Towards Modelling Skeletal Human Motions via Diffusion Models

Ziyi Chang

Towards Modelling Skeletal Human Motions via Diffusion Models

Ziyi Chang

A Thesis presented for the degree of
Doctor of Philosophy

Department of Computer Science
Durham University
United Kingdom
March 30, 2026

Abstract

Skeleton-based human motion modelling has been a long-standing challenge and continues to attract strong academic and industrial interest. Human motions exhibit substantial diversity arising from inter-class variation (e.g., differences in action semantics and dynamics) and intra-class variation (e.g., differences in spatial extent and temporal pace). These complementary factors are deeply entangled, and conventional approaches struggle with limited capacity and mode coverage when trying to model human motions. This thesis aims to investigate human motion modelling problems within a unified generative framework through the lens of inter-class and intra-class variations. To examine this perspective in practice, the thesis considers three representative tasks: styled motion generation, adversarial motion generation, and multi-character interaction generation.

Styled motion generation aims to synthesizes different motion contents under diverse styles where contents are treated as inter-class features and styles are treated as intra-class features to account for the variations of human motions. In this thesis, the term *style* is employed in a motion-specific context. Motion styles denote attributes that yield systematic variations in the execution of an action, despite the underlying action class remaining the same. For instance, emotional states or age-related characteristics can introduce distinct spatial and temporal nuances in the performance of an otherwise identical action. This motion-centric definition differentiates our use of style from its broader interpretations in other disciplines, such as visual aesthetics or linguistic expression. Within this formulation, content represents inter-class semantics and dynamics, whereas style encapsulates fine-grained intra-class variations. Previous methods cannot generate styled motions in an end-to-end manner, either requiring to specify contents and/or styles to reduce the demand of jointly modelling inter-class and intra-class variations. To facilitate the end-to-end styled motion generation, a denoising diffusion probabilistic model is proposed where action classes are recognised as contents and action executions are recognised as styles. Different contents and styles are modelled jointly in the same diffusion latent space. This results in an integrated, end-to-end trained pipeline that facilitates

the generation of stylized motion.

Adversarial motion generation aims to synthesize motions to mislead a system by treating high-level action semantics as inter-class features and low-level execution details as intra-class features. In terms of human motions, a classical scenario is to mislead action recognition systems for their wide applications. A diffusion model is proposed for the generation of adversarial human motions against human action recognition. The variations modelled by the diffusion model facilitates to generate adversarial motions to reveal the adversarial robustness of human action recognition. The diffusion model generates the adversarial motions from the stochastic diffusion latent space and the distributional knowledge captured by the diffusion model.

Multi-character interaction generation aims to synthesize interactions of a large number of characters, treating high-level interaction semantics as inter-class variations and coordination as intra-class variations under the context of interactions. Different from previous interaction modelling approaches that mainly focus on two characters, the coordination between multiple characters is recognised as an unique intra-class variation in the context of multi-character interactions, allowing characters to change the interaction partners. A conditional diffusion model is proposed with reinforcement learning as a framework for the generation of multi-character interactions without any multi-character dataset. The framework comprising a coordinatable multi-character interaction space for interaction synthesis and a transition planning network for coordination. The two component advances the modelling of inter-class and intra-class variations for multi-character interactions, facilitating the generation of realistic, dynamic interactions among multiple characters.

By respectively integrating diffusion models for modelling inter-class and intra-class variations of human skeletal motions through stylized motions, adversarial motions, and interactive motions, this thesis demonstrates significant improvements in unleashing the potential of generative diffusion models in human motion modelling. The demonstrated results hold promising potential for further applications in diverse human-centric motion-based artificial intelligence such as behaviour diagnostics, physical rehabilitation, and responsible systems. Most of the works have been recognized in peer-reviewed conferences, underscoring their impacts and contributions to the field.

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.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Hubert P. H. Shum, for guiding me throughout my PhD journey at Durham and helping me grow from a novice researcher into a more mature and independent one.

I would also like to thank my co-supervisor, Dr George Alex Koulieris, whose guidance had a significant impact on my research journey and helped me develop valuable research skills that have greatly benefited me.

