as it is conditioned on demographic and physiological factors such as age, sex, and body habitus. Image-only models are inherently

blind to distinctions in these aspects and therefore have the risk of misclassifying benign variations as pathological. Diff3M (Kim et al., 2025) addresses this limitation by introducing multimodal conditioning into the diffusion process. Specifically, it integrates structured electronic health record (EHR) information, such as age, sex, and basic clinical indicators, through an Image-EHR Cross-Attention (IECA) mechanism. This design allows the generative model to modulate its internal representation of "normality" based on patient-specific context. By conditioning the diffusion trajectory on both visual and non-visual inputs, Diff3M effectively narrows the distribution of expected healthy appearances for a given individual, thereby reducing false positives driven by demographic variability. A comparison of anomaly detection using Diff3M and DDPM is shown in Fig. 2.26.

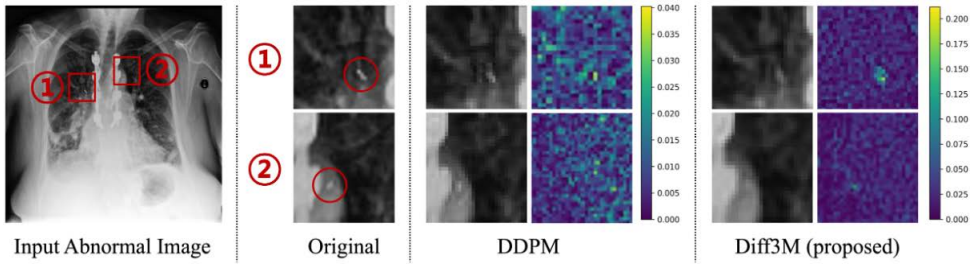


Figure 2.26: Comparison of anomaly detection using Diff3M and DDPM. (Adapted from Kim et al. (2025))

Anatomical decomposition and transformation

Apart from the generation of synthetic images and unsupervised anomaly detection, DMs have been adopted for anatomical transformation and representation refinement in CXR. In contrast to conventional generative applications, these methods operate directly on real clinical images with the goal of selectively suppressing, enhancing, or restructuring specific anatomical components. Such transformations aim to improve diagnostic visibility, mitigate dataset bias, and enhance downstream model robustness.

Bone suppression is a long-standing challenge in chest radiography, as bony structures reduce the detectability of pulmonary nodules and subtle parenchymal abnormalities. Traditional bone suppression techniques require specialized hardware and expose patients to additional radiation. To address this, BS-Diff (Chen et al., 2024c)

introduces a conditional diffusion framework for digital bone suppression in CXR. Instead of directly predicting a bone-suppressed image in a single forward pass, BS-Diff formulates the task as a conditional denoising process. A U-Net-based diffusion model is trained to progressively transform an input radiograph into a soft-tissue-dominant representation, conditioned explicitly on the original image. This conditioning allows the model to preserve patient-specific anatomical structure while selectively attenuating high-contrast bony elements. Training is performed using paired datasets, where bone-suppressed images serve as ground truth. Compared to GAN-based bone suppression approaches, which frequently suffer from ghosting artifacts and discontinuous vascular structures, BS-Diff demonstrates superior structural coherence and anatomical plausibility.

Another class of transformation targets spurious correlations embedded in medical imaging datasets. DL models trained on large-scale CXR collections often exploit shortcuts, such as the presence of medical devices, hospital-specific markers, or acquisition artifacts, instead of learning true pathological features, leading to poor generalization under domain shift and undermining clinical trust. MaskMedPaint (Caron et al., 2021) addresses this problem by leveraging diffusion-based inpainting to explicitly remove confounding visual cues. The core idea is to identify regions corresponding to non-biological or spurious features and replace them with anatomically plausible content generated by a DM. By conditioning the inpainting process on the surrounding tissue context, MaskMedPaint ensures that the filled regions conform to realistic anatomical statistics rather than introducing synthetic artifacts.

DMs have also been extended to structured prediction tasks such as medical image segmentation. DiNO-Diffusion (Jimenez-Perez et al., 2024) is introduced to scale diffusion models via self-supervised pre-training on unlabeled CXR. The generative model is conditioned on learned image embeddings derived from a self-supervised vision transformer DiNO (Caron et al., 2021; Oquab et al., 2023). DiNO-Diffusion demonstrates strong performance in zero-shot segmentation of lung lobes by achieving Dice scores up to 84.4%. This capability emerges from the model's intrinsic understanding of anatomical structure encoded during self-supervised training, allowing segmentation masks to be inferred via iterative attention map aggregation

and semantic alignment of latent representations.

The self-attention mechanisms embedded in modern diffusion backbones capture global context and semantic coherence, properties that can be exploited for delineating anatomical regions from complex radiographic projections. Built upon generalized self-attention diffusion systems, ADZUS (Hamrani and Godavarty, 2025) has been evaluated across multiple medical imaging modalities, including infection region segmentation in CXR. ADZUS underscores the capacity of self-attention-driven diffusion mechanisms to produce segmentation masks in a zero-shot or unsupervised manner, achieving Dice scores in the high 80s across tasks without annotation supervision.

2.4.2 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is one of the most versatile and information-rich medical imaging modalities. Nonetheless, it is constrained by long acquisition times, sensitivity to motion, and complex physical acquisition processes (Zaitsev et al., 2015; Godenschweger et al., 2016). These challenges make MRI an especially challenging domain for diffusion-based generative modeling. Recent advances have demonstrated that DMs can effectively address a wide range of MRI-specific tasks, including reconstruction, cross-modality synthesis, as well as 3D generation and super-resolution. They have become prominent tools for bridging the gap between acquisition efficiency and image quality in modern MRI pipelines.

Datasets

Among the most influential datasets supporting diffusion-based MRI research are fastMRI and BraTS, which collectively cover both low-level reconstruction tasks and high-level semantic modeling. The fastMRI dataset (Zbontar et al., 2018) has become the fundamental benchmark for MRI reconstruction research. It comprises over 1,500 knee scans and approximately 7,000 brain scans acquired using multi-coil MRI systems, providing raw k-space measurements rather than post-processed magnitude images. In MRI, k-space refers to the frequency-domain representation of the acquired signal, from which the final spatial-domain image is reconstructed

through inverse Fourier transformation. This characteristic is particularly significant for diffusion-based reconstruction models, as it allows learning the full forward and inverse physics of MRI acquisition rather than operating solely in image space. fastMRI includes standardized acquisition protocols such as coronal proton-density (PD) sequences for the knee and axial T1-weighted, T2-weighted, and FLAIR-weighted sequences for the brain, captured across both 1.5T and 3T scanners. Examples of axial brain MRI images with different contrasts can be found in Fig. 2.27. Moreover, the availability of raw multi-coil data has enabled the development of diffusion-based reconstruction models that operate either directly in k-space domains, facilitating principled modeling of noise, aliasing, and undersampling artifacts.

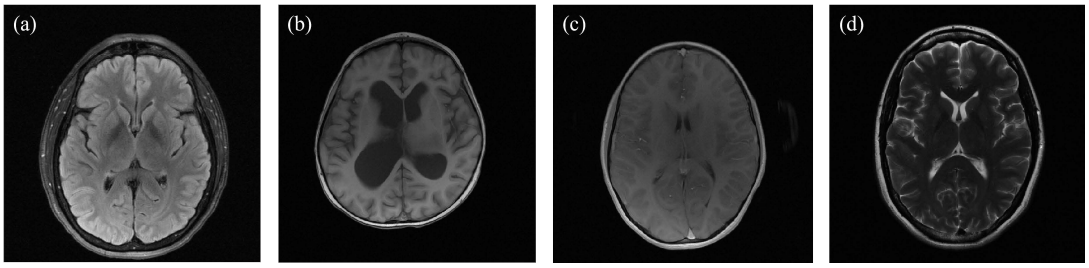


Figure 2.27: Example of axial brain MRI images from fastMRI dataset with different contrasts: (a) FLAIR; (b) T1 weighted; (c) T1 weighted with contrast agent (T1 POST); (d) T2 weighted (Adapted from Zbontar et al. (2018))

In contrast, the BraTS benchmark has played a foundational role in advancing diffusion-based synthesis and anomaly modeling. The BraTS dataset aggregates multi-institutional, multi-scanner MRI volumes of glioma patients and provides voxel-level annotations for tumor subregions, including enhancing tumor, tumor core, and peritumoral edema. Each subject includes co-registered multi-parametric MRI sequences, which are typically T1, contrast-enhanced T1, T2, and FLAIR, making BraTS particularly suitable for studying cross-modality synthesis and conditional generation.

Accelerated Reconstruction

Accelerated MRI reconstruction is one of the most mature and influential domain in MRI imaging analysis. The central challenge in accelerated MRI lies in recov-

ering high-fidelity images from undersampled k-space measurements. Traditional reconstruction approaches rely on compressed sensing and handcrafted regularization, whereas modern DL methods aim to learn powerful priors over anatomical structure. DMs offer a principled probabilistic framework for this problem, enabling reconstruction through iterative sampling from learned image distributions while explicitly incorporating measurement physics.

A central methodological in diffusion-based MRI reconstruction comes from the class of score-based generative models, which extend traditional denoising diffusion frameworks to the Bayesian solution of inverse problems. Song et al. (2021) introduce a general framework in which a DM is trained to estimate the score function of the clean image distribution and is subsequently used as a powerful prior for solving inverse problems such as accelerated MRI. During inference, reconstruction is performed by iteratively sampling from the posterior distribution $p(x|y)$, where the reverse diffusion process is guided jointly by the learned score function and a data-consistency term derived from the physical forward model of MRI acquisition. This formulation enables principled integration of measurement fidelity without requiring paired training data, allowing the model to generalize across different sampling masks, acceleration factors, and acquisition protocols. Subsequent work has extended this paradigm to complex-valued MRI and parallel imaging settings, demonstrating that score-based DMs can achieve competitive or superior reconstruction quality while naturally providing uncertainty estimates (Chung and Ye, 2022).

On the other hand, score-based methods remain computationally expensive due to the iterative nature of reverse-time SDE sampling, PaDIS-MRI (Sanda et al., 2025) introduces a localized diffusion prior tailored to the complex-valued nature of MRI data. Rather than modeling entire images globally, PaDIS-MRI operates on overlapping patches, learning a local prior that captures fine-grained anatomical textures. These patch-level predictions are then globally reconciled through the forward MRI operator, ensuring consistency with k-space measurements. This hybrid strategy allows the model to better handle the spatially structured aliasing artifacts induced by undersampling, while maintaining computational efficiency. Moreover, Res-SRDiff (Safari et al., 2025) reframes the reconstruction problem by

modeling the residual between a low-quality zero-filled reconstruction and the fully sampled target image. This residual exhibits significantly lower complexity than the full image distribution, enabling the diffusion process to converge in a dramatically reduced number of steps. By learning to denoise residuals rather than images, Res-SRDiff achieves high-fidelity reconstructions in as few as four diffusion iterations, representing a substantial improvement in efficiency without sacrificing perceptual or quantitative quality.

In addition, A-DPS (Aali et al., 2024) extends the diffusion framework to scenarios where observations are indirect or corrupted versions of the underlying signal. In this setting, the model is trained not to reconstruct a clean image directly, but to model the distribution of measurements conditioned on latent clean images.

Cross-modality synthesis

MRI protocols usually acquire multiple contrast-weighted sequences, such as T1-weighted, T2-weighted, and FLAIR-weighted images, to capture complementary tissue characteristics. However, in clinical practice, full protocol acquisition is often infeasible due to time constraints, patient motion, cost, or contraindications. Consequently, many real-world datasets are incomplete, motivating the development of cross-modality synthesis methods that can reliably infer missing contrasts from available ones. DMs have emerged as an effective framework for this task, owing to their strong generative capacity and robustness to distributional uncertainty.

Missing modality imputation is an important cross-modality synthesis task, whose objective is to generate a target contrast (e.g., T2-weighted) conditioned on one or more observed modalities (e.g., T1-weighted), so that downstream tasks can proceed without performance degradation. Conditional DMs model the full conditional distribution $p(x_{target}|x_{source})$ through a stochastic denoising process. In CG-DDPM (Hu et al., 2025), the reverse diffusion trajectory is explicitly conditioned on the observed modality, allowing the model to iteratively refine a noisy sample toward a plausible reconstruction consistent with the source image. CG-DDPM has achieved faithful cross-modality synthesis, as shown in Fig. 2.28.

On the other hand, most existing studies in cross-modality synthesis rely on fixed

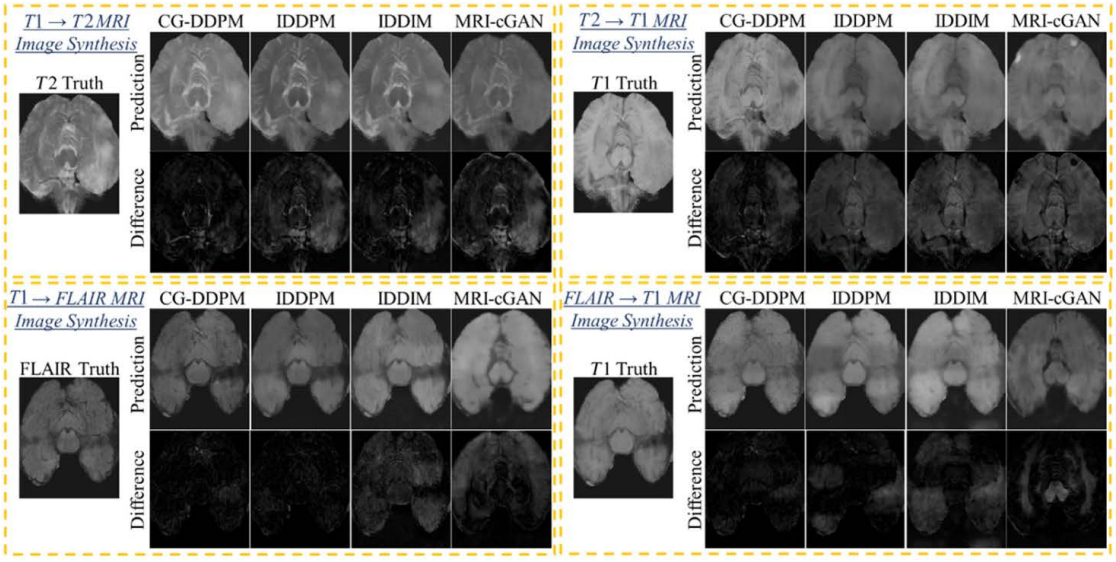


Figure 2.28: Examples of ground truth and synthetic MRIs for cross-modality synthesis task (Adapted from Hu et al. (2025))

modality pairs, requiring separate models for each translation direction (e.g., T1 → T2, T1 → FLAIR). To address this scalability issue, AMM-Diff (Kebaili et al., 2025) proposes a unified architecture capable of handling arbitrary combinations of missing and available modalities. It introduces a shared latent representation in which information from observed modalities is fused via attention-based or frequency-aware modules. During training, random subsets of modalities are masked, forcing the model to learn robust cross-modal correspondences. This design enables a single model to flexibly impute any missing contrast at inference time, significantly improving practicality for clinical deployment where acquisition protocols vary widely.

Additionally, cross-domain translation also draws substantial attention, as acquiring co-registered scans across different scanners, field strengths, or institutions is operationally challenging. SynDiff (Özbey et al., 2023) integrate adversarial learning with stochastic denoising processes, and employs coupled DMs to translate between domains while maintaining anatomical consistency through shared latent representations and diffusion-based regularization.

3D volumetric generation and super-resolution

MRI is inherently volumetric, capturing three-dimensional anatomical structures with rich spatial continuity across slices. However, most early approaches for MRI

analysis have relied on two-dimensional slice-wise processing due to prohibitive memory and computational costs associated with volumetric modeling, resulting in spatial incoherence and loss of contextual anatomical information. DMs have facilitated substantial progress toward true 3D volumetric generation and completion, addressing long-standing limitations of slice-based methods.

Directly modeling a full-resolution 3D MRI volume (e.g., a 256^3 voxel brain scan) requires orders of magnitude more memory and computation than 2D processing. To alleviate this issue, cDPM (Peng et al., 2023) generates MRI slices sequentially, conditioning the generation of slice z on its neighboring slices, thereby enforcing local continuity while retaining tractable computational complexity. To move beyond conditional approximations, PatchDDM (Bieder et al., 2024) decomposes a full 3D volume into overlapping sub-volumes that can be processed independently by a 3D DM. During inference, overlapping predictions are fused through averaging or weighted blending to reconstruct the full volume, substantially reducing memory consumption while preserving local 3D context.

Moreover, 3D DMs act a impactful role in anisotropic super-resolution. PDM (Zhao et al., 2024) framework formulates anisotropic MRI super-resolution as a conditional denoising process, where only the unknown high-resolution components are treated as stochastic variables while the observed low-resolution measurements are preserved throughout sampling. Partial diffusion injects noise selectively along the low-resolution axis, thereby maintaining fidelity to acquired data while hallucinating anatomically plausible details in the missing dimensions. This design significantly reduces computational cost and improves stability, while also preventing over-smoothing commonly observed in deterministic super-resolution networks. By explicitly disentangling observed and unobserved components, partial diffusion provides a principled probabilistic framework for anisotropic MRI enhancement that aligns naturally with the broader diffusion paradigm.

2.4.3 Computed Tomography

Computed tomography (CT) is a fundamental medical imaging modality that utilizes X-ray technology to visualize internal anatomical structures. By rotating an

X-ray source and detector around the patient, the scanner acquires a series of projection data from multiple angles (Seeram, 2010). CT has resulted in better surgery, better diagnosis and treatment of cancer, better treatment after injury and major trauma, better treatment of stroke and better treatment of cardiac conditions (Aali et al., 2024). Despite its widespread clinical adoption, CT image reconstruction remains fundamentally constrained by the physical and practical limitations of data acquisition. Conventional analytical reconstruction methods rely on idealized assumptions and exhibit degraded performance under realistic conditions including sparse-view sampling, low-dose acquisition, and noisy measurements (Szczykutowicz et al., 2022). As a result, CT reconstruction is inherently an ill-posed inverse problem (Demircan-Tureyen and Kamasak, 2017). DMs have rapidly developed to studies these problems, and generate samples that are not only perceptually indistinguishable from high-dose clinical scans but also statistically consistent with the complex noise properties of CT physics.

Datasets

The validity and clinical relevance of DL modeling for CT critically depend on the quality, diversity, and standardization of the datasets used for training and evaluation. CT data present unique challenges arising from their high dynamic range, volumetric structure, and strong physical grounding in X-ray attenuation physics.

In the domain of thoracic imaging, the LIDC-IDRI (Armato III et al., 2011) dataset is the gold standard for lung nodule analysis. Comprising 1,018 thoracic CT scans, it is unique in its rigorous annotation protocol, which addresses the inherent subjectivity of radiological interpretation. Each scan was reviewed independently by four experienced thoracic radiologists, who provided not only segmentation masks but also subjective ratings on malignancy and subtlety. LIDC-IDRI captures the inter-observer variability inherent in clinical practice, requiring algorithms to navigate conflicting expert opinions.

Expanding the scope to acute cardiopulmonary emergencies, the RSNA Pulmonary Embolism (PE) CT dataset (Colak et al., 2021) represents a significant

leap in data volume and complexity. As the largest publicly available annotated dataset of its kind, it contains over 12,000 CT studies and approximately 3 million images collected from five international institutions. The dataset challenges automated systems to go beyond simple binary detection; it requires the characterization of emboli in terms of acuteness (acute vs. chronic) and severity, including the assessment of right ventricular strain. The annotations are generated by a team of subspecialty thoracic radiologists, providing a granular hierarchy of labels at both the examination and image levels. This dataset addresses the critical clinical need for prioritizing positive studies in high-volume emergency settings, where rapid diagnosis of pulmonary embolisms is vital for patient survival. Several examples from the RSNA PE CT dataset are illustrated in Fig. 2.29.

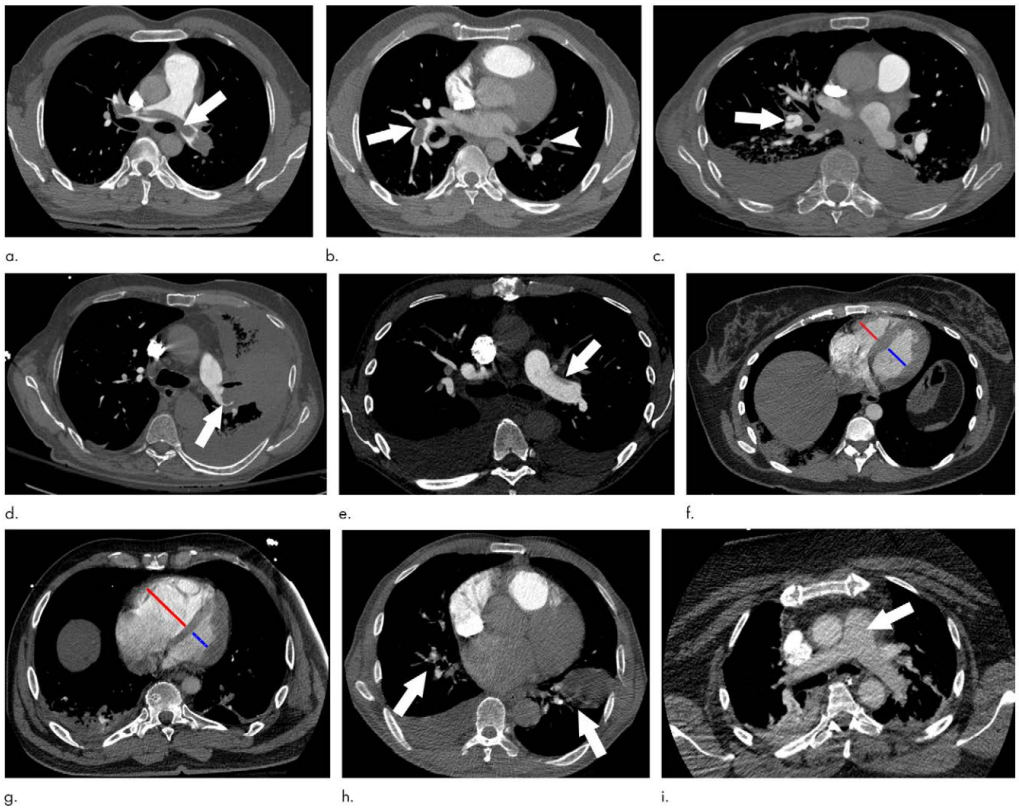


Figure 2.29: Examples from RSNA PE CT dataset: (a) central PE; (b) right-sided PE and left-sided PE; (c) chronic PE; (d) true filling defect not PE; (e) flow artifact; (f) RV/LV ratio; (g) RV/LV ratio; (h) QA-motion; (i) QA-contrast. LV = left ventricle, QA = quality assurance RV = right ventricle (Adapted from Colak et al. (2021))

A similar CT dataset for neurological emergencies is the RSNA Intracranial Hem-

orrhage CT dataset (Flanders et al., 2020). Curated to address the detection and classification of acute brain bleeds, this collection comprises more than 25,000 cranial CT exams containing over 870,000 images. The dataset is annotated by a volunteer cohort of more than 60 neuroradiologists. A key contribution of this dataset is its detailed categorization of hemorrhage subtypes, requiring models to distinguish between intraparenchymal, intraventricular, subdural, extradural, and subarachnoid hemorrhages. By capturing a diverse range of scanner protocols and patient pathologies from multiple centers, the RSNA Intracranial Hemorrhage dataset provides the heterogeneity necessary to train robust models capable of generalizing across different clinical environments, thereby mitigating the domain shift issues often seen in smaller, single-center studies.

Low-dose CT denoising

Reducing radiation exposure remains a central objective in clinical CT. Lowering the tube current or exposure time, however, introduces severe image degradation due to increased quantum noise and electronic noise. Noise in low-dose CT (LDCT) arises from the stochastic nature of X-ray photon emission and detection, resulting in complex, spatially varying noise patterns that significantly impair the visibility of low-contrast anatomical structures and limit diagnostic confidence. Consequently, LDCT denoising has become a critical application domain for DL-based reconstruction methods.

NEED (Gao et al., 2025) proposes a dual-stage framework that explicitly models noise generation and removal across both sinogram and image spaces. In its first stage, NEED employs a shifted Poisson diffusion model to operate directly on pre-log projection data. It incorporates a Poisson-based noise model consistent with photon-counting statistics, allowing more faithful separation of signal and noise at the source of corruption. In the second stage, a Doubly Guided Diffusion process refines the reconstructed image by conditioning on both the low-dose reconstruction and the intermediate output of the sinogram-domain diffusion. A qualitative result is presented in Fig. 2.30. This dual guidance mechanism enables the model to correct physics-induced distortions while simultaneously enforcing anatomical plausibility.

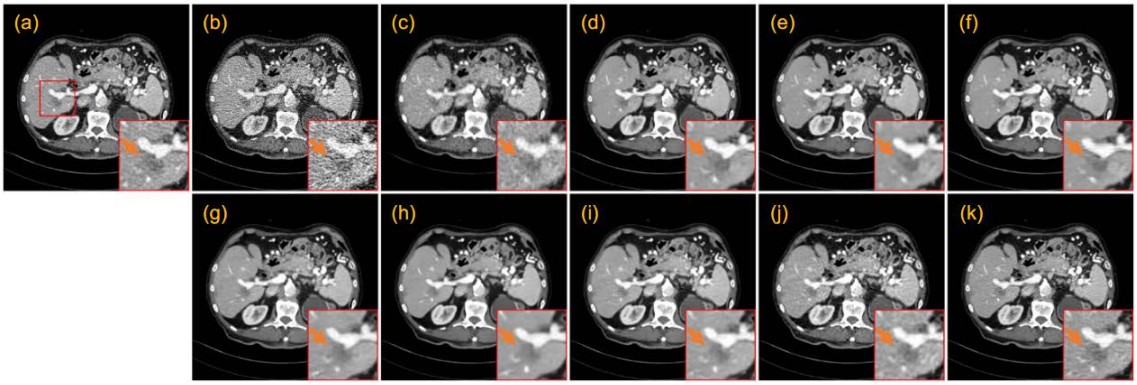


Figure 2.30: Example of qualitative results of a 25% does CT image (a) ground truth; (b) FBP; (c) PWLS; (d) Noise2Noise; (e) Noise2Sim; (f) SSDDNe; (g) DR2; (h) GDP; (i) Dn-Dp; (j) SPDiff; (k) NEED (Adapted from Gao et al. (2025))

In addition, SADiff (Niknejad Mazandarani et al., 2025) introduces a condition-guided diffusion tailored to medical imaging by incorporating a CT Prompt (CTP) module. Rather than relying on textual conditioning as in T2I models, CTP extracts semantic and structural embeddings directly from the input LDCT image. These embeddings guide the denoising trajectory, ensuring that patient-specific anatomical characteristics are preserved throughout the diffusion process. Moreover, SADiff integrates a CT-specific conditioning (CTC) module that injects sinogram-domain priors into the image-domain diffusion network. This cross-domain conditioning allows the model to simultaneously account for measurement physics and spatial anatomy, mitigating the risk of hallucinated structures while maintaining high-frequency fidelity. The resulting framework demonstrates improved robustness across varying dose levels and anatomical regions.

Sparse-view and limited-angle CT reconstruction

Sparse-view and limited-angle CT reconstruction is one of the most ill-posed inverse problems in medical imaging. In sparse-view CT, the number of projection angles is drastically reduced for lower radiation dose or faster acquisition. In limited-angle CT, data are missing over a continuous angular range due to physical or geometric constraints. In both cases, classical analytical reconstruction methods face severe streaking artifacts, loss of structural continuity, and anisotropic resolution degradation.

The reconstruction task is typically expressed as solving the ill-posed inverse problem $y = Ax + \eta$, where A denotes the discrete Radon transform, x is the unknown clean image, and η is the noise. Diffusion posterior sampling (Li et al., 2024) reformulates the reverse diffusion process as a sequence of constrained denoising steps that jointly enforce data fidelity and learned image priors. At each timestep, the update rule is given by:

$$x_{t-1} \leftarrow \text{DiffusionStep}(x_t) - \zeta \nabla_{x_t} \|y - A(\hat{x}_0(x_t))\|^2 \tag{2.42}$$

where $\hat{x}_0(x_t)$ is the estimated clean image at step t derived from Tweedie's formula. The gradient term $\nabla \|y - A(\hat{x}_0)\|^2$ pulls the generation toward the manifold of images that are consistent with the observed sinogram y. This formulation ensures that while the DM explores the learned manifold of anatomically plausible CT images, it remains anchored to physically valid solutions consistent with the projection data. As a result, diffusion posterior sampling significantly mitigates hallucination artifacts, while outperforming classical iterative reconstruction and plug-and-play denoisers in sparse-view regimes.

Furthermore, physics-informed diffusion framework PSDM integrate classical optimization principles directly into the generative process (Han et al., 2024). PSDM embeds the primal–dual hybrid gradient (PDHG) algorithm within the reverse diffusion dynamics, which is a well-established convex optimization method for solving inverse problems with explicit data-fidelity and regularization terms. In PSDM, each denoising step alternates between enforcing consistency with the forward operator A through primal-dual updates and refining image realism via score-based denoising, effectively unifying model-based reconstruction and data-driven priors into a single iterative process. PSDM yields reconstructions with improved anatomical continuity and texture realism, as shown in Fig. 2.31.

3D volumetric generation

The generation of fully CT imaging is challenging, as it is constrained by the curse of dimensionality. A standard diagnostic CT volume with dimensions of $512 \times 512 \times 200$ contains over 50 million voxels, rendering naive diffusion-based generation

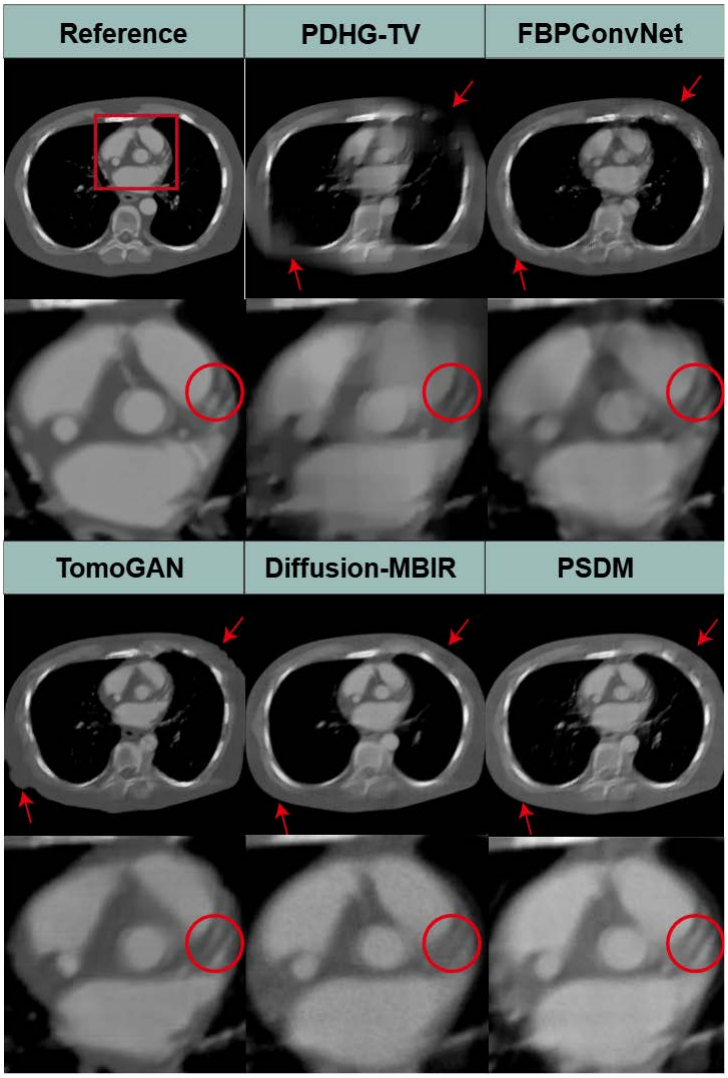


Figure 2.31: Reconstruction results from simulated projections for different methods with a scanning angular range of 90° (Adapted from Han et al. (2024))

computationally infeasible. To address this,

LDM-based 3D generation framworks (Mahdi et al., 2025) compress high-resolution 3D volumes into a lower-dimensional latent representation, achieving several orders of magnitude reduction in memory while preserving semantically meaningful anatomical structure. Diffusion is then performed in compressed latent space, where each denoising step is significantly cheaper and more stable. This paradigm has proven particularly effective in medical imaging, where global anatomical coherence is often more important than pixel-level fidelity.

On the other hand, LDMs introduce an implicit information bottleneck due to compression, which may be undesirable in applications requiring voxel-level fidelity.

Alternative representations have been proposed that trade global compression for structured factorization of the spatial domain. As patch-based DM, PatchDDM (Bieder et al., 2024) decomposes a 3D volume into smaller overlapping sub-volumes augmented with spatial coordinate embeddings. This approach reduces memory complexity, while maintaining anatomical fidelity. A more structurally elegant alternative is the tri-plane representation, inspired by neural radiance fields (NeRFs). In methods such as Blaze3DM and DiffuX2CT, the diffusion model does not generate volumetric voxels directly. Instead, it synthesizes three orthogonal 2D feature planes corresponding to the axial, coronal, and sagittal views. Furthermore, methods based on tri-plane representation are proposed (He et al., 2024; Liu et al., 2024b). In this line, DMs synthesize three orthogonal 2D feature planes corresponding to the axial, coronal, and sagittal views. A lightweight decoder subsequently queries these planes to reconstruct any 3D coordinate via feature interpolation. This factorization reduces computational complexity from cubic $O(N^3)$ to quadratic $O(N^2)$, enabling efficient high-resolution synthesis. Examples of volumetric generation by Blaze3DM (He et al., 2024) on MRI and CT is shown in Fig. 2.32.

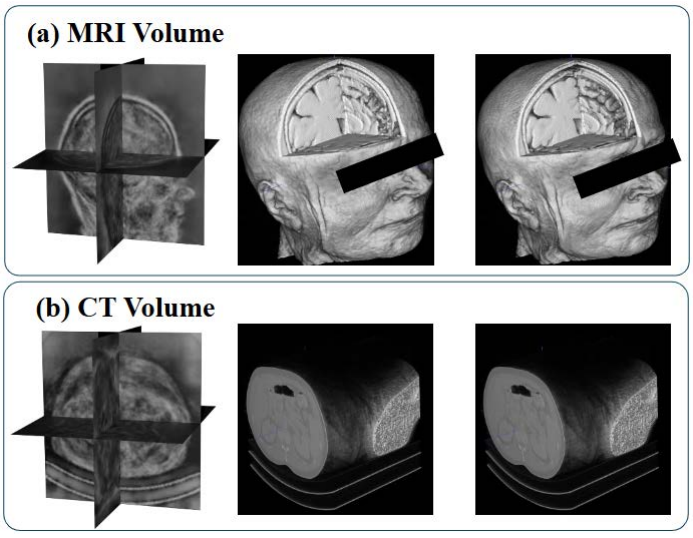


Figure 2.32: Examples of volumetric generation on MRI and CT (Adapted from He et al. (2024))

Additionally, recent work has explored conditioning volumetric DMs on semantic information, enabling controllable and task-specific generation. In CT-RATE (Molino et al., 2025), radiology reports are embedded into a shared semantic space

and used as conditioning signals for diffusion-based generation, allowing users to specify high-level clinical attributes, and synthesize anatomically plausible CT volumes consistent with the textual description. These capabilities open new avenues for rare disease augmentation and scenario-based training.

2.4.4 Dermoscopic Imaging

Dermoscopic imaging is a non-invasive imaging technique widely used in the clinical assessment of pigmented skin lesions, significantly improving diagnostic accuracy for melanoma and other skin cancers compared to visual inspection (Kittler et al., 2002). By revealing subsurface morphological structures such as pigment networks, globules, and vascular patterns, dermoscopic imaging has become a cornerstone of modern dermatological screening and computer-aided diagnosis systems. However, dermoscopic image analysis remains challenging due to substantial intra-class variability, high inter-class similarity, and severe class imbalance (Wang et al., 2023a; Adepu et al., 2023). These issues often limit the generalization ability of discriminative models trained on imbalanced or visually biased datasets.

DMs offer a promising framework to address several of these challenges. By learning the underlying data distribution through iterative denoising, diffusion-based approaches enable high-fidelity image synthesis, controllable data augmentation, and structure-preserving image restoration. As a result, DMs are increasingly being explored as a powerful tool for improving robustness, fairness, and interpretability in dermoscopic image analysis.

Datasets

To support the development and evaluation of learning-based methods for dermoscopic image analysis, several publicly available datasets have been curated over the past decade. Large-scale repositories and carefully curated subsets have played a crucial role in enabling data-driven approaches, including recent diffusion-based methods for synthesis and augmentation. The International Skin Imaging Collaboration (ISIC) Archive (Codella et al., 2019; Gutman et al., 2016; Codella et al., 2018; Rotemberg et al., 2021) is the largest and most comprehensive public repos-

itory for dermoscopic images, designed to support research in skin lesion analysis and melanoma detection. It aggregates data from multiple clinical centers and has served as the foundation for a series of annual ISIC challenges, providing standardized benchmarks for tasks such as lesion classification, segmentation, and attribute detection. The dataset contains tens of thousands of dermoscopic images with varying degrees of annotation, including diagnostic labels, lesion masks, and clinical metadata. Owing to its scale, diversity, and continual expansion, the ISIC Archive has become the primary benchmark for evaluating learning-based methods in dermoscopic imaging. A count of images contained in ISIC 2016-2020 datasets is listed in Table 2.1, and examples of dermoscopic images from ISIC datasets are presented in Fig. 2.33.

Table 2.1: Summary of image numbers in the ISIC 2016-2020 datasets

Dataset	Train	Test	Total
ISIC 2016	900	379	1279
ISIC 2017	2000	600	2600
ISIC 2018	10015	1512	11527
ISIC 2019	25331	8238	33569
ISIC 2020	33126	10982	44108

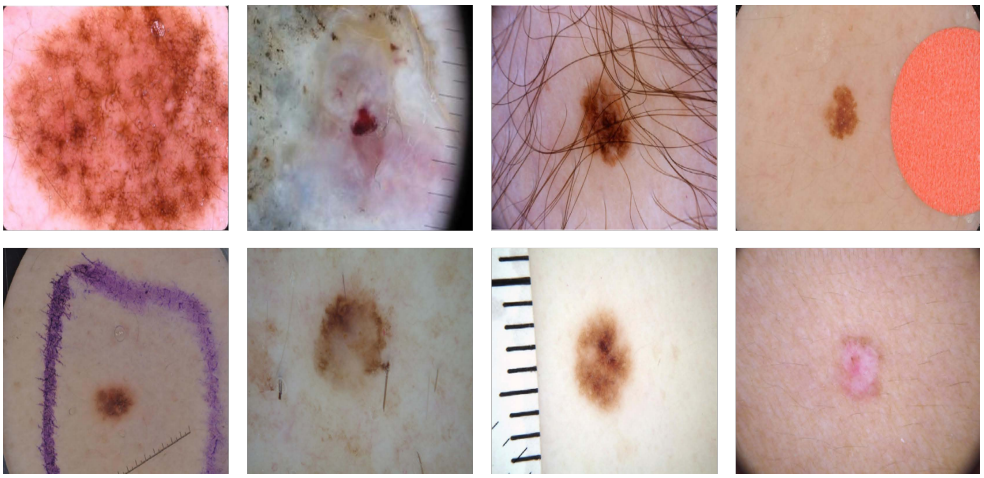


Figure 2.33: Examples of dermoscopic images from the ISIC datasets

Moreover, the HAM10000 dataset (Tschandl et al., 2018) is an independently curated subset of the ISIC Archive designed to provide a well-annotated benchmark for skin lesion classification. It consists of 10,015 dermoscopic images spanning

seven diagnostic categories, with careful curation to reduce noise and annotation ambiguity. Compared to the broader ISIC collection, HAM10000 offers improved label consistency and cleaner image quality, making it particularly suitable for controlled experimental evaluation. As a result, it is widely adopted in studies on data augmentation, class imbalance mitigation, and generative modeling.

Additionally, the PH2 dataset (Mendonça et al., 2013) is a smaller but high-quality dermoscopic dataset consisting of 200 images acquired under controlled conditions, each accompanied by expert annotations and precise lesion segmentation. Although limited in scale, PH2 is frequently used as a benchmark for segmentation accuracy and structural analysis due to its high annotation fidelity and minimal acquisition noise. In the context of image generation research, PH2 is often employed for evaluating fine-grained image reconstruction, lesion boundary preservation, and the effectiveness of generative models in low-data regimes.

Image generation

Class imbalance remains a primary challenge in computational dermatology. Publicly available datasets such as ISIC are inherently skewed toward benign or normal lesions, while clinically critical categories are severely underrepresented. For example, in the ISIC 2018 dataset, benign nevi account for over 65% of samples. This imbalance leads to biased decision boundaries in supervised learning, where classifiers achieve high overall accuracy but demonstrate poor sensitivity to rare but clinically significant malignancies. As a consequence, conventional discriminative models often fail in scenarios where reliability is most critical.

Diffusion-based generative models offer a reliable solution for the class imbalance problem. Derm-T2IM (Farooq et al., 2024) adapts large-scale T2I DM for dermatological image synthesis. It finetunes a pretrained SD backbone using the dermoscopic images paired with textual descriptors, leveraging the strong semantic priors learned from large-scale natural image-text corpora while adapting the visual representations to the medical domain. The model can synthesize clinically plausible yet previously unseen dermoscopic images with specific diagnoses, as presented in Fig. 2.34, effectively expanding the support of the training distribution.

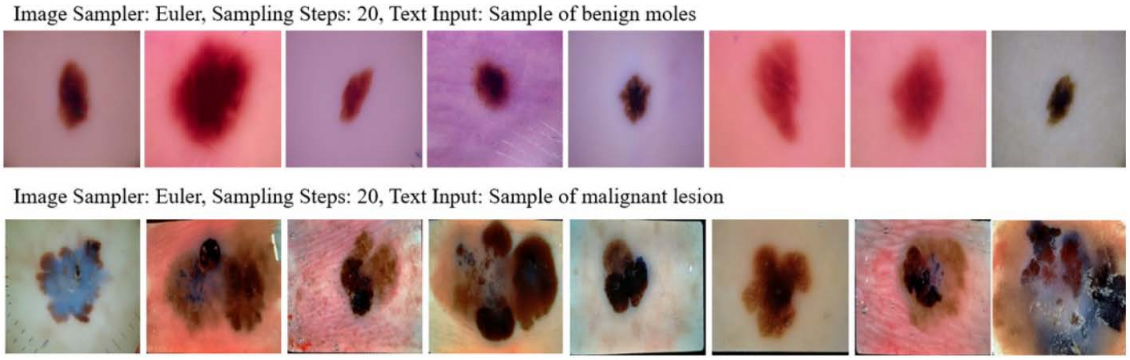


Figure 2.34: Examples of newly generated benign and malignant skin mole data using Derm-T2IM (Adapted from Farooq et al. (2024))

Furthermore, extreme class imbalance poses additional challenges that require models capable of rapid adaptation from very few examples. DAME (Ali et al., 2025) addresses this regime by integrating diffusion-based synthesis with meta-learning principles. It trains a DDPM to generate high-fidelity synthetic samples for minority classes, while optimizing a meta-learning classifier to rapidly adapt to new class distributions. The central hypothesis underlying DAME is that diffusion-generated samples populate intermediate regions of the class manifold rather than duplicating existing observations. This effectively smooths the decision boundary and reduces overfitting to noisy or atypical samples. Experimental results on the HAM10000 dataset demonstrate substantial performance gains.

Beyond dataset augmentation, the explainable image generation also draws attention. Diffusion-based models introduce an explainable paradigm by enabling counterfactual reasoning. CDLC (Varshney et al., 2025) introduces a diffusion-based framework for generating clinically meaningful counterfactual explanations in dermatological image analysis. By leveraging an LDM, CDLC encodes dermoscopic images into a structured latent space and guides the reverse diffusion process toward a target diagnostic class using classifier gradients. Rather than performing naive pixel-level perturbations, CDLC analyzes large collections of factual-counterfactual latent pairs and clusters their difference vectors to identify dominant semantic directions. The method reveals both medically valid features and spurious dataset biases, enabling deeper insight into model behavior. By grounding explanations in realistic image transformations rather than saliency heuristics, CDLC provides a principled

and clinically interpretable mechanism for understanding decision boundaries in dermatological classifiers.

Artifact removal

Dermoscopic images acquired in clinical practice frequently contain a variety of acquisition-induced artifacts, including hairs, ruler markings used for scale reference, air bubbles from immersion fluids, and surgical ink annotations. These artifacts pose a dual challenge for automated analysis. On the one hand, they may partially occlude diagnostically critical structures such as pigment networks, globules, or streaks. On the other hand, they can act as spurious correlates that bias data-driven models, degrading their reliability and clinical trustworthiness.

Artifact removal in dermoscopic imaging is naturally formulated as an inpainting task, where corrupted regions are masked and reconstructed based on the surrounding context. MaskMedPaint (Jin et al., 2024) applies DMs for medical image inpainting, conditioning generation on masked dermatological images and leveraging the DM's learned semantic structure to reconstruct anatomically coherent skin patterns. Crucially, it preserves diagnostically meaningful microstructures, such as pigment networks or rulers, which are often destroyed by traditional techniques. By generating diverse yet realistic restorations, MaskMedPaint also mitigates overfitting to a single plausible completion. Figure 2.35 illustrates that MaskMedPaint can arbitrarily add or remove artifacts like ruler and hair.

Complementing this general-purpose approach, DM-AHR (Benjdira et al., 2024) proposes a task-specific diffusion architecture designed to disentangle hair-like structures from dermoscopic features. DM-AHR learns to differentiate between linear artifacts (e.g., hair shafts) and biologically meaningful linear patterns such as vessels or streaks. Evaluations demonstrate that DM-AHR substantially improves downstream lesion classification accuracy compared to both traditional preprocessing pipelines and GAN-based inpainting, highlighting the value of diffusion models in preserving diagnostically relevant texture while removing confounders.

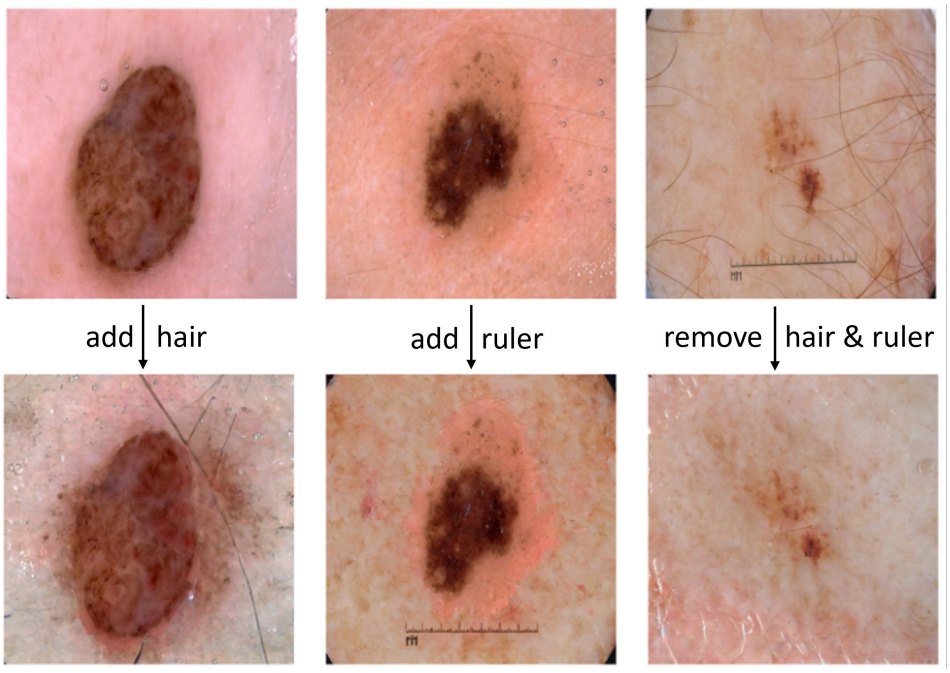


Figure 2.35: Illustration of adding or removing artifacts from dermoscopic images using MaskMedPaint (Adapted from Jin et al. (2024))

Image segmentation

Accurate segmentation of skin lesions is a foundational requirement for automated dermatological diagnosis, as it directly supports the extraction of clinically meaningful descriptors. Conventional architectures, most notably U-Net and its numerous variants, have long dominated this task due to their strong inductive bias for spatial localization. However, these models often struggle in clinically realistic scenarios where lesion boundaries are ill-defined, gradually fading into surrounding healthy tissue. Beyond generation, DMs have also emerged as a new paradigm for dermoscopic imaging segmentation.

Unlike discriminative segmentation networks that produce a single deterministic output, diffusion-based segmentation models learn to generate segmentation masks through an iterative denoising process conditioned on the input image. Building upon this idea, DermoSegDiff (Bozorgpour et al., 2023) incorporates a boundary-aware training objective that assigns higher importance to pixels near lesion edges, encouraging the denoising process to focus on regions of diagnostic uncertainty. By progressively refining boundary structures through the diffusion steps, DermoSegDiff achieves substantially improved boundary accuracy. A qualitative comparison is

presented in Fig. 2.36. It is worth noting that the approach aligns well with clinical reasoning, as diagnostic uncertainty is often concentrated at lesion margins rather than within homogeneous interior regions.

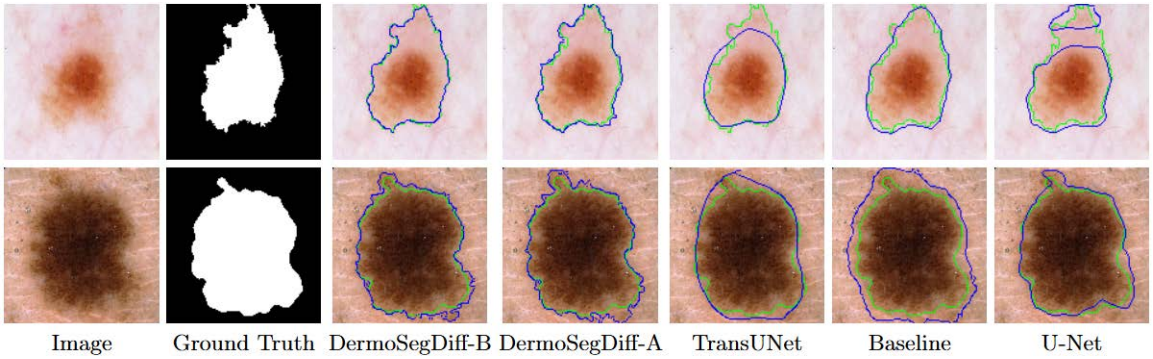


Figure 2.36: Qualitative comparisons of different methods on dermoscopic images (Adapted from Bozorgpour et al. (2023))

A closely related line of work is represented by the MB-Diff (Wang et al., 2023b), which explicitly models boundary uncertainty by treating lesion contours themselves as diffusion targets. MB-Diff learns a probabilistic boundary representation that evolves during the reverse diffusion process. This design proves particularly effective for extremely challenging cases, such as low-contrast or partially occluded lesions, where classical edge-detection or region-growing strategies fail. To further enhance representational capacity, MedSegDiff-V2 (Wu et al., 2024b) integrates Transformer-based modules into the diffusion denoising network. These components enable the model to capture long-range spatial dependencies and global lesion morphology while the diffusion process refines fine-grained boundary details. Compared with transformer-based discriminative models such as TransUNet (Chen et al., 2021), MedSegDiff-V2 demonstrates superior performance in multiple segmentation tasks including dermoscopic imaging, underscoring the benefit of coupling global context modeling with probabilistic generation.

CHAPTER 3

Attention-based Disentanglement of Multiple Concepts for Text-to-Image Customization

Following the comprehensive review of deep visual generative models, attention mechanisms, and controllability strategies in Chapter 2, this chapter presents the first technical investigation of the thesis. Chapter 2 demonstrates that attention plays a central role in governing semantic alignment and spatial structure in T2I DMs, and that attention-based guidance has become a key mechanism for controllable generation. This chapter builds directly upon these theoretical foundations and seeks to translate the understanding of attention mechanisms into a practical framework for fine-grained control in general-purpose image customization.

The primary motivation of this chapter is to address the challenge of learning and controlling multiple novel concepts from limited data, particularly from a single image. While recent customization methods enable DMs to adapt to new concepts, they often suffer from feature fusion and asynchronous learning when multiple concepts are learned simultaneously, resulting in degraded fidelity and poor controllability. This chapter aims to investigate how internal attention signals can be exploited to disentangle multiple concepts, synchronize their learning dynamics, and achieve reliable customization without requiring manual masks or specialized

auxiliary models.

To this end, this chapter introduces *AttenCraft*, an attention-based multi-concept disentanglement framework for T2I customization. By leveraging attention maps to autonomously derive concept-specific masks and by introducing an adaptive sampling strategy guided by attention scores, *AttenCraft* effectively mitigates feature entanglement and imbalanced optimization across concepts. A feature-retaining training scheme further enhances concept isolation and visual fidelity. Beyond demonstrating state-of-the-art performance in multi-concept customization, the insights and techniques developed in this chapter form the methodological basis for subsequent chapters. Furthermore, the attention-driven control mechanisms explored in this chapter directly inspire the controllable generation frameworks for dermoscopic image synthesis and segmentation presented in Chapter 4 and Chapter 5, where attention is further leveraged to guide region-specific and style-specific generation in medical imaging contexts.

3.1 Introduction

Diffusion models have shown exceptional capabilities in generating high-quality and diverse images (Ho et al., 2020; Dhariwal and Nichol, 2021). T2I diffusion models, in particular, display notable proficiency in producing images aligned with natural language prompts (Rombach et al., 2022; Nichol et al., 2021; Ramesh et al., 2022; Gu et al., 2022). However, incorporating new concepts absent from pre-training datasets remains a challenge (Gal et al., 2022). Studies on "customizing" T2I models for generalization to new concepts suggested finetuning pre-trained models using a few or even a single image of the target object, resulting in subject-driven T2I models (Gal et al., 2022; Ruiz et al., 2023a; Kumari et al., 2023; Li et al., 2023b; Jia et al., 2023; Gal et al., 2023a; Arar et al., 2023; Ruiz et al., 2023b; Ma et al., 2023b). In subject-driven T2I learning, the visual representation is mapped to an identifier token [V] via the cross-attention mechanism and is generalized to diverse contexts (Gal et al., 2022). Nonetheless, existing subject-driven T2I models are primarily designed to learn from images containing a single new concept (Kumari

et al., 2023), struggling to learn multiple concepts from one image, as shown in the results of *Custom Diffusion* (*CusDiff*) in Fig. 3.1.

Several studies have explored learning multiple concepts from a single image or localized regions of the image (Avrahami et al., 2023b; Jin et al., 2023; Rahman et al., 2024; Safaee et al., 2023; Zhang et al., 2024). Two main strategies for disentangling multiple concepts have been identified. The first strategy uses masks (Avrahami et al., 2023b; Jin et al., 2023; Rahman et al., 2024; Safaee et al., 2023) to guide cross-attention activation during training, represented by *Break-a-scene* (*BAS*) (Avrahami et al., 2023b); while the second directly adjusts cross-attention to focus on different concepts in the given image, represented by *DisenDiff* (Zhang et al., 2024). However, *BAS* depends on masks provided by specialized segmentation models (e.g., SAM (Kirillov et al., 2023)) or human input, while *DisenDiff* struggles to remove background features from the target concepts. More importantly, two key issues that deteriorate the results of concept disentanglement emerge, as presented in Fig. 3.1. First, baseline models may present feature fusion when learning multiple concepts (e.g., the human haircuts and faces in Fig. 3.1(a)). Second, an asynchronous learning across different concepts happens in baseline models, as reflected by the "corruption" shown in Fig. 3.1(b). The corruption manifests as noisy patches, which indicates overfitting (Wu et al., 2024c) of the corresponding concept. The asynchronous learning can be observed between the single concept and concept group (*DisenDiff*), and between different single concepts (*BAS*), depending on specific model settings. A detailed analysis will be presented in Chapter 3.3.2.

In this study, I propose *AttenCraft*, a novel method for disentangling multiple concepts from a single image in subject-driven T2I generation. Specifically, I adopt the mask-based strategy for disentanglement, using self-attention and cross-attention maps to generate accurate masks for each target concept in a single step, without the need for specialized segmentation models or human input. These masks guide cross-attention activation for disentanglement during training. Aligning the cross-attention map of the identifier token [V] with the corresponding mask establishes an explicit connection between [V] and the visual representation of the target concept. I also investigate the relationship between feature acquisition and the

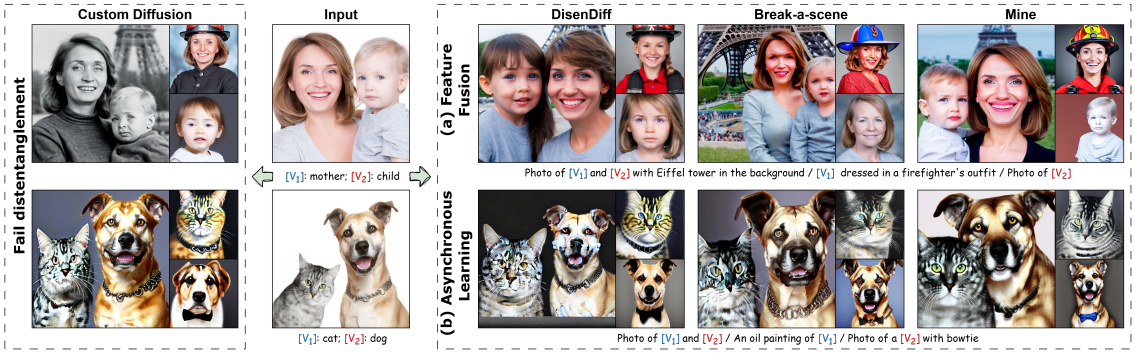


Figure 3.1: I propose *AttenCraft*, an optimized method for disentangling multiple concepts in a single image. Baseline models present two key issues: (a) feature fusion; (b) asynchronous learning. My method significantly mitigates these issues and realizes robust concept disentanglement and feature learning.

initialization of [V], proposing an adaptive algorithm that automatically estimates the sampling ratio of multiple concepts based on cross-attention scores. This approach mitigates asynchronous learning and enhances learning quality. Furthermore, I demonstrate that back-propagating reconstruction loss during multiple-concept sampling is a primary cause of feature fusion. Thus, I optimize the training framework by introducing different loss functions for sampled subsets with varying sizes.

3.2 Related Work

3.2.1 Diffusion models and T2I customization

By utilizing pre-trained text encoders (Vaswani et al., 2017; Ramesh et al., 2021), diffusion models implement the T2I diffusion model in pixel space under classifier-free guidance (Ho and Salimans, 2022; Saharia et al., 2022; Nichol et al., 2021). SD (Rombach et al., 2022) trains the denoising U-Net (Ronneberger et al., 2015) in latent space by applying a VAE (Kingma and Welling, 2013) and the text encoder of Contrastive Language-Image Pre-training CLIP (Radford et al., 2021) model. Furthermore, subject-driven T2I models (Gal et al., 2022; Ruiz et al., 2023a) learn a new concept from several images and reverse to an identifier token [V]. In addition, parameter-efficient tuning (PEFT) (Cao et al., 2025) is employed to minimize training time by utilizing a smaller set of trainable parameters. These include cross-

attention layers (Kumari et al., 2023; Zhang et al., 2024, 2023c; Cai et al., 2024), Low-rank Adaptation (LoRA (Hu et al., 2021)) parameters (Ruiz et al., 2023b; Gu et al., 2024; Chen et al., 2023a; Yang et al., 2024), and supplementary components such as an encoder, adapter, or weight offset (Gal et al., 2023a; Hao et al., 2023; Liu et al., 2023b). Moreover, some studies pre-train a universal encoder capable of directly encoding input images (Li et al., 2023b; Ma et al., 2023b; Shi et al., 2023; Wei et al., 2023; Chen et al., 2024d). However, the majority of subject-driven T2I models focus on input images containing a single concept, neglecting the exploration of extracting multiple concepts from a single image.

3.2.2 Application of attention in diffusion models

The Attention mechanism manipulates feature dependencies during T2I generation. Guided by cross-attention, pre-trained diffusion models exhibit superior semantic alignment with provided text prompts (Chefer et al., 2023; Feng et al., 2022; Wang et al., 2023c; Phung et al., 2023), achieve image editing (Hertz et al., 2022; Nguyen et al., 2024a), and provide positional control (Liu et al., 2023b; Phung et al., 2023; Ma et al., 2023c; He et al., 2023; Chen et al., 2023b). Moreover, cross-attention guidance is applied during model training to eliminate background interference or concentrate on specific regions in input images using provided masks (Ma et al., 2023b; Avrahami et al., 2023b; Safaee et al., 2023; Chen et al., 2023a; Hao et al., 2023; Wei et al., 2023; Shentu et al., 2024). Meanwhile, self-attention can promote subject consistency across different contexts (Tewel et al., 2024) or facilitate subject swaps while preserving style consistency (Jeong et al., 2024b). Furthermore, the self-attention and cross-attention maps are applied to achieve unsupervised segmentation (Tian et al., 2023) and augmentation of the segmentation datasets (Wu et al., 2023a; Nguyen et al., 2024b; Marcos-Manchón et al., 2024).

3.2.3 Disentangling multiple concepts from a single image

BAS disentangles multiple concepts from a single image by applying masks provided by users or specialized segmentation models (Kirillov et al., 2023) to guide cross-

attention activation. Meanwhile, Safaee et al. (2023) adopt automatically identified masks to learn a given concept and apply them to edit other images. Jin et al. (2023) apply a fixed threshold on the cross-attention maps to obtain the mask. Furthermore, Rahman et al. (2024) utilize dense conditional random field (CRF) (Krähenbühl and Koltun, 2011), and Hao et al. (2023) applies Otsu thresholding (Otsu et al., 1975), to obtain masks from cross-attention maps. However, these automatic masks are typically coarse (Jin et al., 2023; Rahman et al., 2024), time-consuming (Safaee et al., 2023), and often fail to separate different concepts (Hao et al., 2023). *DisenDiff* (Zhang et al., 2024) calibrates cross-attention to encourage the model to separate its attention and achieve disentanglement without masks, but fails to exclude the background. My proposed approach efficiently disentangles multiple concepts and backgrounds from a single input image using self-generated accurate masks guided by the attention mechanism.

3.3 Proposed Method

In this section, I carefully introduce my method, which includes mask auto-creation guided by attention maps, adaptive estimation of sampling ratios of different concepts, and a dedicated training framework to prevent feature fusion across concepts. An illustration of my method is presented in Fig. 3.2. Algorithm 1 summarizes the overall pipeline, including the pre-processing stage for mask creation and adaptive sampling ratio estimation, and the feature-retaining training stage for multi-concept disentanglement.

3.3.1 Attention-guided mask creation

The U-Net in the SD incorporates self-attention and cross-attention layers to capture the dependencies within the input data (Rombach et al., 2022; Vaswani et al., 2017). The self-attention layers capture the global attention within the image while the cross-attention layers learn to attend between the image and text prompts. The cross-attention map A_C and self-attention map A_S can be calculated as follows:

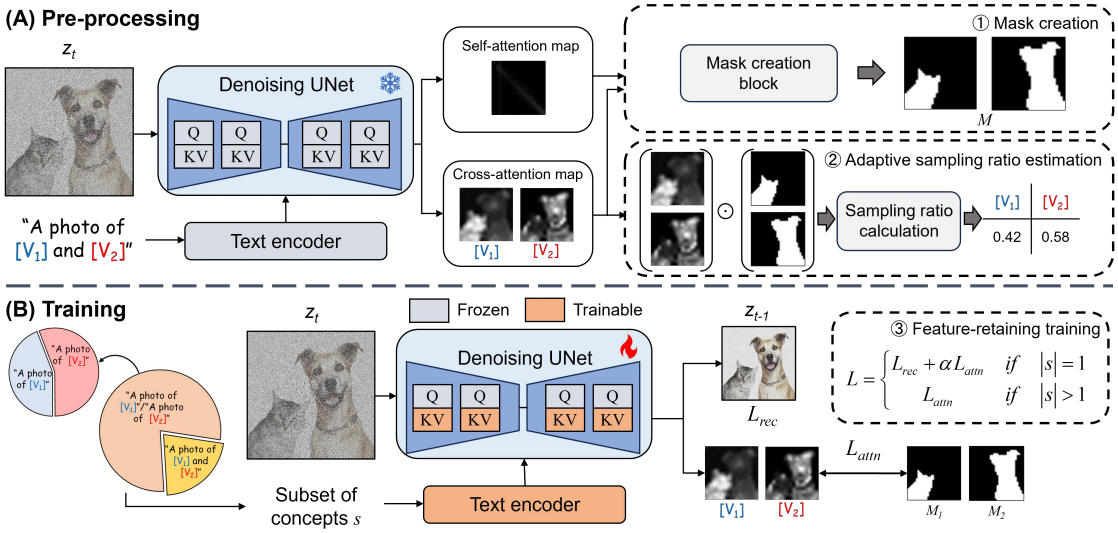


Figure 3.2: Method overview. Given an image with multiple concepts, within a few steps in the pre-processing stage, I create accurate masks for each concept and adaptively estimate the sampling ratio for multiple concepts to enhance learning synchronicity. I also propose an optimized training framework by introducing different loss functions for sampled subsets of varying sizes to prevent feature fusion.

$$A_C = softmax\left(Q_I K_T^\top / \sqrt{d}\right) \tag{3.1a}$$

$$A_S = softmax\left(Q_I K_I^\top / \sqrt{d}\right) \tag{3.1b}$$

where Q_I, K_I, K_T are the query matrix, key matrix of z_t, and key matrix of $\tau_\theta(y)$, respectively.

The cross-attention map A_C outlines the location and shape of the target concept. However, it often displays coarse granularity and noise, leading to two main challenges in mask creation: (1) strong attention activation is shown within the target region, but weak activation occurs in other areas; (2) attention activation is unevenly distributed, leading to an incomplete representation, as shown in Fig. 3.3. To address the first challenge, I apply *Cross-attention suppression* (Zhang et al., 2024) following the left part of Eq. (3.2):

$$\hat{A}_C = (A_C)^v, A_C^S = \hat{A}_C \otimes (A_S)^\tau \tag{3.2}$$

The activation values of the attention map, generated through a *Softmax* oper-

Algorithm 1 Overall pipeline of *AttenCraft*

Require: Input image x containing multiple concepts $\{c_1, c_2, \ldots, c_N\}$; identifier tokens $\{[V_1], [V_2], \ldots, [V_N]\}$; pretrained SD model
Ensure: Finetuned SD model for disentangled multi-concept customization

Stage I: Pre-processing
Encode the input image and the prompt containing all identifier tokens
Run one denoising step and extract attention maps
for each concept c_i **do**
    Generate mask M_i by refining cross-attention with self-attention and suppressing non-target regions
    Compute attention score S_i from the masked responses across selected timesteps
end for
Convert $\{S_i\}$ into adaptive single-concept sampling ratios $\{r_i\}$
Stage II: Feature-retaining training
for each training step **do**
    Sample a concept subset s using $\{r_i\}$ for single-concept sampling and a preset proportion for multi-concept sampling
    Construct the prompt y_s and mask set M_s for s
    Run the denoising UNet conditioned on y_s
    Compute the cross-attention alignment loss L_{attn}
    if $|s| = 1$ **then**
        Compute L_{rec} and optimize with $L = L_{\text{rec}} + \alpha L_{\text{attn}}$
    else
        Optimize with $L = L_{\text{attn}}$
    end if
end for
return finetuned SD model

ation, range from 0 to 1. Consequently, element-wise exponentiation of A_C by υ can reduce weak activation in non-target regions but amplifies uneven activation. To address this, I use *Self-attention enhancement* (Nguyen et al., 2024b), which multiplies $\hat{A}_C$ by A_S^τ to enhance the smoothness and precision of $\hat{A}_C$, as depicted in the right part of Eq. (3.2). A_S captures pairwise correlations among patches in z_t, allowing attention activation to spread to related regions while reducing activation elsewhere. Similarly, element-wise exponentiation of A_S by τ decreases correlations between patches of different concepts. With A_C^S, I observe that the attention activation of different tokens emphasizes distinct regions in the attention map. Thus, masks can be inferred from activation differences. For the target concept i, I compute the maximum difference between its processed attention map $A_{C_i}^S$ and that of

another concept j ($A^S_{C_j}$, $i \neq j$), setting the mask value M_i to 1 if it exceeds a preset threshold γ. I term this process *Delta masking*, and it is defined by the following:

$$M_i = \begin{cases} True & if \max(A^S_{C_i} - A^S_{C_j}) > \gamma, i \neq j \\ False & Otherwise \end{cases} \tag{3.3}$$

Attention-guided mask creation is performed within the mask creation block shown in Fig. 3.2. This process requires only a single step, where the noisy latent z_t is sampled from $t \in [0, 300]$ in the DDPM noise schedule (Ho et al., 2020) since z_t retains finer semantic details at this stage (Nguyen et al., 2024a). Details of the mask creation process are depicted in Fig. 3.3.

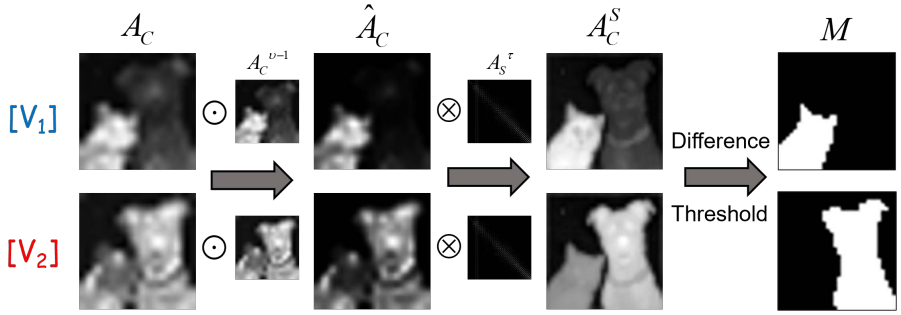


Figure 3.3: Process of attention-guided mask creation. By applying the cross-attention and self-attention maps, precise masks can be created without specialized models or human inputs.

3.3.2 Adaptive sampling ratio estimation based on attention scores

The issue of asynchronous learning is illustrated in Fig. 3.1. Wu et al. (2024c) note that the corruption is caused by a narrowed learning distribution when applying few-shot or one-shot learning, creating a limited window between underfitting and overfitting. The adverse effects of this limited window are pronounced when learning multiple concepts, as the learning windows for different concepts may not align perfectly, leading to asynchronous learning. Baseline models use a fixed sampling scheme during training, which cannot adapt to varied inputs. Specifically, *Disen-Diff* utilizes a consistent text prompt encompassing all target concepts throughout

the training process, resulting in the overfitting of the concept group when single concepts are properly learned; while *BAS* employs a union sampling scheme that randomly selects a subset of multiple target concepts to form the text prompt, achieving comparatively better learning synchronicity than *DisenDiff*. However, the sampling ratio for each single concept in *BAS* remains identical, which still raises the asynchronous learning issue since the learning steps required for different concepts vary. Thus, an optimized sampling scheme with an adaptive sampling ratio for different concepts is required.

Identifier token initialization I first investigate the relationship between feature acquisition and identifier token initialization through a preliminary experiment, where *BAS* is deployed to learn multiple concepts from 10 datasets (Zhang et al., 2024), each containing two concepts, over 1000 training steps. Before training, identifier tokens $[V_1]$ and $[V_2]$ are initialized by text embeddings of existing tokens. For each dataset, I apply three token initialization patterns using text embeddings of the precise class (dubbed as P) and the general category (dubbed as G), resulting in P-P, P-G, and G-P. The CLIP-I scores of the generated images are assessed to reflect the feature acquisition of each concept (detailed in Chapter 3.4.1). Note that this experiment assesses single-concept generation, meaning that the initialization pattern varies relative to concepts within the same dataset [1]. The variation in CLIP-I scores over training steps is shown in Fig. 3.4(a). The model begins at a higher initial point when [V] is initialized with a precise class P but tends to degrade after 300 steps. Conversely, when initialized with the general category G, the model starts lower but continues to learn until the end of training. These results indicate that the initialization of [V] significantly impacts the feature acquisition.

Attention activation and sampling ratio The difference between P and G for learning lies in their semantic connection to the target concepts, which can be reflected in the cross-attention scores. To provide supporting evidence for this inter-

[1]For example, in the "cat & dog" dataset, I set the triplet "cat-dog", "cat-animal", and "animal-dog" as different initialization patterns. These correspond to P-P, P-G, and G-P when evaluating the "cat", while P-P, G-P, and P-G when evaluating the "dog".

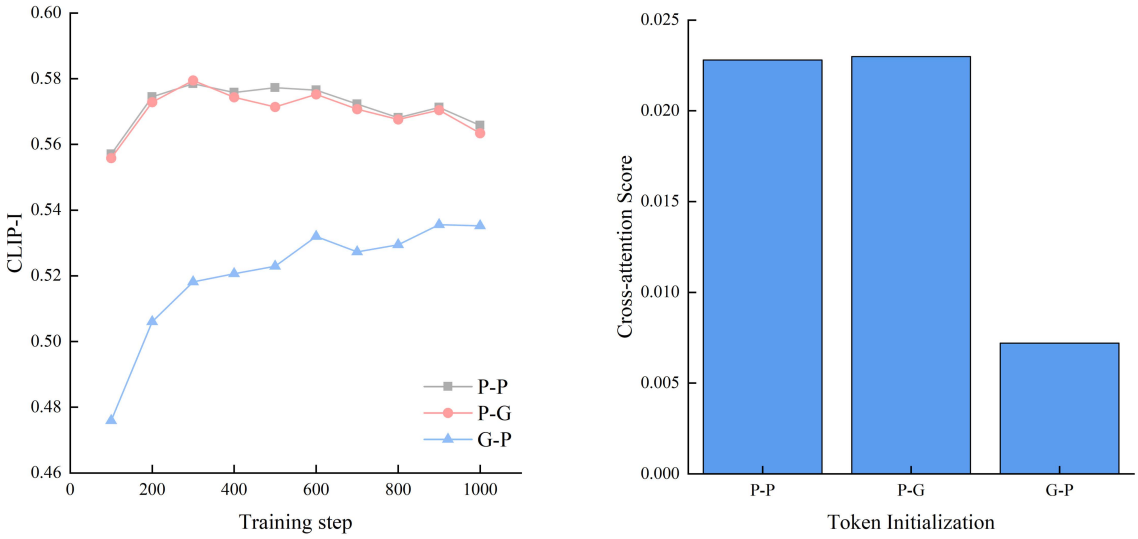


Figure 3.4: Results of the token initialization experiment. (a) Variation of single-concept CLIP-I scores with training step; (b) The highest cross-attention score of [V] concerning different initialization patterns.

pretation, I extract the highest activation score from the cross-attention map of each [V], with the result presented in Fig. 3.4(b). The results show that cross-attention scores are significantly higher when a specific concept (e.g., [V_1]) is initialized with *P*, while the initialization of the other concept [V_2] has negligible effects on [V_1]'s activation score. This observation is further supported by the results in Fig. 3.4(a), where the CLIP-I scores for *P*-*G* and *P*-*P* show only marginal divergence. In conclusion, an identifier token [V] initialized with a less semantically rich embedding requires more steps for feature learning and should be assigned a larger sample ratio to achieve more balanced and synchronized feature acquisition, where the implicit semantic connection can be explicitly reflected by cross-attention scores.

Adaptive sampling ratio estimation I propose an attention-based algorithm for an adaptive sampling ratio estimation, grounded in experimental results. Specifically, I first apply self-created masks (see Chapter 3.3.1) on cross-attention maps following $A_M = A_C \odot M$ to eliminate the noise outside the target region, and then extract the highest activation score S from the masked maps. To mitigate contingency, I average the n highest activation scores across m denoising timesteps, as expressed in:

$$S = \frac{1}{m} \sum_{t \in \mathbb{T}} \frac{1}{n} \sum_{i=1}^{n} \max A_{M_t}^{(k)} \tag{3.4}$$

where $\max A_{M_t}^{(k)}$ denotes the k-th maximum element in A_{M_t} from timestep t. $\mathbb{T}$ is a set of t, and has $m = |\mathbb{T}|$. With N concepts, I normalize the highest activation score S_i of each [$\mathrm{V_i}$] by $\bar{S}_i = S_i / \sum_{j=1}^{N} S_j$, and apply a *Softmax* operation to $\bar{S}_i$ to obtain the sampling ratio r_i:

$$r_i = 1 - \frac{e^{\bar{S}_i}}{\sum_{j=1}^{N} e^{\bar{S}_j}} \tag{3.5}$$

Since the initialization of [V] induces differences in attention activation and requires varying training steps for each concept, the proposed adaptive sampling ratio, r_i, can appropriately adjust the sampling frequency based on activation scores, thereby improving synchronicity.

3.3.3 Feature-retaining training framework

The goal of multiple-concept disentanglement is to learn multiple concepts from a single image and sample individual concepts or concept groups with minimal distortion. By applying a mask for each concept, multiple concepts can be disentangled through the combination of a masked reconstruction loss and a cross-attention loss, expressed as:

$$L_{rec} = \mathbb{E}_{z, y_s, t, \varepsilon} \left[\| (\varepsilon - \varepsilon_\theta (z_t, t, \tau_\theta(y_s))) \odot M_s \|_2^2 \right] \tag{3.6}$$

$$L_{attn} = \frac{1}{|s|} \sum_{i \in s} \| A_C (v_i, z_t) - M_i \|_2^2 \tag{3.7}$$

where y_s and M_s are text prompts and masks for a sampled subset s, and $A_C (v_i, z_t)$ denotes the cross-attention map between [$\mathrm{V_i}$] in s and z_t. $M_i \in M_s$ is the corresponding mask. L_{rec} promotes the model to learn features of target concepts, and L_{attn} helps disentangle concepts (Avrahami et al., 2023b).

Nevertheless, when $|s| > 1$, the back-propagation of L_{rec}, which contains features from multiple concepts, may induce feature fusion. Therefore, I propose an

optimized feature-retaining framework for multiple-concept disentanglement by introducing different training objectives for s of varying sizes. Concretely, for $|s| = 1$, both L_{rec} and L_{attn} are back-propagated to jointly learn visual features and establish explicit connections between [V] and the visual features. In contrast, only L_{attn} is back-propagated when $|s| > 1$, preventing feature fusion and enforcing the model to disentangle the cross-attention of multiple concepts while learning their joint presence. The general loss function is formulated as follows:

$$L = \begin{cases} L_{rec} + \alpha L_{attn} & if |s| = 1 \\ L_{attn} & if |s| > 1 \end{cases} \tag{3.8}$$

where α is a scaling coefficient. Moreover, I design a hyperparameter ω as the proportion of multiple-concept sampling steps (i.e., $|s| > 1$), and apply the adaptive sampling ratio of each concept (see Chapter 3.3.1) when $|s| = 1$ to facilitate synchronized feature learning across different concepts. This forms the training pipeline of *AttenCraft*, as shown at the bottom of Fig. 3.2.

3.4 Experiments

3.4.1 Experimental settings

Datasets and baseline. I conduct experiments on 16 datasets across various categories, including human, animal, and object. I collect 10 datasets with relatively simple backgrounds from *DisenDiff* (Zhang et al., 2024). I also synthesize 6 datasets using *Gen4Gen* (Yeh et al., 2024), which combines multiple personalized concepts into complex backgrounds sourced from copyright-free platforms [2], where the concepts are curated from the *DreamBooth* (Ruiz et al., 2023a) and *CustomConcept101* (Kumari et al., 2023). I compare my method with *BAS* (Avrahami et al., 2023b) and *DisenDiff* (Zhang et al., 2024). Additionally, I implement *CusDiff* (Kumari et al., 2023) to demonstrate the disentanglement capability of general subject-driven models.

[2] `https://unsplash.com`

Evaluation metrics. Following baseline models, I calculate the CLIP-I and CLIP-T scores to assess image fidelity and prompt fidelity. Specifically, CLIP-I represents the cosine similarity between the CLIP-ViT-L/14 embeddings of generated and input images, while CLIP-T measures that between generated images and text prompts. In addition, I calculate the DINO score, which is the cosine similarity between the ViT-B/16 DINO-V2 embeddings of the generated and input images, to reveal how much the model preserves the concept identity. Depending on the dataset and concept evaluation scope, CLIP-I and DINO scores use different target references: *DisenDiff* datasets uses cropped input images for concept subsets and the original input for all concepts, while *Gen4Gen* datasets uses original single-concept images for subsets and the generated composite input image for all concepts. Moreover, I evaluate the learning synchronicity using CLIP-I-sync, which is the absolute difference between CLIP-I scores of single concepts in the same dataset. For each dataset, I prepare 10 text prompts for single concepts and the concept group, respectively. I generate 10 images for each text prompt using 50 steps of the PNDM scheduler (Liu et al., 2022) with a guidance scale of 7.5, resulting in an evaluation set consisting of 300 images.

Implementation details. I use SD v2.1 trained on the LAION-5B dataset (Schuhmann et al., 2022) as the base model. I initialize each identifier token [V] with the text embedding of the corresponding class name. I extract cross-attention and self-attention maps from the attention layers, with dimensions of 16×16 and 32×32, respectively. These maps contain abundant semantic and visual information (Hertz et al., 2022; Nguyen et al., 2024b). I set the powers υ and τ to 2 and 4, respectively (Zhang et al., 2024; Nguyen et al., 2024b). The threshold γ for *Delta masking* is empirically set to 0.1. An illustration of the initialized masks is provided in the supplement. When extracting attention scores as described in Eq. (3.4), I set $n = 5$ and $\mathbb{T} = \{0, 20, 40, 60, 80\}$. Moreover, I set the scaling coefficient $\alpha = 0.01$, and the ratio $\omega = 0.3$. All experiments are conducted on an NVIDIA A100 GPU with a single input image, a batch size of 1, and a learning rate of 1×10^{-4} for 300 steps. To reduce computational costs, only the W_k and W_v matrices in the cross-attention

layers are trained (Kumari et al., 2023). Implementation details of baseline models are provided in the supplement.

3.4.2 Qualitative comparisons

I present a qualitative comparison between my method and baseline models in Fig. 3.5 and Fig. 3.6. Figure 3.5 presents the generated images of single concepts and concept groups from different datasets to illustrate the models' performance in disentangling multiple concepts, and the presence of feature fusion. Upon examination, *CusDiff* struggles to disentangle multiple concepts, as the generated images for single concepts show distinct features from the input images. The concept group images generated by *CusDiff* present all target concepts, but the feature fusion can be spotted from the color of the concepts. On the other hand, *DisenDiff* and *BAS* present disentangling capability, but the problems of feature fusion still stand out. *DisenDiff* shows blended features in both single concept and concept group images (e.g., color of the bird, dog, and car for single concept; necklace of the cat and dog, color of the horse and dog for concept group). Also, both *CusDiff* and *DisenDiff* present background features in the "horse & dog" dataset, as the grass appears with the target concepts, indicating that the model fails to detach background features from the target concept. While *BAS* exhibits fewer blended features than *DisenDiff*, the feature fusion can still be observed (e.g., color of the car for the single concept; ear shape of the dog, necklace and color of the cat and dog for concept group). In contrast, my method shows clear disentanglement across multiple concepts and background information, and the features from each target concept are well-retained without blending and fusion.

Furthermore, I analyze the learning synchronicity across multiple concepts of models capable of disentangling and learning them separately. Figure 3.6 presents examples of image triplets consisting of single concept and concept group images generated by the model undergoing the same training step for a fair comparison of learning synchronicity. Different forms of asynchronous learning can be observed from *DisenDiff* and *BAS*. Specifically, *DisenDiff* tends to show asynchronous learning between single concepts and concept groups, and overfits the concept group

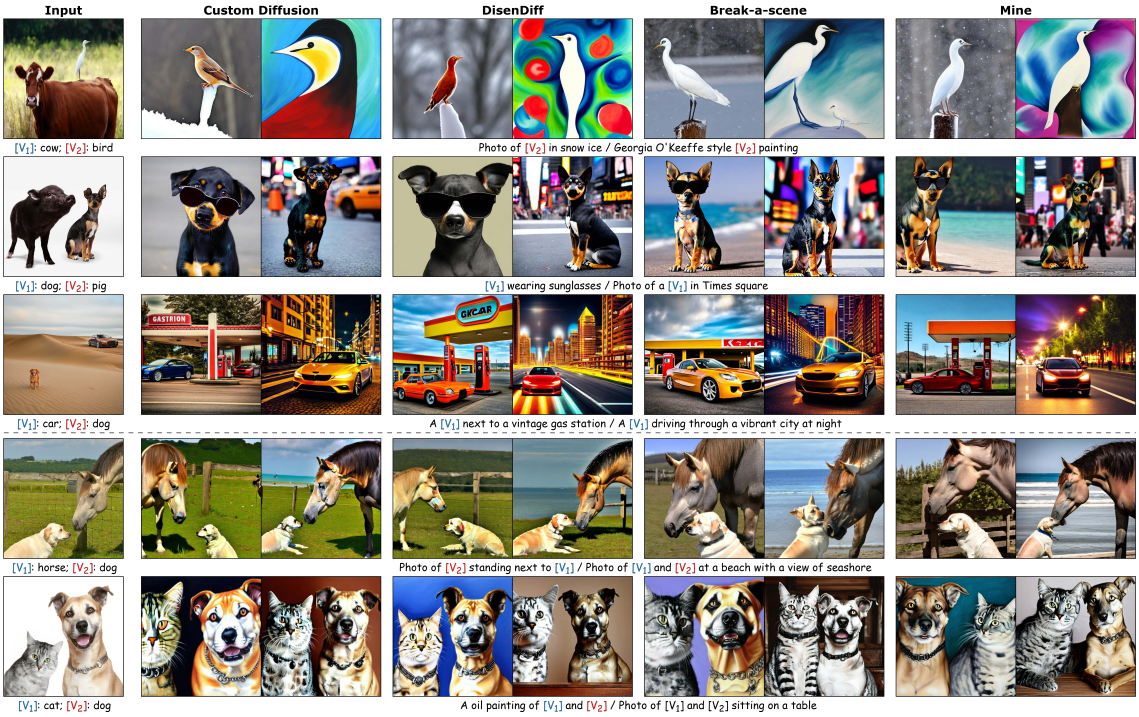


Figure 3.5: Qualitative results for concept disentanglement and feature fusion. *CusDiff* cannot disentangle multiple concepts, and both *DisenDiff* and *BAS* present feature fusion. My method not only disentangles the target concepts, but also mitigates the feature fusion problems .

before single concepts (manifest by the corruption in specific regions). In the "baby & toy" and "toy & vase" datasets, the concept groups are overfit while the target toys are not fully learned. On the other hand, *BAS* usually presents asynchronous learning across single concepts, and overfits one of the target concepts. My method exhibits better learning synchronicity compared to the baseline models, as reflected by the results.

In addition, I assess the models' disentanglement capabilities by visualizing the cross-attention maps of each [V]. As demonstrated in Fig. 3.7, although all models generate both target concepts, *CusDiff* fails to show appropriate attention activation fitting the concepts, indicating that it does not disentangle them. Moreover, *DisenDiff* displays attention activation on the background for $[V_1]$ apart from the horse, suggesting that it struggles to eliminate the background. While *BAS* demonstrates attention activation consistent with the concepts, it fails to accurately depict the dog's appearance as pointy ears are observed. My model shows strong consistency between the attention maps and concepts, effectively highlighting the concept

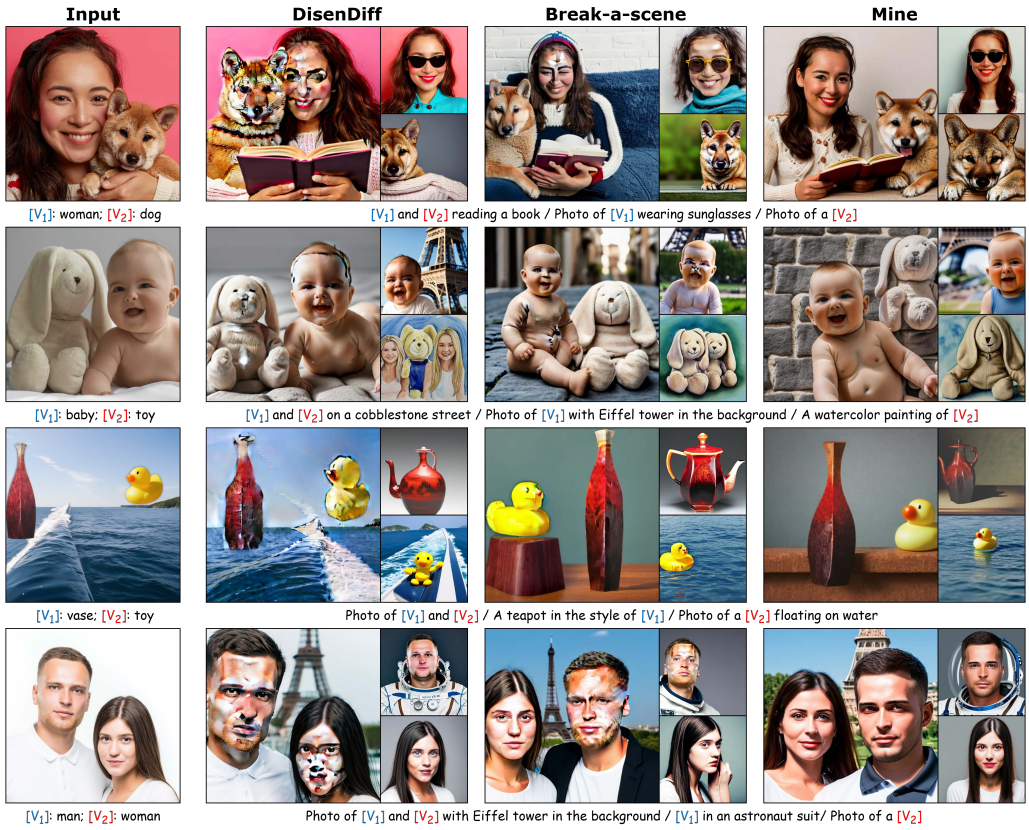


Figure 3.6: Qualitative results for learning synchronicity. *DisenDiff* and *BAS* show asynchronous learning in different forms, while my method achieves a more synchronous feature learning.

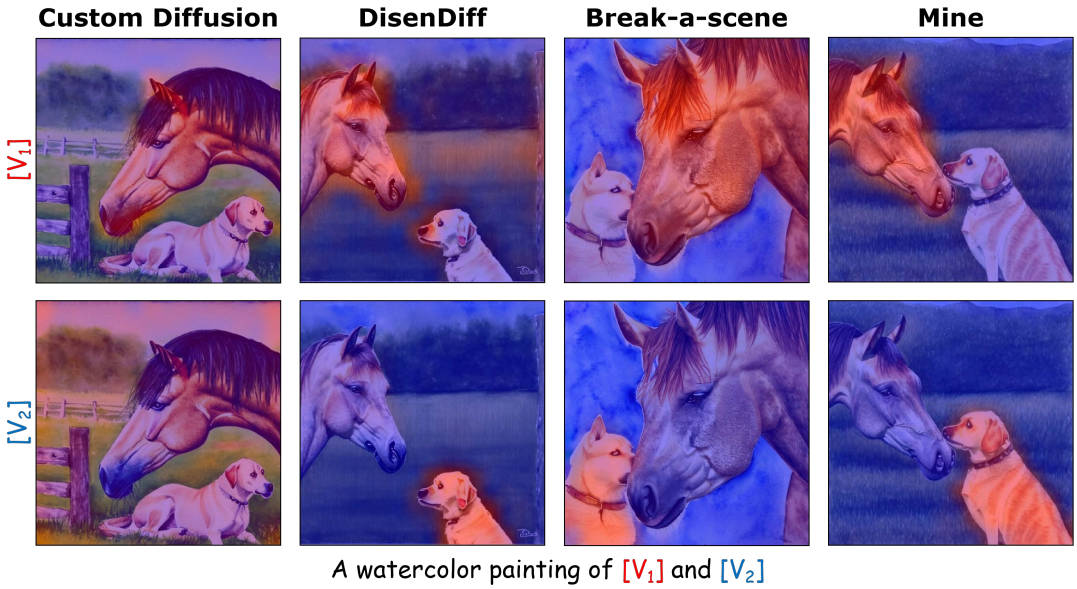


Figure 3.7: Visualization of cross-attention maps. My method presents proper attention activation for multiple conceptions.

features.

3.4.3 Quantitative comparisons

Quantitative comparisons between my method and baseline models are presented in Table 3.1. Despite *CusDiff* achieving the second-highest score in concept-group CLIP-I, it has been shown to be incapable of disentangling multiple concepts, resulting in the lowest single-concept CLIP-I score. *BAS* displays higher CLIP-I scores but lower CLIP-T and group DINO scores compared to *DisenDiff*. Notably, my method surpasses all baseline models in the CLIP-I and DINO scores across both scenarios. Regarding the CLIP-T score, my method ranks second in concept-group generation with only marginal differences compared to baseline models. Remarkably, my method records the lowest CLIP-I-sync score, demonstrating its improvement in learning synchronicity. Although *CusDiff* has the second-lowest CLIP-I-sync score, the high synchronicity stems from its failure to learn single concepts effectively. *BAS*, benefiting from the union sampling scheme, achieves a better CLIP-I-sync score than *DisenDiff*.

Table 3.1: Results of quantitative comparisons

Metrics	CusDiff	DisenDiff	BAS	Mine
Single concept				
CLIP-I↑	0.531	0.554	<u>0.563</u>	**0.576**
CLIP-T↑	**0.186**	<u>0.185</u>	0.178	0.184
DINO↑	0.653	0.657	<u>0.666</u>	**0.703**
Concept group				
CLIP-I↑	<u>0.567</u>	0.553	0.557	**0.584**
CLIP-T↑	**0.213**	0.202	0.201	<u>0.209</u>
DINO↑	0.654	<u>0.684</u>	0.654	**0.723**
CLIP-I-sync↓	<u>0.089</u>	0.097	0.091	**0.062**

3.4.4 Ablation studies

Attention-guided mask creation First, I ablate the attention-guided mask creation process by individually disabling each of the three key techniques. I find that disabling *Cross-attention suppression* permits weak attention activations outside the concept region, resulting in fragmented mask activations. Moreover, omitting *Self-attention enhancement* results in uneven and unsmooth attention distributions

within the target region, producing low-quality masks. Furthermore, I substitute *Delta masking* with Otsu thresholding and observe that the latter often fails to separate masks of different concepts, leading to incorrect associations between identifier tokens and corresponding visual features. Representative cases are presented in Fig. 3.8. Therefore, combining the three key techniques ensures the creation of high-quality masks, which guide the precise disentanglement of multiple concepts.

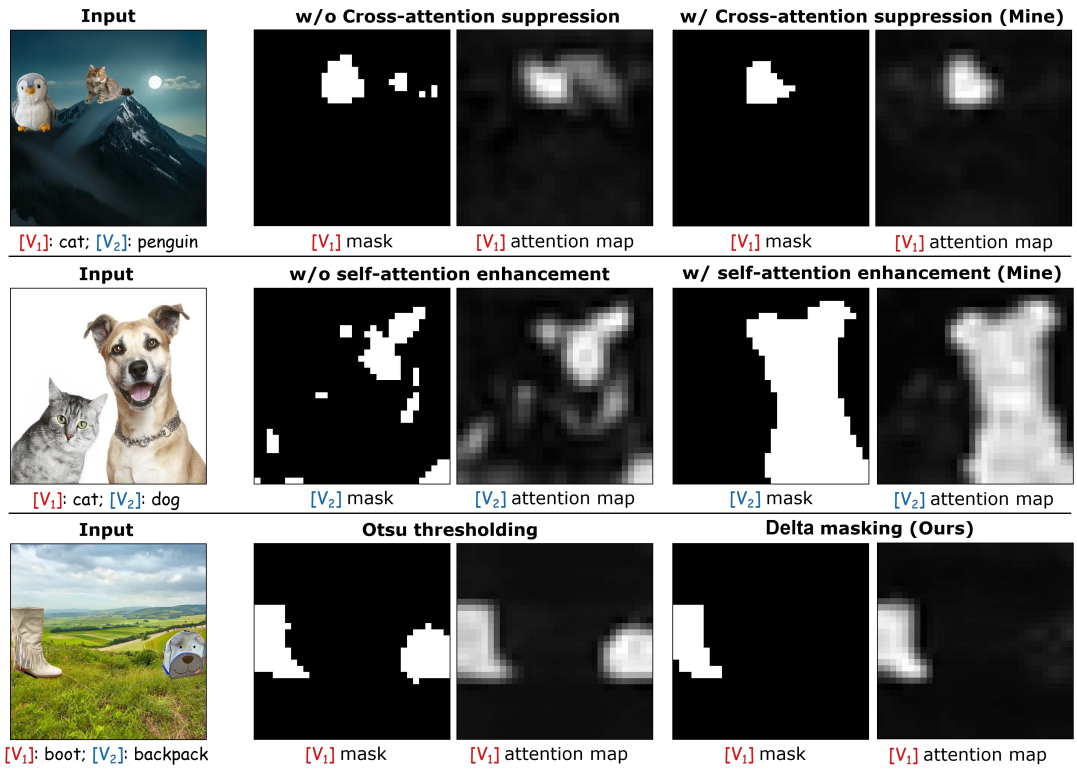


Figure 3.8: Qualitative results for ablating attention-guided mask creation. All three techniques are vital for mask creation, and disabling them will cause failure in certain datasets.

Feature-retaining training framework I propose the feature-retaining training framework by applying different loss functions to sampled subsets s of varying sizes. To validate the optimized framework, I compare my method with a variant that back-propagates L_{rec} when $|s| > 1$. A quantitative result comparison is presented in Fig. 3.9, revealing that back-propagating L_{rec} can increase the risk of feature fusion, as evidenced by the color of bird and the hairstyle of baby. Quantitative results in Table 3.2 indicate that back-propagating L_{rec} reduces CLIP-I and DINO scores while slightly increasing single-concept CLIP-T scores. In addition, I investigate

the value of ω, as over-sampling single concepts would impair the model's ability to generate multiple concepts, whereas over-sampling multiple concepts would delay feature learning. Model performance with ω ranging from 0.1 to 0.5 is illustrated in Table 3.2. My method with $\omega = 0.3$ achieves the highest CLIP-I and DINO scores, with only a marginal difference in CLIP-T score.

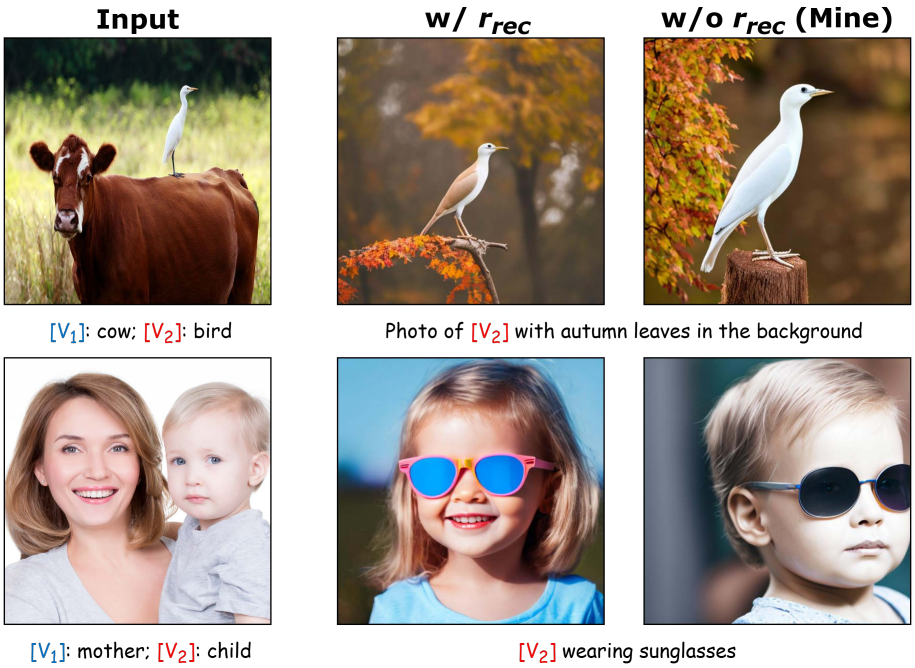


Figure 3.9: Qualitative results of ablation studies on feature-retaining training framework. My proposed framework can effectively prevent feature fusion during training.

Table 3.2: Ablation results of feature-retaining training framework

Metrics	Single concep			Concept group		
	CLIP-T	CLIP-I	DINO	CLIP-T	CLIP-I	DINO
ω=0.1	**0.185**	0.558	0.695	0.206	0.559	0.698
ω=0.2	0.183	0.562	0.691	0.202	0.563	0.693
Mine	0.184	**0.576**	**0.703**	0.208	**0.584**	**0.723**
ω=0.4	0.182	0.564	0.692	**0.210**	0.580	0.707
ω=0.5	0.184	0.565	0.696	0.208	0.573	0.720
w/ L_{rec}	**0.185**	0.569	0.680	0.204	0.571	0.670

Adaptive sampling ratio estimation As shown in Eq. (3.5), the adaptive sampling ratio is estimated using attention activation scores through normalization and

Softmax operations. Thus, I evaluate the model performance under two modifications: (1) applying an equal sampling ratio (0.5-0.5) to single concepts and (2) disabling the *Softmax* to validate my design. I select six datasets in which the difference in the estimated sampling ratio between the two concepts exceeds 0.5 to allow for an apparent comparison. Table 3.3 lists the evaluation results, which indicate that using an equal sampling ratio or disabling the *Softmax* operation degrades the fidelity of generated images and increases the disparity in learning synchronicity across different concepts. A qualitative comparison is presented in Fig. 3.10.

Table 3.3: Ablation results of adaptive sampling ratio estimation

Metrics	Equal ratio	w/o *softmax*	Mine
Single concept			
CLIP-I↑	0.580	<u>0.581</u>	**0.585**
CLIP-T↑	<u>0.182</u>	**0.183**	<u>0.182</u>
DINO↑	0.673	<u>0.680</u>	**0.688**
Concept group			
CLIP-I↑	<u>0.581</u>	0.578	**0.607**
CLIP-T↑	0.207	<u>0.209</u>	**0.211**
DINO↑	<u>0.715</u>	0.704	**0.723**
CLIP-I-sync↓	<u>0.046</u>	0.053	**0.041**

3.4.5 Generalizing to more concepts

While my main experiments focus on two-concept cases for fair comparison with baselines, the proposed innovations (i.e., attention-guided mask generation, adaptive sampling ratio estimation, and feature-retaining training) in *AttenCraft* are inherently scalable to images with more than two concepts. Each module can operate on arbitrary numbers of concepts (via multi-mask generation, normalized attention-based ratios, and concept-specific loss design), making the method directly applicable beyond two-concept scenarios.

To further support my claim, I conduct additional experiments using images that contain more than two concepts as inputs to my method. The generated results are shown in Fig. 3.11. As observed, even with more complex inputs, my method successfully disentangles the concepts and produces coherent generations for both

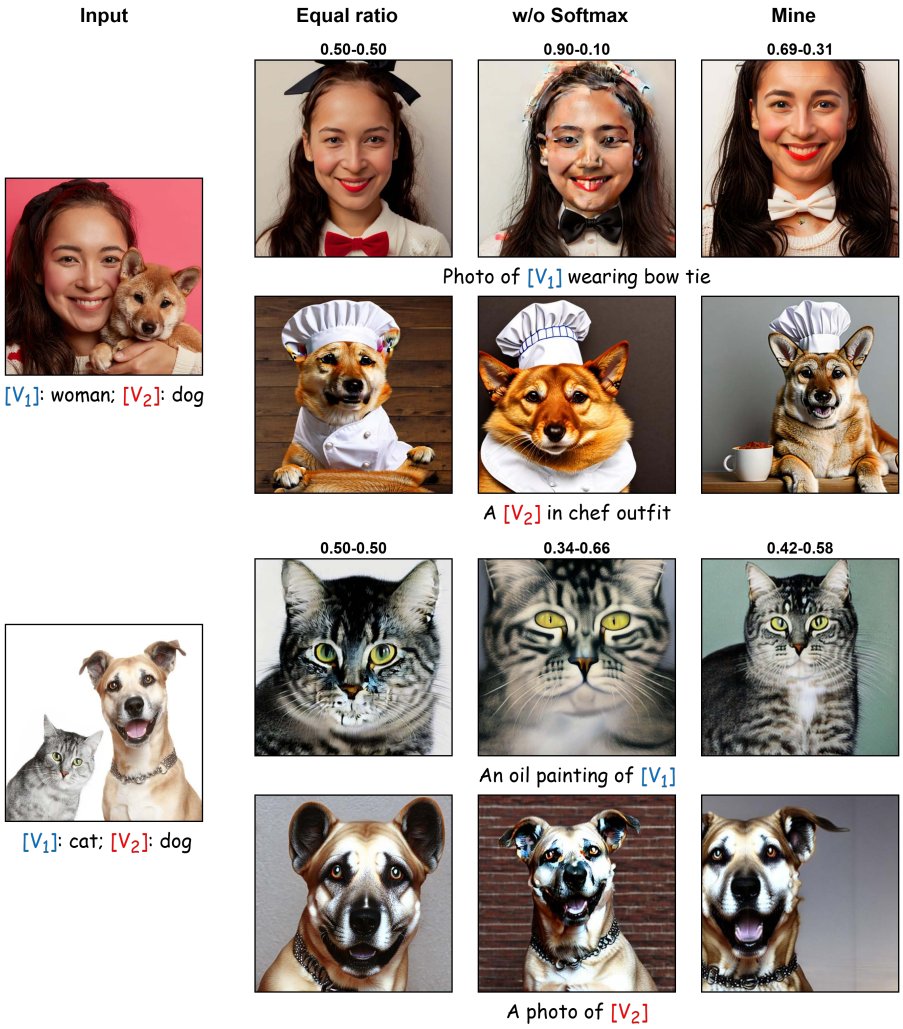


Figure 3.10: Qualitative results of ablation studies on adaptive estimation of sampling ratio. The numbers on images denote the sampling ratio determined by the method.

single concepts and concept groups. It should be noted that, given that the lamp's base in the second row was occluded in the input image, my method is nonetheless able to synthesize a new base. The lamp shade, however, is learned precisely and is rendered accurately.

3.5 Conclusion

In this paper, I identify two key issues in diffusion-based T2I models designed to disentangle multiple concepts from a single input image for T2I customization: feature fusion and asynchronous learning. To mitigate them, I propose a novel attention-

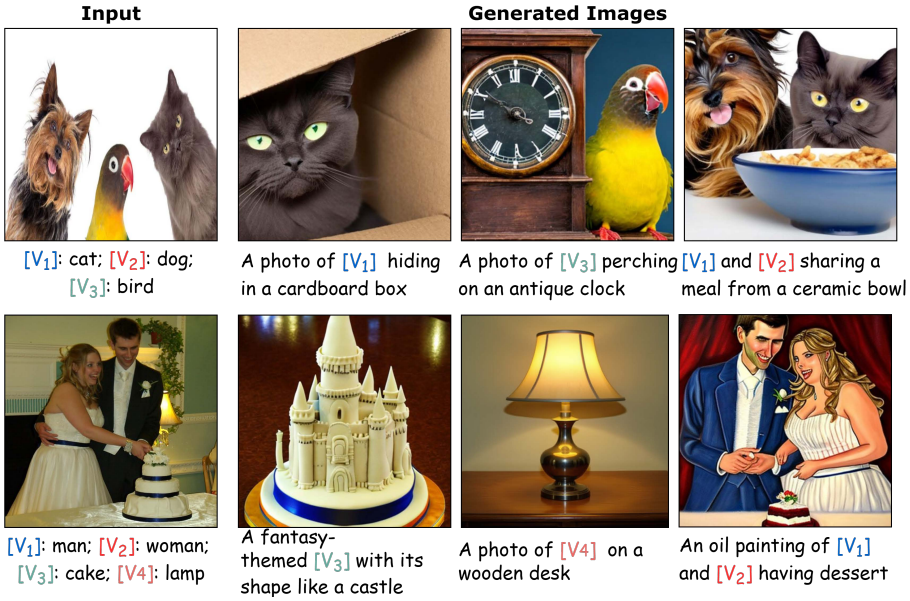


Figure 3.11: Qualitative results for *AttenCraft* applied on input images containing more than two concepts. My proposed method can be seamlessly applied to input images containing more than two concepts.

based method named *AttenCraft* as an optimized solution. I investigate the relationship between feature acquisition and identifier token initialization, and introduce an adaptive algorithm based on cross-attention scores for automatically estimating the sampling ratio of multiple concepts to mitigate asynchronous learning. Moreover, I optimize the training framework by introducing different loss functions for sampled subsets of varying sizes, retaining concept features and preventing feature fusion. In addition, I utilize attention maps to create accurate masks for each concept to guide disentanglement within a single step, without using specialized models or human inputs.

3.6 Appendix

3.6.1 Datasets

In this study, I curate 16 datasets for experiment and evaluation. I include 10 datasets introduced by *DisenDiff* (Zhang et al., 2024), generally featuring simple backgrounds. Additionally, I utilized the Gen4Gen dataset creation pipeline (Yeh et al., 2024) to amalgamate personalized concepts into complex backgrounds (e.g.,

fields, mountains, forests) sourced from copyright-free platforms, resulting in 6 synthetic datasets. The personalized concepts used for dataset synthesis were collected from the *DreamBooth* dataset (Ruiz et al., 2023a) and *CustomConcept101* (Kumari et al., 2023). All datasets are presented in Fig. 3.12, and the class names for each dataset are: (1) baby & toy; (2) cat & dog; (3) chair & lamp; (4) chair & vase; (5) cow & bird; (6) dog & pig; (7) horse & dog; (8) man & woman; (9) mother & child; (10) woman & dog; (11) boot & backpack; (12) car & dog; (13) cat & penguin; (14) dog & bear; (15) backpack & toy; (16) vase & toy.

Figure 3.12: Illustration of datasets

3.6.2 Additional details for preliminary experiment

I introduce a preliminary experiment in Chapter 3.3.2 to evaluate how the initialization of the identifier token [V] affects feature acquisition. The experiment utilizes the dataset (1)-(10) introduced in Chapter 3.6.1, employing *BAS* to disentangle multiple concepts and learn each one individually. The learning rate is set to 5×10^{-5}, the maximum training steps to 1000, and the first textual inversion phase is disabled

to improve efficiency. I design three initialization patterns for the token, consisting of combinations of the text embeddings of the precise class (dubbed as P) and the general category (dubbed as G) of the target concept. For each dataset, two target concepts are initialized by a triplet of *P-P*, *P-G*, and *G-P*, respectively. The complete list of triplets for all datasets is provided in Table 3.4.

Table 3.4: Patterns for initialization of identifier tokens

Dataset	*P-P*	*P-G*	*G-P*
baby & toy	baby-rabbit	baby-toy	human-rabbit
cat & dog	cat-dog	cat-animal	animal-dog
chair & lamp	chair-lamp	chair-lighting	furniture-lamp
chair & vase	chair-vase	chair-decor	furniture-vase
cow & bird	cow-bird	cow-animal	animal-bird
dog & pig	dog-pig	dog-animal	animal-pig
horse & dog	horse-dog	horse-animal	animal-dog
man & woman	man-woman	man-human	human-woman
mother & child	mother-child	mother-human	human-child
woman & dog	woman-dog	woman-animal	human-dog

Table 3.5: Highest cross-attention scores of [V] using different initialization patterns

Dataset	Concept	*P-P*	*P-G*	*G-P*
baby & toy	baby	0.020	0.020	0.004
	toy	0.026	0.015	0.026
cat & dog	cat	0.039	0.042	0.010
	dog	0.020	0.005	0.023
chair & lamp	chair	0.011	0.011	0.005
	lamp	0.012	0.006	0.014
chair & vase	chair	0.013	0.014	0.006
	vase	0.011	0.005	0.012
cow & bird	cow	0.046	0.045	0.010
	bird	0.027	0.005	0.014
dog & pig	dog	0.036	0.040	0.013
	pig	0.052	0.005	0.062
horse & dog	horse	0.031	0.032	0.008
	dog	0.031	0.009	0.032
man & woman	man	0.003	0.003	0.003
	woman	0.003	0.002	0.004
mother & child	mother	0.007	0.010	0.002
	child	0.012	0.003	0.012
woman & dog	woman	0.005	0.004	0.003
	dog	0.043	0.023	0.036

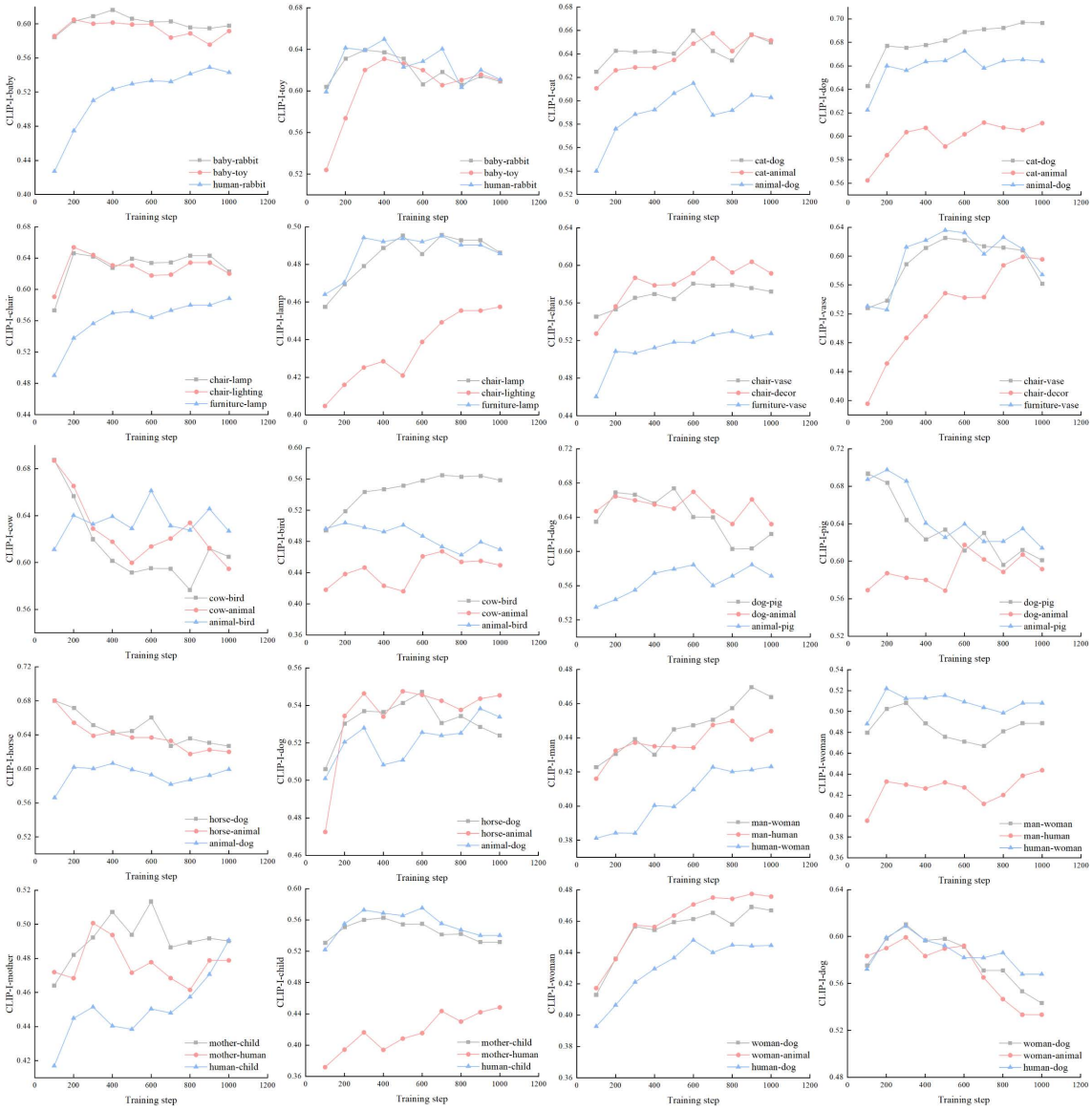


Figure 3.13: Variation of single concept CLIP-I scores with training step

I evaluate the single-concept CLIP-I scores of images generated by *BAS* at intervals of 100 training steps, following the evaluation pipeline described in Chapter 3.4.1. Detailed results for all datasets are presented in Fig. 3.13. Notably, the initialization pattern varies relative to concepts within the same dataset. For example, I initialize $[V_1]$ and $[V_2]$ in the "cat & dog" dataset by "cat-dog", "cat-animal", and "animal-dog", corresponding to *P-P*, *P-G*, and *G-P*, respectively. The initialization 'cat-animal' functions as *P-G* for assessing the concept "cat", but serves as *G-P* for assessing the concept "dog", and a similar relationship applies to "animal-dog". For most target concepts in the datasets, the CLIP-I score starts higher when

[V] is initialized with P compared to G, highlighting the significant impact of the semantic information in [V] on feature acquisition. However, when initialized with P, the CLIP-I score tends to decrease with additional training steps, indicating potential overfitting and corruption. In contrast, when initialized with G, the CLIP-I score generally increases throughout training. Additionally, for a specific identifier token [V], the initialization of another token in the same dataset has minimal impact on its feature acquisition. The average variation in the CLIP-I score is shown in Fig. 3.4(a).

Furthermore, I analyze the highest cross-attention scores extracted from the cross-attention maps for each [V], as detailed in Table 3.5. The cross-attention activation under different initialization patterns shows similar trends to feature acquisition, with higher scores observed when [V] is initialized using P rather than G. Similarly, the initialization of other [V] tokens within the same dataset has negligible effects on the cross-attention score. The average cross-attention scores for each initialization pattern are displayed in Fig. 3.4(b).

3.6.3 Additional details for main experiment

Implementation details

Custom Diffusion. I utilize the official implementation of Custom Diffusion from the HuggingFace platform (von Platen et al., 2022) with 200 training steps, a batch size of 1, and a learning rate of 5×10^{-5}. An AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is applied. During training, the text prompt is "[V_1] [$Class_1$] and [V_2] [$Class_2$]" to fit the original design of *CusDiff*. The same prompt design is also employed during inference. The identifier tokens are initialized by rare token embeddings. PEFT is applied in *CusDiff* so that only the W_k and W_v matrices in cross-attention layers of U-Net are optimized.

DisenDiff. I implement DisenDiff based on the official implementation with 250 training steps, a batch size of 1, and a learning rate of 5×10^{-5}. The optimizer is the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Similar to *CusDiff*, the design of text prompt "[V_1] [$Class_1$] and [V_2] [$Class_2$]" is applied in *DisenDiff* and the

identifier tokens are initialized by rare token embeddings. Also, *DisenDiff* follows the selection of trainable parameters of *CusDiff*.

Break-a-scene. I combine the official implementation of *BAS* with the implementation presented in *Textual Localization* (Shentu et al., 2024). Since the original implementation of *BAS* optimizes the whole U-Net following *DreamBooth* (Ruiz et al., 2023a), while *Textual Localization* presents a similar method with PEFT by only optimizing the W_k and W_v matrices in cross-attention layers of U-Net, following *CusDiff*. To ensure a fair comparison, I adapt the implementation of *BAS* with PEFT. I optimize the text embeddings of identifier tokens with a high learning rate of 5×10^{-4} for 400 steps in the first training stage, and train the text encoders and W_k and W_v matrices in cross-attention layers with a low learning rate of 5×10^{-5} for 200 steps, with a batch size of 1 applied for both stages. An AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is applied for both stages. The masks are created by jointly using the Grounding DINO (Liu et al., 2023c) and SAM (Kirillov et al., 2023). Moreover, the design of the text prompt is "[V_1] and [V_2]" where the identifier tokens are initialized by corresponding class name embeddings.

AttenCraft (Mine). I detail the implementation of my method in Chapter 3.4.1. An AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is applied.

For completeness, I note that I performed systematic experiments across multiple learning rates (1×10^{-5}, 5×10^{-5}, and 1×10^{-4}) and a range of finetuning steps for each baseline and my method. I then reported the best-performing trial for each method in the main results (Table 3.1). Moreover, to ensure fairness, I also conducted harmonized comparisons where all methods were trained under the same learning rate 1×10^{-4}) as presented in Table 3.6, and my method consistently outperformed the baselines. These additional results confirm that my reported superiority is not attributable to hyperparameter selection.

In my method, I use a single step to initialize the masks for each concept in the dataset. To present the effectiveness of my method on mask initialization, I present the initial masks for each dataset in Fig. 3.14.

Table 3.6: Performance of adopted methods under a learning rate value of 1×10^{-4}

Metrics	CusDiff	DisenDiff	BAS	Mine
Single concept				
CLIP-I↑	0.528	0.551	<u>0.564</u>	**0.576**
CLIP-T↑	**0.185**	<u>0.184</u>	0.177	<u>0.184</u>
DINO↑	0.646	0.654	<u>0.662</u>	**0.703**
Concept group				
CLIP-I↑	<u>0.565</u>	0.555	0.547	**0.584**
CLIP-T↑	**0.211**	0.201	0.195	<u>0.209</u>
DINO↑	0.648	<u>0.667</u>	0.652	**0.723**

Mask initialization and sampling ratio determination

My method tends to present stronger attention on the humans themselves than other parts of humans, such as the long hair and cloth, as illustrated by the woman's mask in the "woman & dog" dataset and the mother's mask in the "mother & child" dataset. Aside from these, my method successfully creates accurate masks for other datasets.

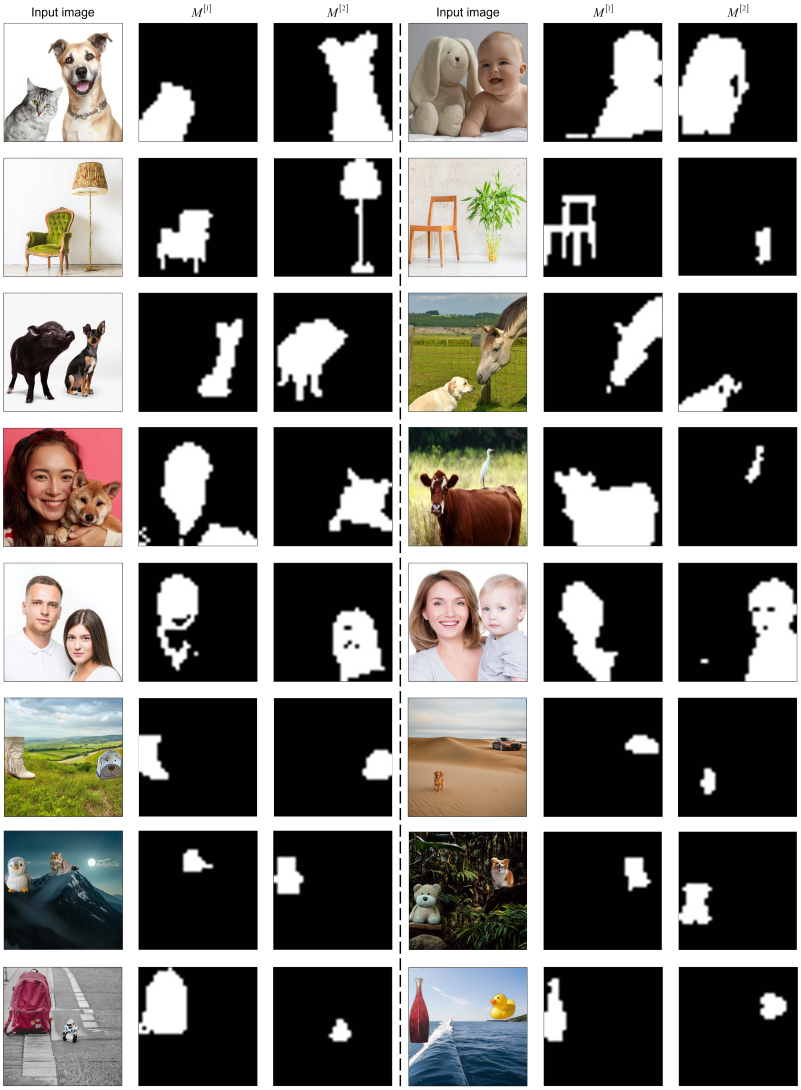


Figure 3.14: Initial masks for each dataset created by my method

CHAPTER 4

Controllable Synthesis of Dermoscopic Images for Enhanced Computer Aided Diagnosis and Detection

Building upon the attention-based controllability mechanisms established in Chapter 3, this chapter transitions from general-purpose T2I DMs to a domain-specific investigation in medical imaging, focusing on dermoscopic image synthesis. While Chapter 3 demonstrates how internal attention representations can be exploited to disentangle and control multiple visual concepts, Chapter 2 highlights that medical imaging presents additional challenges, including data scarcity, severe class imbalance, and the need for precise spatial correspondence between images and annotations. This chapter represents the first application-oriented extension of the thesis, examining how controllable DMs can be adapted to address these challenges in a clinically relevant setting.

The motivation of this chapter arises from the limitations of existing data augmentation strategies for skin lesion analysis. Traditional augmentation techniques and GAN-based generative models often fail to capture the diversity and fine-grained attributes of dermoscopic images, while recent diffusion-based methods typically lack effective mechanisms for semantic and regional control. For computer aided diagnosis/detection (CAD) systems, it is not sufficient to generate visually plausi-

ble images alone; controllability over lesion attributes and the ability to synthesize aligned lesion-mask pairs are essential for improving downstream classification and segmentation performance. This chapter aims to develop a controllable diffusion-based synthesis framework that enhances both visual-textual alignment and spatial consistency, while remaining efficient and scalable.

To this end, this chapter introduces DiDGen, a diffusion-based dermoscopic image generation framework designed to support controllable and task-oriented data augmentation. The proposed approach incorporates a dynamic prompting strategy, DermPrompt, which leverages large language models to construct attribute-rich textual descriptions, and a region-aware fine-tuning mechanism that strengthens local visual-textual correspondence. In addition, a training-free pipeline is proposed to enable simultaneous generation of dermoscopic images and their corresponding lesion masks. Extensive experiments demonstrate that the synthesized data significantly improves the performance of downstream CAD systems in both classification and segmentation tasks. Beyond its immediate contributions, this chapter establishes a practical foundation for more fine-grained and targeted controllable generation. The techniques for region-aware control and paired image-mask synthesis are further extended in Chapter 5, where controllable diffusion models are leveraged to address clinically significant bias arising from low-contrast lesions in skin lesion segmentation.

4.1 Introduction

As one of the most prevalent cancer diagnoses, skin cancer has become a major public health problem (Watson et al., 2015). Melanoma is the most lethal skin cancer, with 8290 deaths estimated in the United States in 2024 (Siegel et al., 2024). However, when detected early survival rates for melanoma can reach over 95% (Xie et al., 2016). Dermoscopy is a noninvasive, common imaging technique for skin cancer diagnosis, providing a detailed view of the skin lesion, and enhances the performance of skin lesion diagnosis compared to evaluation by the naked eye (Vestergaard et al., 2008; Silveira et al., 2009). However, the visual inspection of

dermoscopic images is time-consuming for dermatologists, making it less practical given the severe shortage of dermatologists (Freeman, 2023). Also, the diagnosis is frequently affected by subjective bias (Xu et al., 2024), resulting in a diagnostic accuracy of about 60%, even among experienced dermatologists (Qin et al., 2020). To this end, CAD systems that can steadily and efficiently assist diagnosis are required to continue to ensure the timely detection of skin cancer.

CAD systems have been deployed to offer a "second opinion" for clinicians to help make the final decision. Early attempts of CAD systems applied traditional machine learning models, such as support vector machine (SVM) and k-nearest neighbor (K-NN), based on features extracted by manually designed algorithms, presenting limitations in the generalizability due to these handcrafted features (Doi, 2007; Giger et al., 2008; Kim et al., 2011). Thanks to the success of deep learning techniques in image processing and pattern recognition, there is growing interest in also applying them to CAD systems. Deep learning-based CAD systems have been investigated in various modalities of medical images, including but not limited to chest X-rays, CT scans, and MRI (Kang et al., 2022; Alshayeji et al., 2022; Hammouda et al., 2021). In skin lesion detection and diagnosis, CAD systems are commonly applied in two tasks: segmentation and classification (Hasan et al., 2023). Lesion segmentation serves as a preprocessing step for feature extraction and classification, and segmented masks and classification results can be applied for lesion detection and recognition (Hasan et al., 2023). Currently, most CAD systems designed for skin lesion diagnosis rely on deep learning-based methods requiring large amounts of faithful data to achieve high performance. Nevertheless, publicly available dermoscopic datasets remain limited in annotated samples and imbalanced across different diagnoses and populations, restricting the improvement of CAD models (Qin et al., 2020), raising an urgent requirement for dataset augmentation.

Conventional augmentation methods for image data comprise geometric and intensity transformations, such as flipping, rotating, and color/brightness adjustments. However, the additional images produced by these methods still fall into a similar distribution to the original images, resulting in limited performance improvement for downstream tasks (Abdelhalim et al., 2021). Moreover, these augmentation methods

are already commonly applied for model training in downstream tasks in dermatology (Esteva et al., 2017). To this end, deep generative models that are capable of synthesizing images that do not exist in the original dataset have become promising solutions for image dataset augmentation. Similar techniques have been shown to have significant promise in other medical imaging domains(Bamoriya et al., 2022; Weber et al., 2023), but the proposed models are dedicated to specific domains, which cannot be directly adapted to dermoscopic image generation tasks.

Dermoscopic image generation models based on GANs have been proposed to augment dermoscopic datasets and enhance classifier performance. However, GAN-based models suffer from inherent limitations such as training instability, difficulties in synthesizing high-fidelity images with fine-grained clinical details, and lack of controllability (Qin et al., 2020; Ren et al., 2021; Bisla et al., 2019). More recently, while DMs have demonstrated unprecedented capabilities in high-fidelity image synthesis and text-guided controllability (Rombach et al., 2022), their adaptation to dermoscopic image generation remains underexplored. Recent studies finetune pre-trained DMs on skin lesion datasets but rely on simplistic text prompts (e.g., fixed templates like "a dermoscopic image of melanoma") that fail to capture the nuanced visual-textual relationships critical for medical imaging(Farooq et al., 2024; Shavlokhova et al., 2023). This approach underutilizes the semantic potential and controllability of DMs, producing generic images lacking diagnostically relevant diversity (e.g., variations in asymmetry, border irregularity, or color variegation). Consequently, downstream models trained on such data risk inheriting biases and incomplete feature representations.

Alongside classification, semantic synthesis of dermoscopic lesion-mask pairs is crucial for segmentation tasks. Existing methods employ GANs or DMs designed specifically for mask-to-image generation (Abhishek and Hamarneh, 2019; Du et al., 2024), but these require task-specific architectures and extensive training, limiting cross-task generalization and increasing computational overhead. Moreover, current approaches lack unified frameworks for multi-level controllability including textual guidance, spatial layout control, and flexible image editing, which is essential for generating comprehensive datasets tailored to diverse clinical needs.

To address these limitations, I propose a novel method for dermoscopic image generation based on T2I SD, termed DiDGen. My method is a controllable image generation pipeline that realizes controllability at different levels, including (1) text-guided attribute customization that allows users to specify features, such as skin tone and lesion type, through natural language; (2) layout-guided semantic synthesis that dictates the spatial arrangement of the generated images based on structural conditions; (3) other controllable applications such as image editing that provides post-generation refinement to the real and generated images. These levels of controllability enable the intuitive and precise generation of dermoscopic images, enhancing the utility and accuracy of synthetic images in downstream tasks. Specifically, I propose DermPrompt, a new framework that leverages large language models (LLMs) to generate captions containing attribute-rich details as dynamic text prompts for SD. The attribute-rich information generated by DermPrompt guides SD to learn fine-grained visual representations, and enables textual controls on the contents of the generated dermoscopic images. To enable layout-guided semantic synthesis, I propose a novel two-stage paradigm. This first builds semantic visual-textual alignment through region-aware finetuning based on the cross-attention mechanism; the built semantic alignments lay a foundation for layout-guided generation. My proposed technique then generates dermoscopic images with the corresponding masks using a training-free pipeline based on test-time layout guidance with attention-based annotation techniques. Significantly, the entire workflow of my method only requires finetuning the SD once while other existing methods require multiple training or finetuning to realize these functions, making it an efficient and useful tool for dermoscopic dataset augmentation. To the best of my knowledge, this is the first study that comprehensively explores the controllable generation of domain-specific dermoscopic images.

4.2 Related Works

4.2.1 CAD systems for skin lesion diagnosis

CAD systems for skin lesion diagnosis primarily comprise a workflow of image acquisition, lesion segmentation, feature extraction, and lesion classification, with lesion segmentation and classification being the main focus of these techniques (Oliveira et al., 2018).

Segmentation methods for skin lesions traditionally use a U-Net-based architecture. Ashraf et al. (2022) utilized U-Net, ResUNet, and ResUNet++ to assist the segmentation of skin lesions. Araújo et al. (2022) combined U-Net and LinkNet with transfer learning, and Nawaz et al. (2022) proposed a DenseNet77-based U-Net for the purpose of improving the segmentation accuracy. Thomas et al. (2021b) introduced an interpretable system for non-melanoma skin lesion segmentation and classification based on U-Net. Moreover, U-Net-based segmentation models designed for other modalities of medical images can also be adapted for skin lesion images (Xu et al., 2023). Transformer-based models have recently been used as a promising tool for skin lesion segmentation. Wang et al. (2021, 2023d) proposed a boundary-aware transformer-based model XBound-Former for skin lesion segmentation, and Chen et al. (2021) introduced TransUNet, a mixed model of a U-Net and a transformer, for segmenting medical images, which can be applied to skin lesion images. With the development of DMs, DM-based segmentation methods were proposed. Vu Quoc et al. (2023) proposed a latent DM for medical image segmentation, and Rahman et al. (2023) introduced a DM-based approach to produce multiple plausible segmentation masks via distribution learning.

In terms of skin lesion classification, conventional convolutional neural network (CNN)-based classifiers are commonly used for feature extraction and lesion diagnosis. Chaturvedi et al. (2020) leveraged various CNN-based classifiers including InceptionV3 and ResNeXt101, for skin lesion diagnosis. Shetty et al. (2022) built a CNN classifier with an architecture optimized for multi-class skin lesion classification. Wang et al. (2023a) proposed an algorithm to increase the intra-class consistency and inter-class discrimination of learned features in skin lesion classi-

fication by combining CNNs with the class activation mapping (CAM) algorithm. Moreover, Ayas (2023) applied a Swin transformer model, and Lungu-Stan et al. (2023) applied a vision transformer model, for skin lesion classification. In addition, DM-based models were introduced to realize general medical image classification by eliminating unexpected noise in images and robustly capturing semantic representation (Yang et al., 2023).

4.2.2 Dataset augmentation for skin lesion images

Conventional dataset augmentation methods including image flipping, cropping, shifting, and rotating are reported to lead to a limited performance increase, as the transformed images inherently have a similar distribution to the original dataset (Abdelhalim et al., 2021). To this end, images generated by generative models have become a new solution for dataset augmentation. In the field of the generation of skin lesion images, existing methods mainly focus on two modalities: macroscopic images and dermoscopic images. Macroscopic images are mostly taken by patients themselves using non-specialized devices, such as smartphones. Akrout et al. (2023) leveraged DM-based textual inversion method to generate macroscopic skin lesion images with different diagnoses for augmenting the classification dataset. Considering the difference in skin tone, Sagers et al. (2023) combined the DM-based DreamBooth with inpainting/outpainting generation to generate skin lesion images with distinct diagnoses and skin tones for training classifiers. Similarly, Wang et al. (2024) proposed a pipeline consisting of DreamBooth, Low-Rank Adaptation (LoRA) and image-to-image translation to augment the groups with underrepresented diagnoses and skin tones for training classifiers. Furthermore, Lin et al. (2024) proposed to use diagnosis results from a vision language model (VLM) as input for the DMs to generate corresponding skin lesion images, thereby enhancing the visual explanability for users.

Generative models for dermoscopic images based on GANs are already prevalent. Various variants of GANs, including PGAN (Abdelhalim et al., 2021; Baur et al., 2018; Bissoto et al., 2018), StyleGAN (Qin et al., 2020; Ren et al., 2021), DCGAN (Bisla et al., 2019; Pollastri et al., 2020), and cascaded GAN (Shahsavari et al.,

2021), have been applied to generate dermoscopic images with different diagnoses, or reconstruct super-resolution images. However, GAN-based methods for generating dermoscopic images are usually unconditional or simply conditioned on class, limiting the models' capability for generalization. DM-based models have demonstrated an ability to realize more flexible and controllable generation via text-to-image generation. Shavlokhova et al. (2023) finetuned the pretrained GLIDE to generate dermoscopic images with different diagnoses and achieved classification accuracy comparable to dermatologists. Farooq et al. (2024) leveraged DreamBooth to generate dermoscopic images with few-shot learning, and augment the classification dataset with synthesized images.

In addition, semantic generation that uses masks as conditions have been proposed to generate dermoscopic images, yielding mask-image pairs for augmenting the segmentation dataset. Abhishek and Hamarneh (2019) leveraged the Pix2Pix model trained on real mask-lesion pairs to generate new skin lesion images given arbitrary masks. Zhang et al. (2023d) proposed a GAN-based image translation model adapted for datasets with insufficient training data, such as medical images. Moreover, Du et al. (2024) applied the DM-based ControlNet model on dermoscopic images to realize mask-to-lesion generation, enhancing the quality of generated images and boosting the accuracy of segmentation models above that of the GANs model.

However, most existing studies propose dedicated models for specific tasks in augmenting skin lesion image datasets, leaving a gap in the design of a versatile method that is adaptable to multiple augmentation tasks. Additionally, the controllability of the model requested for multi-functional generation has not been fully investigated.

4.2.3 Controllable generation with diffusion models

DMs have achieved significant advances in image generation as they present remarkable capability in synthesizing high-fidelity and high-diversity images. More importantly, benefiting from its generation process in the form of Markov chain and text-guided generative functions, it exhibits enormous potential in allowing precise

control over the generated content to fulfill complex and diverse needs, enabling "controllable generation" (Cao et al., 2025). Based on basic text-to-image generation, two paradigms of controllable generation methods have been proposed: model-based conditional generation and training-free conditional generation. Model-based conditional generation employs an additional encoder to align novel conditions (e.g., semantic maps, depth maps, sketches) with textual semantics, enabling multi-modal control through learned feature fusion.

However, model-based conditional generation methods rely on annotated data and extra computational resources, thereby motivating the emergence of training-free conditional generation that bypasses additional training and directly introduces the controlling conditions to the generation process through the intrinsic ability of the structure of U-Net. Chefer et al. (2023); Rassin et al. (2023) strengthened the linguistic bindings between the generated objects and their attributes by regularizing cross-attention maps during generation. Similarly, Xie et al. (2023); Chen et al. (2024a) explored the spatial control of the generated layout by comparing cross-attention maps with bounding-box annotations, and Couairon et al. (2023); Balaji et al. (2022); Kim et al. (2023) proposed to modulate the cross-attention map using semantic maps to realize spatial-conditioned semantic generation. Moreover, Phung et al. (2024) combined the regularization from both the cross-attention maps and self-attention maps to improve the spatial controllability of text-to-image generation, achieving a more precise and robust control for generation. In addition, Hertz et al. (2024) introduced features of the reference image via self-attention to control the style of generated images.

In this study, I propose an image generation pipeline by exploiting the controllable generation of pretrained DMs. The controllability is reflected in multiple input modalities, including text prompts, spatial layout information, and domain-specific visual features.

4.3 Proposed Method

In this section, I begin by providing a brief overview of the diffusion model, and then introduce my novel techniques. An overall illustration of my method is presented in Fig. 4.1. Algorithm 2 summarizes the overall pipeline of DiDGen, including attribute-aware prompt construction, region-aware finetuning, and training-free generation of dermoscopic lesion-mask pairs.

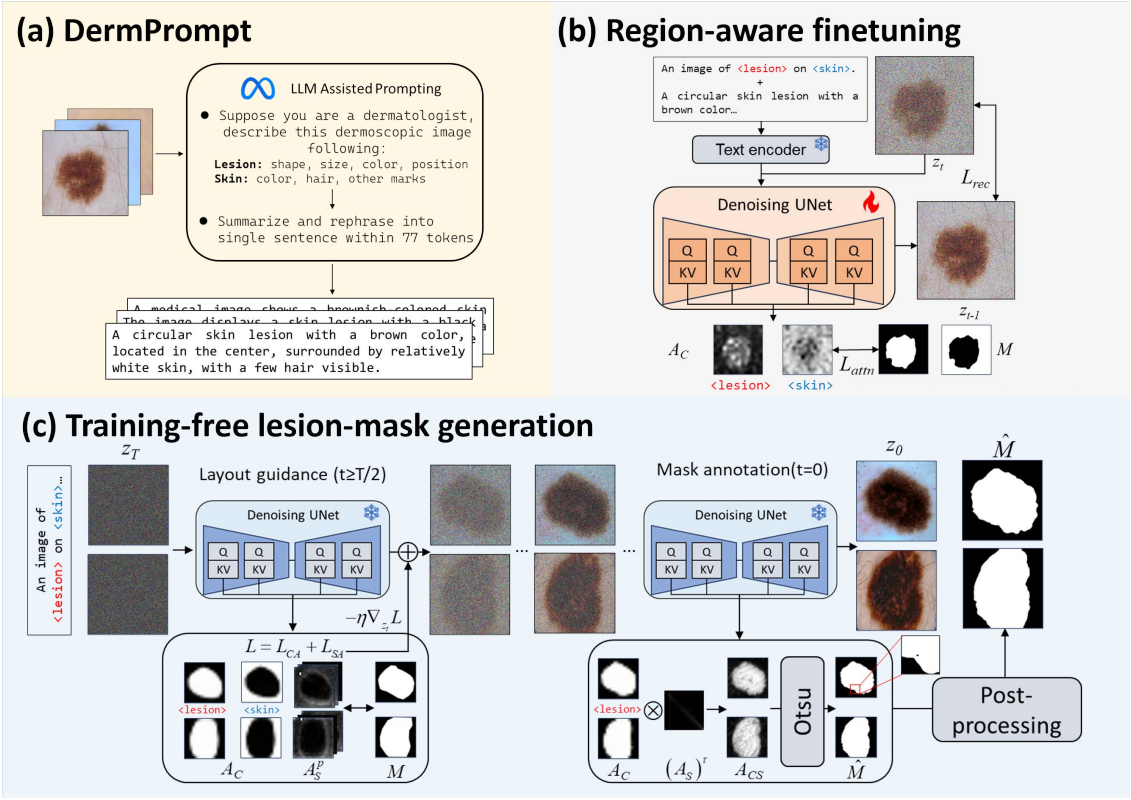


Figure 4.1: Overview of my proposed method DiDGen including three novel technical contributions: (a) DermPrompt for producing attribute-rich text prompts; (b) region-aware finetuning that facilitate the establishment of the semantic visual-textual alignments between text prompts and visual representations, and (c) training-free lesion-mask generation pipeline for synthesizing high-quality images of lesion-mask pairs

Algorithm 2 Overall pipeline of DiDGen

Require: Dermoscopic images $\{x\}$, lesion masks $\{M\}$, pretrained SD model, Llama-3.2 Vision model

Ensure: Finetuned SD model; generated dermoscopic images or lesion-mask pairs

Stage I: Attribute-aware DermPrompt
for each dermoscopic training image **do**
    Extract lesion and skin attributes using Llama-3.2 Vision model
    Rephrase the attributes into an attribute-rich text prompt y
end for
Stage II: Region-aware finetuning
for each training image x and lesion mask M **do**
    Build the prompt "An image of <lesion> on <skin>" + DermPrompt
    Initialize <lesion> and <skin> as placeholder tokens
    Run the denoising UNet and extract cross-attention maps
    Compute L_{rec} and region-aware attention loss L_{attn}
    Optimize with $L = L_{\text{rec}} + \alpha L_{\text{attn}}$
end for
Stage III: Training-free lesion-mask generation
Initialize noisy latent z_T from a text prompt and lesion layout mask
for early denoising timesteps **do**
    Extract cross-attention and self-attention maps
    Compute the layout guidance losses to refine the latent
    Continue denoising with the refined latent
end for
Extract final lesion attention maps and combine cross-/self-attention responses
Obtain the lesion mask via Otsu thresholding and post-processing
return generated dermoscopic image and corresponding lesion mask

4.3.1 Preliminary

For an input image $x \in \mathbb{R}^{H\times W\times 3}$, SD first projects x into a latent representation $z \in \mathbb{R}^{h\times w\times c}$ via a VAE $\mathcal{E}$ (Kingma and Welling, 2013), where c is the latent feature

dimension. Then it produces a Markov chain $z_1, \ldots, z_T$ by progressively adding Gaussian noise:

$$q\left(z_t|z_{t-1}\right) := \mathcal{N}\left(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t \mathbf{I}\right) \tag{4.1}$$

where β_t is the variance hyperparameter of the noise at timestep t. The reversed process of Eq. (4.1) can be approximated by a neural network as follows:

$$p_\theta\left(z_{t-1}|z_t\right) := \mathcal{N}\left(z_{t-1}; \mu_\theta\left(z_t, t\right), \sum_\theta\left(z_t, t\right)\right) \tag{4.2}$$

where the neural network parameterizes the mean $\mu_\theta\left(z_t, t\right)$ and variance $\sum_\theta\left(z_t, t\right)$. This is commonly achieved using a U-Net (Ronneberger et al., 2015) architecture. During the training of U-Net, the text prompts y are projected into text embeddings by a pre-trained text encoder τ_θ, and U-Net is trained to predict the random noise ε added to the noisy latent z_t:

$$L_{rec} = \mathbb{E}_{z,y,t,\varepsilon}\left[\|\varepsilon - \varepsilon_\theta\left(z_t, t, \tau_\theta(y)\right)\|_2^2\right] \tag{4.3}$$

where ε and ε_θ are standard Gaussian noise and predicted noise residual, respectively.

The U-Net incorporates self-attention and cross-attention layers to capture the dependencies within the input data (Rombach et al., 2022; Vaswani et al., 2017). The self-attention layers capture the global attention within the image while the cross-attention layers learn to attend between the image and text prompts. The cross-attention map A_C and self-attention map A_S can be calculated as follows:

$$\begin{aligned} A_C &= softmax\left(Q_I K_T^\top / \sqrt{d}\right) \\ A_S &= softmax\left(Q_I K_I^\top / \sqrt{d}\right) \end{aligned} \tag{4.4}$$

where Q_I, K_I, K_T are the query matrix, key matrix of z_t, and query matrix of $\tau_\theta(y)$, respectively. d denotes the latent dimension.

4.3.2 Attribute-aware DermPrompt

Existing studies in T2I synthesis for dermoscopic image generation often rely on simplistic and fixed text prompts (e.g., *"An image of skin lesion."*) during training and inference. While effective for basic generation, these prompts lack the granularity to capture intricate morphological and contextual attributes of dermatological conditions, limiting the model's ability to synthesize clinically relevant details. To address this, I propose DermPrompt, a framework that leverages LLMs to generate dense, attribute-rich captions as dynamic text prompts for T2I models. Specifically, I utilize the state-of-the-art vision-language model (VLM) Llama 3.2 Vision model (Dubey et al., 2024) to perform visual attribute extraction from dermoscopic images, decomposing each image into fine-grained descriptors, including explicit annotations of:

- **Lesion characteristics**: shape, color, size, position
- **Skin characteristics**: color, presence of hair, other artificial markers

To assess the reliability of the generated prompts by Llama-3.2 Vision model, I conduct a human evaluation on 200 randomly sampled training images. Three open-source VLMs, which are LLaVA-1.5, Phi-3.5 Vision, and Llama-3.2 Vision, are prompted identically to generate one caption per image. Three volunteers independently ranked each caption from 1 (worst) to 3 (best) according to correctness, completeness, and clinical plausibility. The averaged rankings are shown in Fig. 4.2.

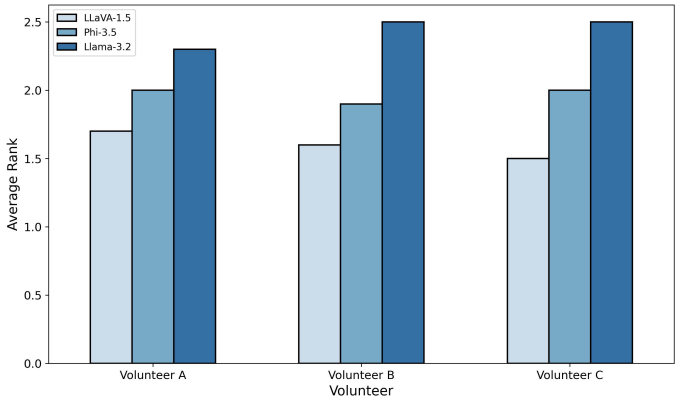


Figure 4.2: Average volunteer rankings of captions generated by different VLMs

Across all models, Llama-3.2 Vision model consistently achieved the highest rankings, demonstrating clear advantages in descriptive accuracy and attribute completeness. I then apply Llama to rephrase these attributes into a concise, structured sentence of less than 77 tokens to conform to the input capacity of the standard text encoders used in SD, while retaining critical clinical semantics. The rich-context DermPrompt framework facilitates pixel-level semantic grounding by transforming generic T2I training into a dermatology-specific, attribute-aware synthesis process. Unlike conventional prompts that weakly associate broad terms like "lesion" with generic visual patterns, DermPrompt's lexically dense captions explicitly bind domain-specific descriptions to localized image regions, thereby reduce ambiguity during inference. To enhance the fidelity, diversity, and granularity of synthetic dermoscopic images, I deploy Llama to create new prompts based on existing DermPrompt and apply these new prompts during inference.

4.3.3 Region-aware Finetuning of SD

Although DermPrompt facilitates a more diverse and controllable generation of dermoscopic images, it still presents limitations in customized spatial controlling required for precise semantic generation, as illustrated in Fig. 4.5. To address this limitation, I propose a new paradigm for generating dermoscopic lesion-mask pairs using a two-stage process, as shown in Fig. 4.1. The first stage involves region-aware finetuning of SD, where the finetuning process of the model is specifically designed to optimize both the reconstruction loss (Eq. (4.3)) and an additional region-aware attention loss. This dual-objective optimization encourages the model to capture both image features and spatial characteristics, ensuring that both lesion and skin regions are accurately represented.

Prompt formulation

I begin by designing a structured text prompt that can effectively guide SD during the finetuning process. Specifically, I use my proposed DermPropmt as the text prompt y, prefixed with the phrase " *An image of <lesion> on <skin>*". In this formulation, the tokens "*<lesion>*" and "*<skin>*" are special tokens that are added

to the SD vocabulary (Gal et al., 2023b). These tokens, named Placeholder Tokens (P-Tokens), serve as anchors to establish robust visual-textual alignment between the textual description and corresponding image regions. The P-Tokens are initialized with the text embeddings of *lesion* and *skin*, and their embeddings remain fixed during finetuning. This strategy preserves the inherent semantic meaning of the tokens while allowing the model to learn detailed spatial correspondences.

Region-aware attention loss

A significant strength of SD is its ability to establish semantic alignment between visual representations and text prompts, a property primarily achieved via the cross-attention mechanism (Chefer et al., 2023; Feng et al., 2023). Traditionally, these alignments are developed through extensive pre-training on vast datasets (e.g., LAION-5B (Schuhmann et al., 2022) that contains 5.85 billion data). However, when adapting SD to a new domain with limited data, such as dermoscopic images, the model may struggle to form accurate semantic links due to the comparatively limited amount of data available. To overcome this challenge, I introduce a region-aware attention loss that directly guides the cross-attention mechanism towards the desired image regions. The loss function is defined as follows:

$$L_{attn} = \frac{1}{2} \sum_{i \in \{l,s\}} \| A_C\left(v_i, z_t\right) - M_i \|_2^2 \tag{4.5}$$

where $A_C\left(v_i, z_t\right)$ is the cross-attention map for P-Token v_i extracted from layers of latent scale 16×16 (following the implementation in (Hertz et al., 2023)). The indices l, s denotes the P-Tokens "*<lesion>*" and "*<skin>*", respectively. Similarly, M_l is the ground-truth mask for the lesion region adopted from the dataset, and M_s is computed as $M_s = 1 - M_l$. This loss penalizes discrepancies between the predicted attention maps and the true region masks, effectively enforcing the association between each P-Token and its corresponding image region. The overall loss function for the finetuning stage is a weighted combination of the reconstruction loss and the region-aware attention loss:

$$L = L_{rec} + \alpha L_{attn} \tag{4.6}$$

where α is a scaling coefficient that balances the weight between the two loss items. An excessively high α would impact the quality of feature learning, while an overly low α would delay the attention association, and I empirically find that setting $\alpha = 0.1$ reaches an equilibrium between them. This composite loss ensures that the model not only reproduces high-quality images but also aligns the generated content with the desired semantic regions, laying the foundation for the faithful generation of dermoscopic lesion-mask pairs introduced in the second stage.

4.3.4 Training-free Pipeline for Dermoscopic Lesion-mask Pair Generation

For the controlled semantic generation of dermoscopic images, in contrast to traditional image translation approaches that are both time-consuming and computationally expensive, I introduce a dedicated training-free pipeline that simultaneously generates dermoscopic image-mask pairs. My method leverages the semantic alignments built through the region-aware attention loss, combining test-time layout guidance with attention-based annotation to achieve diverse image generation and accurate mask annotation (Nguyen et al., 2023; Xie et al., 2023; Couairon et al., 2023; Wu et al., 2023b). A schematic diagram is presented in Fig. 4.1(c).

Layout guidance

The sampling process in SD can be effectively steered by classifier guidance, wherein the generation process is conditioned on a label c, enabling test-time conditional sampling (Dhariwal and Nichol, 2021; Song et al., 2020b). By replacing the label c with a mask M, and leveraging the cross-attention maps to manipulate spatial attributes such as position, size, and shape, the concept of classifier guidance can be extended to layout guidance (Xie et al., 2023; Couairon et al., 2023; Phung et al., 2024).

For skin lesion generation, my method integrates both cross-attention and self-attention regularization. For cross-attention regularization, I first extract the cross-attention maps of the P-Tokens in the prefix, apply a *Softmax* function to normalize

these maps, and then apply Gaussian Smoothing to refine them. The regularization for the cross-attention guides the layout of the generated lesion, and is performed on the P-Token *<lesion>*, named as L_{CA}. This loss function enforces the lesion's spatial configuration, including its position, size, and shape, to closely match that of the input mask M, expressed as:

$$L_{CA} = \left(1 - \frac{\sum \left[A_C\left(v_l, z_t\right) \cdot M_l\right]}{\sum M_l}\right) + \frac{\sum \left[A_C\left(v_l, z_t\right) \cdot (1 - M_l)\right]}{\sum (1 - M_l)} \tag{4.7}$$

where the first item enhances the cross-attention activation within the mask region, while the second item penalizes the cross-attention activation outside the mask region. In addition to cross-attention, the self-attention maps, which capture pixel-wise alignments on the latent feature map z_t, provide complementary spatial constraints. To harness this, I extract the self-attention map A_S^p for each pixel p highlighted in M from a latent scale of 32×32, following (Nguyen et al., 2023). I then define the background component $A_S^{p,B}$ of the self-attention map for pixel p as:

$$A_S^{p,B} = A_S^p \cdot (1 - M) \tag{4.8}$$

Inspired by Phung et al. (2024), I penalize excessive attention alignments between pixels inside and outside the region defined by M. The resulting self-attention regularization is expressed as:

$$L_{SA} = \frac{\sum A_S^{p,B}}{\sum (1 - M)} \tag{4.9}$$

This regularization not only reinforces the spatial layout during generation but also contributes to more accurate mask annotation, as validated by my ablation study (see Chapter 4.4.5). During the sampling process, given the noisy latent z_t at timestep t, layout guidance refines the latent by minimizing the combined losses through gradient descent:

$$\hat{z}_t \leftarrow z_t - \eta \bigtriangledown_{z_t} \left(L_{CA} + L_{SA}\right) \tag{4.10}$$

where η is a learning rate controlling the guidance effect, which is empirically set to

20. Since classifier guidance has been shown to be most effective during the early stages of sampling (Couairon et al., 2023), I apply this optimization only in the first 50% of the timesteps. After optimization, $\hat{z}_t$ replaces z_t in Eq. (4.2) to infer z_{t-1} in the subsequent denoising step.

Mask annotation

While layout guidance effectively controls the spatial configuration of the generated dermoscopic images based on the input mask, it does not guarantee a perfect pixel-level match between the generated image and the original mask (see Chapter 4.4.4 for illustration), as the shape and size of the lesion are also semantically affected by the text prompt. To address this limitation, I propose to simultaneously generate dermoscopic lesion-mask pairs by exploiting both cross-attention and self-attention maps during sampling.

My approach builds on the observation that self-attention maps can sharpen the boundaries defined by cross-attention maps (Nguyen et al., 2023; Khani et al., 2024). To extract the mask corresponding to the generated lesion, I first compute an intermediate attention map A_{CS} by multiplying the processed self-attention map with the cross-attention map, then applying Otsu's thresholding method (Otsu et al., 1975):

$$A_{CS} = (A_{S_{\hat{t}}})^{\tau} \cdot A_{C_{\hat{t}}}, \hat{M} = Otsu\,(A_{CS}) \tag{4.11}$$

where $A_{C_{\hat{t}}}$ is the cross-attention map for the P-Token *<lesion>*, and $A_{S_{\hat{t}}}$ is the corresponding self-attention map, both extracted at timestep $\hat{t}$. The exponent τ, set to 4, is used to sharpen the self-attention map and enhance contrast, as supported by (Nguyen et al., 2023). I set $\hat{t}$ to the final sampling step (i.e., $\hat{t} = 0$) since semantic and spatial information accumulates progressively during the denoising process (Nguyen et al., 2024a). Finally, I apply post-processing operations—dilation followed by erosion—to the mask $\hat{M}$ to fill any small holes and smooth the boundaries, as illustrated in Fig. 4.1(c).

4.4 Experiments and Results

To evaluate the effectiveness of my proposed method, I perform a series of experiments that address three key aspects: (1) the general quality of the generated images, (2) the impact of generated images on downstream classification tasks, and (3) the utility of generated lesion-mask pairs for segmentation performance. I demonstrate the setup and results of each experiment in the corresponding sections, and provide details of implementation information and evaluation metrics in the supplementary materials.

4.4.1 Dataset

For both training and evaluation, I utilize the publicly available International Skin Imaging Collaboration (ISIC) 2018 dataset (Codella et al., 2019). This dataset comprises 2594 annotated lesion-mask pairs (Task 1), which I use for both region-aware finetuning and training of segmentation models. In addition, the dataset includes 10015 dermoscopic images spanning seven diagnostic categories (Task 3), also known as HAM10000 (Tschandl et al., 2018). The included skin lesion diagnoses consist of melanoma (MEL), nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AKIEC), benign keratosis (BKL), dermatofibroma (DF), and vascular lesion (VASC). These images enable us to assess the diversity of generated images across different diagnoses and to train robust multi-class classifiers.

In addition, to evaluate the generalization of downstream classifiers, I leverage the Derm7pt dataset (Kawahara et al., 2018), which comprises 1011 dermoscopic images encompassing both melanoma and nevus cases. For segmentation assessment, I employ the PH2 dataset (Mendonça et al., 2013) that consists of 200 dermoscopic images with lesion mask annotations. These external datasets allow us to robustly measure model performance across different benchmarks, thereby demonstrating the reliability of my synthetic data.

4.4.2 General generation quality

I first evaluate the general quality of the generated dermoscopic images using standard visual metrics, including the Fréchet Inception Distance (FID), Multi-scale Structural Similarity Index (MS-SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). Specifically, I compute the FID and MS-SSIM between the generated images and real images from the testing set of Task 1 to assess generation fidelity, while LPIPS is used to capture generation diversity by measuring perceptual differences within the generated set.

For comparison, I adopt the SL-StyleGAN (Qin et al., 2020), PGAN (Baur et al., 2018), DreamBooth (Ruiz et al., 2023a), and traditional SD-finetuning techniques as baselines. All models are trained on the training set of Task 1. After training, I synthesize 1000 images per model for evaluation. For DreamBooth and SD-finetuning, I use the prefix "An image of lesion on skin." as the text prompt. I deploy my model, which is trained with my proposed DermPrompt, to generate images using the new DermPrompt and the prefix, respectively. This comparison allows us to verify that the integration of DermPrompt significantly enhances the representations learned by SD and reduces ambiguity.

Table 4.1: General generation quality of different methods

Methods	FID ↓	MS-SSIM ↑	LPIPS ↑
SL-StyleGAN	9.691	0.345	0.561
PGAN	86.174	0.371	0.365
SD-Finetune	12.764	0.393	0.548
DreamBooth	29.042	**0.403**	0.478
Mine-prefix	<u>9.267</u>	0.381	<u>0.569</u>
Mine-DermPrompt	**8.964**	<u>0.396</u>	**0.587**

Figure 4.3 shows representative qualitative results from each model, and quantitative evaluation results are summarized in Table 4.1. Notably, the images generated by GAN-based models often suffer from blurriness and lack structural details, which is reflected in their relatively lower MS-SSIM scores. Traditional SD-finetuning achieves higher MS-SSIM values but inferior FID and LPIPS scores compared to SL-StyleGAN; such performance is hypothesized as the result of the limitations of simple text prompts. The DreamBooth attains the highest MS-SSIM score, albeit

at the expense of FID and LPIPS due to its design focus on preserving common features. A similar observation was made in other medical image modalities (Chambon et al., 2022a). In contrast, my method, which incorporates DermPrompt in both training and sampling, achieves the best FID and LPIPS scores and the second-best MS-SSIM score with only a marginal gap. Furthermore, when sampling with the prefix, my method still outperforms both DreamBooth and SD-finetuning, demonstrating that training with DermPrompt significantly enhances the semantic representations learned by the model. It is important to note that in this experiment, only text inputs are used as conditions.

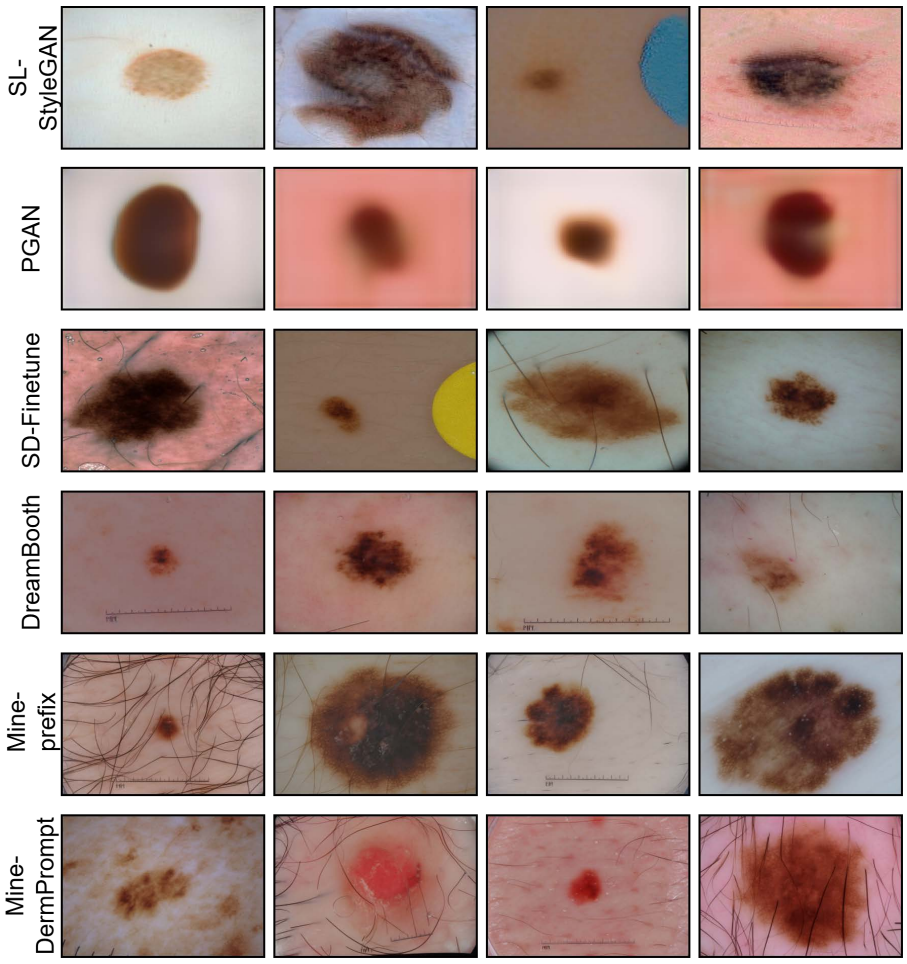


Figure 4.3: Samples of synthetic dermoscopic images generated by different methods

4.4.3 Dataset Augmentation for Multi-class Classification using Controllable Generation

In this experiment, I evaluate the ability of generative models to capture and reproduce the diagnostic characteristics of dermoscopic images, thereby enhancing multi-class classification performance. To this end, I train my proposed generative method and the baselines introduced in Chapter 4.4.2 using the HAM10000 training set. The training of generative models focuses on learning the diagnostic features of skin lesions across different diagnoses. Note that I train the adopted generative model to generate synthetic images for the six minority categories (MEL, BCC, AKIEC, BKL, DF, and VASC), while excluding the NV class because NV samples already dominate the HAM10000 dataset, constituting 66.9% of all images. my generative pipeline targets minority-class imbalance, and further augmenting an already overrepresented class would not benefit downstream classification. A complementary experiment are conducted to confirm this argument, showing that adding synthetic NV images does not improve classifier performance (detailed in the Supplementary Materials). Since PGAN and SL-StyleGAN are unconditional, I train them separately in each category. In contrast, DM-based models are trained jointly across all six categories in a T2I manner, thereby realizing text-conditioned controllable generation for different categories. After training, I synthesize 1000 images per class using each model, forming an augmentation set comprising a total of 6000 synthetic images.

Representative samples for each diagnostic class generated by different models are presented in Fig. 4.4. Qualitative observations reveal that GAN-based models struggle to generate high-fidelity images, and the diagnostic characteristics are not fully presented. Moreover, the bubble-like textures present in MEL and DF images generated by PGAN appear to replicate bubbles in the training data, indicating that the model failed to correctly recognize diagnostic features from training data. On the other hand, images generated by DM-based models demonstrate high fidelity, in which the class-specific lesion characteristics are better presented compared to those generated by GAN-based models.

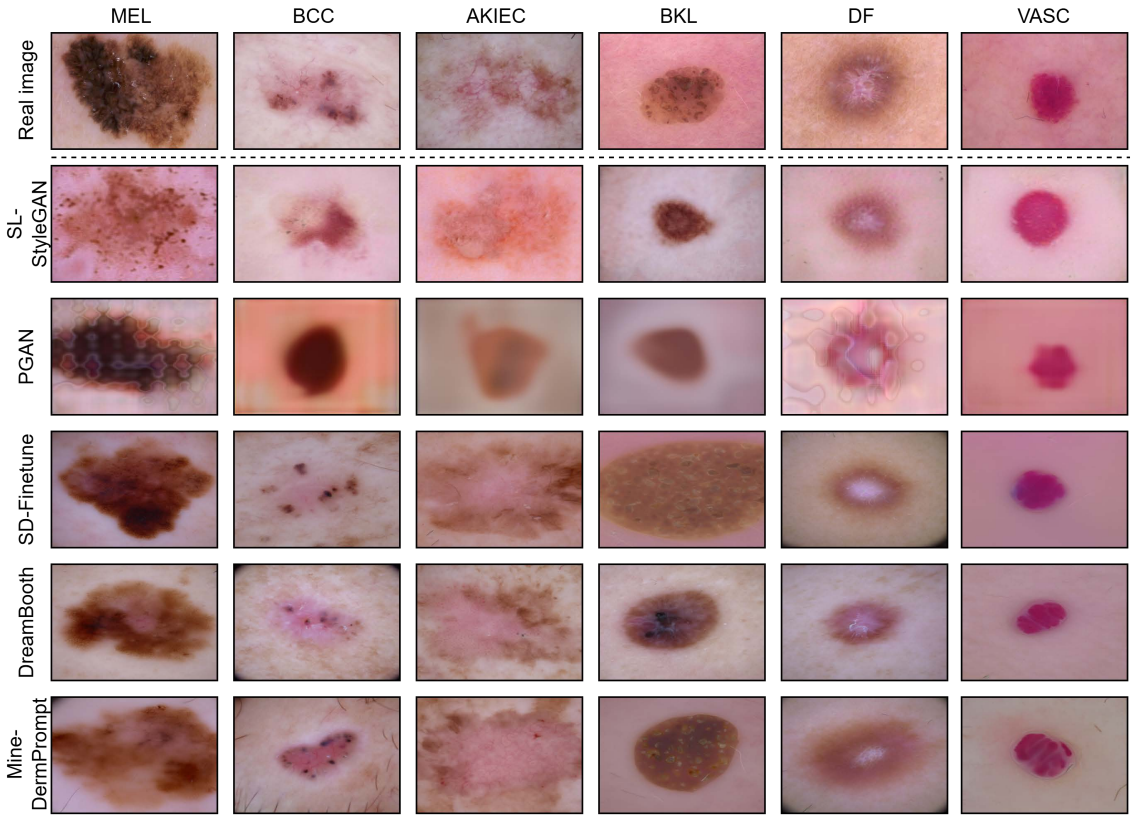


Figure 4.4: Qualitative comparison of synthetic dermoscopic images for different diagnostic categories generated by different methods

To quantify the impact of the augmented data on classification performance, I train four classifiers-VGG16, ResNet18, DenseNet121, and vision transformers (ViT)-using the original training set of HAM10000 (10015 images) and then again with the training set augmented by the synthetic set (total size 16015 images). All classification backbones were initialized with ImageNet-pretrained weights from Torchvision, with only the final classification layer reinitialized for the 7-class dermoscopic task. Classifier performance is evaluated on the HAM10000 testing set using micro-scale precision, recall, and F1 score. The results summarized in Table 4.2 indicate that the augmented datasets improve performance across all classifiers. Among them, the dataset augmented by my method presents the best diagnostic characteristics, as the ResNet18, DenseNet121, and ViT achieve the highest precision, recall, and F1 scores, while VGG16 falls behind the one trained on the dataset augmented by traditional SD-finetuning by a minor gap. I hypothesize that the simplistic design of VGG16 has constrained the utilization of the additional variation and complexity

introduced by the augmented data, resulting in the gap between my method and SD-finetuning. On average, my method enhances classifier performance by 2.03% in precision, 2.61% in recall, and 2.32% in F1 score.

In addition, Table 4.3 provides a class-wise comparison of the F1 scores between classifiers trained on the original dataset and the enlarged dataset augmented by my method. In comparison with the original training set, when trained on the augmented dataset, VGG16 exhibits improvement in five categories, ResNet18 and DenseNet121 present improvement in 6 categories, and ViT shows improvement in six categories while maintaining parity in the VASC class. Significantly, all classifiers display increased F1 scores in the MEL class, the most lethal skin cancer, with the highest improvement of 12.92% demonstrated by VGG16. Furthermore, the F1 scores of all classifiers in the NV class also increase despite the absence of synthetic images in this class in the augmented set, suggesting that augmenting with high-quality synthetic images for specific classes can have a positive transfer learning-esque effect on overall classification performance.

To further assess the generalizability of the incorporated classifiers trained on the original HAM10000 dataset and various augmented versions, I apply these pre-trained classifiers directly on the Derm7pt dataset. Note that I exclude images from the melanosis and miscellaneous classes since they lack counterparts in HAM10000, resulting in 987 images for evaluation. The average performance across all four classifiers is presented in Table 4.4. The results indicate that augmenting the training data reliably enhances classifier performance on the external dataset. Notably, the dataset augmented by my method achieves the highest score across all three metrics, with an increment of 5.18% in precision, 3.52% in recall, and 6.29% in F1 score compared to the original dataset, underscoring the robust generalizability. These results demonstrate that the synthetic images produced by my method capture a broad range of diagnostically relevant features and thereby support the development of more robust and widely applicable classification models.

Table 4.2: Performance of classifiers trained on original dataset and datasets augmented by different methods on the ISIC 2018 dataset

Metrics	Classifier	Original	SL-StyleGAN	PGAN	SD-Finetune	DreamBooth	Mine-DermPrompt
Precision↑	VGG16	0.782	0.790	0.793	**0.807**	0.797	0.805
	ResNet18	0.797	0.796	0.800	0.803	0.802	**0.812**
	DenseNet121	0.811	0.801	0.817	0.823	0.798	**0.826**
	ViT	0.834	0.831	0.838	0.840	0.825	**0.846**
Recall↑	VGG16	0.775	0.792	0.791	**0.808**	0.795	0.805
	ResNet18	0.791	0.794	0.799	0.805	0.806	**0.816**
	DenseNet121	0.813	0.804	0.822	0.822	0.802	**0.828**
	ViT	0.834	0.832	0.841	0.843	0.831	**0.847**
F1 Score↑	VGG16	0.775	0.788	0.786	**0.803**	0.793	0.801
	ResNet18	0.792	0.794	0.797	0.802	0.802	**0.811**
	DenseNet121	0.811	0.804	0.816	0.819	0.798	**0.825**
	ViT	0.830	0.832	0.837	0.840	0.824	**0.845**

Table 4.3: Class-wise F1 score of various classifiers trained on the original dataset and dataset augmented by my method

Classifier	Training data	MEL	NV	BCC	AKIEC	BKL	DF	VASC
VGG16	Original	0.534	0.878	0.689	**0.564**	0.686	0.583	**0.592**
	Augmented	**0.603**	**0.901**	**0.701**	0.558	**0.707**	**0.610**	0.582
ResNet18	Original	0.609	0.886	0.705	**0.588**	0.674	0.641	0.644
	Augmented	**0.622**	**0.902**	**0.718**	0.561	**0.720**	**0.674**	**0.667**
DenseNet121	Original	0.608	0.898	0.725	0.552	0.720	0.717	**0.754**
	Augmented	**0.657**	**0.910**	**0.749**	**0.585**	**0.723**	**0.720**	0.712
ViT	Original	0.647	0.910	0.721	0.567	0.767	0.783	0.733
	Augmented	**0.671**	**0.919**	**0.746**	**0.576**	**0.781**	**0.868**	0.733

Table 4.4: Averaged performance of classifiers trained on original dataset and datasets augmented by different methods on the Derm7pt dataset

Methods	Precision↑	Recall↑	F1 Score↑
Original	0.676	0.653	0.604
SL-StyleGAN	0.696	0.648	0.614
PGAN	0.693	0.661	0.610
SD-Finetune	<u>0.710</u>	<u>0.662</u>	<u>0.618</u>
DreamBooth	0.708	<u>0.662</u>	0.609
Mine-DermPrompt	**0.711**	**0.676**	**0.642**

4.4.4 Dataset Augmentation for Segmentation using Controllable Generation

My method enables SD to learn multi-scale representations from dermoscopic images and achieve a controllable generation. Although training with my proposed DermPrompt allows SD to synthesize images guided by high-level semantic conditions, the granularity of control remains limited by the inherent expressiveness of text. As illustrated in Fig. 4.5, while my method can generate a heart-shaped lesion according to the given DermPrompt, it fails to capture all the fine-grained shape characteristics present in the real image. This limitation arises because a text prompt cannot encode the full complexity of spatial details. To overcome this issue, I incorporate semantic generation to achieve fine-grained control over the layout of the generated dermoscopic images. In my pipeline, synthetic lesion-mask pairs are produced simultaneously, and these pairs are subsequently used to augment the training dataset for skin lesion segmentation. Augmenting the segmentation dataset with high-quality synthetic samples helps improve the performance of segmentation models by increasing variability and reducing overfitting.

I compare the controllable semantic generation ability of my proposed method with four baselines, divided into two categories. The first category comprises image translation models that require paired training data, for which I adopt Pix2PixHD (Wang et al., 2018) and ControlNet (Zhang et al., 2023a) as representatives. The second category includes training-free methods that use masks as guidance during sampling in text-to-image diffusion models, in which I consider ZestGuide (Couairon et al., 2023) and Attn-Refocus (Phung et al., 2024). For the training-based baselines,

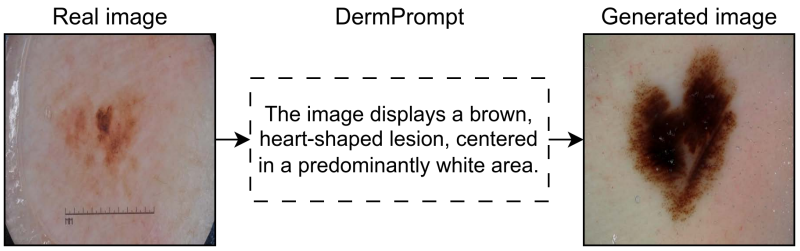


Figure 4.5: Limitation of DermPrompt in layout-guided generation. The intrinsic limitation of the text guidance cannot provide fine-grained layout guidance, meaning that the layout of the generated image does not resemble the real image.

models are trained on the training set of Task 1. In contrast, the training-free methods are directly applied to the SD model fine-tuned with my proposed region-aware attention loss.

A key advantage of my method is that it simultaneously generates lesion-mask pairs. Unlike baselines that require a prepared mask as an input condition, my approach produces masks that not only mimic the input masks but also introduce variations through the inherent variability during the denoising process. This ability allows my method to create novel mask configurations that do not exist in the original dataset. To leverage this advantage, I condition my model on real masks from the training set of Task 1 and generate 2500 synthetic lesion-mask pairs. I then use these new masks as conditions for the baseline methods to generate additional dermoscopic images. This design helps prevent information leakage, as training-based models might otherwise simply replicate training images when given their corresponding masks during inference.

Figure 4.6 shows qualitative comparisons of the synthetic lesion-mask pairs generated by different methods. Notably, only my method produces both new images and corresponding masks, whereas the other models generate new images conditioned on pre-existing masks. Among the training-free methods, ZestGuide and Attn-Refocus do not adhere strictly to the mask layout, resulting in minor shape disparities in the generated lesions. Although the training-based models (Pix2PixHD and ControlNet) capture layout details and produce lesions with accurate shapes, they require significant training time and computational resources. In contrast, my method efficiently produces accurate lesion-mask pairs in a training-free manner, making it an

attractive solution for segmentation dataset augmentation.

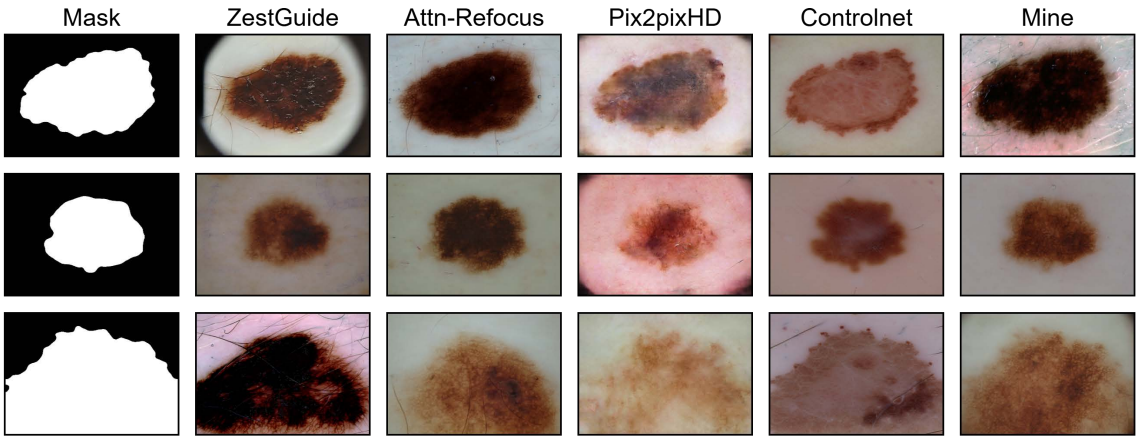


Figure 4.6: Qualitative comparison of different methods in the generation of dermoscopic lesion-mask pairs

I further quantify the quality of synthesized lesion-mask pairs by training four segmentation models, U-Net (Ronneberger et al., 2015), AttenU-Net (Oktay et al., 2018), DCSAU-Net (Xu et al., 2023) and XBound-Former (Wang et al., 2023d), on the original training set of ISIC 2018 Task 1 and the training sets augmented by synthetic lesion-mask pairs. The performances of the segmentation models are evaluated by Dice coefficient (Dice) and Intersection over Union (IoU) on the test set of ISIC 2018 Task 1, and the evaluation results are listed in Table 4.5. To verify the effect of dataset size on augmentation benefits, experiments are performed on two scales: one with 1000 randomly sampled images (denoted as S_{1k}) and another using the full training set (denoted as $S_{2.5k}$). It is worth noting that the averaged pairwise IoU score between the mask produced by my method and the real mask taken as conditioning information is 0.67 ± 0.21, indicating structural similarity while avoiding exact replication. The evaluation results in Table 4.5 show that all segmentation models benefit from the augmented datasets. Remarkably, the datasets augmented by my proposed method yield the most significant performance improvements. In particular, U-Net, AttenU-Net, and DCSAU-Net achieve the highest segmentation scores when trained on the augmented datasets at both scales. The XBound-Former shows the highest score on S_{1k} augmented by the ControlNet. I hypothesize that this behavior is induced by the design of XBound-former, which focuses on edge detection, while ControlNet presents strict adherence to the mask layout, resulting

Table 4.5: Performance of segmentation models trained on the original and augmented datasets on the ISIC 2018 dataset

Metrics & Base Dataset	Segmentation model	Original	Pix2PixHD	ControlNet	Zestguide	Attn-refocus	Mine
IoU↑&S_{1k}	U-Net	0.716	<u>0.733</u>	0.722	0.720	0.729	**0.737**
	AttenU-Net	0.719	0.726	<u>0.743</u>	0.730	0.738	**0.745**
	DCSAU-Net	0.723	0.731	<u>0.742</u>	0.727	0.740	**0.756**
	XBound-Former	0.773	<u>0.790</u>	**0.796**	0.776	0.784	0.788
IoU↑&$S_{2.5k}$	U-Net	0.731	0.749	<u>0.754</u>	0.742	0.753	**0.760**
	AttenU-Net	0.733	0.736	<u>0.754</u>	0.738	0.752	**0.763**
	DCSAU-Net	0.768	<u>0.778</u>	0.775	0.764	0.772	**0.788**
	XBound-Former	0.810	0.819	<u>0.823</u>	0.816	0.811	**0.826**
Dice↑&S_{1k}	U-Net	0.831	<u>0.844</u>	0.835	0.842	0.841	**0.845**
	AttenU-Net	0.833	0.838	<u>0.850</u>	0.840	0.847	**0.851**
	DCSAU-Net	0.836	0.841	<u>0.848</u>	0.839	0.847	**0.858**
	XBound-Former	0.868	<u>0.881</u>	**0.884**	0.872	0.876	0.880
Dice↑&$S_{2.5k}$	U-Net	0.841	0.854	<u>0.857</u>	0.848	<u>0.857</u>	**0.861**
	AttenU-Net	0.843	0.845	<u>0.857</u>	0.846	0.856	**0.864**
	DCSAU-Net	0.867	<u>0.874</u>	0.871	0.865	0.869	**0.880**
	XBound-Former	0.884	0.890	<u>0.891</u>	0.883	0.885	**0.895**
Training-free		-	×	×	√	√	√

in the benefits to edge-sensitive architectures of XBound-former. On the other hand, the XBound-Former presents the best performance on $S_{2.5k}$ augmented by my method. Overall, my method presents the best effect on the augmentation of the segmentation dataset for dermoscopic images over all the investigated methods, improving the performance of downstream segmentation models on both scales. Across all segmentation models, my method improves IoU and Dice scores by 3.26% and 1.96% on S_{1k}, and by 3.16% and 1.90% on $S_{2.5k}$, respectively.

In addition, I validate the robustness and generalizability of my synthetic lesion-mask pairs by extending my evaluation to the external PH2 dataset. The four adopted segmentation models, pretrained on the original ISIC 2018 dataset at the $S_{2.5k}$ scale and various augmented versions, are directly applied to the PH2 dataset to assess their performance. The averaged IoU and Dice scores across the four segmentation models are presented in Table 4.6. The results consistently demonstrate that the augmented dataset enhances the generalization performance of segmentation models. Impressively, models trained with the dataset augmented by my method achieve the best performance, with gains of 3.17% in IoU and 1.90% in Dice compared with models trained solely on the original dataset. These findings underscore the ability of my synthetic data to capture diverse lesion characteristics and enhance model generalization across different acquisition environments.

Table 4.6: Averaged performance of segmentation models on the PH2 dataset

Methods	IoU↑	Dice↑
Original	0.821	0.896
Pix2PixHD	0.838	0.907
ControlNet	0.842	0.909
Zestguide	0.827	0.900
Attn-refocus	0.831	0.903
Mine	**0.847**	**0.913**

4.4.5 Ablation study

I ablate the contributions of the various design choices, including each of my novel technical contributions, in my method by performing extensive ablation studies. The effect of DermPrompt on the quality of the representations learned by SD is dis-

cussed in Chapter 4.4.2. In this section, I further analyze the contributions of other components, including the region-aware finetuning and the guidance provided by the cross-attention and self-attention losses in the lesion-mask generation pipeline.

Region-aware finetuning

The core distinction between my proposed region-aware finetuning and the traditional finetuning of SD lies in the incorporation of a region-aware loss that leverages the cross-attention mechanism. To enable a more controllable generation of dermoscopic images, it is crucial that SD develops a nuanced semantic understanding of the image components during finetuning. In T2I generation, such semantic understanding can be reflected by the cross-attention maps between the text tokens and latent feature maps (Tang et al., 2023b). Therefore, I extract and visualize the cross-attention maps between the P-Tokens and latent feature maps corresponding to the dermoscopic images synthesized by SD undergoing region-aware finetuning and traditional finetuning in Fig. 4.7(a), respectively. While both methods generate high-fidelity dermoscopic images, the one finetuned without region-aware loss fails to highlight the regions corresponding to *<lesion>* and *<skin>* in the cross-attention maps, indicating the semantic alignment is not constructed. In contrast, the cross-attention maps explicitly display these regions when the region-aware finetuning is applied. Moreover, in the cross-attention maps of the P-Token *<lesion>*, attention values are high in the central region, where the lesion color is deep in the generated image; while attention values become lower in the near-central region, and the lesion shows a lighter color in the corresponding region, presenting a transition region between lesion and skin. An inverted value distribution can be observed in the cross-attention map of *<skin>*. These observations indicate the model has developed a semantic recognition of the generated content, and the built semantic alignments also lay a foundation for controllable generation, as shown in Chapter 4.4.4 and Chapter 4.4.6.

Cross-attention guidance

The cross-attention guidance in the lesion-mask generation pipeline contributed to the diversity of generated images by providing rich layout information from real masks. A highly diverse dataset can benefit the training of segmentation models as they can learn abundant geometric patterns of lesions. I verify the effect of cross-attention guidance on generation diversity by calculating the LPIPS score within the generated set of masks. As listed in Table 4.7, the LPIPS score increases from 0.172 to 0.403 when the cross-attention is applied during inference, indicating that the generated masks are more perceptually diverse. Moreover, I train the four segmentation models introduced in Chapter 4.4.4 on the whole training set $S_{2.5k}$ augmented by the variant of my method without cross-attention guidance. The results in Table 4.7 reveal that the segmentation models can benefit more from datasets with higher diversity, as all the segmentation models show higher scores when trained on the datasets augmented by my method with cross-attention guidance.

Table 4.7: Ablation results of the effects of cross-attention guidance in the lesion-mask generation pipeline

Metrics	Segmentation models	w/o guidance	w/ guidance (Mine)
IoU↑	U-Net	0.745	**0.760**
	AttenU-Net	0.745	**0.763**
	DCSAU-Net	0.771	**0.788**
	XBound-Former	0.821	**0.826**
Dice↑	U-Net	0.851	**0.861**
	AttenU-Net	0.851	**0.864**
	DCSAU-Net	0.868	**0.880**
	XBound-Former	0.889	**0.895**
LPIPS↑		0.172	**0.403**

Self-attention guidance

The self-attention guidance applied in the lesion-mask generation pipeline does not present a direct impact on the synthetic images. Instead, I find it contributes to preventing fails in mask thresholding, thereby increasing the accuracy of generated masks. Since my training-free lesion-mask generation pipeline aims for a high-efficiency and lightweight generation, I utilize the Otsu thresholding method that

automatically thresholds images according to pixel values instead of deploying an additional segmentation model. Benefiting from the processed attention map A_{CS} with clear contrast and sharp edges, the Otsu thresholding method can produce accurate masks in most cases. However, it faces challenges in some complex situations. As shown in the first row of Fig. 4.7(b), when the model generates an artificial mark, the A_{CS} map of $<lesion>$ presents high values in the lesion region, lower values in the skin region, and the lowest values in the mark region, respectively. Such distributional disparity originates from the self-attention attention maps, since the pixels in the lesion region have different self-attention values between pixels in the skin region and mark region. The disparity can be easily perceived by humans, but is challenging for the thresholding algorithm. To address this, I find applying self-attention guidance can regularize the self-attention maps and make the A_{CS} map more cohesive, thereby producing accurate masks, as illustrated in the second row of Fig. 4.7(b).

4.4.6 Further applications in controllable generation

My proposed method establishes a pipeline for the controllable generation of dermoscopic images by introducing the DermPrompt and region-aware finetuning, with its application in the semantic generation discussed in Chapter 4.4.4. Apart from semantic generation, I highlight that my model is a versatile controllable generator and its controllability is reflected in multiple aspects, boosting the applicability of my proposed method for various tasks related to dermoscopic images and making it more generalizable than existing methods.

Text-guided attribute customization. I have verified that finetuning pretrained SD with my proposed attribute-rich DermPrompt can enhance the representation learning of the model. Although the DermPrompt cannot offer fine-grained layout guidance for semantic generation due to the intrinsic limitations of textual guidance, it can provide high-level guidance to generate dermoscopic images with customized attributes. Figure 4.8 presents the guidance from the DermPrompt in different aspects, including skin color, lesion color, and other marks such as dyes. The generated

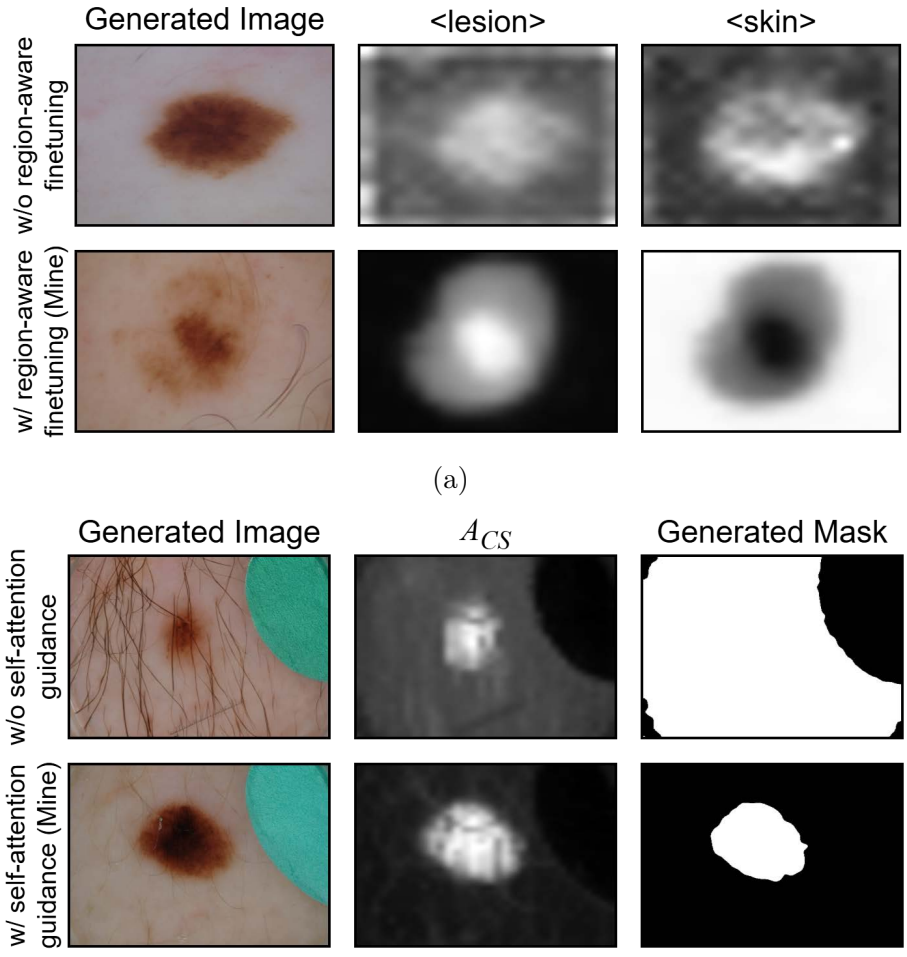


Figure 4.7: Qualitative results for ablation studies: (a) visualization of generated images and cross-attention maps of the P-Tokens from models finetuned with/without region-aware finetuning; (b) visualization of generated images and processed attention maps A_{CS} of P-Token *<lesion>* from models with/without self-attention guidance.

images with customized images can act as supplementary data for the training of various downstream models, such as skin tone classifiers, and hair/marker removers.

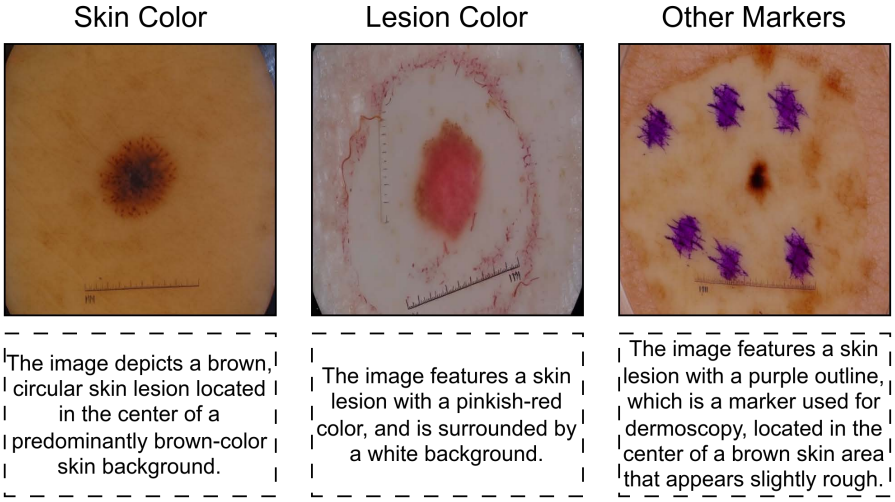


Figure 4.8: Text-guided attribute customization. My proposed DermPrompt can achieve high-level controllable generation including but not limited to skin color, lesion color, and other marks.

Bounding-box conditioned layout guidance. Region-aware finetuning helps establish the semantic visual-textual alignments between the P-Tokens (see Chapter 4.3.3 for details) and visual representations, enabling spatially controlled generation using other training-free pipelines based on DMs without requiring further finetuning. Alongside the semantic masks, spatial information can also be given by other modalities such as bounding boxes. I combine my finetuned SD with a training-free bounding-box-conditioned pipeline, BoxDiff (Xie et al., 2023), to synthesize dermoscopic images where the size and position of the lesion are constrained by the layout of bounding boxes, as shown in Fig. 4.9. The synthetic dermoscopic images with annotated bounding boxes can then support the training of downstream bounding-box detectors such as YOLO (Tian et al., 2025).

Domain-specific image editing. The established semantic alignments can not only support layout-guided controllable generation, but also facilitate the model's semantic understanding of the content in real domain-specific images, thereby enabling the editing of real dermoscopic images. As shown in Fig. 4.10, given some

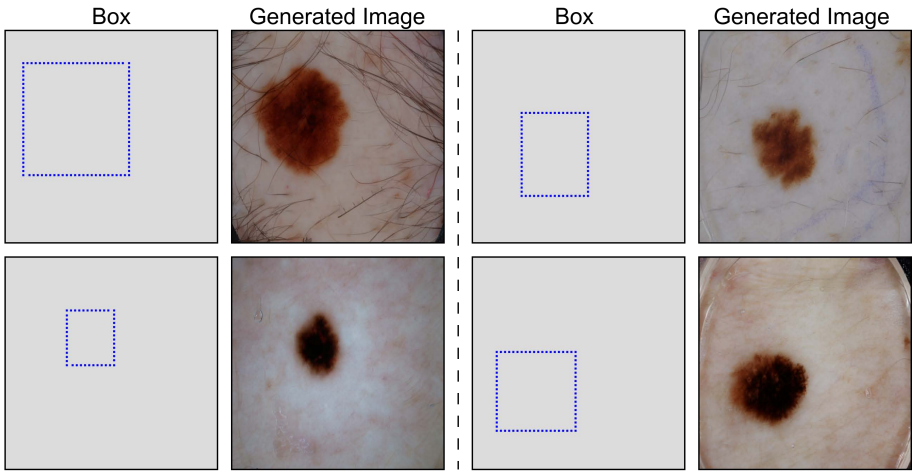


Figure 4.9: Bounding-box guided generation. My finetuned SD can be combined with the training-free pipeline BoxDiff (Xie et al., 2023) to generate dermoscopic images wherein the lesions comply with the layout of bounding boxes.

real dermoscopic images with benign lesions, I can edit the color and texture of the lesions by combining my finetuned SD model with the image editing method Null-text Inversion (Mokady et al., 2023), and render them to present the characteristics of melanoma lesions learned by the SD through finetuning. Notably, the success of Null-text Inversion is contingent on the model's semantic understanding of the given content, underscoring the robust semantic alignments built by my proposed method. Adapting image editing techniques to dermoscopic images offers a promising way to create images that combine multiple target features. For instance, one can first generate dermoscopic images with a desired skin tone and then edit the lesion to display a specific type. However, achieving this kind of compositional generation remains a significant challenge for current generative models that are designed exclusively for image generation.

4.5 Discussion

In this study, I propose a novel framework for the controllable generation of dermoscopic images based on the pretrained SD. My method demonstrates superior generation quality and generalizability. I conducted comparative experiments to evaluate the general quality of the synthetic dermoscopic images. The results indicate that my synthesized images exhibit fine-grained details, strong structural coherence, and

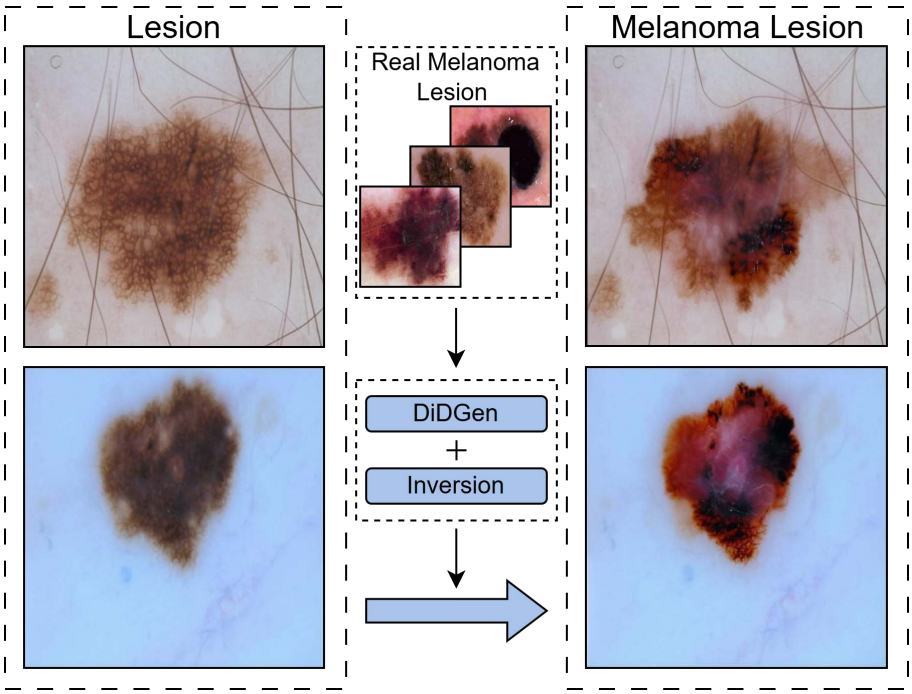


Figure 4.10: Domain-specific image editing. My finetuned SD can be combined with attention-based image editing methods, such as Null-text Inversion (Mokady et al., 2023), and edit real domain-specific dermoscopic images. For example, changing the color and texture of benign lesions to resemble the characteristics of melanoma lesions.

substantial diversity, which are critical for reducing dataset bias, enhancing model generalizability, and improving the detection of diagnostically relevant features. The improved image quality is largely attributable to the use of my DermPrompt technique, which encourages the model to establish nuanced visual-textual alignments, thereby enabling diverse text guidance as shown in Fig. 4.8. Moreover, the enhanced capturing of class-specific features benefits downstream classifiers; as these models rely on distinct features for decision-making, the precise extraction of such features by my method promotes classification accuracy, while the diversity of the synthesized images enhances classification robustness. The results listed in Table 4.2 and Table 4.4 confirm the effectiveness of my method on multi-class classifiers. The class-wise improvements of F1 scores listed in Table 4.5 reveal that my method offers robust augmentation for dermoscopic images in different categories. Across the seven categories, my method enhances each classifier's performance in at least five categories compared to the original training set. Note that the categories showcas-

ing negative effects contain limited samples for evaluation (AKIEC: 43; VASC: 35). Overall, the comprehensive enhancements provided by my method can reduce bias in underrepresented categories and contribute to the development of more robust CAD systems for use by dermatologists.

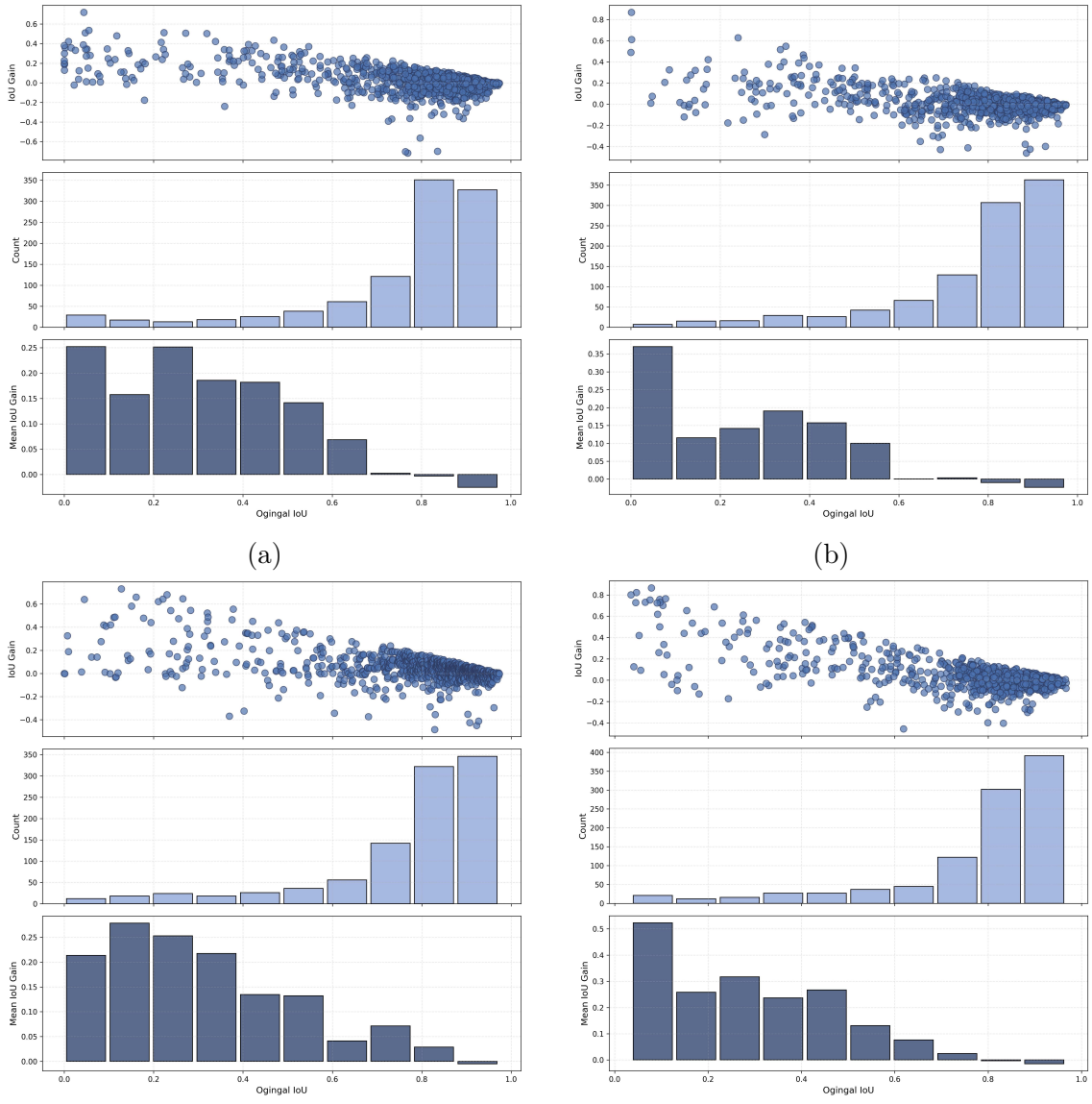


Figure 4.11: Distribution of per-model sample-wise IoU gains versus original IoU, with sample counts and mean IoU gains by interval, collected from the results of (a) U-Net; (b) AttenU-Net; (c) DCSAU-Net; (d) XBound-Former, highlighting that significant improvements occur on the poorly-segmented cases.

In addition, my method investigates the controllable generation of dermoscopic images. Beyond text-guided generation, I also explored layout-guided generation in

depth. Existing studies typically employ dedicated models for mask-to-image generation to produce lesion-mask pairs. In contrast, my approach leverages the attention mechanism to produce masks corresponding to the generated images, offering three main advantages: (1) a training-free pipeline that directly utilizes the finetuned SD model without additional adaptation or training for layout-guided generation; (2) the generation of new masks based solely on synthetic images rather than relying on existing dataset masks; and (3) the ability to generate multiple lesion-mask pairs using a single reference mask by adjusting hyperparameters during generation, demonstrating significant potential for dataset augmentation. Comparative results in Table 4.3 Table 4.6 confirm the effectiveness of my method for segmentation models. The success of my training-free pipeline for lesion-mask pair synthesis is rooted in the semantic visual-textual alignments established during my proposed region-aware finetuning. This lays the foundations for various forms of controllable dermoscopic image generation, including bounding-box-guided generation and dermoscopic image editing as shown in Chapter 4.4.6, making my method a versatile tool for generating dermoscopic images.

To emphasize that the average metric gains reported in Table 4.2 may understate the true benefits of data augmentation, I examined performance improvements within specific subsets of interest. For example, when trained on the dataset augmented by my method, an average F1 score increase of 6.71% in the MEL class across the four classifiers can be observed in Table 4.3. Compared with the overall F1 score increase of 2.32%, this larger gain for the most lethal form of skin cancer highlights the practical value of my approach in enhancing recognition of cases that demand the greatest clinical attention. Moreover, I carry out a detailed analysis of segmentation performance gains for each model when trained on my augmented dataset. Figure 4.11 presents, from top to bottom, the per-sample distribution of IoU gains, the number of samples falling into each IoU-gain interval, and the mean IoU gain per interval, with the horizontal axis representing each sample's baseline IoU under training on the original data. Although the average IoU increase across all samples appears modest, the most pronounced improvements occur on the challenging cases that the original models segmented poorly. Despite the minor performance degrada-

tion happening to samples already well-segmented, improving segmentation on these challenging lesions is critical for reducing false negatives, since under-segmentation in these cases can lead to missed or delayed diagnoses. Additionally, I perform paired t-tests comparing model predictions trained on the original dataset versus those trained on my augmented dataset to assess the statistical significance of my augmentation. The test results present a p-value of 1.35×10^{-3} for classification and a p-value of 6.64×10^{-29} for segmentation. Both p-values are less than the conventional 0.05 threshold, indicating the robustness of my augmentation.

In general, my method has markedly improved the efficiency of dermoscopic image dataset augmentation by requiring only a single finetuning process to support the generation of multiple visual modalities including image, image-mask pair, image-box pair, and edited image, reducing the need for labor-intensive annotations and accelerating dataset curation. Through the control over lesion type and skin tone, my approach also presents the potential to help mitigate bias in dermoscopic CAD systems by synthesizing data with infrequent patterns and representations, and correcting the skewed distribution of the original dataset. Concretely, supplementing synthetic images with rare lesion types can enhance the prediction accuracy of downstream classifiers on these specific symptoms; and additional synthetic images with brown or dark skin tones can help alleviate the skin tone bias and improve segmentation accuracy across various populations (Benčević et al., 2024). However, several limitations remain in this study. Firstly, the representation learned by the DMs still falls within the representation distribution of the training data, restricting the ability of my method to synthesize images not exist in the training data. For example, my method struggles to generate images jointly featured by a melanoma lesion and dark skin, since such images are not included in the training dataset. Moreover, the computational resource requirement for deploying DMs remains high, which might constrain accessibility to clinicians or researchers with limited resources.

In future work, I will extend the application of my method to the balance of skin tone representation for training data. My future experiments will further exploit the controllability introduced in this study to synthesize images featuring brown or dark skin tones. Considering the differences in lesion presentation across skin tones

(Thompson et al., 2023), it is necessary to curate additional dermoscopic images annotated with diverse lesion types and skin tones to provide authentic representations for my method, ensuring more reliable and faithful generation results. Moreover, I will explore the distillation of DMs to provide a lightweight and fast version of my method while retaining its generation performance (Kim et al., 2024; Zhou et al., 2024b). Finally, I aim to extend my method to user-taken images, such as those in the Fitzpatrick 17k dataset (Groh et al., 2021), which are more challenging to generate due to their varied structures and backgrounds. Adapting my method to user-taken images will further enhance its generalizability, thereby facilitating more robust diagnostic support in diverse clinical scenarios.

4.6 Conclusion

In this study, I present DiDGen, a novel and efficient framework for the controllable synthesis of dermoscopic images using pretrained T2I SD, tailored to enhance CAD of skin lesions. By integrating DermPrompt, which employs LLMs to generate attribute-rich and clinical-related text prompts, and a two-stage semantic synthesis pipeline featuring region-aware finetuning and a training-free layout-guided generation process, my method overcomes the shortcomings of prior generative techniques, such as poor generalizability and limited controllability. Experimental evaluations demonstrate that DiDGen produces synthetic images of superior quality, diversity, and diagnostic relevance, substantially boosting the performance of downstream classification and segmentation models. Requiring only one finetuning process, DiDGen stands as a practical and powerful tool for augmenting dermoscopic datasets.

4.7 Appendix

4.7.1 Evaluation Metrics

In this study, I evaluate the generated dermoscopic images regarding the general and clinical quality with a suite of well-established metrics. To quantify the fidelity of generated images, I compute their FID and MS-SSIM scores. The FID score

measures the distance between the Inception-net feature distributions of generated and real images:

$$FID = \|\mu_r - \mu_g\|_2^2 + Tr\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right) \tag{4.12}$$

where μ and Σ are the mean and covariance of image features extracted by a CNN model Inception-v3, subscriptions r and g denote real and generated images.

While the FID evaluates the overall similarities between distributions of real and generated images, MS-SSIM captures the structural similarity across multiple scales. For a pair of real and generated images r, g, the MS-SSIM score is calculated by:

$$MS - SSIM(r,g) = [l_M(r,g)^{\alpha_M}] \prod_{j=1}^{M} [c_j(r,g)\, s_j(r,g)]^{\beta_j} \tag{4.13}$$

where the luminance term l_M, contrast term c_j, and structure term s_j are given by:

$$L_M(r,g) = \frac{2\mu_{r,M}\mu_{g,M} + C_1}{\mu_{r,M}^2 + \mu_{g,M}^2 + C_1} \tag{4.14}$$

$$c_j(r,g) = \frac{2\sigma_{r,j}\sigma_{g,j} + C_2}{\sigma_{r,j}^2 + \sigma_{g,j}^2 + C_2} \tag{4.15}$$

$$s_j(r,g) = \frac{\sigma_{rg,j} + C_3}{\sigma_{r,j}\sigma_{g,j} + C_3} \tag{4.16}$$

where $\mu_{\cdot,j}$ and $\sigma_{\cdot,j}$ are mean and standard deviation of an image at scale j, and $\sigma_{rg,j}$ denotes the covariance between r and g at scale j. M is the coarsest scale. α_M and β_j are weights. C_1, C_2, and C_3 are small constants to stabilize division.

I also adopt LPIPS values between the generated images to evaluate their diversity. LPIPS measures the perceptual similarity between images using features extracted by deep neural networks (e.g., AlexNet):

$$LPIPS(r,g) = \sum_l \frac{\sum_{h,w} \|w_l \odot (\hat{r}_{l,h,w} - \hat{g}_{l,h,w})\|_2^2}{H_l W_l} \tag{4.17}$$

where $\hat{r}_{l,h,w}$ and $\hat{g}_{l,h,w}$ are normalized activation at location (h,w) of the feature maps extracted from layer l of the neural networks. The size of the feature maps is $H_l \times W_l$. w_l is the learned weight for layer l.

In addition, I conduct experiments for downstream tasks to evaluate the clinical quality of the generated images. I adopt the Precision, Recall, and F1 score to describe the performance of classifiers. Precision reflects the accuracy of positive predictions and penalizes false positives, Recall captures the model's ability to detect all relevant cases and penalizes false negatives, and F1 score balances both false positives and false negatives.

$$Precision = \frac{TP}{TP + FP} \tag{4.18}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.19}$$

$$F1 = 2\frac{Precision \cdot Recall}{Precision + Recall} \tag{4.20}$$

For segmentation, I utilize the Dice and IoU to evaluate the overlap between two binary masks, reflecting the performance of segmentation models.

$$Dice = \frac{2\,|A \cap B|}{|A| + |B|} \tag{4.21}$$

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{4.22}$$

where A and B are predicted foreground region and ground truth foreground region, respectively.

4.7.2 Implementation details

All the experiments are conducted on a computer cluster with an NVIDIA A100 GPU. Implementation details of model training and evaluation presented in each experiment are listed as follows.

General generation quality.

I utilize images from ISIC 2018 Task 1 as training data, and resize them to various sizes for training different deep generative models. For DM-based model, input images are resized to 512×512 pixels. I adopt the pre-trained SD v2.1 checkpoint as the backbone model, and finetune all the models for 20,000 steps with a batch

size of 4 and a learning rate of 1×10^{-5}, using the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay=1×10^{-2}). For GAN-based models, the input images are resized to 256×256 pixels. I also train the model for 20,000 steps with a batch size of 32 and a learning rate of 1×10^{-3}, using the same optimizers as the DM-based models. During evaluation, synthetic images generated by different models are resized to 256×256 pixels

Classification Augmentation.

I train the baseline models for generating dermoscopic images in the same setting as the first experiment, but using the HAM10000 dataset for training data. Then I evaluated the effect of synthetic images on six lesion classes using four pretrained classifiers: VGG16, ResNet18, DenseNet121, and ViT, with a unified input image size of 224×224. The ViT is configured in its "Base" variant with a 16×16 patch size. All the classifiers are pretrained on the ImageNet dataset, and the weights are taken from the PyTorch platform (https://pytorch.org). During the training of all classifiers, I set a batch size of 128, an SGD optimizer with an initial learning rate of 1×10^{-3}, and a momentum of 0.9. An early stopping with a patience of 10 epochs is set concerning the accuracy on the validation set.

Segmentation Augmentation.

I adopt four baseline models for the semantic generation of dermoscopic images. Among them, ZestGuide and Attn-Refocus are training-free methods, while Pix2PixHD and ControlNet require dedicated training processes. Thus, I utilize the image-mask pairs in ISIC 2018 Task 1 dataset for training. For evaluation, I train four segmentation models: U-Net, AttenU-Net, DCSAU-Net, and XBound-Former on the original and augmented datasets. Models are trained with a batch size of 16, a learning rate of 1×10^{-3}, using the Adam optimizer. An early stopping with a patience of 20 epochs is set concerning the accuracy on the validation set.

CHAPTER 5

Mitigating Low-Contrast Bias in Skin Lesion Segmentation using a Dual-Branch Controllable Diffusion Model

Following the controllable dermoscopic image synthesis framework introduced in Chapter 4, this chapter further deepens the application of controllable DMs to address a clinically meaningful and previously underexplored problem: the source of bias in skin lesion segmentation. Chapter 4 demonstrated that diffusion-based models can effectively alleviate data scarcity and class imbalance in dermoscopic datasets when equipped with region-aware control and paired image-mask synthesis capabilities. However, beyond overall data insufficiency, segmentation performance is often degraded on specific challenging subsets of images. This chapter builds upon the controllable generation mechanisms developed in Chapter 3 and Chapter 4 and investigates how they can be leveraged not merely for data augmentation, but as a principled tool for bias analysis and mitigation.

The motivation of this chapter stems from the observation that performance disparities in skin lesion segmentation are frequently attributed to skin tone variations, without adequately examining the underlying visual factors that directly impair model predictions. Through a finer-grained analysis, this chapter identifies low color contrast between lesions and surrounding skin as a dominant contributor to

segmentation bias. Such low-contrast cases are underrepresented in existing datasets and are difficult to address using conventional augmentation techniques. The goal of this chapter is therefore to develop a controllable generative framework capable of precisely manipulating lesion structure and appearance, enabling the targeted synthesis of challenging low-contrast samples for robust model training.

To achieve this goal, this chapter proposes a dual-branch controllable DM that disentangles lesion layout and style, allowing fine-grained and independent control over structural and style-related attributes. By integrating this model with the simultaneous image-mask generation strategy introduced in Chapter 4, the proposed framework efficiently produces realistic, paired dermoscopic samples tailored to low-contrast scenarios. Targeted finetuning using the synthesized data yields substantial improvements in segmentation accuracy on low-contrast subsets while preserving overall performance on standard test sets. This chapter demonstrates how controllable diffusion models can be used not only to augment data, but also to systematically diagnose and mitigate bias in medical imaging systems, reinforcing the broader theme of this thesis that fine-grained controllability is essential for reliable and clinically impactful generative modeling.

5.1 Introduction

Skin cancer represents a significant global public health challenge. Malignant melanoma, a particularly deadly form, accounted for 331,722 new cases and 58,667 deaths worldwide in 2022 (World Cancer Research Fund (WCRF), 2022). In the U.S., skin cancer is the most commonly diagnosed cancer (Guy Jr et al., 2015a,b). 107,240 new cases of melanoma are estimated in the U.S. for 2025 (Siegel et al., 2025). The associated healthcare expenditures are substantial, with an average cost of $19,427 per individual patient (Olateju et al., 2024).

Early detection is critical for improving patient outcomes. The five-year survival rate for early-stage skin cancer is 99%, dropping to 32% when diagnosed at an advanced stage (American Cancer Society (ACS), 2023). Dermoscopy is a non-invasive medical imaging technique used for skin cancer diagnosis. This technique

plays a pivotal role by providing a detailed view of skin lesions and visualizing subsurface structures, thereby enhancing the accuracy of naked-eye diagnosis (Kittler et al., 2002; Vestergaard et al., 2008). Nevertheless, diagnostic accuracy remains highly dependent on the expertise of dermatologists and clinicians, with reported rates varying from 24% to 77% (Tran et al., 2005). To address this variability, CAD systems have emerged as valuable tools. CAD systems aim to provide clinicians with more accurate, robust, and reliable diagnoses, particularly for processing large volumes of dermoscopic images efficiently (Hasan et al., 2021).

Modern DL-driven CAD systems for automated skin cancer diagnosis involve a multi-stage process, including skin lesion segmentation, feature extraction, and model-assisted diagnosis. Segmentation is a vital and challenging operation, enabling the cropping of skin lesion images, the tracking of lesion evolution, and the removal of imaging artifacts (Mirikharaji et al., 2023). However, the data-driven nature of DL models makes them prone to exhibit bias against certain input characteristics. Skin tone, in particular, has been identified as a significant source of this segmentation bias (Benčević et al., 2024). Several studies have found that segmentation and classification models trained predominantly on images of light skin often perform poorly on images featuring dark skin (Groh et al., 2021; Daneshjou et al., 2022; Bevan and Atapour-Abarghouei, 2022; Corbin and Marques, 2023; Morales-Forero et al., 2025).

Nevertheless, most existing studies confine their discussion of segmentation bias to skin tone, which may limit a deeper understanding of the problem. In this study, I conduct a comprehensive investigation of bias in skin lesion segmentation. I begin by conducting a rigorous statistical analysis to pinpoint specific sources of bias. Concretely, I move beyond skin tone categories and annotate the ISIC 2018 dataset with a comprehensive set of attributes, including fine-grained color metrics like the Individual Typology Angle (ITA) for both skin and lesion regions, as well as structural attributes like lesion area and circularity. By pre-training several representative DL segmentation models and correlating their performance with these attributes, I demonstrate that the dominant source of bias is not skin tone itself, but rather the low color contrast between the lesion and the surrounding

skin, which I quantify as the ITA Difference.

To mitigate this identified low-contrast bias, I propose a novel dual-branch controllable diffusion model. My proposed model can generate realistic dermoscopic images and their segmentation masks, uniquely disentangling and controlling lesion layout (structure) and style (color, texture) from reference images in a single efficient forward pass. Using my proposed method, I generate a synthetic dataset of high-fidelity, low-contrast samples, and use it to finetune the pre-trained, biased segmentation models. Extensive experiments show significant segmentation accuracy gains on these challenging samples from segmentation models finetuned on my synthetic dataset, and the performance improvements are further validated on the external HAM10000 dataset.

5.2 Related Works

5.2.1 Bias in segmentation of dermoscopic images

The rapid integration of AI into dermatology, particularly for the classification and segmentation of skin lesions, holds great promise for enhancing diagnostic accuracy and accessibility (Chinta et al., 2024). However, an increasing body of evidence demonstrates that DL-based models for dermatological analysis exhibit substantial performance disparities across skin tones, with markedly lower segmentation accuracy for lesions on darker skin (Benčević et al., 2024; Corbin and Marques, 2023). Segmentation inaccuracy is especially critical because it serves as a fundamental preprocessing step for diagnostic classification, where errors may propagate and amplify throughout the diagnostic pipeline, elevating the risk of misdiagnosis. Data scarcity and imbalance commonly drive DL-based models toward biased decision-making (Paproki et al., 2024). A general consensus attributes the primary cause of this bias to the severe underrepresentation of darker skin tones in training datasets, leading to the inheritance of dataset bias by downstream models (Mikołajczyk et al., 2022). In this study, I conduct a comprehensive investigation of bias in skin lesion segmentation, considering not only skin tone categories but also fine-grained attributes related to color and spatial layout in dermoscopic images.

5.2.2 Skin tone analysis and annotation

The skin tone bias embedded in foundational datasets used for model training fundamentally undermines the efficacy and fairness of DL tools for dermatological diagnosis (Benčević et al., 2024; Daneshjou et al., 2022). However, most publicly available dermoscopic image datasets lack explicit skin tone annotations, limiting the exploration of bias in downstream tasks (Morales-Forero et al., 2024). To effectively identify and mitigate bias related to skin tone, accurate and consistent measurement and categorization are essential. The ITA is a widely used computational metric for quantifying skin tone (Groh et al., 2021). Notably, the Fitzpatrick Skin Type (FST) can be derived from ITA values: FST I (ITA $>$ 55); FST II (55 $\geq$ ITA $>$ 41); FST III (41 $\geq$ ITA $>$ 28); FST IV (28 $\geq$ ITA $>$ 19); FST V (19 $\geq$ ITA $>$ 10); FST VI (ITA $\leq$ 10).

Beyond the standard ITA calculation, Bevan and Atapour-Abarghouei (2022) proposed a variant that computes ITA over eight 20×20 pixel patches along image edges, selecting the patch with the highest ITA as the estimated skin tone. Furthermore, Tadesse et al. (2023) demonstrated that pretrained CNNs, such as ResNet-18, can estimate skin tone more accurately than traditional machine learning models like Random Forest or AdaBoost. In this study, I develop an ensemble-based annotation model that integrates multiple approaches to provide accurate and robust skin tone labels for dermoscopic images.

5.2.3 Generative models for dermoscopic images

Data augmentation is a vital strategy for enhancing model performance and mitigating dataset-driven bias (Ktena et al., 2024). Traditional augmentation techniques, including geometric transformations (e.g., rotation, flipping, scaling) and photometric transformations (e.g., brightness and contrast adjustments), improve model robustness by artificially expanding dataset diversity. However, these methods are often insufficient to address severe representation gaps (Perez et al., 2018; Behara et al., 2023). Deep generative models offer a major advancement by synthesizing realistic images that capture the underlying data distribution. Variants of GANs

have been used to generate dermoscopic images of specific lesion types (Qin et al., 2020; Bisla et al., 2019) or to transfer image styles (Mikołajczyk and Grochowski, 2019). More recently, diffusion models have been adopted for dermoscopic image synthesis due to their superior image quality (Farooq et al., 2024; Dhariwal and Nichol, 2021; Shavlokhova et al., 2023).

However, augmentation for skin lesion segmentation requires not only realistic images but also corresponding masks, posing specific challenges for generative model design. A practical solution is to use masks as conditioning inputs to generate images with lesion layouts consistent with the mask (Abhishek and Hamarneh, 2019; Du et al., 2024). This study extends the strategy with a dual-branch controllable diffusion model that enables fine-grained disentangled control over stylistic and structural features, and ensures layout consistency between the generated image-mask pairs, which is critical for mitigating the identified segmentation bias.

5.3 Methods

5.3.1 Workflow overview

In this section, I introduce the overall workflow of bias identification and mitigation in skin lesion segmentation using DL-based models. As illustrated in Fig. 5.1, I first annotate multiple attributes for dermoscopic images used in both training and evaluation. These attributes include categorical skin tone labels and fine-grained continuous variables such as ITA values and lesion size. DL-based segmentation models are trained on dermoscopic image datasets annotated with corresponding mask labels. By combining the segmentation results of pretrained models with the annotated attributes, I perform statistical analyses to identify the primary sources of bias in these models.

After pinpointing the bias sources, I pick out minority samples that share similar attributes with poorly segmented samples from the training set. Using my proposed dual-branch diffusion model, I then generate new samples with distributions similar to those of the filtered data. The dual-branch diffusion model is characterized by its multifaceted controllability, enabling separate manipulation of the structural and

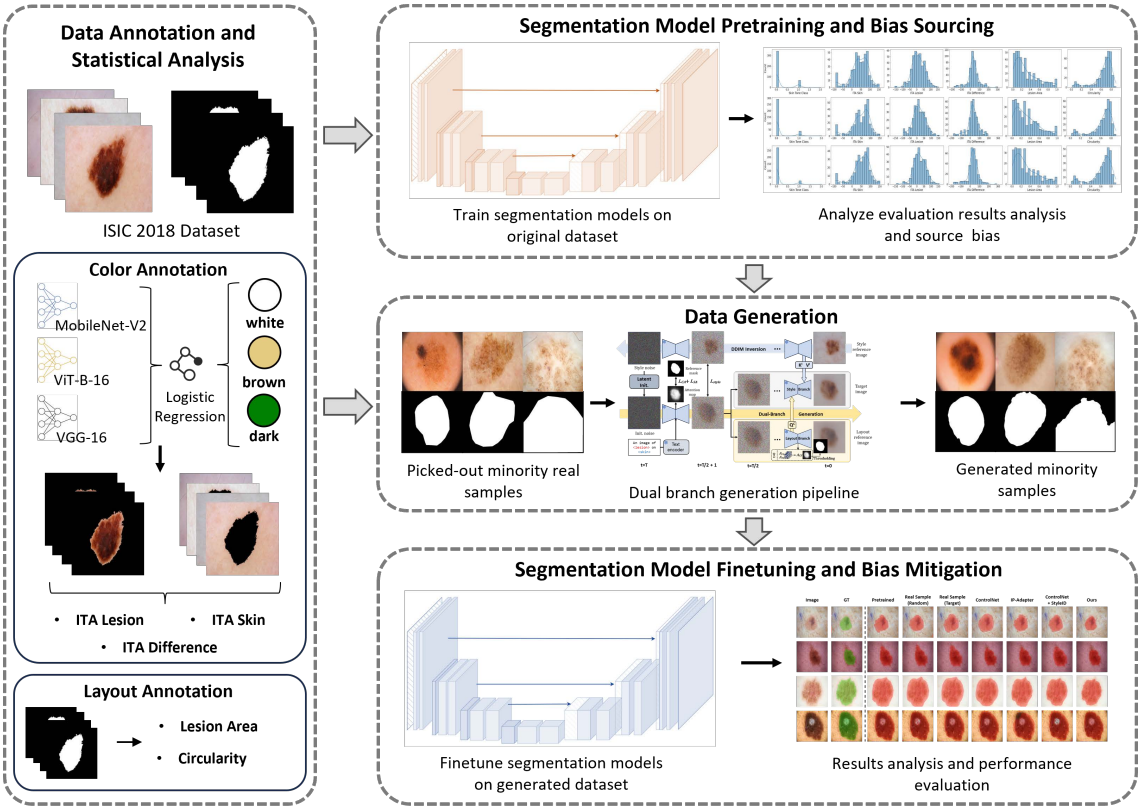


Figure 5.1: The proposed workflow for identifying and mitigating low-contrast bias in skin lesion segmentation. Dermoscopic images are first annotated with skin tone, ITA-based color attributes, and structural attributes; segmentation models are then trained and their errors analyzed with respect to these attributes to identify the main bias source; finally, targeted training samples are selected as input to synthesize image-mask pairs with the proposed dual-branch diffusion model, and these are used to finetune the segmentation models to mitigate bias.

stylistic aspects of generated dermoscopic images. It can also generate paired image-mask samples through an attention-based mechanism. The resulting synthetic images constitute a new dataset for finetuning pretrained segmentation models, thereby mitigating bias.

5.3.2 Attribute annotation of dermoscopic images

To conduct a comprehensive bias analysis from multiple perspectives, I annotate not only skin tone labels but also other fine-grained continuous attributes that describe dermoscopic images from both color and structural viewpoints. For skin tone annotation, I adopt an ensemble approach that integrates three base models as supports, which are VGG-16, MobileNet-V2, and ViT-B-16. Their prediction outputs

are aggregated using a logistic regression ensemble, which classifies skin tone into three categories: FST I-II, FST III-IV, and FST V-VI.

Furthermore, I quantify the color of both skin and lesion regions by computing their respective ITA values, defined as:

$$ITA = arctan\left(\frac{L^* - 50}{B^*}\right) \cdot \frac{180}{\pi} \tag{5.1}$$

where L^* and B^* are the lightness and blue-yellow opponents of the CIELAB colorspace, respectively. I overlay masks (or their inverses) on corresponding images to compute ITA values for the skin and lesion regions separately. In addition, I quantify the contrast between skin and lesion regions by calculating the ITA difference between the corresponding regions (ITA Difference = ITA Skin - ITA Lesion).

Apart from the color-related attributes, I also investigate structural attributes. Specifically, I quantify the size and circularity of lesions in images. Lesion size is measured as the ratio of lesion pixels to total image pixels. Circularity is computed as:

$$\text{Circularity} = \frac{4\pi \cdot S}{P^2} \tag{5.2}$$

where S is the number of pixels in the lesion region, and P represents the perimeter length of the lesion boundary. A Circularity value of 1 indicates a perfect circle, whereas values approaching 0 denote highly irregular shapes. An illustration of attribute calculation is shown in Fig. 5.1.

5.3.3 Dual-branch controllable dermoscopic generation model

Dataset augmentation with synthetic data generated by deep generative models is a widely adopted and effective strategy for improving and debiasing downstream models (Ktena et al., 2024). In segmentation tasks, however, the generated images must also be accompanied by accurate mask annotations. While diffusion-based models capable of conditioning on semantic masks provide a potential solution for generating image-mask pairs, addressing dataset bias introduces additional challenges.

As the segmentation bias identified in Chapter 5.4.2 originates primarily from color-related characteristics of dermoscopic images, the generated data should share

similar color distributions with minority samples to achieve effective debiasing. Since color is a key stylistic attribute, a natural approach is to design a two-stage generation pipeline that sequentially applies a spatially controllable model and a style transfer model based on diffusion models to jointly control image structure and color. However, such a pipeline is computationally complex, involving two forward-backward processes in each pass.

To overcome this limitation, I propose a dual-branch controllable dermoscopic generation model that integrates both style injection and mask generation within a single architecture. Crucially, my approach requires only one forward-backward process, significantly simplifying the generation procedure. In the following sections, I will shortly introduce the proposed region-aware finetuning strategy (details can be referred to in Chapter 4.3.3), and carefully describe the dual-branch generation design that jointly performs structure guidance, style injection, and mask annotation in a single pass in Chapter 5.3.3. An overview of the proposed method is presented in Fig. 5.2, and the overall pipeline is summarized in Algorithm 3.

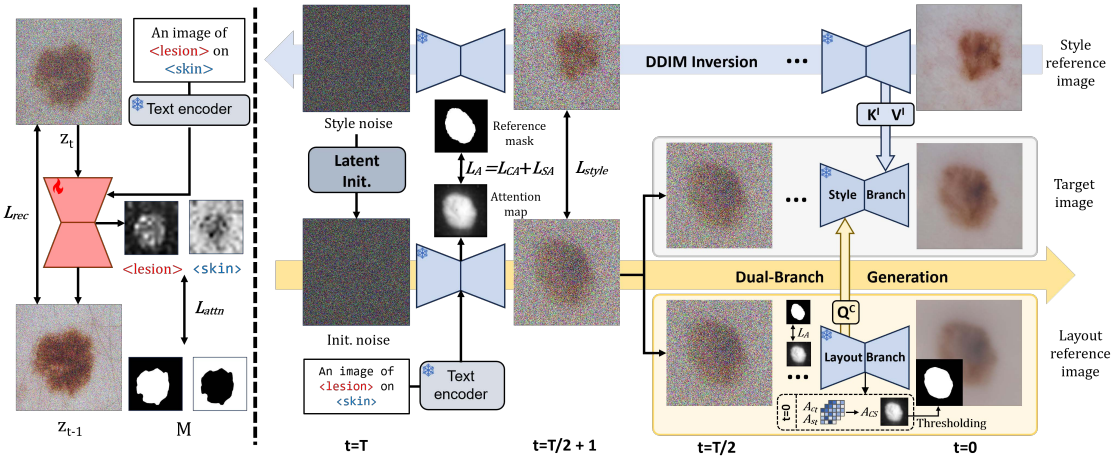


Figure 5.2: Illustration of the dual-branch controllable diffusion model, detailing the (left) region-aware finetuning strategy and (right) dual-branch generation pipeline. In the region-aware finetuning stage, special *<lesion>* and *<skin>* tokens are aligned with lesion and skin regions via an attention loss that matches cross-attention maps to the corresponding masks. In the dual-branch generation pipeline, a style reference image is first inverted to obtain its latent and attention statistics through a DDIM inversion, the final noisy latent is used as the initial latent for generation. Then the latent is jointly optimized for layout and style before splitting into a layout branch (driving spatial structure and mask extraction) and a style branch (injecting style via attention), yielding aligned lesion images and masks in a single denoising pass.

Algorithm 3 Overall pipeline of the dual-branch controllable generation model

Require: Training images $\{x\}$ and masks $\{M\}$; pretrained SD model; style reference image x_{style}; target layout mask M

Ensure: Generated dermoscopic image and aligned lesion mask

Stage I: Region-aware finetuning
for each image-mask pair (x, M) **do**
    Construct the prompt "An image of <lesion> on <skin>"
    Align <lesion> and <skin> cross-attention maps with lesion and skin regions
    Finetune the SD model using $L = L_{\text{rec}} + \alpha L_{\text{attn}}$
end for
Stage II: Dual-branch generation
Invert x_{style} to obtain style latents and cached self-attention keys/values
Initialize generation from the terminal style latent
for early denoising timesteps **do**
    Refine the latent using layout guidance and style consistency guidance
end for
Split denoising into a Layout Branch and a Style Branch
for later denoising timesteps **do**
    Use the Layout Branch to preserve lesion structure
    Use the Style Branch to inject style with layout queries and cached style keys/values
end for
Stage III: Mask annotation
Extract final attention maps from the Layout Branch
Combine attention maps and apply Otsu thresholding to obtain the lesion mask
return generated dermoscopic image and corresponding lesion mask

Region-aware finetuning

Publicly available SD backbones are generally trained on large-scale, natural image datasets, which limits their generalizability to domain-specific contexts such as dermoscopic imaging. Consequently, finetuning these models is essential for generating

high-fidelity, clinically plausible images. Unlike conventional approaches that adapt SD to the dermoscopic domain through standard text-to-image (T2I) finetuning, I propose a region-aware finetuning strategy that simultaneously adapts SD to dermoscopic data and establishes a foundation for spatially controllable generation during inference.

My region-aware finetuning consists of two key components: a structured prompt design and an attention-based spatial alignment constraint. Specifically, I formulate the training prompt as *"An image of <lesion> on <skin>"*, where the tokens *<lesion>* and *<skin>* are special placeholder tokens (P-Tokens) appended to the tokenizer's vocabulary (Gal et al., 2023b). These P-Tokens serve as semantic anchors, enabling stable visual-textual correspondence between the text concepts and their spatial regions in the image. Their embeddings are initialized using the original embeddings of "lesion" and "skin" and remain frozen throughout finetuning, preserving their semantic integrity while encouraging spatial specialization. To explicitly guide the model's cross-attention mechanism toward the desired spatial regions, I introduce a region-aware attention loss:

$$L_{attn} = \frac{1}{2} \sum_{i \in \{l,s\}} \| A_C\left(v_i, z_t\right) - M_i \|_2^2 \tag{5.3}$$

where $A_C(v_i, z_t)$ denotes the cross-attention map for P-Token v_i (*<lesion>* or *<skin>*), and M_i represents the corresponding region mask, with $M_s = 1 - M_l$. This loss term enforces a tight coupling between the P-Tokens and their spatial counterparts, encouraging the model to assign attention to semantically relevant image regions. The overall finetuning objective combines the standard reconstruction loss with the proposed attention alignment term:

$$L = L_{rec} + \alpha L_{attn} \tag{5.4}$$

where α balances the trade-off between reconstruction fidelity and spatial precision. Empirically, setting $\alpha = 0.1$ provides a good equilibrium between visual quality and semantic alignment.

Through this region-aware finetuning, the model learns to generate dermoscopic images that are not only realistic but also spatially interpretable and controllable, enabling fine-grained manipulation of lesion regions in downstream generation and segmentation tasks.

Dual-branch controllable generation pipeline

To achieve style injection and mask annotation within a single forward-backward pass, I introduce a unified diffusion framework featuring a dual-branch generation process. This framework integrates layout optimization and style modulation within a single denoising trajectory, leveraging a bifurcated architecture to ensure both spatial and stylistic consistency. The process begins with style-aware initialization, proceeds through a joint optimization phase for layout and style, and culminates in a dual-branch generation stage for efficient style injection.

Style Reference Inversion and Initialization The inference process begins with the style reference image. I apply DDIM inversion (Song et al., 2020a) to recover its latent trajectory across all T timesteps, $z_T^{style}, z_{T-1}^{style}, ..., z_0^{style}$. During inversion, I extract and store the self-attention keys K_s^{style} and values V_s^{style} from the U-Net decoder blocks at each timestep t. The generation process is then initialized using the terminal latent z_T^{style}, ensuring that the subsequent denoising starts from a noise distribution aligned with the target style domain.

Dual-branch layout-style control During inference, both spatial and style control are applied to guide the generation process. Among them, spatial control is primarily applied in the early stage, where most structural information is formed (Liu et al., 2024c). I implement spatial guidance through attention regularization that acts on both cross- and self-attention layers. For cross-attention regularization, I extract the cross-attention maps A_C of the P-Token *<lesion>*, apply Softmax normalization and Gaussian smoothing, and compute the loss L_{CA}:

$$L_{CA} = \left(1 - \frac{\sum \left[A_C\left(v_l, z_t\right) \cdot M_l\right]}{\sum M_l}\right) + \frac{\sum \left[A_C\left(v_l, z_t\right) \cdot (1 - M_l)\right]}{\sum (1 - M_l)} \tag{5.5}$$

where the first item encourages activation within the mask and the second penalizes activation outside it. Complementary constraints are provided by self-attention regularization, which discourages unwanted coupling between lesion and background pixels.For each pixel p in M, I extract its self-attention map A_S^p and define its background component $A_S^{p,B}$ as:

$$A_S^{p,B} = A_S^p \cdot (1 - M) \tag{5.6}$$

The self-attention regularization loss is then given by (Phung et al., 2024):

$$L_{SA} = \frac{\sum A_S^{p,B}}{\sum (1 - M)} \tag{5.7}$$

This constraint reduces spurious attention alignments between the foreground lesion and its background. Note that the cross-attention maps A_C and the self-attention map A_S are extracted from attention layers with latent sizes of 16×16 and 32×32, respectively (Nguyen et al., 2023; Hertz et al., 2023).

Although spatial control enforces precise structure, it inevitably perturbs the latent trajectory, causing stylistic degradation relative to the reference image. To counter this, I introduce a style consistency loss L_{style} between the current latent z_t and the corresponding style latent z_t^{style}, aligning their channel-wise mean and standard deviation:

$$L_{style} = \left\| \mu(z_t) - \mu(z_t^{style}) \right\|_2^2 + \left\| \sigma(z_t), \sigma(z_t^{style}) \right\|_2^2 \tag{5.8}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ compute the mean and standard deviation across the spatial dimensions (H, W). At each timestep t, I refine the latent z_t through gradient descent:

$$\hat{z}_t \leftarrow z_t - \eta \bigtriangledown_{z_t} \left[\lambda_{attn}(L_{CA} + L_{SA}) + \lambda_{style} L_{Style} \right] \tag{5.9}$$

where η is the guidance learning rate. λ_{attn} and λ_{style} weight the contributions of attention and style losses (empirically set to 1 and 50, respectively). This optimization is applied during the first 50% of timesteps, after which the dual-branch scheme is introduced.

At the bifurcation step ($t = T/2$), the optimized latent $\hat{z}_t$ serves as a shared ancestor for two parallel denoising branches. The **Layout Branch** continues the original trajectory with further gradient updates to refine spatial structures and serve as the master branch for spatial semantics:

$$\hat{z}_t \leftarrow z_t - \eta \bigtriangledown_{z_t} [\lambda_{attn}(L_{CA} + L_{SA})] \tag{5.10}$$

The **Style Branch**, initialized with the same latent, performs attention-space style injection. At each self-attention layer, it receives layout queries $Q_s^{content}$ from the Layout Branch and combines them with the stored style keys K_s^{style} and values V_s^{style} extracted during inversion.

By enforcing shared spatial queries, the Style Branch applies stylistic modulation over a spatially synchronized feature map, ensuring pixel-wise alignment between the stylized and layout-guided outputs. This unified dual-branch design achieves quality comparable to traditional two-stage approaches while halving inference cost, since style injection occurs in parallel rather than as a separate generation pass. A schematic illustration of this process is shown in Fig. 5.3.

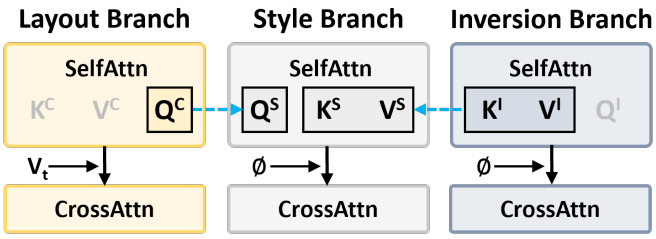


Figure 5.3: Attention-based mechanism for layout and style control within the dual-branch framework. Queries from the layout branch encode spatial structure, while keys and values from the style reference control appearance. By reusing layout queries when combining them with style keys and values, the model enforces that the stylized output follows the lesion layout of the layout branch while adopting the color and texture statistics of the style image.

Mask Annotation While layout guidance ensures coarse spatial alignment, the final lesion boundaries may still deviate slightly from the input mask M due to the stochastic process of denoising. Thus, input masks might not perfectly fit generated images. To produce precise image-mask pairs, I extract a refined mask $\hat{M}$ using

attention-based post-processing from the Layout Branch at the final sampling step ($t = 0$). This builds on the observation that self-attention maps can sharpen the structural boundaries suggested by cross-attention maps (Nguyen et al., 2023; Khani et al., 2024).

I first compute an intermediate attention map A_{CS} by multiplying the processed self-attention map with the cross-attention map and applying Otsu's thresholding (Otsu et al., 1975):

$$A_{CS} = (A_{S_t})^{\tau} \cdot A_{C_t}, \hat{M} = \text{Otsu}\,(A_{CS}) \tag{5.11}$$

where A_{C_t} and A_{S_t} are the cross-attention and self-attention maps of the P-Token $<lesion>$ at $t = 0$. The exponent τ (set to 4) sharpens self-attention contrast (Khani et al., 2024), enhancing boundary clarity. The resulting $\hat{M}$ provides a high-fidelity lesion mask accurately aligned with the generated image, ensuring both visual and structural consistency across the layout-guided and stylized outputs. An illustration of the mask annotation at $t = 0$ is presented in Fig. 5.2.

5.4 Bias analysis of skin lesion segmentation

5.4.1 Skin tone annotation of dermoscopic images

The publicly available dermoscopic datasets commonly lack annotations for skin tone. Therefore, developing a skin tone annotation method that can accurately and robustly label skin tone categories for dermoscopic images is of great significance for bias identification. To this end, I adopt clinical images that include skin tone annotations for training and evaluation of skin tone classifiers. Specifically, I collect clinical images from the Fitzpatrick 17k dataset (Groh et al., 2021) and the Diverse Dermatology Images (DDI) dataset (Daneshjou et al., 2022).

The Fitzpatrick 17k dataset contains 16,577 clinical images systematically annotated with Fitzpatrick skin types (I-VI) by two professional labeling services. The DDI dataset comprises 656 clinical images with expert-annotated skin tone labels. Notably, DDI follows a rigorous labeling protocol and maintains a bal-

anced distribution across three aggregated Fitzpatrick categories: I-II (light), III-IV (medium/brown), and V-VI (dark).

For the skin tone classifier, I leverage an efficient yet effective ensemble method as presented in Fig. 5.1. First, I employ multiple pretrained CNNs downloaded from the Pytorch platform [1], including Resnet-18, VGG-16, DenseNet-121, EfficientNet-V2, MobileNet-V2, and ViT-B-16. Each model is finetuned on Fitzpatrick 17k with a training/validation split of 80%/20%, using a learning rate of 1×10^{-3} and a batch size of 64. Training is stopped early if validation performance does not improve for 10 consecutive epochs.

Table 5.1: Performance comparison of skin tone annotation methods on the DDI dataset

Methods	Precision ↑	Recall ↑	F1 Score ↑
ITA-based methods			
ITA w/ YCbCr	0.435	0.422	0.393
8-patch ITA	0.457	0.391	0.334
Base models			
ResNet-18	0.632	0.587	0.588
VGG-16	<u>0.664</u>	<u>0.610</u>	<u>0.612</u>
DenseNet-121	0.627	0.568	0.569
EfficientNet-V2	0.606	0.543	0.543
MobileNet-V2	0.649	**0.612**	**0.615**
ViT-B-16	**0.668**	0.585	0.584
Ensemble models			
Random Forest	0.654	0.628	0.633
Logistic Regression	**0.667**	**0.644**	**0.648**
SVM	0.657	0.630	0.634
Decision Tree	0.654	0.627	0.632
Extra Tree	0.653	0.626	0.631
AdaBoost	0.658	0.631	0.635
Gradient Boosting	<u>0.662</u>	<u>0.640</u>	<u>0.644</u>

DDI, being a class-balanced dataset, is used exclusively for testing. To comply with the annotation protocol in DDI, all models are trained to classify images into three aggregated Fitzpatrick tone groups (i.e, FST I-II, FST III-IV, and FST V-VI). Additionally, I implemented traditional ITA-based methods, including ITA with YCbCr masking (Groh et al., 2021) and the 8-patch ITA approach (Bevan

[1] https://pytorch.org/

and Atapour-Abarghouei, 2022). As shown in Table 5.1, DL-based models significantly outperform ITA-based baselines across all metrics. Among the base models, MobileNet-V2 achieves the highest recall and F1 scores, ViT-B-16 yields the highest precision, and VGG-16 provides the second-best performance across all metrics.

Based on these results, I select MobileNet-V2, ViT-B-16, and VGG-16 as the foundation for an ensemble model. Multiple ensemble strategies are tested, using predictions from these three base models as inputs. As shown in Table 5.1, Logistic Regression achieves the best overall performance. Consequently, I design a two-stage skin tone annotation pipeline in which the three base models first generate individual predictions, which are then aggregated using a Logistic Regression ensemble to assign the final skin tone label for each dermoscopic image.

5.4.2 Identifying bias in skin lesion segmentation

Although categorical skin tone annotations for dermoscopic images provide a useful basis for bias analysis, a single attribute alone is insufficient for an in-depth investigation of bias sources in skin lesion segmentation. To achieve a more fine-grained analysis, I further extract quantitative image-level attributes. Specifically, I compute color-related attributes, including the ITA value of the skin region (ITA Skin), the ITA value of the lesion region (ITA Lesion), and their difference (ITA Difference = ITA Skin - ITA Lesion). In addition, I calculate layout-related attributes such as lesion area and circularity to characterize the structural properties of the lesions.

I select the publicly available and widely used dermoscopic image dataset International Skin Imaging Collaboration (ISIC) 2018 for segmentation bias identification. It comprises 2,594 dermoscopic images with mask annotations indicating the regions of lesions in its Task 1 subset. Afterwards, I evaluate the segmentation model on the Task 1 test set, which contains 1,000 dermoscopic images with mask annotations. The overall distributions of the annotated attributes for the test set are presented in Fig. 5.4.

To analyze potential segmentation bias, I evaluate three representative deep learning-based segmentation models with distinct architectural designs: Attention U-Net (Oktay et al., 2018), DCSAU-Net (Xu et al., 2023), and FAT-Net (Wu et al.,

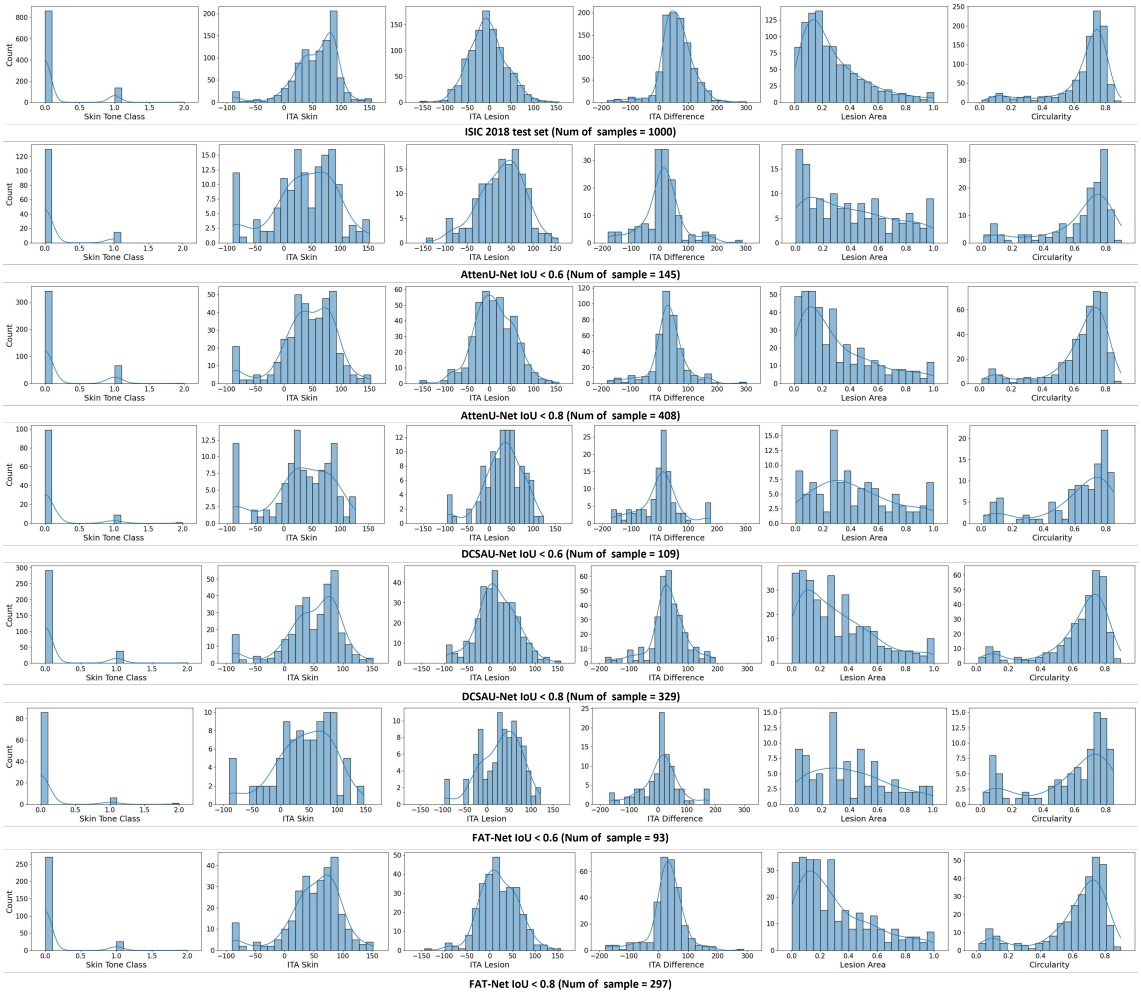


Figure 5.4: Distributions of ISIC 2018 test set and poorly segmentation samples on various image attributes. The distributions over skin tone and geometric attributes remain broadly similar, whereas poorly segmented cases are strongly concentrated at small or negative ITA Difference values, highlighting low lesion-skin color contrast as the dominant and consistent source of segmentation bias.

2022). Each model is trained on the ISIC 2018 training set using a batch size of 24 and a learning rate of 1×10^{-3}. Early stopping with a patience of 20 epochs is applied. The best-performing checkpoint, determined by validation performance, is used for evaluation on the test set. Intersection-over-Union (IoU) and Dice scores are employed as the quantitative metrics to measure segmentation accuracy. The overall performance of each model is summarized in Table 5.3.

I hypothesize that an unbiased segmentation model should exhibit no statistical dependence between segmentation performance and any specific attribute. In other words, the attribute distribution among poorly segmented samples should be sta-

tistically indistinguishable from that of the full test set. Significant deviations from this baseline indicate attribute-related bias. To verify this, I analyze two levels of poorly segmented samples for each model: (1) samples with IoU < 0.8 (denoted as $S_{0.8}$), which fall below the approximate mean IoU, and (2) samples with IoU < 0.6 (denoted as $S_{0.6}$), which correspond to severely under-segmented cases. The corresponding attribute distributions are shown in Fig. 5.4.

From Fig. 5.4, I observe that the distributions of poorly segmented samples with respect to the categorical skin tone labels largely mirror those of the full test set across all models. This suggests that the coarse, three-class skin tone annotation alone is insufficient to capture the underlying bias. For layout-related attributes (lesion area and circularity), the distributions of $S_{0.8}$ samples remain close to those of the test set, while noticeable deviations appear only in the more severely under-segmented subset ($S_{0.6}$) with respect to lesion area. This pattern indicates that lesion geometry does not induce systematic bias: if segmentation models were inherently biased toward lesion size or shape, consistent deviations would appear in both $S_{0.8}$ and $S_{0.6}$ subsets.

In contrast, the color-related attributes reveal clearer and more consistent patterns. Both ITA Skin and ITA Lesion show partial deviations between the poorly segmented subsets and the entire test set, yet their trends differ across models, implying that the sensitivity to absolute color values is model-dependent. The most prominent and consistent bias emerges in the ITA Difference, which measures color contrast between the lesion and surrounding skin. Poorly segmented samples are strongly concentrated in regions with small or negative ITA Difference values, corresponding to low contrast between lesion and skin. As illustrated in Fig. 5.5, DCSAU-Net achieves accurate segmentation for an image with a large ITA Difference (row 1) but performs poorly when the contrast is small (row 2).

These findings indicate that segmentation bias in dermoscopic images primarily arises from insufficient color contrast between lesion and surrounding skin, rather than from absolute skin tone or geometric factors. In summary, my analysis indicates that the dominant source of bias in skin lesion segmentation arises from the ITA difference between the lesion and the surrounding skin.

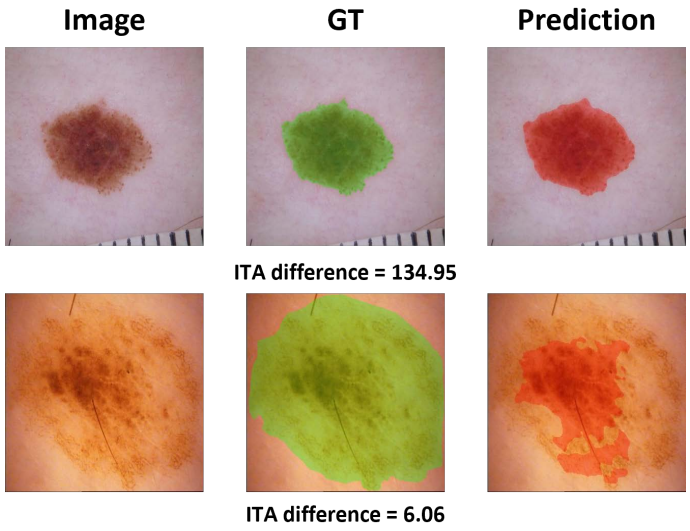


Figure 5.5: Examples of dermoscopic images, ground-truth masks, and predictions from DCSAU-Net for cases with different ITA Differences. The top example, with a large positive ITA Difference (high-contrast), is segmented accurately, while the bottom example, with a small ITA Difference (low-contrast), is severely under-segmented, visually demonstrating how insufficient color contrast between lesion and surrounding skin leads to segmentation failure.

5.5 Synthetic dataset for segmentation bias mitigation

In Chapter 5.4.2, I have identified that the ITA difference between the skin and lesion regions is the dominant factor contributing to bias in skin lesion segmentation. To mitigate this bias, I propose to employ synthetic dermoscopic images generated by deep generative models to finetune the pre-trained and biased segmentation models.

5.5.1 Experimental setup

Data preparation A detailed examination of the ITA difference distributions within the poorly segmented subsets ($S_{0.6}$ and $S_{0.8}$) reveals a consistent trend across all evaluated models: most poorly segmented samples are concentrated in the low ITA difference range, specifically when the ITA difference is below 30. This range corresponds to images where the lesion and surrounding skin exhibit low color contrast, thereby posing challenges to segmentation models.

To directly target this bias-prone region, I extract from the ISIC 2018 Task 1

training set all samples with ITA difference values below 30. This subset comprises 569 dermoscopic images, denoted as the low-contrast subset. These images effectively represent the region associated with bias identified in Chapter 5.4.2 and are thus prioritized for debiasing through synthetic augmentation.

For evaluation, I test the debiased segmentation models on two datasets: the ISIC 2018 Task 1 test set and the HAM10000 dataset (Tschandl et al., 2018). The HAM10000 dataset contains 10,015 dermoscopic images spanning seven common lesion categories. Although the original HAM10000 release lacked mask annotations, subsequent studies provided expert-segmented masks (Tschandl et al., 2020).

Baselines To assess the effectiveness of synthetic data in mitigating segmentation bias, I adopt several diffusion-based generative models as baselines. Given that segmentation tasks require paired image-mask data, I focus on models capable of mask-conditioned image generation, including ControlNet (Zhang et al., 2023a), IP-Adapter (Ye et al., 2023), and ControlNet+StyleID (Chung et al., 2024). These models represent diverse conditioning and style-control mechanisms within diffusion frameworks.

ControlNet enables explicit spatial control via conditioning masks and has been explored for segmentation data augmentation (Du et al., 2024). IP-Adapter introduces a lightweight image encoder that injects visual features into diffusion backbones through decoupled cross-attention, achieving simultaneous spatial and stylistic control. StyleID, a recent diffusion-based framework, focuses on high-fidelity style transfer at inference time. By sequentially combining ControlNet and StyleID, I achieve robust structure-style disentanglement, as the two models operate without mutual interference. For fair comparison, all generative models are initialized from the publicly available Stable Diffusion v2.1 backbone.

Generation details In generating synthetic dermoscopic images, my method first synthesizes image-mask pairs conditioned on real samples, where the real image provides style information and the real mask guides the spatial layout. To increase variability and prevent duplication, I apply geometric augmentations to real masks, including random expansion, compression, flipping, and rotation.

For a fair comparison, I use the masks generated by my method as inputs to ControlNet and IP-Adapter. Using the original training masks for these models could lead to image reproduction, as the models may memorize the original image-mask pairs during finetuning. Replacing them with newly generated masks prevents this issue and promotes genuine data synthesis.

Training details I finetune all generative models, including my method, ControlNet, and IP-Adapter, on the ISIC 2018 Task 1 training set. ControlNet is conditioned directly on ground-truth masks, while my method incorporates an additional attention-based loss that aligns the generated attention maps with target masks. IP-Adapter learns from both mask and image inputs to jointly capture structural and stylistic information. For the two-stage ControlNet + StyleID configuration, I employ the finetuned ControlNet for structure generation and apply StyleID for style transfer during inference.

All models are trained using the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$), a learning rate of 1×10^{-5}, and a batch size of 4 for 20,000 steps. For segmentation model finetuning, I use a consistent learning rate of 1×10^{-4} and a batch size of 24. All experiments are conducted on an NVIDIA RTX A800 GPU.

5.5.2 General quality of generated images

I evaluate the quality of generated dermoscopic images both qualitatively and quantitatively. For the identified low-contrast subset comprising 569 real samples, a new image is generated for each sample using my method and all baselines. The visual results are shown in Fig. 5.6, while quantitative comparisons based on Fr'echet Inception Distance (FID), Learned Perceptual Image Patch Similarity (LPIPS), and DINO scores are summarized in Table 5.2. Specifically, FID measures the feature distribution similarity between generated and real images; LPIPS evaluates perceptual similarity in terms of lesion shape, color, and texture using human-aligned CNN features; and the DINO score quantifies feature-level alignment and consistency based on the pre-trained DINOv2 model (Oquab et al., 2023).

Qualitatively, Fig. 5.6 demonstrates that my dual-branch generation pipeline

consistently produces high-fidelity dermoscopic images that faithfully adhere to the target style. The generated images exhibit realistic appearance and capture complex stylistic characteristics such as skin tone gradients, textural patterns, and illumination from the reference image. Furthermore, my pipeline generates corresponding mask annotations that are precisely aligned with the lesion regions, as illustrated in Fig. 5.6. In contrast, the baseline methods reveal clear limitations. ControlNet accurately follows the mask layout but lacks style awareness, producing overly generic lesions. IP-Adapter attempts style transfer but often introduces significant artifacts, such as erroneously copying hair from the style reference onto a different skin patch, and the style features are not fully reflected in some samples. The two-stage ControlNet+StyleID pipeline achieves more plausible visual quality and serves as a strong baseline, yet my method produces finer structural and textural fidelity, exemplified by the more faithful reproduction of local lesion details (e.g., the darker subregion in row 2).

Quantitatively, the results in Table 5.2 align with the qualitative observations. Although ControlNet + StyleID attains the lowest FID, indicating high overall distributional similarity, my method achieves a comparably low FID with only a marginal gap, outperforming both ControlNet and IP-Adapter. More importantly, my approach substantially surpasses all baselines in perceptual and feature-level metrics. It achieves a markedly lower LPIPS score, suggesting that my generated images are perceptually much closer to the reference styles, and the highest DINO score, confirming superior semantic and structural feature alignment. These results jointly validate that my dual-branch generation framework effectively balances realism, structural fidelity, and stylistic precision, making it particularly suitable for constructing high-quality synthetic datasets aimed at debiasing segmentation models.

5.5.3 Mitigating segmentation bias with generated data

To mitigate identified bias, I utilize the synthetic images generated from the low-contrast subset containing 569 real samples using different generative models. The pretrained segmentation models are subsequently finetuned on these generated im-

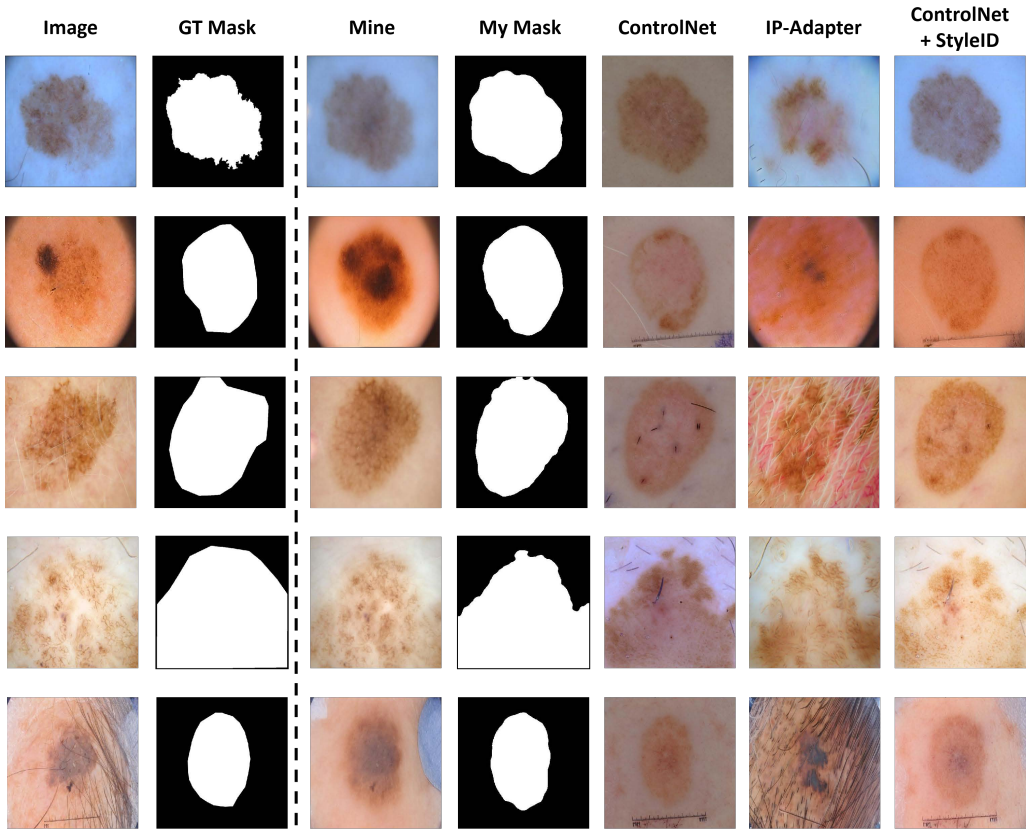


Figure 5.6: Qualitative comparison of synthetic images generated by my proposed method versus baseline models, and mask annotation from my method. My dual-branch framework generates dermoscopic images that simultaneously preserve realistic lesion structure and faithfully inherit stylistic features from the reference, while producing masks that tightly follow lesion boundaries, whereas baselines either lack style control, introduce artifacts, or misalign structure and style.

ages to enhance robustness to low-contrast cases.

I adopt a finetuning strategy instead of augmenting the pretraining dataset because the pretrained models already achieve strong performance on the overall test set. Mixing synthetic and real data during pretraining would obscure the isolated effects of debiasing and confound them with general data expansion benefits. Furthermore, I restrict the number of synthetic samples to one per real image ($n = 1$) to ensure balanced adaptation. Generating too many variants for each sample could cause the segmentation models to overfit to the synthetic domain, introducing new distributional biases and harming generalization to real dermoscopic images. A detailed ablation study of different generation ratios ($n = 2, 3, 4$) is provided in Chapter 5.5.5.

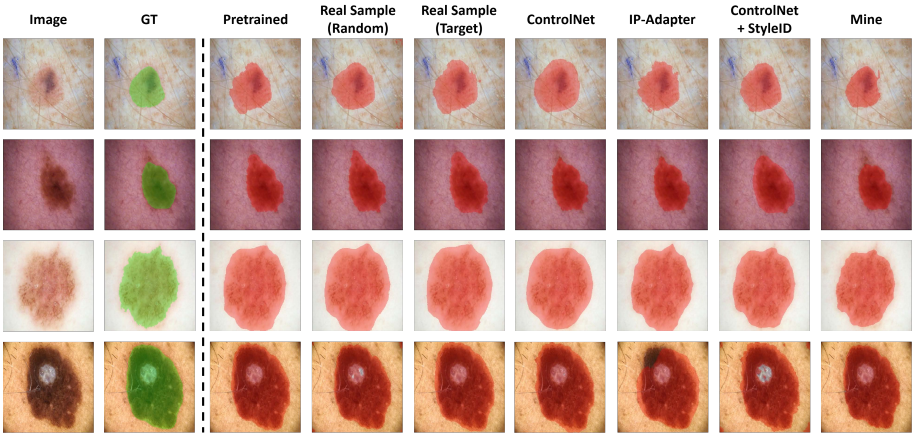
Table 5.2: Quantitative evaluation of general synthetic image quality for the proposed method and baselines

Methods	FID ↓	LPIPS ↓	DINO ↑
ControlNet	118.949	0.583	0.529
IP-Adapter	135.051	<u>0.527</u>	<u>0.587</u>
ControlNet + StyleID	**110.593**	0.599	0.516
Mine	<u>118.037</u>	**0.384**	**0.643**

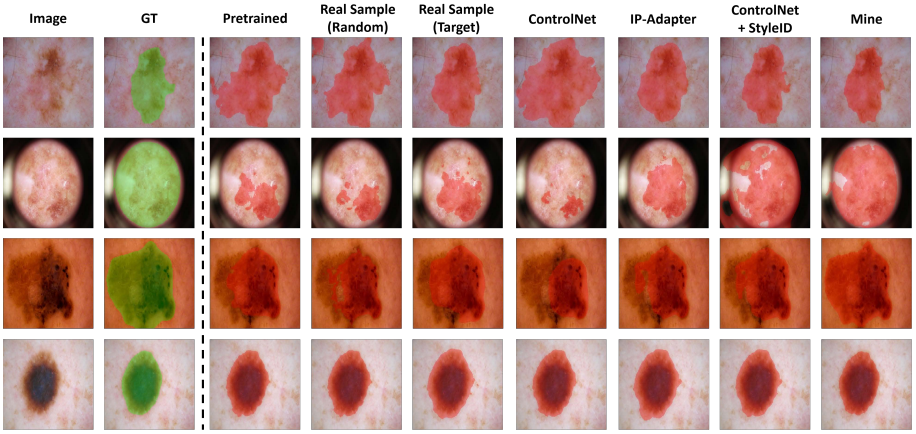
The qualitative and quantitative results of this bias-mitigation experiment are shown in Fig. 5.7 and Table 5.3, respectively. I evaluate all three segmentation models after finetuning on data from different sources. To isolate the debiasing effect, I include a control group ("Training Sample (Random)") composed of 569 randomly selected training samples. This setup helps determine whether performance gains stem from targeted debiasing or generic finetuning. Moreover, I include the low-contrast subset containing the selected 569 real samples as a comparison, and denote it as "Training Sample (Target)". Comparisons with the real low-constrast subset can help reveal the difference and effect of using synthetic model for bias mitigation.

The results show that finetuning on my generated data yields the most significant and balanced performance gains. Across all augmentation methods, my method achieves the highest IoU and Dice scores on the poorly segmented subset $S_{0.6}$, with average improvements of 18.09% and 12.31%, respectively. This demonstrates the effectiveness of my targeted debiasing approach. As shown in the first three rows of each figure in Fig. 5.7, segmentation models finetuned on my generated data produce notably more accurate predictions for low-contrast samples with blurry lesion boundaries and similar skin-lesion colors. Importantly, my method is the only generative approach that also maintains or even slightly enhances model performance on the entire test set, where both IoU and Dice scores increase consistently across all architectures. As shown in the last rows of Fig. 5.7, my finetuned models retain accurate predictions for samples that were already well-segmented before finetuning. This key observation indicates that my approach not only mitigates segmentation bias in low-contrast cases but also preserves the models' generalization ability to the full data distribution.

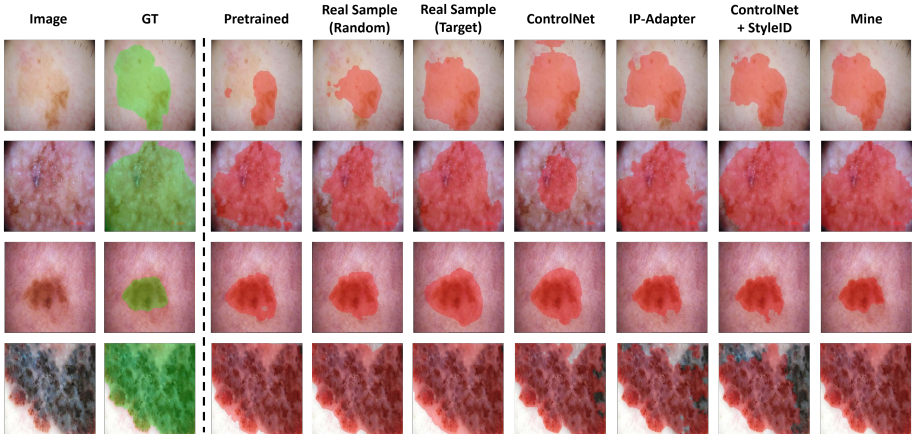
By contrast, the "Training Sample (Random)" yields marginal or inconsistent im-



(a) AttenU-Net



(b) DCSAU-Net



(c) FAT-Net

Figure 5.7: Qualitative results showing ground-truth and predicted masks from different segmentation models before and after finetuning on various data sources for challenging low-contrast samples (first three rows) and an originally well-segmented sample (last row). They illustrate that finetuning on my synthetic low-contrast data consistently improves prediction accuracy on difficult cases, while preserving good performance on easy cases.

Table 5.3: Bias mitigation results on ISIC 2018, comparing segmentation performance after finetuning on different data sources

Segmentation model	Data source	Poorly segmented samples $S_{0.6}$		All samples	
		IoU ↑	Dice ↑	IoU ↑	Dice ↑
AttenU-Net	Pretrained	0.381	0.530	<u>0.769</u>	<u>0.853</u>
	Training Sample (Random)	0.398	0.546	0.763	0.848
	Training Sample (Target)	<u>0.414</u>	0.531	0.704	0.812
	ControlNet	0.401	0.531	0.675	0.780
	IP-Adapter	0.378	0.512	0.407	0.526
	ControlNet+StyleID	0.413	<u>0.553</u>	0.634	0.756
	Mine	**0.447**	**0.595**	**0.791**	**0.867**
DCSAU-Net	Pretrained	0.376	0.518	<u>0.792</u>	<u>0.870</u>
	Training Sample (Random)	0.365	0.494	0.788	0.864
	Training Sample (Target)	**0.448**	**0.581**	0.697	0.808
	ControlNet	0.368	0.493	0.78	0.866
	IP-Adapter	0.410	0.550	0.685	0.807
	ControlNet+StyleID	0.423	0.557	0.714	0.709
	Mine	<u>0.437</u>	<u>0.575</u>	**0.803**	**0.875**
FAT-Net	Pretrained	0.391	0.541	<u>0.812</u>	<u>0.884</u>
	Training Sample (Random)	0.420	0.569	0.795	0.875
	Training Sample (Target)	<u>0.469</u>	<u>0.610</u>	0.754	0.848
	ControlNet	0.416	0.560	0.665	0.781
	IP-Adapter	0.443	0.590	0.795	0.874
	ControlNet+StyleID	0.459	0.595	0.502	0.638
	Mine	**0.472**	**0.615**	**0.817**	**0.887**

provements on $S_{0.6}$. Similar patterns appear for ControlNet-generated data, which sometimes causes noticeable degradation on the full test set while offering unstable or negligible gains on $S_{0.6}$. This is because ControlNet lacks explicit style conditioning, resulting in a mismatch between synthetic and real distributions. Moreover, while IP-Adapter and ControlNet+StyleID show some improvement on the $S_{0.6}$ sets, as they specifically generate data with similar low-contrast characteristics, this comes at the cost of a drastic drop in performance on the whole test set, indicating overfitting to the low-contrast domain. The quantitative conclusions are also reflected by the qualitative results. In the first three rows of each figure in Fig. 5.7, models finetuned on images generated by IP-Adapter or ControlNet+StyleID perform well on some low-contrast samples; however, they often present degraded predictions in the last row, confirming their limited generalizability. Furthermore, it is worth noting that the "Training Sample (Target)" also enhances segmentation models' performance on $S_{0.6}$ sets, and helps DCSAU-Net achieves the best accuracy on $S_{0.6}$. Nonetheless, the overfitting of segmentation models on the low-contrast subset is obvious as the performance on the whole set is significantly degraded.

The superior performance of my method arises from the domain fidelity and target fidelity of my generated data. My proposed dual-branch diffusion framework produces synthetic samples that specifically address the low-contrast domain (target fidelity) while maintaining high visual and feature consistency with real dermoscopic data (domain fidelity), as demonstrated in Fig. 5.7 and Table 5.2. This combination enables the segmentation models to learn contrast-invariant and semantically robust representations without drifting from the true data manifold. Consequently, the models not only overcome the low-contrast segmentation bias but also generalize to the entire distribution of real-world dermoscopic images.

5.5.4 External validation on HAM10000

After investigating bias identification and mitigation on the ISIC 2018 Task 1 dataset, I further validate my findings on the HAM10000 dataset to examine their generalizability. I first verify my conclusion regarding the segmentation bias induced by low color contrast between lesion and surrounding skin. Specifically, I directly test

the segmentation models pretrained on ISIC 2018 on the HAM10000 dataset and analyze the distributions of poorly segmented samples with respect to color-related attributes. As illustrated in Fig. 5.8, I report results at three performance thresholds, $S_{0.4}$, $S_{0.6}$, and $S_{0.8}$, where $S_{0.4}$ becomes statistically significant due to HAM10000's larger scale.

It can be observed in Fig. 5.8 that the distributions of poorly segmented samples exhibit partial deviations in ITA Skin and clear shifts in ITA Lesion. More importantly, across all segmentation models, these samples share similar distribution patterns on both ITA Skin and ITA Lesion, despite the original distributions of the entire dataset being distinct. This consistency supports my earlier finding that color contrast is the primary factor contributing to segmentation bias. Furthermore, the distributions of poorly segmented samples in ITA Difference deviate substantially from the original distribution and are concentrated around an ITA Difference value of 0, further confirming the strong relationship between segmentation difficulty and insufficient skin-lesion contrast.

Subsequently, I validate the effect of bias mitigation through finetuning on my generated data by evaluating model performance on the poorly segmented subsets $S_{0.4}$, $S_{0.6}$, as well as the entire HAM10000 dataset. The results are summarized in Table 5.4. My method demonstrates robust and consistent improvements on the poorly segmented subsets. Specifically, finetuning with my data yields the highest IoU and Dice scores on the most challenging subsets ($S_{0.4}$ and $S_{0.6}$) for DCSAU-Net and FAT-Net; the best $S_{0.6}$ and second-best $S_{0.4}$ results for AttenU-Net. On average across all models, my method improves IoU by 59.30% and 30.96% and Dice by 18.76% and 12.41% for $S_{0.4}$ and $S_{0.6}$, respectively.

Although IP-Adapter and ControlNet+StyleID also achieve partial gains on the biased subsets, they sometimes suffer from the same catastrophic overfitting observed on ISIC 2018, leading to a severe drop in overall test performance. In addition, the low-contrast subset "Training Sample (Target)" not only improves performance on the poorly segmented subsets $S_{0.4}$ and $S_{0.6}$ but also maintains better performance than synthetic images across the full HAM10000 dataset as it comprises real samples. However, its improvements on these two subsets are less significant, and the

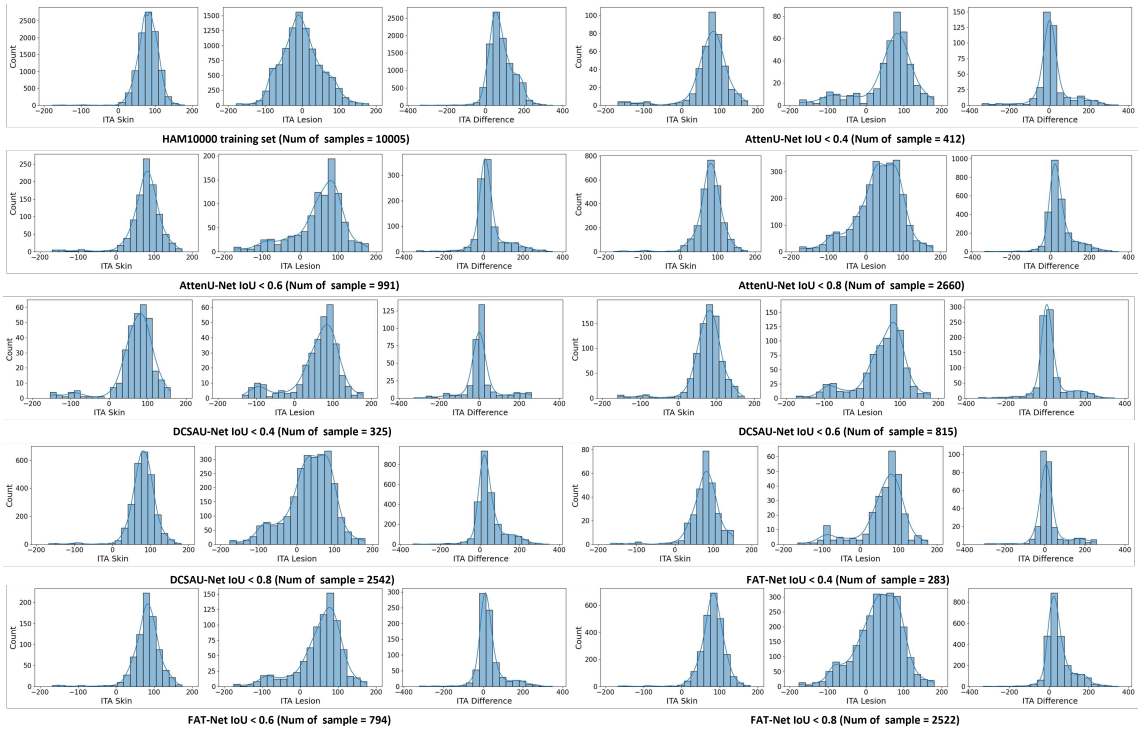


Figure 5.8: Validation of bias source on the HAM10000 dataset, showing the distributions of ITA Skin, ITA Lesion, and ITA Difference for the full HAM10000 dataset and for poorly segmented subsets at three IoU thresholds ($S_{0.4}$, $S_{0.6}$, $S_{0.8}$) across all segmentation models. As in ISIC 2018, the poorly segmented samples cluster around ITA Difference values near zero, confirming on an external dataset that low lesion-skin contrast is the main factor driving segmentation errors.

improvement is particularly modest on $S_{0.4}$, thereby limiting its effectiveness in bias mitigation. In contrast, my method, while not improving the overall scores in this scenario, achieves the best performance across all deep generative approaches on the whole HAM10000 dataset. It successfully mitigates bias in the poorly segmented subsets while maintaining generalization to the majority data. This balanced performance demonstrates that my generated data enable the segmentation models to adapt effectively to the low-contrast domain without "forgetting" the representation learned from normal cases.

Table 5.4: External validation of bias mitigation on the HAM10000 dataset

Segmentation model	Data source	Poorly segmented samples $S_{0.4}$		Poorly segmented samples $S_{0.6}$		All samples	
		IoU ↑	Dice ↑	IoU ↑	Dice ↑	IoU ↑	Dice ↑
AttenU-Net	Pretrained	0.243	0.375	0.401	0.551	**0.826**	**0.892**
	Training Sample (Random)	0.269	0.404	0.431	0.576	0.820	0.888
	Training Sample (Target)	0.321	0.462	0.464	0.608	0.816	0.887
	ControlNet	0.262	0.378	0.393	0.521	0.768	0.846
	IP-Adapter	0.346	0.487	0.459	0.602	0.776	0.848
	ControlNet+StyleID	**0.388**	**0.530**	0.464	0.615	0.720	0.723
	Mine	0.373	0.521	**0.476**	**0.625**	0.791	0.873
DCSAU-Net	Pretrained	0.245	0.381	0.409	0.561	**0.833**	**0.898**
	Training Sample (Random)	0.260	0.387	0.408	0.549	0.816	0.885
	Training Sample (Target)	0.314	0.445	0.464	0.601	0.793	0.873
	ControlNet	0.247	0.367	0.393	0.530	0.781	0.861
	IP-Adapter	0.377	0.520	0.456	0.599	0.772	0.859
	ControlNet+StyleID	0.379	0.518	0.469	0.611	0.758	0.840
	Mine	**0.389**	**0.527**	**0.471**	**0.613**	0.787	0.868
FAT-Net	Pretrained	0.256	0.410	0.428	0.584	**0.830**	**0.898**
	Training Sample (Random)	0.303	0.447	0.442	0.593	0.828	0.889
	Training Sample (Target)	0.379	0.519	0.517	0.656	0.811	0.874
	ControlNet	0.355	0.493	0.438	0.583	0.746	0.843
	IP-Adapter	0.401	0.542	0.453	0.605	0.537	0.676
	ControlNet+StyleID	0.405	0.543	0.455	0.610	0.522	0.656
	Mine	**0.424**	**0.569**	**0.524**	**0.669**	0.779	0.861

Table 5.5: Ablation study on the impact of layout guidance on mask diversity and downstream segmentation performance

Method	Diversity ↑	AttenU-Net		DCSAU-Net		FAT-Net	
		IoU ($S_{0.6}$ ↑)	IoU (All ↑)	IoU ($S_{0.6}$ ↑)	IoU (All ↑)	IoU ($S_{0.6}$ ↑)	IoU (All ↑)
w/o layout guidance	0.165	0.401	0.769	0.380	0.790	0.381	0.788
w/ layout guidance (mine)	0.388	**0.447**	**0.791**	**0.437**	**0.803**	**0.472**	**0.817**

5.5.5 Ablation studies

Impact of the dual-branch architecture

A key component of my proposed framework is the dual-branch generation pipeline, which integrates layout guidance and style injection within a single denoising process. I hypothesize that this bifurcated design is crucial for achieving high-fidelity results, as the representations of style and structure are often entangled. Enforcing strong structural control in a single branch may suppress stylistic features, leading to suboptimal visual quality. To verify the effectiveness of the dual-branch architecture, I implement a "single-branch" variant for comparison. This baseline removes the bifurcation at $t = T/2$ and instead performs both layout guidance (using $L_{CA} + L_{SA}$) and style injection (via K_s^{style} and V_s^{style} from the style reference) within the same denoising trajectory. As shown in Fig. 5.9, the single-branch model fails to maintain a proper balance between structure and style, producing images with desaturated textures and diminished color fidelity relative to the style reference.

In addition to superior generation quality, the dual-branch pipeline offers substantial computational efficiency compared to the strong baseline ControlNet+StyleID. The ControlNet+StyleID framework executes two full T-step denoising processes sequentially: one for layout-conditioned generation (ControlNet) and another for subsequent style transfer (StyleID), resulting in a total cost of approximately $2T$ denoising steps. In contrast, my dual-branch framework unifies these two processes into a single T-step generation trajectory. Although a T-step DDIM inversion is still required on the style reference to extract latent and attention statistics (as in StyleID), this is a precomputation step, and the forward generation remains a single-pass process.

Effect of the data volume on finetuning

I finetune the segmentation models using a synthetic dataset with a 1:1 generation ratio ($n = 1$), meaning that one new image is generated for each of the 569 samples in the low-contrast subset. To assess whether this data volume is sufficient, I perform an ablation on the generation ratio by synthesizing additional datasets with $n = 2, 3,$

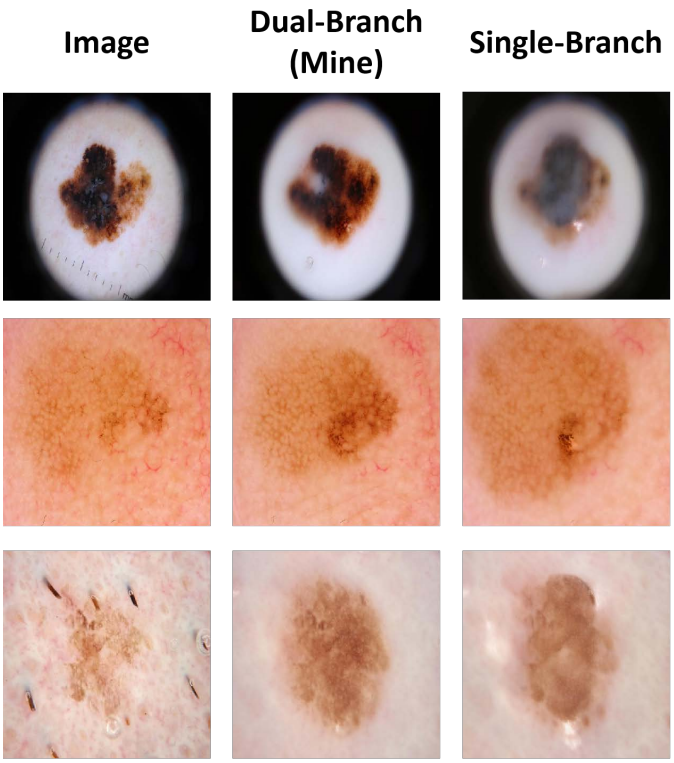


Figure 5.9: Qualitative comparison of the visual fidelity of the proposed dual-branch architecture and a single-branch variant under the same style reference and layout conditions. My dual-branch model preserves both lesion geometry and rich stylistic properties such as color saturation and fine texture, whereas the single-branch model produces more washed-out, structurally less faithful results, visually demonstrating the benefit of separating layout and style into two coordinated branches.

and 4, respectively, using different random seeds. Each dataset is then used to finetune a pretrained DCSAU-Net model, and performance is compared against the $n = 1$ setting.

The results, presented in Fig. 5.10, reveal a trade-off between the IoU scores on the poorly segmented subset $S_{0.6}$ and those on the full test set. As n increases, the IoU on $S_{0.6}$ improves slightly, whereas the IoU on the full set declines. Considering the diminishing gains on $S_{0.6}$, the trade-off in generalization, and the additional computational cost, I conclude that setting $n = 1$ achieves an optimal balance and is sufficient for my study.

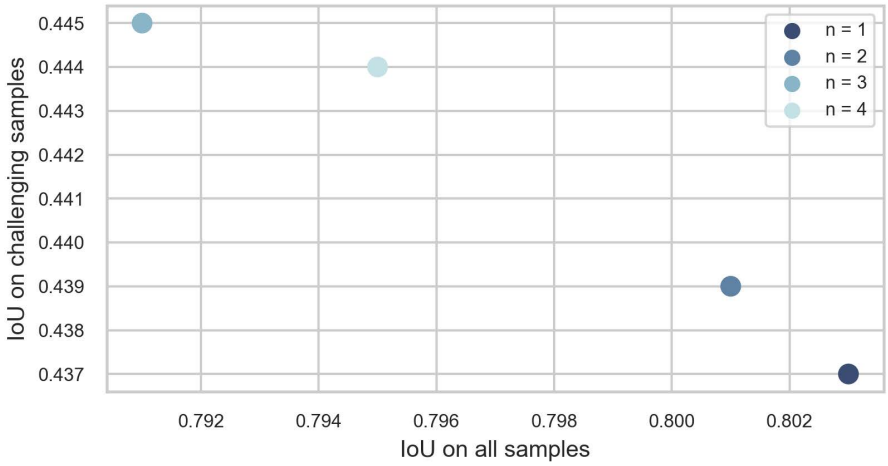


Figure 5.10: Comparison of model performance for DCSAU-Net finetuned with different synthetic generation ratios ($n = 1, 2, 3, 4$) per low-contrast real sample. It shows that increasing the number of synthetic images yields only modest gains on challenging samples but progressively harms performance on the full test set, supporting the choice of a 1:1 generation ratio as a good trade-off between targeted debiasing and overall generalization under the consideration of computational cost.

The Role of layout guidance

I further analyze the role of my layout guidance mechanism, enforced by L_{CA} and L_{SA}. Although the primary source of segmentation bias arises from color attributes, the layout guidance mainly enhances the geometric diversity of generated lesions. By conditioning on preprocessed real masks, the model learns to produce lesions with varied shapes and scales, rather than converging to an average lesion geometry.

To validate this, I compare my full model against a variant without layout guidance during inference (i.e., excluding L_{CA} and L_{SA}). As shown in Table 5.5, I first evaluate the diversity of the generated masks. By comparing the LPIPS scores within the generated masks, I find that the set of masks from the model with guidance exhibits a higher LPIPS score, confirming greater geometric diversity. Moreover, segmentation models finetuned on data generated without layout guidance consistently underperform across all three architectures and on both the challenging $S_{0.6}$ subset and the full test set. This indicates that geometric diversity, encouraged by my layout guidance, provides beneficial variation that improves segmentation robustness.

5.6 Conclusion

In this study, I investigate the underlying sources of bias in deep learning-based skin lesion segmentation and identify low color contrast between lesions and surrounding skin as the primary and consistent factor contributing to segmentation errors. To address this issue, I propose a novel dual-branch controllable diffusion framework that simultaneously generates high-fidelity dermoscopic images and corresponding segmentation masks in a single pass. The proposed model uniquely disentangles and independently controls lesion layout and style, enabling precise manipulation of structural and visual attributes. By leveraging this capability, I generate targeted low-contrast synthetic samples for finetuning existing segmentation models. The resulting models exhibit substantial improvements in accuracy on challenging low-contrast cases while maintaining performance on the overall test sets.

CHAPTER 6

Controllable Generation of Clinically Accurate Chest X-Ray Image-Report Pairs using an Integrated Vision-Language Model

While Chapter 3 to Chapter 5 primarily investigated controllable generation in visual DMs, this chapter extends the scope of the thesis to controllable generation of multimodal medical data. In medical practices, clinically relevant tasks involve not only images but also accompanying textual descriptions, such as radiology reports. As discussed in Chapter 2, controllable generation in multimodal settings introduces additional challenges, including cross-modal alignment, semantic consistency, and clinical coherence. Therefore, this chapter explores controllable generation beyond the visual domain by jointly modeling medical images and language.

The motivation of this chapter arises from the limitations of existing synthetic data generation methods for CXR analysis. Prior work has largely focused on generating either images or reports alone, which restricts their utility for training multimodal vision–language models. In clinical practice, CXR images and radiology reports are inherently coupled, with reports encoding structured observations, findings, and impressions grounded in the visual content of the image. The goal of this chapter is to develop a unified generative framework that can synthesize clinically coherent CXR image–report pairs, while enabling effective control over both visual

and textual content to enhance downstream medical applications.

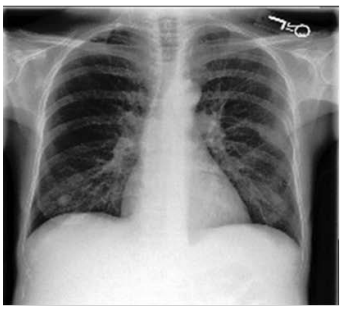
To this end, this chapter introduces *CXR-IRGen*, an integrated vision–language model for the controllable generation of CXR image–report pairs. The proposed framework adopts a modular design, consisting of a diffusion-based vision module and a language module tailored for radiology report generation. A novel prompt formulation is proposed for the vision module, which combines textual embeddings with visual embeddings from a reference image to improve generation quality and clinical relevance. In parallel, a self-supervised report generation strategy leveraging large language models is developed to produce reports that are both clinically accurate and well aligned with the generated images. Experimental results demonstrate that the proposed approach improves both perceptual realism and clinical efficacy of synthetic data, while achieving strong cross-modal alignment. This chapter concludes the technical contributions of the thesis by illustrating how controllable generative modeling can be extended to multimodal medical data, reinforcing the broader theme that reliable controllability is essential for scalable and clinically meaningful generative systems.

6.1 Introduction

Medical imaging plays a crucial role in medical practice by providing spatially resolved information about organs, tissues, and bones. The CXR image is the most common medical image due to its cost-effectiveness and low radiation dose. Notably, on average, 238 CXR images are acquired per 1000 of the population annually in industrialized countries, with 129 million CXR images acquired in the United States in 2006 (Çallı et al., 2021). However, the large number of CXR images increases the workload and diagnosis time, posing a challenge for radiologists. DL techniques provide huge support to this issue by demonstrating promising performance in AI-assisted medical applications, including segmentation and diagnosis (Ronneberger et al., 2015; Liu et al., 2023a). Nonetheless, the availability of high-quality medical data is still limited due to privacy protocols and imbalanced data distribution, which further constrains the deployment of DL models in the medical field (Torfi and Fox,

2020; Loey et al., 2020; Karbhari et al., 2021).

For this purpose, deep generative models are utilized to augment the CXR image dataset. Previous studies have demonstrated the generation of CXR images using deep generative models, including GANs and DMs (Chambon et al., 2022a,b; Weber et al., 2023; Loey et al., 2020; Karbhari et al., 2021; Motamed et al., 2021; Buragadda et al., 2022; Kora Venu and Ravula, 2020; Bhagat and Bhaumik, 2019). CXR images are typically annotated with radiology reports detailing clinical observations made by radiologists, as depicted in Fig. 6.1. However, the majority of previous studies have primarily focused on generating high-quality CXR images, overlooking the importance of paired radiology reports. To the best of my knowledge, no study has yet addressed the feasibility of generating paired CXR images and radiology reports in a unified workflow. The generated CXR image-report pairs can significantly extend the applications of the augmented dataset and provide substantial support for training DL models that handle data from various modalities.



FINDINGS:
There is no focal consolidation, pleural effusion or pneumothorax. Bilateral nodular opacities that most likely represent nipple shadows. The cardiomediastinal silhouette is normal. Clips project over the left lung, potentially within the breast. The imaged upper abdomen is unremarkable. Chronic deformity of the posterior left sixth and seventh ribs are noted.

IMPRESSION:
No acute cardiopulmonary process.

Figure 6.1: CXR image with radiology report

This work introduces *Chest X-Ray-Image Report Generation*(*CXR-IRGen*), an integrated model designed to generate CXR image-report pairs. In detail, *CXR-IRGen* is modularized and consists of a vision module and a language module (Fig. 6.2), providing high flexibility in generating multimodal CXR image-report pairs or unimodal images or reports. Furthermore, I evaluate the performance of *CXR-IRGen* on the test split of MIMIC-CXR dataset (Johnson et al., 2019a,b) and compare it with the baseline models concerning the general quality and clinical accuracy of the generated CXR image and report. Experimental results demonstrate that *CXR-IRGen* surpasses the baseline models in generating high-quality and clin-

ically accurate CXR images and reports, while ensuring clinical alignment of the generated image-report pairs.

6.2 Related Work

6.2.1 Generative models for CXR image generation

In recent years, GANs are frequently adopted for generating CXR images, and promising results were attained (Loey et al., 2020; Karbhari et al., 2021; Motamed et al., 2021; Buragadda et al., 2022; Kora Venu and Ravula, 2020; Bhagat and Bhaumik, 2019; Sirazitdinov et al., 2019; Zhang et al., 2019). Nonetheless, GANs exhibit problems including mode collapse and training instabilities, which increase training difficulties, and degrade generation quality. On the other hand, denoising diffusion models are proposed recently, which avoid these problems by adopting likelihood-based models and have been verified to outperform GANs in terms of the generation quality in general fields (Ho et al., 2020; Dhariwal and Nichol, 2021; Nichol et al., 2021; Ramesh et al., 2022). In the medical domain, Chambon et al. (2022a,b) sought the feasibility of adapting a pre-trained latent LDM for generating CXR images, finding that finetuning the U-Net component of the LDM enables the domain adaptation of a pre-trained LDM. They presented the *RoentGen* that can generate high-fidelity and diverse CXR images with radiology-specific text prompts. Packhäuser et al. (2022) verified the performance of LDM in generating high-quality CXR images, and found that the images generated by LDM outperform those by PGGAN in an abnormality identification task. Weber et al. (2023) proposed a cascaded LDM *Cheff* that can generate high-quality CXR images on a 1-megapixel scale. Based on the conclusions drawn by Chambon et al. (2022a,b), I adopt a pre-trained LDM as the backbone of the vision module, and attempt methods to further improve generation quality.

6.2.2 Generation of CXR reports

Many prior studies treat the generation of CXR reports as an image captioning task that generates natural language text conditioned on image input (Liu et al., 2019). Image captioning models adopt an image encoder to extract information from the input image and a text decoder to synthesize corresponding text conditioned on the extracted vision information (Vinyals et al., 2015; Xu et al., 2015). Jing et al. (2017) leveraged a CNN-RNN structure with a hierarchical LSTM (Krause et al., 2017) being the text decoder to generate corresponding descriptions and localize sub-regions. Xue et al. (2018) used a stacked LSTM decoder in the CNN-RNN structure. Liu et al. (2019) introduced a hierarchical generation strategy for CNN-RNN-RNN architecture, which enables the model to look at different parts of the image and enhance captioning accuracy. Ma et al. (2021) introduced the contrastive attention mechanism that can better represent the visual features of abnormal regions. Chen et al. (2020) proposed the memory-driven Transformer that uses transformers as backbones of the encoder and decoder. Based on Meshed-Memory Transformer ($\mathcal{M}^2$ *Trans*) (Cornia et al., 2020), Miura et al. (2020) proposed two new rewards for capturing the factual completeness and report consistency, and optimized these rewards via reinforcement learning.

On the other hand, the presence of medically inconsistent and incoherent reports can still be frequently found in the reports generated by image captioning models (Jeong et al., 2023). Endo et al. (2021) developed a retrieval-based CXR report generation method *CXR-RePaiR* that uses a CLIP (Radford et al., 2021) model to retrieve the report with the highest similarity score. *CXR-RePaiR* gets a higher F1 score than the baseline models, but much lower natural language metrics. Jeong et al. (2023) also introduced a retrieval-based method *X-REM* that uses a novel image-text match score. My work takes advantage of both the image captioning model and retrieval-based model, and applies a two-stage CXR report generation method in the language module, which further improves generation quality compared to the aforementioned models.

6.3 Method

The inference process of *CXR-IRGen* is depicted in Fig. 6.2. *CXR-IRGen* accomplishes a "label-to-image & report" task, taking the label from the MIMIC-CXR dataset as input, which alleviates the difficulties and complexities of input preparation. The input labels are subsequently converted into simple text to leverage the capabilities of the pre-trained CLIP text encoder. Simultaneously, a reference image with the same label is selected from the training set and encoded by a pre-trained CLIP image encoder. By combining the CLIP text and image embeddings, I obtain the conditional information for LDM sampling. The image embedding produced by the denoising backbone serves two purposes. First, it is decoded into the pixel space to create human-perceptible images. Additionally, it is projected into text embedding by a prior model for report generation. Consequently, I can obtain clinically accurate and aligned CXR image-report pairs by inputting simple labels.

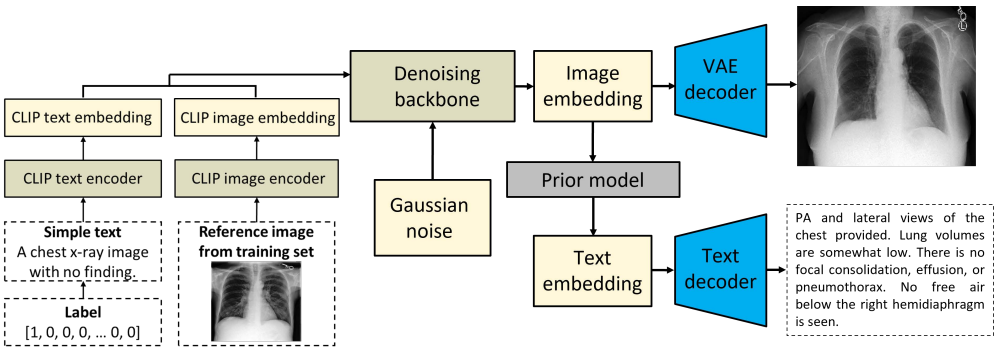


Figure 6.2: An overview of the inference process of *CXR-IRGen*

Algorithm 4 summarizes the overall pipeline of *CXR-IRGen*, including the training of the vision and language modules and the inference process for generating clinically aligned CXR image-report pairs.

6.3.1 Text-to-image generation and optimization with the diffusion model

In the vanilla LDM, the denoising backbone is a CNN-based U-Net consisting of down-sampling blocks and up-sampling blocks with skip connections between them. Besides, the feasibility of replacing the CNN layers with ViT (Dosovitskiy, 2020) was

Algorithm 4 Overall pipeline of *CXR-IRGen*

Require: CXR images, radiology reports, pathology labels; pretrained SD model; pretrained encoder-decoder language model and prior model
Ensure: Generated CXR image and corresponding radiology report

Stage I: Vision module training
for each training sample **do**
    Convert the pathology label into a semi-structured text prompt
    Select a reference CXR image with the same or closest pathology label
    Encode the text prompt and reference image using CLIP encoders
    Combine the text and image embeddings as the diffusion condition
    Finetune the SD model under the combined condition
end for
Stage II: Language module training
for each radiology report **do**
    Encode the report into a representative text embedding
    Train the language decoder to reconstruct the report from this embedding
end for
for each paired CXR image and report **do**
    Extract the image embedding from the vision module
    Train the prior model to map the image embedding to the corresponding text embedding
end for
Stage III: Image-report pair generation
Convert the input pathology label into a text prompt
Select and encode a reference image with a matching or similar label
Combine the CLIP text and image embeddings as the generation condition
Sample and decode a CXR image using the finetuned SD model
Project the generated image embedding into a text embedding using the prior model
Decode the projected text embedding into a radiology report
return generated CXR image and corresponding radiology report

discussed, and a ViT-based backbone named U-ViT was proposed (Bao et al., 2023). Following the conclusions drawn by Chambon et al. (2022a,b), I finetune the LDM on CXR images using a text-to-image approach to evaluate its domain-adapting performance. Both the U-Net and U-ViT backbones are involved and analyzed. To leverage the powerful capabilities of the pre-trained CLIP text encoder, I transform input labels into semi-structured text using the format of "*A chest X-Ray image with ..., without..., and unclear about ...*", where the three blanks are filled by pathology marked as 1.0, 0.0, and -1.0 in the label, respectively.

In T2I generation, the text prompts are projected into text embedding, and I

additionally combine the CLIP reference image embedding of an image that shares the same label as the input label with the CLIP text embedding. I hypothesize that the inclusion of an additional reference image embedding is beneficial for generating high-quality CXR images, as the model can access more structural and semantic information from the input. Therefore, the optimization objective can be expressed as:

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x),y,\varepsilon\sim\mathcal{N}(0,1),t}\left[\|\varepsilon - \varepsilon_\theta\left(z_t, t, \tau_t(y_t), \tau_i(y_i)\right)\|_2^2\right] \tag{6.1}$$

where y_t represents the input text, and y_i represents the reference image. τ_t and τ_i denote the CLIP text encoder and CLIP image encoder, respectively. Due to the difference in model architecture, the combination of the CLIP reference image embedding with the CLIP text embedding varies. For the U-Net backbone, I concatenate the image embedding and text embedding, while for the U-ViT backbone, I take the average value of them. Moreover, during the preparation of the reference image, I first search for a reference image with the same label in the training set. If none is found, then I search for an image with the same positive elements (marked as 1.0) but different negative elements (marked as -1.0) as the reference image.

The finetuning process follows the standard design of LDM finetuning and domain-adaption (Chambon et al., 2022a,b) with the exception of the input design, as depicted in Fig. 6.3. I use a pre-trained SD model (checkpoint v1.4 (Rombach et al., 2022)) with the U-Net backbone as the LDM, and a pre-trained U-ViT backbone (Bao et al., 2023).

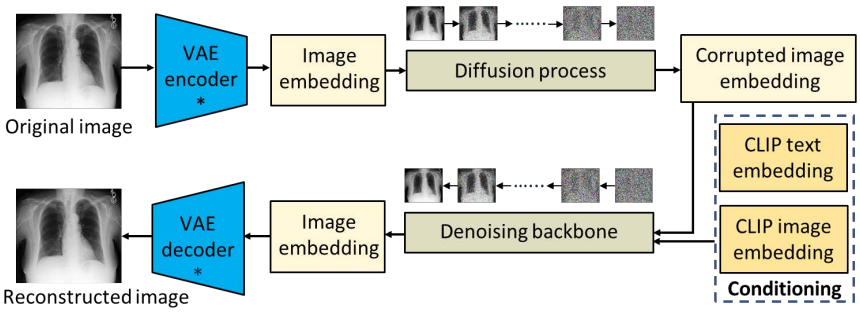


Figure 6.3: Illustration of the training process of the vision module (* denotes the frozen part)

6.3.2 CXR report generation with self-supervised learning

Image captioning models often exhibit inconsistency and incoherence between input images and generated reports, whereas retrieval-based models prioritize clinical accuracy, overlooking the consistency between retrieved and original reports (Endo et al., 2021). I propose a two-stage CXR report generation method in the language module of *CXR-IRGen* that integrates the strengths of both models. In the first stage, I utilize a pre-trained large language model with an encoder-decoder architecture to process the CXR reports. Specifically, I encode the text into a sequence of text embedding and obtain the average value of all text embedding in the sequence as a representative text embedding. Subsequently, I use this representative text embedding as the prompt for the decoder to reconstruct the input text. The loss function is calculated as the cross-entropy between the original and reconstructed text, expressed as:

$$L_{CE} = -\sum_{i=1}^{n} t_i log(p_i) \tag{6.2}$$

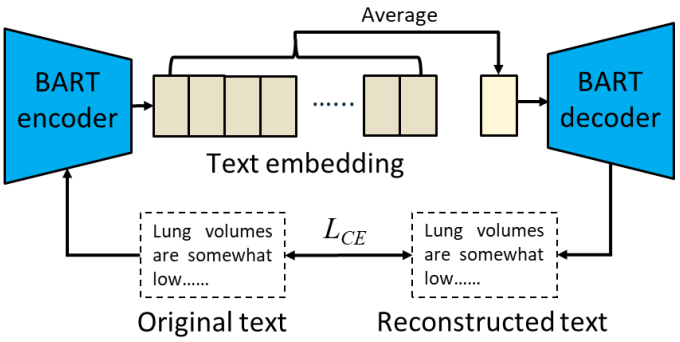
where t_i and p_i are ith elements of the original and reconstructed text, respectively. n denotes the total sequence length.

In the second stage, a prior model is employed to project the image embedding produced by the vision module into the corresponding text embedding. The training objective is to minimize the mean squared error and maximize the cosine similarity between the original and reconstructed text embeddings, which is given by:

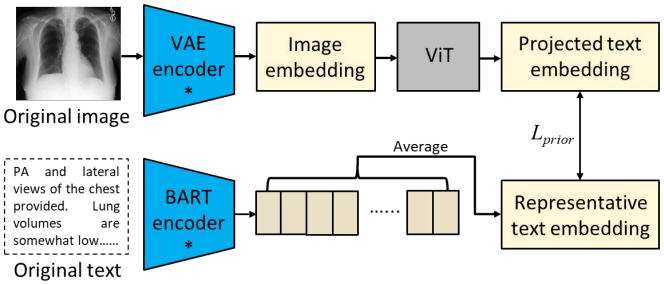
$$L_{prior} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 + \lambda\left[1 - \frac{\sum_{i=1}^{m} y_i\hat{y}_i}{\sum_{i=1}^{m}(y_i)^2\sum_{i=1}^{m}(\hat{y}_i)^2}\right] \tag{6.3}$$

where y_i represents the text embedding projected by the prior model, and $\hat{y}_i$ represents the text embedding encoded from the input text. m is the dimension of text embedding, and λ is a scaling coefficient that aligns the magnitude of the cosine similarity with that of the mean squared error, set at 0.01 for this study. Other options for L_{prior} will be discussed in the ablation tests in Chapter 6.5.3.

Specifically, the first stage resembles the image captioning models that recurrently produce a sequence of text. However, in my approach, I utilize highly sum-



(a) First stage



(b) Second stage

Figure 6.4: Illustration of the training process of the language module during (a) first stage and (b) second stage (* denotes the frozen part)

marized text information from the representative text embedding as a prompt for the decoder, rather than using vision information extracted from images. This design enhances the consistency between the original and generated reports compared to retrieval-based models. Similar to contrastive learning, which is commonly used in retrieval-based models, the second stage operates on image and text embeddings. Both the image encoder and text encoder are pre-trained and frozen. Instead of comparing the image embedding and text embedding based on cosine similarity, I employ a prior model to directly project and match the image and text embedding pair using a novel loss function Eq. (6.3) under self-supervised learning, thereby strengthening their alignment. This approach ensures that the generated report exhibits high consistency with both the image and the original report.

For the large language model, I select Bidirectional and Auto-Regressive Transformers (BART (Lewis et al., 2019)) as the backbone, and for the prior model, I utilize ViT as the backbone. The training process of both stages is depicted in Fig. 6.4.

6.4 Experiments

6.4.1 Dataset

In this study, I use MIMIC-CXR (Johnson et al., 2019a,b) for training and evaluation. MIMIC-CXR is a publicly available large-scale dataset consisting of 377,110 images and 227,943 reports from 225,000 studies. Following Chambon et al. (2022a), I extract images in the "PA" (postero-anterior) view position from the training set to finetune the vision module. For training the language module, I extract the findings and impression sections separately from all reports in the first stage. In the second stage, I select images in the "PA" view position from each study that contains a report, and if "PA" is inapplicable, I consider images in the "AP" (antero-posterior) view position, as they are also taken from a frontal view and present the same content to those in the "PA" view position but in a mirrored position. All the extracted images are matched with the reports to form a dataset of image-report pairs.

For model testing, I utilize the official testing split of the MIMIC-CXR dataset. I randomly extracted 1000 images in the "PA" view position to evaluate the vision module. Subsequently, I select images in the "PA" or "AP" view position that are paired with reports and extract findings and impression sections, resulting in 2608 image-findings/impression pair samples and 1460 image-findings pair samples. The former is adopted to evaluate the clinical efficacy of generated reports, while the latter is employed to evaluate the natural language metrics.

6.4.2 Baselines and evaluation metrics

I conduct a comparative analysis between the vision module of *CXR-IRGen* and the vanilla Stable Diffusion model. Additionally, I compare the effects of different backbones finetuned with and without the CLIP reference image embedding. For the text module of *CXR-IRGen*, I employ three CXR report generation models that have been tested on MIMIC-CXR, including two image captioning models, namely, *R2Gen* (Chen et al., 2020) and $\mathcal{M}^2$ *Trans* (Miura et al., 2020), as well as one retrieval-based model *CXR-RePaiR* (Endo et al., 2021). Particularly, I re-implement *R2Gen* and $\mathcal{M}^2$ *Trans* using publicly available code and checkpoints, and I cite the results

of *CXR-RePaiR* from the original paper.

For the generated CXR images, the general quality is evaluated using image quality metrics, including Fréchet Inception Distance (FID) (Heusel et al., 2017), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) (Wang et al., 2004). The clinical efficacy is assessed by the Area Under the Receiver Operating Characteristic (AUROC) value calculated in binary classification tasks on CXR images with and without specific pathologies.

For the generated CXR reports, I employ both natural language metrics and clinical efficacy metrics. BLEU (Papineni et al., 2002) measures the n-gram precision overlap between a generated report and the corresponding reference report, and therefore reflects whether the generated text uses similar local phrases and terminology to the ground-truth radiology report. ROUGE-L (Lin, 2004) is based on the longest common subsequence between the generated and reference reports, capturing the degree to which the generated report preserves the sequential content and key textual information of the reference report. In the context of the MIMIC-CXR dataset, these two metrics provide a surface-level assessment of linguistic similarity between generated and original radiology reports, including common descriptive patterns in the findings and impression sections. I additionally evaluate the clinical efficacy of generated reports using the F1 score. Specifically, the CheXpert labeler (Irvin et al., 2019) is applied to both the original and generated reports to extract pathology labels. The F1 score is then calculated from these labels, measuring the agreement between generated and reference reports in terms of clinically relevant findings. This metric provides a pathology-level evaluation of whether the generated report preserves the diagnostic information in the original report, complementing BLEU and ROUGE-L by focusing on clinical consistency rather than only textual overlap.

6.4.3 Evaluation of CXR images

The image labels from the testing set are used as input to generate 1000 images for evaluation. The general image quality metrics are presented in Table 6.1. Compared to the vanilla LDM, all three metrics exhibit improvements after finetuning,

confirming that finetuning on domain-specific data contributes to domain adaptation. When solely taking text embedding as input, the U-Net backbone variant finetuned for 5k steps outperforms the one finetuned for 10k steps. In contrast, for the U-ViT backbone, the variant finetuned for 5k steps demonstrates a better FID score but worse PSNR and SSIM scores compared to the variant finetuned for 10k steps. Furthermore, I investigate the effect of the reference image embedding, which shows an overall improvement in the generation quality of the U-Net backbone. As for the U-ViT backbone, the reference image embedding improves the FID score but slightly degrades the PSNR and SSIM scores. These different effects on general metrics could be attributed to the way I combine CLIP text embedding and reference image embedding, as taking the average value of the text embedding and image embedding may induce information loss.

Table 6.1: General metrics of CXR images generated by different models (RIE: reference image embedding)

Model	FID↓	PSNR↑	SSIM↑
Baseline			
Vanilla LDM	303.4451	6.7723	0.9734
LDM with the U-Net backbone			
5k steps without RIE	54.0164	<u>10.9598</u>	<u>0.9889</u>
5k steps with RIE	**49.5479**	**11.2136**	**0.9897**
10k steps without RIE	59.8236	10.3455	0.9873
10k steps with RIE	<u>53.1351</u>	10.4316	0.9875
LDM with the U-ViT backbone			
5k steps without RIE	64.4917	<u>11.1186</u>	<u>0.9896</u>
5k steps with RIE	**43.4003**	10.4192	0.9876
10k steps without RIE	54.5434	**11.1798**	**0.9897**
10k steps with RIE	<u>47.8233</u>	10.5437	0.9878

I employ the U-Net backbone for investigating the clinical efficacy of the generated CXR images, considering its clear tendency and superior robustness in analyzing image general metrics, as elaborated in Table 6.1. To evaluate clinical efficacy, I select five pathologies, namely, *Atelectasis, Cardiomegaly, Lung opacity, Effusion*, and *Pneumonia* as positive labels, while *No finding* serves as the negative label. Each label is used to generate 500 CXR images, which are grouped together, resulting in five sub-testing sets, each containing 500 positive samples and 500 nega-

tive samples. Subsequently, a pre-trained classification model (DenseNet-121, XRV (Cohen et al., 2022)) is applied to perform a binary classification task on each sub-testing set, and the AUROC value is calculated to assess the classification accuracy, with results presented in Table 6.2. It is observed that CXR images generated by vanilla LDM exhibit the worst performance, as all AUROC values are close to 0.5. Following finetuning, the AUROC values for all pathologies improve, and variants finetuned with reference image embedding achieve higher AUROC values than those without reference image embedding by an average value of 1.84%, indicating that the additional CLIP reference image embedding enhances clinical characteristics. Notably, the variant finetuned for 10k steps generates CXR images with higher AUROC scores than the original images extracted from the training set. This implies potential overfitting as the model might learn certain features highly discriminative to the XRV, therefore the training steps should be prudently designed, but the effect of the reference image embedding can still be reflected as the mean AUROC is improved by 1.89% for this variant.

6.4.4 Evaluation of CXR reports

I conduct a performance comparison of the language module of *CXR-IRen* with the baseline models. The evaluation results, presented in Table 6.3, are based on the original CXR images from the testing set. Unless specified otherwise, both the original and generated CXR reports refer to the findings section. I introduce two variants, namely *CXR-IRGen (F)* trained solely on the findings section, and *CXR-IRGen (F+I)* trained jointly on the findings and impression sections. The natural language metrics are evaluated using only the former, while both variants are used for assessing clinical efficacy. In comparison to the retrieval-based model *CXR-RePaiR*, *CXR-IRGen* demonstrates a dramatic improvement in BLEU-2 score. As the CXR reports in the dataset are highly diverse, the reports retrieved by *CXR-RePaiR* are clinically matched with images but usually different from the originals. On the other hand, *CXR-IRGen* learns the common textual description of the images in the same class and achieves good proficiency in generating reports consistent with the originals, resulting in the highest BLEU-1 score among all models, with the other

Table 6.2: AUROC values of the binary classification task on original CXR images and CXR images generated by different models (RIE: reference image embedding)

Source	Atelectasis	Cardiomegaly	Lung opacity	Effusion	Pneumonia	Mean
Baseline						
Original	0.7799	0.8197	0.8081	0.8921	0.7127	0.8025
Vanilla LDM	0.5504	0.5378	0.5876	0.5785	0.5458	0.5600
Proposed approach (U-Net backbone)						
5k steps no RIE	0.6303	0.7284	0.6397	0.8128	0.5769	0.6776
5k steps w/ RIE	0.6470	0.7326	0.6605	0.8126	0.5956	0.6897
10k steps no RIE	**0.8897**	0.9800	0.8938	0.9867	0.8267	0.9150
10k steps w/ RIE	0.8688	**0.9836**	**0.9537**	**0.9953**	**0.8602**	**0.9323**

Table 6.3: Comparison of *CXR-IRGen* and baselines models on original CXR images (Results with * are taken from the original paper (Endo et al., 2021))

Model	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE-L↑	F1 score↑
Baseline						
*CXR-RePaiR-2** (Endo et al., 2021)	-	0.0690	-	-	-	0.2560
*CXR-RePaiR-Select** (Endo et al., 2021)	-	0.0500	-	-	-	0.2740
R2Gen (Chen et al., 2020)	0.2870	0.1651	0.1072	0.0726	0.2093	0.1716
$\mathcal{M}^2$*Trans* (Miura et al., 2020)	0.3174	**0.1917**	**0.1195**	**0.0734**	**0.2252**	0.2665
Proposed approach						
CXR-IRGen (F)	**0.3200**	0.1760	0.1066	0.0669	0.2080	0.2695
CXR-IRGen (F+I)						**0.2930**

four natural language metrics being on par with those of *R2Gen* but slightly below those of $\mathcal{M}^2$ *Trans*. However, it should be noted that $\mathcal{M}^2$ *Trans*'s image encoder is additionally trained on the CheXpert dataset (Irvin et al., 2019), which may enhance CXR report generation quality and lead to unfair comparison. Furthermore, *CXR-IRGen* exhibits exceptional clinical accuracy, with the variant *CXR-IRGen (F+I)* achieving the highest F1 score among all the models.

I also compare the clinical efficacy of all models on the CXR images generated by the vision module of *CXR-IRGen*. I utilize the 3000 generated CXR images introduced in Chapter 6.4.3 for report generation. The evaluation results are provided in Table 6.4. It is evident that *CXR-IRGen (F+I)* outperforms all other models in terms of clinical efficacy on the generated CXR images. While *CXR-IRGen (F)* demonstrates superior clinical efficacy to *R2Gen*, it falls short compared to $\mathcal{M}^2$ *Trans*. This difference can be attributed to the fact that $\mathcal{M}^2$ *Trans* employs an image encoder that is additionally trained on the CheXpert dataset (Irvin et al., 2019), which aids in feature recognition and representation. The impact of reference image embedding on clinical efficacy is also reflected. For $\mathcal{M}^2$ *Trans* and *CXR-IRGen*, the F1 scores are higher on the CXR images generated by the vision module trained with reference image embedding by an average value of 6.58%.

6.5 Ablation

I conduct an analysis of various design choices in *CXR-IRGen* that might affect the generation quality, including (1) the strategy of extracting the representative text embedding; (2) utilizing the image or image embedding for report generation; and (3) different options for L_{prior}. Note that all the variants discussed in this section are trained using the findings section of the CXR report.

6.5.1 Extracting representative text embedding

During the first training stage of the language module in *CXR-IRGen*, I select a representative text embedding from a sequence of text embedding and use this representative embedding as input for the BART decoder. The goal is to ensure

Table 6.4: F1 scores of *CXR-IRGen* and baseline models on CXR images generated by the vision module *CXR-IRGen* (RIE: reference image embedding)

Model	5k steps without RIE	5k steps with RIE	10k steps without RIE	10k steps with RIE
Baseline				
R2Gen (Chen et al., 2020)	0.1095	0.1102	0.1895	0.1794
$\mathcal{M}^2$*Trans* (Miura et al., 2020)	0.2328	0.2347	0.3226	**0.3738**
Proposed approach				
CXR-IRGen (F)	0.2157	0.2280	0.3410	0.3627
CXR-IRGen (F+I)	**0.2390**	**0.2543**	**0.3603**	0.3719

Table 6.5: Comparison of different variants of *CXR-IRGen*

Model	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE-L↑	F1 score↑
Proposed approach						
CXR-IRGen (F)	**0.3200**	**0.1760**	**0.1066**	**0.0669**	**0.2080**	**0.2695**
Change input						
Input image	0.3096	0.1658	0.0975	0.0594	0.1996	0.2617
Change loss function						
Mean square error	0.3070	0.1616	0.0934	0.0558	0.1951	0.2433
Cosine similarity	0.0206	0.0055	0.0022	0.0008	0.0292	0.0696

that the representative text embedding captures as much semantic information as possible. Several strategies for extracting the representative text embedding are considered, including using the text embedding of the `[BOS]` (beginning of sentence) token, the text embedding of the `[EOS]` (end of sentence) token, or the averaged text embedding of all tokens. The reconstruction quality is evaluated using different representative text embeddings, and the results are presented in Fig. 6.5. Notably, the averaged text embedding of all tokens outperforms the other strategies in terms of BLEU and ROUGE-L scores, displaying higher scores and a more consistent increase during the training process.

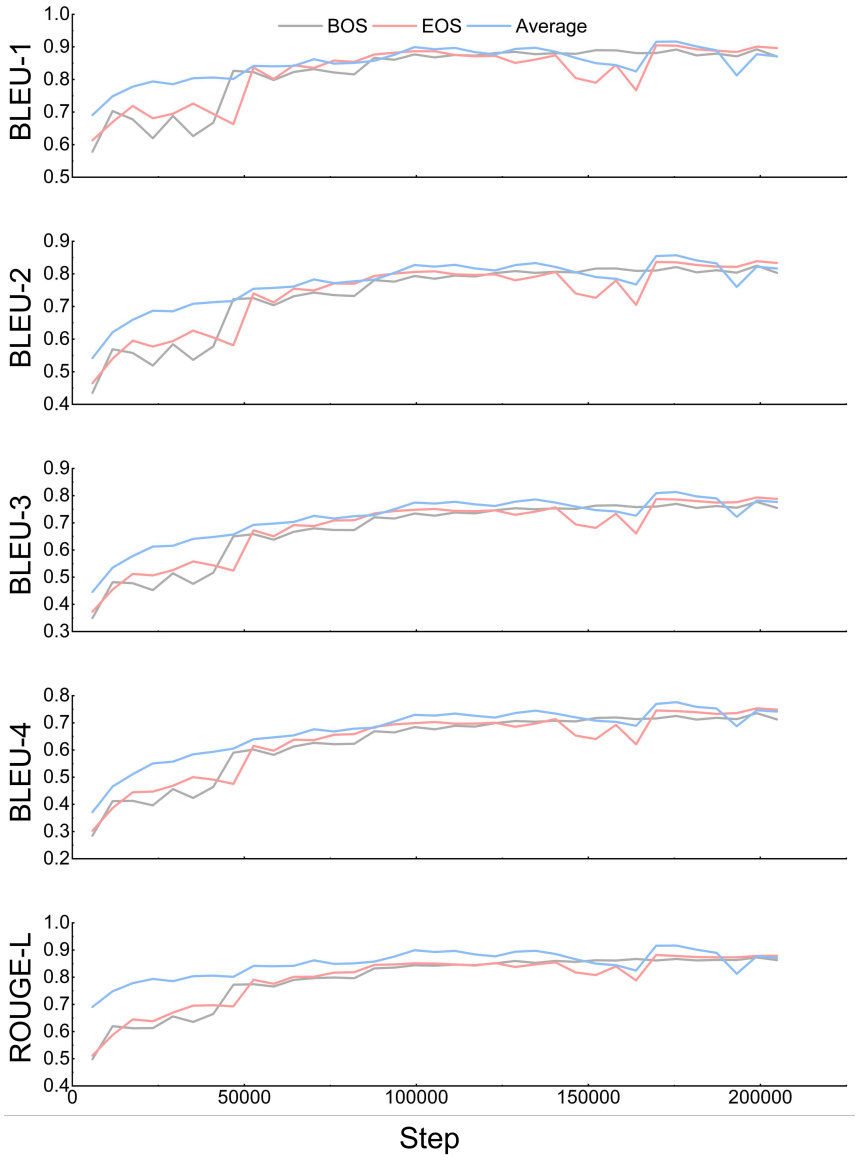


Figure 6.5: Comparison of different representative text embedding

6.5.2 Image vs. Image embedding

In the language module of *CXR-IRGen*, I utilize image embeddings as input for the prior model, whereas image captioning models typically take images directly as input. To compare the generation quality, I evaluate the performance using both image embeddings and images, and the results are detailed in Table 6.5. I observe that employing image embedding as input leads to higher scores in both natural language metrics and clinical efficacy metrics compared to using raw images, suggesting that the encoding process meaningfully compresses image information, emphasizing relevant details crucial for feature extraction and recognition by the prior model. Notably, this is consistent with the observation in image generation tasks reported by Weber et al. (2023), who concluded that semantic features are more beneficial for a cascaded diffusion model in generating high-quality and high-resolution CXR images compared to low-resolution images.

6.5.3 Choice of L_{prior}

The prior model within the language module of *CXR-IRGen* is responsible for learning a projection from image embeddings to text embeddings. To achieve this, it is crucial to minimize the distance between the target embedding space of this projection and the pre-determined text embedding space. In the process of measuring this distance, several options are available, including the mean square error, the cosine similarity, or a combination of both, as shown in Eq. (6.3). The mean square error quantifies the Euclidean distance between two vectors, while the cosine similarity measures the angle between them. I trained the prior model with each of these metrics as loss functions, and the resulting model performances are presented in Table 6.5. The outcomes indicate that combining the mean square error and cosine similarity yields the best result. Particularly, solely using cosine similarity as the loss function severely limits model performance, but its inclusion alongside the mean square error with a coefficient λ that balances their values significantly improves performance.

6.6 Conclusion

In this study, I introduce an integrated model called *CXR-IRGen* designed for generating high-quality CXR image-report pairs. *CXR-IRGen* comprises a vision module for generating CXR images and a language module for generating corresponding reports. These modules can either be utilized together to produce CXR image-report pairs or independently to generate CXR images or reports separately. The vision module incorporates a novel prompt design for the text-to-image LDM by combining text embedding with a reference image embedding, which enhances the general quality and clinical efficacy of the generated CXR images. For the language module, I propose a new CXR report generation model that benefits from both image captioning and retrieval-based approaches, leveraging a large language model and self-supervised learning strategy. The proposed report generation model demonstrates the ability to produce coherent, consistent CXR reports, and it outperforms baseline models in terms of clinical efficacy. Furthermore, the CXR image-report pairs generated by *CXR-IRGen* exhibit a high level of clinical alignment.

CHAPTER 7

Concluding Remarks

7.1 Contributions

This thesis has explored controllable generation with DMs from both methodological and application-oriented perspectives, with a particular emphasis on addressing practical challenges in medical imaging. Motivated by the increasing adoption of DMs as a dominant paradigm for high-fidelity generative modeling, the work set out to investigate how controllability can be made more fine-grained, reliable, and clinically meaningful. Across a sequence of interconnected chapters, this thesis progressively advanced the understanding and application of controllable DMs. Beginning with general-purpose T2I generation, it examined how internal attention mechanisms can be leveraged to disentangle and synchronize multiple visual concepts. Building on these insights, the thesis then demonstrated how controllable generation can be adapted to domain-specific medical imaging tasks, first for dermoscopic image synthesis and segmentation bias mitigation, and subsequently for multimodal CXR image-report generation. Together, these studies illustrate a coherent trajectory from foundational controllability mechanisms to targeted clinical applications, highlighting how principled control over generative processes enables DMs to move

beyond visual realism toward reliability, robustness, and clinical relevance. The main contributions and innovations are summarized by chapter below:

Chapter 3. ***Attention-based Disentanglement of Multiple Concepts for Text-to-Image Customization***

This chapter contributes an attention-based framework for disentangling and customizing multiple visual concepts in T2I generation. It first shows that cross-attention and self-attention maps can be leveraged to create concept-specific masks from a given image within a single initialization step, without relying on specialized segmentation models or manual annotation. This provides an efficient mechanism for localizing different concepts and establishes the basis for more precise concept-level customization.

Building on this attention-based localization mechanism, this chapter further proposes an adaptive sampling strategy that automatically estimates the sampling ratio of multiple concepts according to their cross-attention scores. This strategy addresses the problem of asynchronous learning, where different concepts are learned at different speeds during optimization. By assigning different sampling ratios to different concepts, the method enables more balanced concept learning and reduces the risk that dominant or easier-to-learn concepts suppress weaker ones.

In addition, this chapter introduces a feature-retaining training framework in which different loss functions are applied to sampled subsets of varying sizes. This design helps prevent feature fusion between concepts and improves the quality and independence of concept-specific feature acquisition. Overall, this chapter establishes a controllable customization framework that improves both concept disentanglement and feature preservation in multi-concept T2I generation.

Chapter 4. ***Controllable Synthesis of Dermoscopic Images for Enhanced Computer Aided Diagnosis and Detection***

This chapter extends controllable diffusion generation to dermoscopic image synthesis and its downstream use in computer-aided diagnosis and detection. A key contribution of this chapter is the development of DermPrompt, a text prompting

framework that uses LLMs to generate attribute-rich captions as dynamic prompts. These prompts enhance the representation learning of dermoscopic images in SD and enable text-guided customization of clinically relevant attributes.

This chapter also proposes a two-stage paradigm for semantic dermoscopic image synthesis. In the first stage, a region-aware fine-tuning strategy is introduced to establish robust semantic visual-textual alignment between lesion and skin regions. This alignment provides the foundation for controllable layout-guided generation. In the second stage, test-time layout guidance and attention-based annotation are combined in a training-free pipeline to generate paired dermoscopic images and lesion masks. This design enables efficient synthesis because SD only needs to be fine-tuned once, while still supporting controllable generation and automatic mask production.

Experimental results further demonstrate the effectiveness of the proposed method. On downstream multi-class classification and lesion segmentation tasks, the synthetic images generated by this method improve model performance more effectively than baseline approaches. These results show that controllable dermoscopic image synthesis can not only generate visually plausible data, but also provide useful training samples for clinically relevant computer-aided diagnosis and detection tasks.

Chapter 5. *Mitigating Low-Contrast Bias in Skin Lesion Segmentation using a Dual-Branch Controllable Diffusion Model*

This chapter focuses on the use of controllable generation to mitigate bias in skin lesion segmentation. It first provides a comprehensive and fine-grained analysis of segmentation bias in dermoscopic images. Rather than attributing segmentation bias only to categorical skin tone, the analysis identifies low lesion-skin color contrast as the primary and most consistent source of performance degradation. This finding provides a more precise understanding of bias in dermoscopic segmentation and motivates a targeted data generation strategy for improving model robustness on difficult low-contrast cases.

To address this problem, this chapter proposes a dual-branch controllable diffusion model that generates high-fidelity dermoscopic images with decoupled control

over structural and stylistic features. The model is designed to preserve lesion layout while modifying appearance-related factors such as contrast and skin style. At the same time, it produces corresponding segmentation masks, enabling the generation of image-mask pairs in a unified pipeline. This makes the method particularly suitable for segmentation bias mitigation, where both realistic images and accurate annotations are required.

Extensive experiments show that fine-tuning segmentation models with the synthetic data generated by this approach substantially improves robustness on low-contrast images. Importantly, the results also show that this improvement is achieved without catastrophic overfitting or severe degradation of overall segmentation performance. This chapter therefore demonstrates that controllable DMs can be used not only for data augmentation, but also as targeted tools for addressing clinically meaningful subgroup bias.

Chapter 6. *Controllable Generation of Clinically Accurate Chest X-Ray Image-Report Pairs using an Integrated Vision-Language Model*

This chapter broadens the scope of controllable generation from unimodal medical image synthesis to multimodal CXR image-report generation. The main contribution of this chapter is CXR-IRGen, an integrated vision-language generative framework composed of a vision module and a language module. This modular design supports multiple generation tasks, including unimodal CXR image generation, radiology report generation, and paired image-report synthesis.

Within the vision module, this chapter introduces a new prompting design for the T2I diffusion model by combining text embeddings with image embeddings extracted from a reference image. This enriched prompt representation incorporates both semantic and visual information, thereby improving the quality of generated CXR images across different diffusion backbones. The results indicate that reference image information can provide useful structural and contextual cues for controllable medical image generation.

In parallel, this chapter proposes a CXR report generation model as the language module. This model uses a language module together with a self-supervised learning

strategy to generate radiology reports from visual and textual information. The generated reports achieve promising performance in terms of both conventional natural language metrics and clinical efficacy metrics. Overall, this chapter demonstrates the potential of integrated vision-language generation for producing clinically meaningful CXR image-report pairs, extending controllable diffusion generation toward multimodal medical applications.

7.2 Limitations and Future Work

Despite the results and contributions presented in this thesis, limitations remain in each chapter. Recognizing these constraints not only provides a balanced assessment of the proposed approaches, but also highlights promising directions for future research aimed at improving scalability, generalization, and clinical robustness of controllable generative models.

Chapter 3. The mask created by my method is determined by the text prompt via cross-attention maps. For a composed target (e.g., a vase and plants inside), initializing [V] by "vase" will create a mask only for the vase, resulting in deviations in presenting the plants. Moreover, my method struggles to disentangle concepts in the same category (e.g., two dogs) under its standard settings. Since both [V]s are initialized by "dog", my method fails to create separated masks for them, resulting in a failure in disentanglement. Future work should target these special cases, and optimize the mechanism for prompting, attention extraction, and embedding optimization to extend the generalizability of my method.

Chapter 4. The primary limitations of this study are the model's dependence on the training data distribution, which prevents the synthesis of unseen combinations like melanoma on dark skin, and the high computational cost that restricts accessibility for resource-limited clinicians. To address these issues, future work will focus on curating diverse datasets to enable authentic generation across various skin tones and employing model distillation to create a lightweight, faster architecture. Additionally, the research aims to improve generalizability by extending the method

to handle the complex structures and backgrounds found in user-taken images.

Chapter 5. In the finetuning stage, the pretrained segmentation models are finetuned on synthetic data to mitigate segmentation bias. Ablation studies identify that the scale of the synthetic dataset reveals a trade-off between the model's performance on the low-contrast subset and the entire test set. However, the current experiment does not finetune segmentation models on mixed synthetic datasets consisting of both low-contrast and normal dermoscopic images. Future work could delve into the experiment scale, and seek to improve overall segmentation performance while mitigating bias through more scaled finetuning.

Chapter 6. Although CXR-IRGen is evaluated using both image-level and report-level metrics, these metrics reflect clinical diagnosis to different extents. FID, PSNR, and SSIM mainly measure the visual realism of generated CXR images, but they do not directly verify clinical correctness. AUROC is more diagnosis-oriented because it evaluates whether disease-related patterns in generated images can be recognized by a pathology classifier, although it remains dependent on the reliability of the classifier itself. Similarly, BLEU and ROUGE-L mainly assess lexical or sentence-level similarity between generated and reference reports, whereas the CheXpert-label-based F1 score provides a more clinically relevant proxy by comparing extracted pathology labels. However, these automated metrics still cannot fully capture the diagnostic correctness, uncertainty, or clinical nuance of radiological interpretation. Future work could therefore include radiologist evaluation of generated image-report pairs or downstream diagnostic experiments on external CXR datasets, providing stronger evidence of clinical validity. In addition, for the U-ViT backbone, the current strategy combines text and reference image embeddings by simple averaging, which may cause information loss and partly explain the improvement in FID but degradation in PSNR and SSIM compared with the U-Net backbone. Future work could explore more sophisticated fusion mechanisms to better preserve both semantic and structural information.

7.3 Epilogue

In a rapidly developing AI community, the lifecycle of this thesis has witnessed numerous innovations. From the proposal of DDPMs, to the success of LDMs, these milestone works are consistently shaping the AI techniques and shedding light on future development of AI theories. This thesis, built upon these works, sets out to explore controllable generation with DMs, motivated by the observation that generative performance alone is insufficient for reliable deployment in real-world and clinical settings. Through a progression from foundational mechanisms to domain-specific applications, the work has shown that controllability is not a secondary enhancement, but a central design principle that determines how generative models can be understood, adapted, and trusted. By grounding controllability in internal model mechanisms and by aligning generation objectives with downstream clinical needs, this thesis contributes to a more principled view of DMs as controllable systems rather than black-box generators.

A consistent topic throughout this work is the value of bridging theory and application. Insights into attention-driven concept representation inform practical solutions for data scarcity, bias mitigation, and multimodal synthesis in medical imaging. Conversely, the demands of clinical tasks expose limitations in existing controllability paradigms and motivate the development of more structured and interpretable control mechanisms. This bidirectional interaction underscores an important lesson: advances in generative modeling are most impactful when guided by concrete application constraints, especially in high-stakes domains such as healthcare.

A further lesson emerging from this thesis is that technical evaluation alone is not sufficient to establish the clinical validity of generative models in medical imaging. While metrics such as image fidelity, segmentation accuracy, classification performance, and report similarity provide useful evidence of model quality, they remain indirect indicators of clinical usefulness. For generated medical data to support real diagnostic or decision-making workflows, future experiments should more directly examine whether the generated outputs preserve clinically meaningful findings, reflect realistic disease presentations, and improve performance in external validation settings. This may require evaluation by clinical experts, reader studies,

downstream diagnostic experiments, and validation across multi-centre datasets. Such experiments would help bridge the gap between methodological progress and clinical reliability, ensuring that controllable generation is assessed not only by how realistic its outputs appear, but also by whether they are diagnostically trustworthy and practically useful.

Looking into the future, controllable DMs hold considerable promise as foundational tools for scientific discovery, clinical decision support, and trustworthy AI. As DMs continue to evolve toward greater scale and multimodality, the challenge will shift from generating plausible data to generating data that is purpose-driven, interpretable, and aligned with human intent. It is hoped that the methods and perspectives presented in this thesis will contribute to this broader endeavor, and that they will inspire future research toward generative models that are not only powerful, but also controllable, interpretable, and ultimately beneficial to clinical practice and beyond.

Bibliography

Chieh-Hsin Lai, Yang Song, Dongjun Kim, Yuki Mitsufuji, and Stefano Ermon. The principles of diffusion models. *arXiv preprint arXiv:2510.21890*, 2025. (document), 2.1, 2.2

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. (document), 1, 2.1.3, 2.3, 3.1, 3.3.1, 6.2.1

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. (document), 1, 1.1, 2.1.4, 2.4, 3.1, 3.2.1, 3.3.1, 4.1, 4.3.1, 6.3.1

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. (document), 1, 2.2, 2.5, 2.6, 3.2.1, 3.3.1, 4.3.1

Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7817–7826, 2024a. (document), 2.7

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. (document), 2.2.3, 2.8, 2.3.2, 3.2.2, 3.4.1

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1921–1930, 2023. (document), 2.2.3, 2.9, 2.3.2

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024. (document), 2.2.3, 2.10, 4.2.3

Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *Advances in Neural Information Processing Systems*, 36:76872–76892, 2023. (document), 2.2.3, 2.11, 4.3.4, 4.3.4, 4.3.4, 4.3.4, 5.3.3, 5.3.3

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. (document), 2.3.1, 2.12

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. (document), 2.3.1, 2.13

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023a. (document), 1, 2.3.2, 2.14, 2.15, 4.4.4, 5.5.1

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4296–4304, 2024. (document), 2.3.2, 2.16

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. (document), 2.3.2, 2.17, 2.18, 5.5.1

Anh-Dung Dinh, Daochang Liu, and Chang Xu. Rethinking conditional diffusion sampling with progressive guidance. *Advances in Neural Information Processing Systems*, 36:42285–42297, 2023. (document), 2.3.2, 2.19

Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8795–8805, 2024. (document), 2.2.3, 2.3.2, 2.20, 5.5.1

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023. (document), 2.3.2, 2.21, 3.2.2, 4.2.3, 4.3.3

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7452–7461, 2023. (document), 2.3.2, 2.22, 4.2.3, 4.3.4, 4.3.4, 4.4.6, 4.9

Guillaume Couairon, Marlene Careil, Matthieu Cord, Stéphane Lathuiliere, and Jakob Verbeek. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2174–2183, 2023. (document), 2.3.2, 2.22, 4.2.3, 4.3.4, 4.3.4, 4.3.4, 4.4.4

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42(4):1–11, 2023a. (document), 2.3.2, 2.23

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019a. (document), 2.4.1, 2.24, 6.1, 6.4.1

Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay Chaudhari. Roentgen: vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022a. (document), 2.4.1, 2.25, 4.4.2, 6.1, 6.2.1, 6.3.1, 6.3.1, 6.4.1

Harim Kim, Yuhan Wang, Minkyu Ahn, Heeyoul Choi, Yuyin Zhou, and Charmgil Hong. Harnessing ehrs for diffusion-based anomaly detection on chest x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 235–245. Springer, 2025. (document), 2.4.1, 2.26

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018. (document), 2.4.2, 2.27

Mingzhe Hu, Shaoyan Pan, Chih-Wei Chang, Richard LJ Qiu, Junbo Peng, Tonghe Wang, Justin Roper, Hui Mao, David Yu, and Xiaofeng Yang. Cross-modality 3d mri synthesis via cycle-guided denoising diffusion probability model. *Journal of Medical Imaging*, 12(6):064003–064003, 2025. (document), 2.4.2, 2.28

Errol Colak, Felipe C Kitamura, Stephen B Hobbs, Carol C Wu, Matthew P Lungren, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Robyn L Ball, George Shih, Anouk Stein, et al. The rsna pulmonary embolism ct dataset. *Radiology: Artificial Intelligence*, 3(2):e200254, 2021. (document), 2.4.3, 2.29

Qi Gao, Zhihao Chen, Dong Zeng, Junping Zhang, Jianhua Ma, and Hongming Shan. Noise-inspired diffusion model for generalizable low-dose ct reconstruction. *arXiv preprint arXiv:2506.22012*, 2025. (document), 2.4.3, 2.30

Shuo Han, Yongshun Xu, Dayang Wang, Bahareh Morovati, Li Zhou, Jonathan S Maltz, Ge Wang, and Hengyong Yu. Physics-informed score-based diffusion model for limited-angle reconstruction of cardiac computed tomography. *IEEE Transactions on Medical Imaging*, 2024. (document), 2.4.3, 2.31

Jia He, Bonan Li, Ge Yang, and Ziwen Liu. Blaze3dm: Marry triplane representation with diffusion for 3d medical inverse problem solving. *arXiv preprint arXiv:2405.15241*, 2024. (document), 2.4.3, 2.32

Muhammad Ali Farooq, Wang Yao, Michael Schukat, Mark A Little, and Peter Corcoran. Derm-t2im: Harnessing synthetic skin lesion data via stable diffusion models for enhanced skin disease classification using vit and cnn. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–5. IEEE, 2024. (document), 2.4.4, 2.34, 4.1, 4.2.2, 5.2.3

Qixuan Jin, Walter Gerych, and Marzyeh Ghassemi. Maskmedpaint: Masked medical image inpainting with diffusion models for mitigation of spurious correlations. *arXiv preprint arXiv:2411.10686*, 2024. (document), 2.4.4, 2.35

Afshin Bozorgpour, Yousef Sadegheih, Amirhossein Kazerouni, Reza Azad, and Dorit Merhof. Dermosegdiff: A boundary-aware segmentation diffusion model for skin lesion delineation. In *International workshop on predictive intelligence in medicine*, pages 146–158. Springer, 2023. (document), 2.4.4, 2.36

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023. (document), 2.2.3, 2.3.2, 4.4.6, 4.10

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR, 2021. (document), 6.2.2, 6.3.2, 6.4.2, 6.3

Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018(1):7068349, 2018. 1

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 3.2.1, 4.3.1, 4.4.4, 6.1

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2.2.1, 6.3.1

Mohammad Hosseini and Mahmudul Hasan. Faster, lighter, more accurate: A deep learning ensemble for content moderation. *arXiv preprint arXiv:2309.05150*, 2023. 1

Stephen Balaban. Deep learning and face recognition: the state of the art. *Biometric and surveillance technology for human and activity identification XII*, 9457:68–75, 2015. 1

GSDMA Sreenu and Saleem Durai. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. *Journal of Big Data*, 6(1):1–27, 2019. 1

Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of field robotics*, 37 (3):362–386, 2020. 1

Shijie Hao, Yuan Zhou, and Yanrong Guo. A brief survey on semantic segmentation with deep learning. *Neurocomputing*, 406:302–321, 2020. 1

Shun Yang, Wenshuo Wang, Chang Liu, and Weiwen Deng. Scene understanding in deep learning-based end-to-end controllers for autonomous vehicles. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(1):53–63, 2018. 1

Yang Zhou, Cai Yang, Ping Wang, Chao Wang, Xinhong Wang, and Nguyen Ngoc Van. Vit-fusenet: Multimodal fusion of vision transformer for vehicle-infrastructure cooperative perception. *IEEE Access*, 12:31640–31651, 2024a. 1

Meiqiao Bi, Minghua Wang, Zhi Li, and Danfeng Hong. Vision transformer with contrastive learning for remote sensing image scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:738–749, 2022. 1

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 2.1.1, 2.1.1, 3.2.1, 4.3.1

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

Anoop Cherian and Alan Sullivan. Sem-gan: Semantically-consistent image-to-image translation. In *2019 ieee winter conference on applications of computer vision (wacv)*, pages 1797–1806. IEEE, 2019. 1

Xining Zhu, Lin Zhang, Lijun Zhang, Xiao Liu, Ying Shen, and Shengjie Zhao. Gan-based image super-resolution with a novel quality loss. *Mathematical Problems in Engineering*, 2020(1):5217429, 2020. 1

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1, 2.2.2, 2.3.2, 3.1, 4.3.4, 5.2.3, 6.2.1

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36:16083–16099, 2023a. 1

Tao Geng and Yuxuan Yang. Diffdesign: Controllable diffusion with meta prior for efficient interior design generation. *PloS one*, 20(9):e0331240, 2025. 1

Jiawei Luo, Liren Yang, Yan Liu, Changbao Hu, Grant Wang, Yan Yang, Tie-Lin Yang, and Xiaobo Zhou. Review of diffusion models and its applications in biomedical informatics. *BMC Medical Informatics and Decision Making*, 25(1): 390, 2025. 1, 2.4

Simon M Thomas, James G Lefevre, Glenn Baxter, and Nicholas A Hamilton. Interpretable deep learning systems for multi-class segmentation and classification of non-melanoma skin cancer. *Medical Image Analysis*, 68:101915, 2021a. 1

Ivo M Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific reports*, 9(1):6381, 2019. 1

Xiaohong W Gao, Rui Hui, and Zengmin Tian. Classification of ct brain images based on deep learning networks. *Computer methods and programs in biomedicine*, 138:49–56, 2017. 1

Jin Liu, Yi Pan, Min Li, Ziyue Chen, Lu Tang, Chengqian Lu, and Jianxin Wang. Applications of deep learning to mri images: A survey. *Big Data Mining and Analytics*, 1(1):1–18, 2018. 1

Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. Resvit: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614, 2022. 1

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022. 1

Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21152–21164, 2023a. 1, 6.1

Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *Computerized medical imaging and graphics*, 79:101684, 2020. 1

Pankaj Bamoriya, Gourav Siddhad, Harkeerat Kaur, Pritee Khanna, and Aparajita Ojha. Dsb-gan: Generation of deep learning based synthetic biometric data. *Displays*, 74:102267, 2022. 1, 4.1

Sukun Tian, Renkai Huang, Zhenyang Li, Luca Fiorenza, Ning Dai, Yuchun Sun, and Haifeng Ma. A dual discriminator adversarial learning approach for dental occlusal surface reconstruction. *Journal of Healthcare Engineering*, 2022, 2022. 1

Bulat Maksudov, Kathleen M Curran, and Alessandra Mileo. Anatomy-preserving counterfactual edits in breast mri via guided diffusion. In *Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care*, pages 206–215. Springer, 2025. 1

Minjae Jeong, Hyuna Cho, Sungyoon Jung, and Won Hwa Kim. Uncertainty-aware diffusion-based adversarial attack for realistic colonoscopy image synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 647–658. Springer, 2024a. 1

Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2025. 1, 2.3, 3.2.1, 4.2.3

Alex Ling Yu Hung, Kai Zhao, Haoxin Zheng, Ran Yan, Steven S Raman, Demetri Terzopoulos, and Kyunghyun Sung. Med-cdiff: Conditional medical image generation with diffusion models. *Bioengineering*, 10(11):1258, 2023. 1

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1.1, 3.1, 3.2.1

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023a. 1.1, 3.1, 3.2.1, 3.4.1, 3.6.1, 3.6.3, 4.4.2

Tri Huynh, Aiden Nibali, and Zhen He. Semi-supervised learning for medical image classification using imbalanced training data. *Computer methods and programs in biomedicine*, 216:106628, 2022. 1.1

Ashwini Kumar Upadhyay and Ashish Kumar Bhandari. Advances in deep learning models for resolving medical image segmentation data scarcity problem: a topical review. *Archives of Computational Methods in Engineering*, 31(3):1701–1719, 2024. 1.1

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 1.1, 2.4.4, 4.4.1

Marin Benčević, Marija Habijan, Irena Galić, Danilo Babin, and Aleksandra Pižurica. Understanding skin color bias in deep learning-based skin lesion segmentation. *Computer methods and programs in biomedicine*, 245:108044, 2024. 1.1, 4.5, 5.1, 5.2.1, 5.2.2

Masahiro Suzuki and Yutaka Matsuo. A survey of multimodal deep generative models. *Advanced Robotics*, 36(5-6):261–278, 2022. 2.1

Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49. JMLR Workshop and Conference Proceedings, 2012. 2.1.1

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016. 2.1.1

Anika Shrivastava, Renu Rameshan, and Samar Agnihotri. Latent space characterization of autoencoder variants. *arXiv preprint arXiv:2412.04755*, 2024. 2.1.1

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2.1.2

Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J Doug Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58, 2011. 2.1.2

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017. 2.1.2

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 2.1.2

Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022. 2.1.3

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2.1.3

Stanley Chan et al. Tutorial on diffusion models for imaging and vision. *Foundations and Trends® in Computer Graphics and Vision*, 16(4):322–471, 2024. 2.1.3

Pierre Chambon, Christian Bluethgen, Curtis P Langlotz, and Akshay Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv preprint arXiv:2210.04133*, 2022b. 2.1.4, 6.1, 6.2.1, 6.3.1, 6.3.1

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2.2.1

Jian Li, Yue Wang, Michael R Lyu, and Irwin King. Code completion with neural attention and pointer networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4159–25, 2018. 2.2.1

George Seif and Dimitrios Androutsos. Large receptive field networks for high-scale image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 763–772, 2018. 2.2.1

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2.2.1, 2.2.2, 3.2.1, 6.2.2

Trong-Tung Nguyen, Duc-Anh Nguyen, Anh Tran, and Cuong Pham. Flexedit: Flexible and controllable diffusion-based object-centric image editing. *arXiv preprint arXiv:2403.18605*, 2024a. 2.2.2, 3.2.2, 3.3.1, 4.3.4

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2.2.2, 2.3.1

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 2.2.2

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 2.2.2

Gemma E Moran and Bryon Aragam. Towards interpretable deep generative models via causal representation learning. *arXiv preprint arXiv:2504.11609*, 2025. 2.2.3

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Türe. What the daam: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5644–5659, 2023b. 2.2.3, 2.3.2, 4.4.5

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a. 2.2.3, 5.3.3

Yinwei Wu, Xingyi Yang, and Xinchao Wang. Relation rectification in diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7685–7694, 2024a. 2.2.3

Sunung Mun, Jinhwan Nam, Sunghyun Cho, and Jungseul Ok. Addressing attribute leakages in diffusion-based image editing without training. *arXiv preprint arXiv:2412.04715*, 2024. 2.2.3

Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5343–5353, 2024a. 2.2.3, 4.2.3

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22560–22570, 2023. 2.2.3, 2.2.3, 2.3.2

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2.2.3

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2.2.3

Cusuh Ham, Matthew Fisher, James Hays, Nicholas Kolkin, Yuchen Liu, Richard Zhang, and Tobias Hinz. Personalized residuals for concept-driven text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8186–8195, 2024. 2.2.3

Ryota Yoshihashi, Yuya Otsuka, Kenji Doi, and Tomohiro Tanaka. Attention as annotation: Generating images and pseudo-masks for weakly supervised semantic segmentation with diffusion. *CoRR*, 2023. 2.2.3

Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1206–1217, 2023a. 2.2.3, 3.2.2

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011. 2.2.3, 3.2.3

Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4981–4990, 2018. 2.2.3

Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. *Advances in Neural Information Processing Systems*, 36:54683–54695, 2023b. 2.2.3, 4.3.4

Chaofan Ma, Yuhuan Yang, Chen Ju, Fei Zhang, Jinxiang Liu, Yu Wang, Ya Zhang, and Yanfeng Wang. Diffusionseg: Adapting diffusion towards unsupervised object discovery. *arXiv preprint arXiv:2303.09813*, 2023a. 2.2.3

Aliasghar Khani, Saeid Asgari Taghanaki, Aditya Sanghi, Ali Mahdavi Amiri, and Ghassan Hamarneh. Slime: Segment like me. *arXiv preprint arXiv:2309.03179*, 2023. 2.2.3

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016. 2.3

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017. 2.3.1

Ikram Eddahmani, Chi-Hieu Pham, Thibault Napoleon, Isabelle Badoc, Jean-Rassaire Fouefack, and Marwa El-Bouz. Unsupervised learning of disentangled representation via auto-encoding: A survey. *Sensors*, 23(4):2362, 2023. 2.3.1

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pages 2649–2658. PMLR, 2018. 2.3.1

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 2.3.1, 4.4.4

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2.3.1

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pages 1857–1865. Pmlr, 2017. 2.3.1

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 2.3.1

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2.3.1

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2.3.1

Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, and Timo Bremer. An unsupervised approach to solving inverse problems using generative adversarial networks. *arXiv preprint arXiv:1805.07281*, 2018. 2.3.1

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018. 2.3.1

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2.3.2, 3.2.1

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494, 2022. 2.3.2, 3.2.1

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2.3.2

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b. 2.3.2, 4.3.4

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 843–852, 2023. 2.3.2

Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7280–7290, 2023. 2.3.2

Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 conference papers*, pages 1–12, 2024. 2.3.2

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10146–10156, 2023b. 2.3.2

Ying Hu, Chenyi Zhuang, and Pan Gao. Diffusest: Unleashing the capability of the diffusion model for style transfer. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pages 1–1, 2024. 2.3.2

Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide & bind your attention for improved generative semantic nursing. *arXiv preprint arXiv:2307.10864*, 2023a. 2.3.2

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36:3536–3559, 2023. 2.3.2, 4.2.3

Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7932–7942, 2024. 2.3.2, 4.2.3, 4.3.4, 4.3.4, 4.4.4, 5.3.3

Huancheng Chen, Jingtao Li, Weiming Zhuang, Haris Vikalo, and Lingjuan Lyu. Boundary attention constrained zero-shot layout-to-image generation. *arXiv preprint arXiv:2411.10495*, 2024b. 2.3.2

Jiayu Xiao, Henglei Lv, Liang Li, Shuhui Wang, and Qingming Huang. R&b: Region and boundary aware zero-shot grounded text-to-image generation. *arXiv preprint arXiv:2310.08872*, 2023. 2.3.2

Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023. 2.3.2

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *International Conference on Machine Learning*, pages 1737–1752. PMLR, 2023. 2.3.2

Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475, 2024. 2.3.2

Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*, 2023. 2.3.2

Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. Renoise: Real image inversion through iterative noising. In *European Conference on Computer Vision*, pages 395–413. Springer, 2024. 2.3.2

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical image analysis*, 88:102846, 2023. 2.4

George Webber and Andrew J Reader. Diffusion models for medical image reconstruction. *BJR| Artificial Intelligence*, 1(1):ubae013, 2024. 2.4

Hassan K Ahmad, Michael R Milne, Quinlan D Buchlak, Nalan Ektas, Georgina Sanderson, Hadi Chamtie, Sajith Karunasena, Jason Chiang, Xavier Holt, Cyril HM Tang, et al. Machine learning augmented interpretation of chest x-rays: a systematic review. *Diagnostics*, 13(4):743, 2023. 2.4.1

Lorena Álvarez-Rodríguez, Joaquim de Moura, Jorge Novo, and Marcos Ortega. Does imbalance in chest x-ray datasets produce biased deep learning approaches for covid-19 screening? *BMC Medical Research Methodology*, 22(1):125, 2022. 2.4.1

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019. 2.4.1, 6.4.2, 6.4.4

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2.4.1

Gregory Schuit, Denis Parra, and Cecilia Besa. Perceptual evaluation of gans and diffusion models for generating x-rays. In *International Workshop on Human-AI Collaboration*, pages 93–101. Springer, 2025. 2.4.1

Shemy Syed, R Elakkiya, and Nick Pears. Ssdm: A self supervised diffusion model for lung anomaly detection using chest x-rays. In *2025 International Conference on Communication, Computing, Networking, and Control in Cyber-Physical Systems (CCNCPS)*, pages 1–6. IEEE, 2025. 2.4.1

Zhanghao Chen, Yifei Sun, Ruiquan Ge, Wenjian Qin, Cheng Pan, Wenming Deng, Zhou Liu, Wenwen Min, Ahmed Elazab, Xiang Wan, et al. Bs-diff: Effective bone suppression using conditional diffusion models from chest x-ray images. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024c. 2.4.1

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2.4.1

Guillermo Jimenez-Perez, Pedro Osorio, Josef Cersovsky, Javier Montalt-Tordera, Jens Hooge, Steffen Vogler, and Sadegh Mohammadi. Dino-diffusion. scaling medical diffusion via self-supervised pre-training. *arXiv preprint arXiv:2407.11594*, 2024. 2.4.1

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2.4.1, 5.5.2

Abderrachid Hamrani and Anuradha Godavarty. Self-attention diffusion models for zero-shot biomedical image segmentation: Unlocking new frontiers in medical imaging. *Bioengineering*, 12(10):1036, 2025. 2.4.1

Maxim Zaitsev, Julian Maclaren, and Michael Herbst. Motion artifacts in mri: A complex problem with many partial solutions. *Journal of Magnetic Resonance Imaging*, 42(4):887–901, 2015. 2.4.2

Frank Godenschweger, Urte Kägebein, Daniel Stucht, Uten Yarach, Alessandro Sciarra, Renat Yakupov, Falk Lüsebrink, Peter Schulze, and Oliver Speck. Motion correction in mri of the brain. *Physics in medicine & biology*, 61(5):R32, 2016. 2.4.2

Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021. 2.4.2

Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri. *Medical image analysis*, 80:102479, 2022. 2.4.2

Rohan Sanda, Asad Aali, Andrew Johnston, Eduardo Pontes Reis, Gordon Wetzstein, and Sara Fridovich-Keil. Padis-mri: Patch-based diffusion for data-efficient, radiologist-preferred mri reconstruction. In *Machine Learning for Health 2025*, 2025. 2.4.2

Mojtaba Safari, Shansong Wang, Zach Eidex, Qiang Li, Richard LJ Qiu, Erik H Middlebrooks, David S Yu, and Xiaofeng Yang. Mri super-resolution reconstruction using efficient diffusion probabilistic model with residual shifting. *Physics in Medicine & Biology*, 70(12):125008, 2025. 2.4.2

Asad Aali, Giannis Daras, Brett Levac, Sidharth Kumar, Alexandros G Dimakis, and Jonathan I Tamir. Ambient diffusion posterior sampling: Solving inverse problems with diffusion models trained on corrupted data. *arXiv preprint arXiv:2403.08728*, 2024. 2.4.2, 2.4.3

Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, Pierre Vera, and Su Ruan. Amm-diff: Adaptive multi-modality diffusion network for missing modality imputation. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2025. 2.4.2

Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Cukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42(12): 3524–3539, 2023. 2.4.2

Wei Peng, Ehsan Adeli, Tomas Bosschieter, Sang Hyun Park, Qingyu Zhao, and Kilian M Pohl. Generating realistic brain mris via a conditional diffusion probabilistic model. In *International conference on medical image computing and computer-assisted intervention*, pages 14–24. Springer, 2023. 2.4.2

Florentin Bieder, Julia Wolleb, Alicia Durrer, Robin Sandkuehler, and Philippe C Cattin. Memory-efficient 3d denoising diffusion models for medical image processing. In *Medical Imaging with Deep Learning*, pages 552–567. PMLR, 2024. 2.4.2, 2.4.3

Kai Zhao, Kaifeng Pang, Alex Ling Yu Hung, Haoxin Zheng, Ran Yan, and Kyunghyun Sung. Mri super-resolution with partial diffusion models. *IEEE transactions on medical imaging*, 2024. 2.4.2

Euclid Seeram. Computed tomography: physical principles and recent technical advances. *Journal of Medical Imaging and Radiation Sciences*, 41(2):87–109, 2010. 2.4.3

Timothy P Szczykutowicz, Giuseppe V Toia, Amar Dhanantwari, and Brian Nett. A review of deep learning ct reconstruction: concepts, limitations, and promise in clinical practice. *Current Radiology Reports*, 10(9):101–115, 2022. 2.4.3

Ezgi Demircan-Tureyen and Mustafa E Kamasak. A discretized tomographic image reconstruction based upon total variation regularization. *Biomedical Signal Processing and Control*, 38:44–54, 2017. 2.4.3

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011. 2.4.3

Adam E Flanders, Luciano M Prevedello, George Shih, Safwan S Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T Mongan, Anouk Stein, Felipe C Kitamura, Matthew P Lungren, et al. Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge. *Radiology: Artificial Intelligence*, 2(3):e190211, 2020. 2.4.3

Farzan Niknejad Mazandarani, Paul Babyn, and Javad Alirezaie. Sadiff: A sinogram-aware diffusion model for low-dose ct image denoising. *Journal of Imaging Informatics in Medicine*, pages 1–21, 2025. 2.4.3

Shudong Li, Xiao Jiang, Matthew Tivnan, Grace J Gang, Yuan Shen, and J Webster Stayman. Ct reconstruction using diffusion posterior sampling conditioned on a nonlinear measurement model. *Journal of Medical Imaging*, 11(4):043504–043504, 2024. 2.4.3

Mohammed A Mahdi, Mohammed Al-Shalabi, Ehab T Alnfrawy, Reda Elbarougy, Muhammad Usman Hadi, and Rao Faizan Ali. 3d latent diffusion model for mr-only radiotherapy: Accurate and consistent synthetic ct generation. *Diagnostics*, 15(23):3010, 2025. 2.4.3

Xuhui Liu, Zhi Qiao, Runkun Liu, Hong Li, Juan Zhang, Xiantong Zhen, Zhen Qian, and Baochang Zhang. Diffux2ct: Diffusion learning to reconstruct ct images from biplanar x-rays. In *European conference on computer vision*, pages 458–476. Springer, 2024b. 2.4.3

Daniele Molino, Camillo Maria Caruso, Filippo Ruffini, Paolo Soda, and Valerio Guarrasi. Text-to-ct generation via 3d latent diffusion model with contrastive vision-language pretraining. *arXiv preprint arXiv:2506.00633*, 2025. 2.4.3

Harold Kittler, H Pehamberger, K Wolff, and MJTIO Binder. Diagnostic accuracy of dermoscopy. *The lancet oncology*, 3(3):159–165, 2002. 2.4.4, 5.1

Lituan Wang, Lei Zhang, Xin Shu, and Zhang Yi. Intra-class consistency and inter-class discrimination feature learning for automatic skin lesion classification. *Medical Image Analysis*, 85:102746, 2023a. 2.4.4, 4.2.1

Anil Kumar Adepu, Subin Sahayam, Umarani Jayaraman, and Rashmika Arramraju. Melanoma classification from dermatoscopy images using knowledge distillation for highly imbalanced data. *Computers in Biology and Medicine*, 154:106571, 2023. 2.4.4

David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016. 2.4.4

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018. 2.4.4

Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):34, 2021. 2.4.4

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 2.4.4, 4.4.1, 5.5.1

Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. In

2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC), pages 5437–5440. IEEE, 2013. 2.4.4, 4.4.1

Muhammad Danish Ali, Muhammad Ali Iqbal, Sejong Lee, Xiaoyun Duan, and Soo Kyun Kim. Explainable ai based multi class skin cancer detection enhanced by meta learning with generative ddpm data augmentation. *Applied Sciences*, 15 (21):11689, 2025. 2.4.4

Payal Varshney, Adriano Lucieri, Christoph Balada, Andreas Dengel, and Sheraz Ahmed. Discovering concept directions from diffusion-based counterfactuals via latent clustering. *arXiv preprint arXiv:2505.07073*, 2025. 2.4.4

Bilel Benjdira, Anas M. Ali, Anis Koubaa, Adel Ammar, and Wadii Boulila. Dm–ahr: A self-supervised conditional diffusion model for ai-generated hairless imaging for enhanced skin diagnosis applications. *Cancers*, 16(17):2947, 2024. 2.4.4

Jiacheng Wang, Jing Yang, Qichao Zhou, and Liansheng Wang. Medical boundary diffusion model for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 427–436. Springer, 2023b. 2.4.4

Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 6030–6038, 2024b. 2.4.4

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2.4.4, 4.2.1

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3.1, 3.2.1, 6.2.1

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 3.1, 6.2.1

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 3.1

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 3.1, 3.2.1, 3.4.1, 3.4.1, 3.6.1

Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023b. 3.1, 3.2.1

Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023. 3.1

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023a. 3.1, 3.2.1

Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H Bermano. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06925*, 2023. 3.1

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023b. 3.1, 3.2.1

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*, 2023b. 3.1, 3.2.1, 3.2.2

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023b. 3.1, 3.2.2, 3.3.3, 3.4.1

Chen Jin, Ryutaro Tanno, Amrutha Saseendran, Tom Diethe, and Philip Teare. An image is worth multiple words: Learning object level concepts using multi-concept prompt learning. *arXiv preprint arXiv:2310.12274*, 2023. 3.1, 3.2.3

Tanzila Rahman, Shweta Mahajan, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, and Leonid Sigal. Visual concept-driven image generation with text-to-image diffusion model. *arXiv preprint arXiv:2402.11487*, 2024. 3.1, 3.2.3

Mehdi Safaee, Aryan Mikaeili, Or Patashnik, Daniel Cohen-Or, and Ali Mahdavi-Amiri. Clic: Concept learning in context. *arXiv preprint arXiv:2311.17083*, 2023. 3.1, 3.2.2, 3.2.3

Yanbing Zhang, Mengping Yang, Qin Zhou, and Zhe Wang. Attention calibration for disentangled text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4764–4774, 2024. 3.1, 3.2.1, 3.2.3, 3.3.1, 3.3.2, 3.4.1, 3.4.1, 3.6.1

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3.1, 3.2.3, 3.6.3

Xiaoyu Wu, Jiaru Zhang, Yang Hua, Bohan Lyu, Hao Wang, Tao Song, and Haibing Guan. Exploring diffusion models' corruption stage in few-shot fine-tuning and mitigating with bayesian neural networks. *arXiv preprint arXiv:2405.19931*, 2024c. 3.1, 3.3.2

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3.2.1

Yuxuan Zhang, Jiaming Liu, Yiren Song, Rui Wang, Hao Tang, Jinpeng Yu, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. *arXiv preprint arXiv:2312.16272*, 2023c. 3.2.1

Yufei Cai, Yuxiang Wei, Zhilong Ji, Jinfeng Bai, Hu Han, and Wangmeng Zuo. Decoupled textual embeddings for customized image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 909–917, 2024. 3.2.1

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3.2.1

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 3.2.1

Hong Chen, Yipeng Zhang, Simin Wu, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2023a. 3.2.1, 3.2.2

Yang Yang, Wen Wang, Liang Peng, Chaotian Song, Yao Chen, Hengjia Li, Xiaolong Yang, Qinglin Lu, Deng Cai, Boxi Wu, et al. Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *arXiv preprint arXiv:2403.11627*, 2024. 3.2.1

Shaozhe Hao, Kai Han, Shihao Zhao, and Kwan-Yee K Wong. Vico: Detail-preserving visual condition for personalized text-to-image generation. *arXiv preprint arXiv:2306.00971*, 2023. 3.2.1, 3.2.2, 3.2.3

Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. *arXiv preprint arXiv:2305.19327*, 2023b. 3.2.1, 3.2.2

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 3.2.1

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023. 3.2.1, 3.2.2

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36, 2024d. 3.2.1

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 3.2.2

Ruichen Wang, Zekang Chen, Chen Chen, Jian Ma, Haonan Lu, and Xiaodong Lin. Compositional text-to-image synthesis with attention map control of diffusion models. *arXiv preprint arXiv:2305.13921*, 2023c. 3.2.2

Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. *arXiv preprint arXiv:2306.05427*, 2023. 3.2.2

Wan-Duo Kurt Ma, JP Lewis, W Bastiaan Kleijn, and Thomas Leung. Directed diffusion: Direct control of object placement through attention guidance. *arXiv preprint arXiv:2302.13153*, 2023c. 3.2.2

Yutong He, Ruslan Salakhutdinov, and J Zico Kolter. Localized text-to-image generation for free via cross attention control. *arXiv preprint arXiv:2306.14636*, 2023. 3.2.2

Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023b. 3.2.2

Junjie Shentu, Matthew Watson, and Noura Al Moubayed. Textual localization: Decomposing multi-concept images for subject-driven text-to-image generation. *arXiv preprint arXiv:2402.09966*, 2024. 3.2.2, 3.6.3

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *arXiv preprint arXiv:2402.03286*, 2024. 3.2.2

Jaeseok Jeong, Junho Kim, Yunjey Choi, Gayoung Lee, and Youngjung Uh. Visual style prompting with swapping self-attention. *arXiv preprint arXiv:2402.12974*, 2024b. 3.2.2

Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. *arXiv preprint arXiv:2308.12469*, 2023. 3.2.2

Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *Advances in Neural Information Processing Systems*, 36, 2024b. 3.2.2, 3.3.1, 3.4.1

Pablo Marcos-Manchón, Roberto Alcover-Couso, Juan C SanMiguel, and Jose M Martínez. Open-vocabulary attention maps with token optimization for semantic segmentation in diffusion models. *arXiv preprint arXiv:2403.14291*, 2024. 3.2.2

Nobuyuki Otsu et al. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975. 3.2.3, 4.3.4, 5.3.3

Chun-Hsiao Yeh, Ta-Ying Cheng, He-Yen Hsieh, Chuan-En Lin, Yi Ma, Andrew Markham, Niki Trigoni, HT Kung, and Yubei Chen. Gen4gen: Generative data pipeline for generative multi-concept composition. *arXiv preprint arXiv:2402.15504*, 2024. 3.4.1, 3.6.1

Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 3.4.1

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022. 3.4.1, 4.3.3

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022. 3.6.3

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023c. 3.6.3

Meg Watson, Erin Garnett, Gery P Guy, and Dawn M Holman. The surgeon general's call to action to prevent skin cancer. *International Journal of Cancer Research and Prevention*, 8(1):55, 2015. 4.1

Rebecca L Siegel, Angela N Giaquinto, and Ahmedin Jemal. Cancer statistics, 2024. *CA: a cancer journal for clinicians*, 74(1):12–49, 2024. 4.1

Fengying Xie, Haidi Fan, Yang Li, Zhiguo Jiang, Rusong Meng, and Alan Bovik. Melanoma classification on dermoscopy images using a neural network ensemble model. *IEEE transactions on medical imaging*, 36(3):849–858, 2016. 4.1

ME Vestergaard, PHPM Macaskill, PE Holt, and SW Menzies. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. *British Journal of Dermatology*, 159(3):669–676, 2008. 4.1, 5.1

Margarida Silveira, Jacinto C Nascimento, Jorge S Marques, André RS Marçal, Teresa Mendonça, Syogo Yamauchi, Junji Maeda, and Jorge Rozeira. Comparison of segmentation methods for melanoma diagnosis in dermoscopy images. *IEEE journal of selected topics in signal processing*, 3(1):35–45, 2009. 4.1

Esther E Freeman. Global health dermatology: An emerging field addressing the access to care crisis. *Indian Journal of Dermatology, Venereology and Leprology*, 90(1):3–4, 2023. 4.1

Rongtao Xu, Changwei Wang, Jiguang Zhang, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. Skinformer: Learning statistical texture representation with transformer for skin lesion segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2024. 4.1

Zhiwei Qin, Zhao Liu, Ping Zhu, and Yongbo Xue. A gan-based image synthesis method for skin lesion classification. *Computer Methods and Programs in Biomedicine*, 195:105568, 2020. 4.1, 4.2.2, 4.4.2, 5.2.3

Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4-5): 198–211, 2007. 4.1

Maryellen L Giger, Heang-Ping Chan, and John Boone. Anniversary paper: history and status of cad and quantitative image analysis: the role of medical physics and aapm. *Medical physics*, 35(12):5799–5820, 2008. 4.1

Tae-Yun Kim, Jaebum Son, and Kwang-Gi Kim. The recent progress in quantitative medical image analysis for computer aided diagnosis systems. *Healthcare informatics research*, 17(3):143–149, 2011. 4.1

Minji Kang, Tai Joon An, Deokjae Han, Wan Seo, Kangwon Cho, Shinbum Kim, Jun-Pyo Myong, and Sung Won Han. Development of a multipotent diagnostic tool for chest x-rays by multi-object detection method. *Scientific reports*, 12(1): 19130, 2022. 4.1

Mohammad H Alshayeji, Silpa ChandraBhasi Sindhu, and Sa'ed Abed. Cad systems for covid-19 diagnosis and disease stage classification by segmentation of infected regions from ct images. *BMC bioinformatics*, 23(1):264, 2022. 4.1

Kamal Hammouda, Fahmi Khalifa, Ahmed Soliman, Mohammed Ghazal, Mohamed Abou El-Ghar, Mohammed Ali Badawy, Hanan E Darwish, Adel Khelifi, and Ayman El-Baz. A multiparametric mri-based cad system for accurate diagnosis of bladder cancer staging. *Computerized Medical Imaging and Graphics*, 90:101911, 2021. 4.1

Md Kamrul Hasan, Md Asif Ahamad, Choon Hwai Yap, and Guang Yang. A survey, review, and future trends of skin lesion segmentation and classification. *Computers in Biology and Medicine*, 155:106624, 2023. 4.1

Ibrahim Saad Aly Abdelhalim, Mamdouh Farouk Mohamed, and Yousef Bassyouni Mahdy. Data augmentation for skin lesion using self-attention based progressive generative adversarial network. *Expert Systems with Applications*, 165:113922, 2021. 4.1, 4.2.2

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017. 4.1

Tobias Weber, Michael Ingrisch, Bernd Bischl, and David Rügamer. Cascaded latent diffusion models for high-resolution chest x-ray synthesis. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 180–191. Springer, 2023. 4.1, 6.1, 6.2.1, 6.5.2

Zhihang Ren, Yunhui Guo, X Yu Stella, and David Whitney. Improve image-based skin cancer diagnosis with generative self-supervised learning. In *2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 23–34. IEEE, 2021. 4.1, 4.2.2

Devansh Bisla, Anna Choromanska, Russell S Berman, Jennifer A Stein, and David Polsky. Towards automated melanoma detection with deep learning: Data purification and augmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 4.1, 4.2.2, 5.2.3

Veronika Shavlokhova, Andreas Vollmer, Christos C Zouboulis, Michael Vollmer, Jakob Wollborn, Gernot Lang, Alexander Kübler, Stefan Hartmann, Christian Stoll, Elisabeth Roider, et al. Finetuning of glide stable diffusion model for ai-based text-conditional image synthesis of dermoscopic images. *Frontiers in medicine*, 10:1231436, 2023. 4.1, 4.2.2, 5.2.3

Kumar Abhishek and Ghassan Hamarneh. Mask2lesion: Mask-constrained adversarial skin lesion image synthesis. In *International workshop on simulation and synthesis in medical imaging*, pages 71–80. Springer, 2019. 4.1, 4.2.2, 5.2.3

Shiyi Du, Xiaosong Wang, Yongyi Lu, Yuyin Zhou, Shaoting Zhang, Alan Yuille, Kang Li, and Zongwei Zhou. Boosting dermatoscopic lesion segmentation via diffusion models with visual and textual prompts. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024. 4.1, 4.2.2, 5.2.3, 5.5.1

Roberta B Oliveira, Joao P Papa, Aledir S Pereira, and Joao Manuel RS Tavares. Computational methods for pigmented skin lesion classification in images: review and future trends. *Neural Computing and Applications*, 29:613–636, 2018. 4.2.1

Hassan Ashraf, Asim Waris, Muhammad Fazeel Ghafoor, Syed Omer Gilani, and Imran Khan Niazi. Melanoma segmentation using deep learning with test-time augmentations and conditional random fields. *Scientific Reports*, 12(1):3948, 2022. 4.2.1

Rafael Luz Araújo, Flávio HD de Araújo, and Romuere RV e Silva. Automatic segmentation of melanoma skin cancer using transfer learning and fine-tuning. *Multimedia Systems*, 28(4):1239–1250, 2022. 4.2.1

Marriam Nawaz, Tahira Nazir, Momina Masood, Farooq Ali, Muhammad Attique Khan, Usman Tariq, Naveera Sahar, and Robertas Damaševičius. Melanoma segmentation: A framework of improved DenseNet77 and UNET convolutional neural network. *Int J Imaging Syst Technol*, 2022. doi: 10.1002/ima.22750. URL `https://onlinelibrary.wiley.com/doi/10.1002/ima.22750`. 4.2.1

Simon M Thomas, James G Lefevre, Glenn Baxter, and Nicholas A Hamilton. Interpretable deep learning systems for multi-class segmentation and classification of non-melanoma skin cancer. *Medical Image Analysis*, 68:101915, 2021b. doi: 10.1016/j.media.2020.101915. URL `https://doi.org/10.1016/j.media.2020.101915`. 4.2.1

Qing Xu, Zhicheng Ma, HE Na, and Wenting Duan. Dcsau-net: A deeper and more compact split-attention u-net for medical image segmentation. *Computers in Biology and Medicine*, 154:106626, 2023. 4.2.1, 4.4.4, 5.4.2

Jiacheng Wang, Lan Wei, Liansheng Wang, Qichao Zhou, Lei Zhu, and Jing Qin. Boundary-aware transformers for skin lesion segmentation. In *Medical image computing and computer assisted intervention–mICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part i 24*, pages 206–216. Springer, 2021. 4.2.1

Jiacheng Wang, Fei Chen, Yuxi Ma, Liansheng Wang, Zhaodong Fei, Jianwei Shuai, Xiangdong Tang, Qichao Zhou, and Jing Qin. Xbound-former: Toward cross-scale boundary modeling in transformers. *IEEE Transactions on Medical Imaging*, 42 (6):1735–1745, 2023d. 4.2.1, 4.4.4

Hung Vu Quoc, Thao Tran Le Phuong, Minh Trinh Xuan, and Sang Dinh Viet. Lsegdiff: a latent diffusion model for medical image segmentation. In *Proceedings of the 12th International Symposium on Information and Communication Technology*, pages 456–462, 2023. 4.2.1

Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11536–11546, 2023. 4.2.1

Saket S Chaturvedi, Jitendra V Tembhurne, and Tausif Diwan. A multi-class skin cancer classification using deep convolutional neural networks. *Multimedia Tools and Applications*, 79(39):28477–28498, 2020. 4.2.1

Bhuvaneshwari Shetty, Roshan Fernandes, Anisha P Rodrigues, Rajeswari Chengoden, Sweta Bhattacharya, and Kuruva Lakshmanna. Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. *Scientific Reports*, 12(1):18134, 2022. 4.2.1

Selen Ayas. Multiclass skin lesion classification in dermoscopic images using swin transformer model. *Neural Computing and Applications*, 35(9):6713–6722, 2023. 4.2.1

Vlad-Constantin Lungu-Stan, Dumitru-Clementin Cercel, and Florin Pop. Skindistilvit: Lightweight vision transformer for skin lesion classification. In *International Conference on Artificial Neural Networks*, pages 268–280. Springer, 2023. 4.2.1

Yijun Yang, Huazhu Fu, Angelica I Aviles-Rivero, Carola-Bibiane Schönlieb, and Lei Zhu. Diffmic: Dual-guidance diffusion network for medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 95–105. Springer, 2023. 4.2.1

Mohamed Akrout, Bálint Gyepesi, Péter Holló, Adrienn Poór, Blága Kincső, Stephen Solis, Katrina Cirone, Jeremy Kawahara, Dekker Slade, Latif Abid, et al. Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 99–109. Springer, 2023. 4.2.2

Luke W Sagers, James A Diao, Luke Melas-Kyriazi, Matthew Groh, Pranav Rajpurkar, Adewole S Adamson, Veronica Rotemberg, Roxana Daneshjou, and Arjun K Manrai. Augmenting medical image classifiers with synthetic data from latent diffusion models. *arXiv preprint arXiv:2308.12453*, 2023. 4.2.2

Janet Wang, Yunsung Chung, Zhengming Ding, and Jihun Hamm. From majority to minority: A diffusion-based augmentation for underrepresented groups in skin lesion analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–23. Springer, 2024. 4.2.2

Bo Lin, Yingjing Xu, Xuanwen Bao, Zhou Zhao, Zuyong Zhang, Zhouyang Wang, Jie Zhang, Shuiguang Deng, and Jianwei Yin. Skingen: An explainable dermatology diagnosis-to-generation framework with interactive vision-language models. *arXiv preprint arXiv:2404.14755*, 2024. 4.2.2

Christoph Baur, Shadi Albarqouni, and Nassir Navab. Generating highly realistic images of skin lesions with gans. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis: First International Workshop, OR 2.0 2018, 5th International Workshop, CARE 2018, 7th International Workshop, CLIP 2018, Third International Workshop, ISIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings 5*, pages 260–267. Springer, 2018. 4.2.2, 4.4.2

Alceu Bissoto, Fábio Perez, Eduardo Valle, and Sandra Avila. Skin lesion synthesis with generative adversarial networks. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis: First International Workshop, OR 2.0 2018, 5th International Workshop, CARE 2018, 7th International Workshop, CLIP 2018,*

Third International Workshop, ISIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings 5, pages 294–302. Springer, 2018. 4.2.2

Federico Pollastri, Federico Bolelli, Roberto Paredes, and Costantino Grana. Augmenting data with gans to segment melanoma skin lesions. *Multimedia Tools and Applications*, 79(21):15575–15592, 2020. 4.2.2

Ali Shahsavari, Sima Ranjbari, and Toktam Khatibi. Proposing a novel cascade ensemble super resolution generative adversarial network (cesr-gan) method for the reconstruction of super-resolution skin lesion images. *Informatics in Medicine Unlocked*, 24:100628, 2021. 4.2.2

Yipeng Zhang, Quan Wang, and Bingliang Hu. Minimalgan: diverse medical image synthesis for data augmentation using minimal training data. *Applied Intelligence*, 53(4):3899–3916, 2023d. 4.2.2

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 4.2.3

Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7701–7711, 2023. 4.2.3

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 4.3.2

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023b. 4.3.3, 5.3.3

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. 4.3.3

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. 4.3.3, 5.3.3

Aliasghar Khani, Saeid Asgari Taghanaki, Aditya Sanghi, Ali Mahdavi Amiri, and Ghassan Hamarneh. Slime: Segment like me. *The Twelfth International Conference on Learning Representations*, 2024. 4.3.4, 5.3.3, 5.3.3

Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2):538–546, 2018. 4.4.1

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 4.4.4, 5.4.2

Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025. 4.4.6

Brandon Thompson, Toni Jenkins, John Paul Sánchez, Matthew Frederick, Alba Posligua-Alban, and Naiara Sbroggio Barbosa. Melanoma: Does it present differently in darker skin tones? *MedEdPORTAL*, 19:11311, 2023. 4.5

Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. Bk-sdm: A lightweight, fast, and cheap version of stable diffusion. In *European Conference on Computer Vision*, pages 381–399. Springer, 2024. 4.5

Zhenyu Zhou, Defang Chen, Can Wang, Chun Chen, and Siwei Lyu. Simple and fast distillation of diffusion models. *Advances in Neural Information Processing Systems*, 37:40831–40860, 2024b. 4.5

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1820–1828, 2021. 4.5, 5.1, 5.2.2, 5.4.1, 5.4.1

World Cancer Research Fund (WCRF). Skin cancer statistics, 2022. URL `https://www.wcrf.org/preventing-cancer/cancer-statistics/skin-cancer-statistics/`. 5.1

Gery P Guy Jr, Steven R Machlin, Donatus U Ekwueme, and K Robin Yabroff. Prevalence and costs of skin cancer treatment in the us, 2002- 2006 and 2007-2011. *American journal of preventive medicine*, 48(2):183–187, 2015a. 5.1

Gery P Guy Jr, Cheryll C Thomas, Trevor Thompson, Meg Watson, Greta M Massetti, Lisa C Richardson, Centers for Disease Control, Prevention (CDC), et al. Vital signs: melanoma incidence and mortality trends and projections-united states, 1982-2030. *MMWR Morb Mortal Wkly Rep*, 64(21):591–596, 2015b. 5.1

Rebecca L Siegel, Tyler B Kratzer, Angela N Giaquinto, Hyuna Sung, and Ahmedin Jemal. Cancer statistics, 2025. *Ca*, 75(1):10, 2025. 5.1

Olajumoke A Olateju, Jieni Li, J Douglas Thornton, and Rajender R Aparasu. Marginal health care expenditures for melanoma care in the united states. *Journal of Managed Care & Specialty Pharmacy*, 30(12):1364–1374, 2024. 5.1

American Cancer Society (ACS). 2023 cancer facts and figures, 2023. URL https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2023/2023-cancer-facts-and-figures.pdf. PDF document. 5.1

Hue Tran, Keng Chen, Adrian C Lim, James Jabbour, and Stephen Shumack. Assessing diagnostic skill in dermatology: a comparison between general practitioners and dermatologists. *Australasian journal of dermatology*, 46(4):230–234, 2005. 5.1

Md Kamrul Hasan, Shidhartho Roy, Chayan Mondal, Md Ashraful Alam, Md Toufick E Elahi, Aishwariya Dutta, SM Taslim Uddin Raju, Md Tasnim Jawad, and Mohiuddin Ahmad. Dermo-doctor: A framework for concurrent skin lesion detection and recognition using a deep convolutional neural network with end-to-end dual encoders. *Biomedical Signal Processing and Control*, 68:102661, 2021. 5.1

Zahra Mirikharaji, Kumar Abhishek, Alceu Bissoto, Catarina Barata, Sandra Avila, Eduardo Valle, M Emre Celebi, and Ghassan Hamarneh. A survey on deep learning for skin lesion segmentation. *Medical Image Analysis*, 88:102863, 2023. 5.1

Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31):eabq6147, 2022. 5.1, 5.2.2, 5.4.1

Peter J Bevan and Amir Atapour-Abarghouei. Detecting melanoma fairly: Skin tone detection and debiasing for skin lesion classification. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pages 1–11. Springer, 2022. 5.1, 5.2.2, 5.4.1

Adam Corbin and Oge Marques. Assessing bias in skin lesion classifiers with contemporary deep learning and post-hoc explainability techniques. *IEEE Access*, 11:78339–78352, 2023. 5.1, 5.2.1

Andrés Morales-Forero, Lili J Rueda, Ronald Herrera, Samuel Bassetto, and Eric Coatanea. Predictive representativity: Uncovering racial bias in ai-based skin cancer detection. *arXiv preprint arXiv:2507.14176*, 2025. 5.1

Sribala Vidyadhari Chinta, Zichong Wang, Avash Palikhe, Xingyu Zhang, Ayesha Kashif, Monique Antoinette Smith, Jun Liu, and Wenbin Zhang. Ai-driven healthcare: A review on ensuring fairness and mitigating bias. *arXiv preprint arXiv:2407.19655*, 2024. 5.2.1

Anthony Paproki, Olivier Salvado, and Clinton Fookes. Synthetic data for deep learning in computer vision & medical imaging: A means to reduce data bias. *ACM Computing Surveys*, 56(11):1–37, 2024. 5.2.1

Agnieszka Mikołajczyk, Sylwia Majchrowska, and Sandra Carrasco Limeros. The (de) biasing effect of gan-based augmentation methods on skin lesion images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 437–447. Springer, 2022. 5.2.1

Andres Morales-Forero, Lili Rueda Jaime, Sebastian Ramiro Gil-Quiñones, Marlon Y Barrera Montañez, Samuel Bassetto, and Eric Coatanea. An insight into racial bias in dermoscopy repositories: A ham10000 data set analysis. *JEADV Clinical Practice*, 3(3):836–843, 2024. 5.2.2

Girmaw Abebe Tadesse, Celia Cintas, Kush R Varshney, Peter Staar, Chinyere Agunwa, Skyler Speakman, Justin Jia, Elizabeth E Bailey, Ademide Adelekun, Jules B Lipoff, et al. Skin tone analysis for representation in educational materials (star-ed) using machine learning. *NPJ Digital Medicine*, 6(1):151, 2023. 5.2.2

Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Taylan Cemgil, et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, pages 1–8, 2024. 5.2.3, 5.3.3

Fábio Perez, Cristina Vasconcelos, Sandra Avila, and Eduardo Valle. Data augmentation for skin lesion analysis. In *International Workshop on Computer-Assisted and Robotic Endoscopy*, pages 303–311. Springer, 2018. 5.2.3

Kavita Behara, Ernest Bhero, and John Terhile Agee. Skin lesion synthesis and classification using an improved dcgan classifier. *Diagnostics*, 13(16):2635, 2023. 5.2.3

Agnieszka Mikołajczyk and Michał Grochowski. Style transfer-based image synthesis as an efficient regularization technique in deep learning. In *2019 24Th International conference on methods and models in automation and robotics (MMAR)*, pages 42–47. IEEE, 2019. 5.2.3

Haipeng Liu, Yang Wang, Biao Qian, Meng Wang, and Yong Rui. Structure matters: Tackling the semantic discrepancy in diffusion models for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8038–8047, 2024c. 5.3.3

Huisi Wu, Shihuai Chen, Guilian Chen, Wei Wang, Baiying Lei, and Zhenkun Wen. Fat-net: Feature adaptive transformers for automated skin lesion segmentation. *Medical image analysis*, 76:102327, 2022. 5.4.2

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature medicine*, 26(8):1229–1234, 2020. 5.5.1

Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G van Leeuwen, and Keelin Murphy. Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, 72:102125, 2021. 6.1

Amirsina Torfi and Edward A Fox. Corgan: correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records. *arXiv preprint arXiv:2001.09346*, 2020. 6.1

Mohamed Loey, Florentin Smarandache, and Nour Eldeen M. Khalifa. Within the lack of chest covid-19 x-ray dataset: a novel detection model based on gan and deep transfer learning. *Symmetry*, 12(4):651, 2020. 6.1, 6.2.1

Yash Karbhari, Arpan Basu, Zong Woo Geem, Gi-Tae Han, and Ram Sarkar. Generation of synthetic chest x-ray images and detection of covid-19: A deep learning based approach. *Diagnostics*, 11(5):895, 2021. 6.1, 6.2.1

Saman Motamed, Patrik Rogalla, and Farzad Khalvati. Data augmentation using generative adversarial networks (gans) for gan-based detection of pneumonia and covid-19 in chest x-ray images. *Informatics in Medicine Unlocked*, 27:100779, 2021. 6.1, 6.2.1

Swathi Buragadda, Kodali Sandhya Rani, Sandhya Venu Vasantha, and M Kalyan Chakravarthi. Hcugan: Hybrid cyclic unet gan for generating augmented synthetic images of chest x-ray images for multi classification of lung diseases. *International Journal of Engineering Trends and Technology*, 70(2):229–238, 2022. 6.1, 6.2.1

Sagar Kora Venu and Sridhar Ravula. Evaluation of deep convolutional generative adversarial networks for data augmentation of chest x-ray images. *Future Internet*, 13(1):8, 2020. 6.1, 6.2.1

Vedant Bhagat and Swapnil Bhaumik. Data augmentation using generative adversarial networks for pneumonia classification in chest xrays. In *2019 Fifth International Conference on Image Information Processing (ICIIP)*, pages 574–579. IEEE, 2019. 6.1, 6.2.1

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019b. 6.1, 6.4.1

Ilyas Sirazitdinov, Maksym Kholiavchenko, Ramil Kuleev, and Bulat Ibragimov. Data augmentation for chest pathologies classification. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pages 1216–1219. IEEE, 2019. 6.2.1

Tianyang Zhang, Huazhu Fu, Yitian Zhao, Jun Cheng, Mengjie Guo, Zaiwang Gu, Bing Yang, Yuting Xiao, Shenghua Gao, and Jiang Liu. Skrgan: Sketching-rendering unconditional generative adversarial networks for medical image synthesis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 777–785. Springer, 2019. 6.2.1

Kai Packhäuser, Lukas Folle, Florian Thamm, and Andreas Maier. Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems. *arXiv preprint arXiv:2211.01323*, 2022. 6.2.1

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019. 6.2.2

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 6.2.2

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 6.2.2

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017. 6.2.2

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325, 2017. 6.2.2

Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. Multimodal recurrent model with attention for automated radiology report generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 457–466. Springer, 2018. 6.2.2

Xuewei Ma, Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Yuexian Zou, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest x-ray report generation. *arXiv preprint arXiv:2106.06965*, 2021. 6.2.2

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020. 6.2.2, 6.4.2, 6.3, 6.4

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020. 6.2.2

Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. *arXiv preprint arXiv:2010.10042*, 2020. 6.2.2, 6.4.2, 6.3, 6.4

Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, Subathra Adithan, and Pranav Rajpurkar. Multimodal image-text matching improves retrieval-based chest x-ray report generation. *arXiv preprint arXiv:2303.17579*, 2023. 6.2.2

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22669–22679, 2023. 6.3.1, 6.3.1

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 6.3.2

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6.4.2

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6.4.2

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6.4.2

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6.4.2

Joseph Paul Cohen, Joseph D Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, et al. Torchxrayvision: A library of chest x-ray datasets and models. In *International Conference on Medical Imaging with Deep Learning*, pages 231–249. PMLR, 2022. 6.4.3