I extend my gratitude to Dr He Wang, Dr Hyung Jin Chang, Dr Qianhui Men, and other invaluable collaborators, who have enriched my research skills, communication abilities, and cross-disciplinary experience in many aspects. Additionally, I would like to express my sincere thanks to my PhD progress reviewers, Dr Stamos Katsigiannis and Dr Wanqing Tu, for providing consistent and objective feedback that helped me stay on the right path.

I am also grateful to lab members at Durham, including Haozheng Zhang, Luca Crosato, Manli Zhu, Mridula Vijendran, Ruishen Han, Ruochen Li, Shuang Chen, Tanqiu Qiao, Xiatian Zhang, Xiaotang Zhang, and Yoshiki Kubotani, as well as other friends and colleagues, including but not limited to Abril Corona-Figueroa, Kaili Sun, Li Li, Neelanjan Bhowmik, Jialin Yu, Jiyao Pu, Minye Shao, Xiaoliang Wu, Xingyu Miao, and Zhongtian Sun. They have provided invaluable support during challenging times, helped me improve the quality of my research, and contributed greatly to my personal and academic development.

Finally, I would like to thank my parents sincerely. Their unconditional support and encouragement gave me the strength to keep going, even during difficult times. Without their understanding and care, I could not have completed this journey.

Dedication

To my parents.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Human motions simultaneously exhibit inter-class and intra-class variations. Fig. 1.1 presents the t-SNE visualisation of human motions to illustrate the existence of inter-class and intra-class variations, where colours indicate different contents and marker shapes represent different styles. For instance, running and jumping differ fundamentally in their temporal dynamics and functional semantics, exemplifying inter-class variation. In contrast, within each action cluster, execution details may vary. For example, when individuals of different ages are instructed to perform the same jumping action, their motions can diverge in aspects such as maximum height and jumping distance, as illustrated in Fig. 1.1. These differences demonstrate intra-class variation within a single action class.

Inter-class and intra-class variations jointly contribute to the substantial diversity observed in human motion, yet they arise from fundamentally different sources and manifest at different structural levels. Overall, inter-class variation defines "what" action is being done, while intra-class variation defines "how" it's done. Inter-class variation refers to differences across distinct action categories such as running and jumping, each characterised by unique temporal dynamics, spatial patterns, and functional semantics [1–3]. These differences are usually driven by the global goal

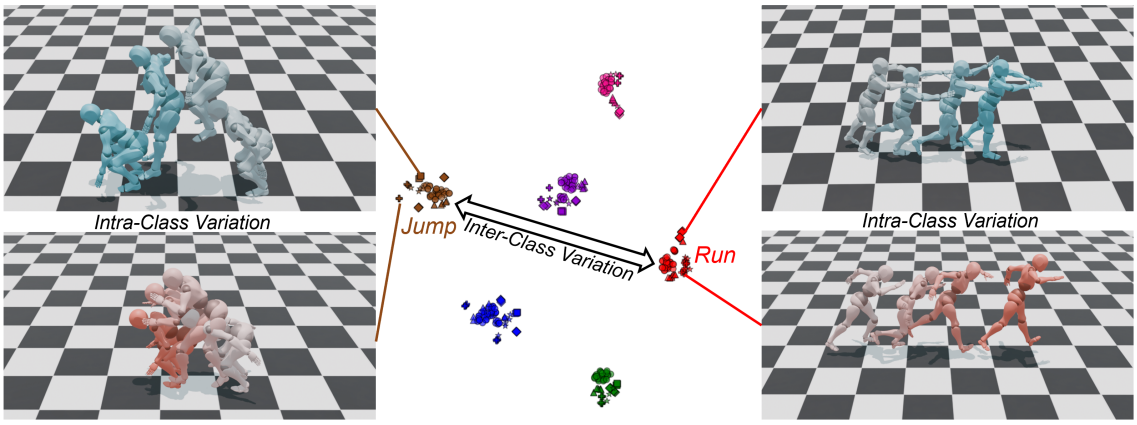

Figure 1.1: Inter-class and intra-class variations in human motion by tSNE and corresponding examples. Inter-class variations are illustrated by different action categories (e.g., running vs. jumping), while intra-class variation is shown by different executions of the same action (blue vs. red). For tSNE, different colors represent different contents and different shapes represent different styles. For motion examples, colour saturation increases with time, i.e., higher colour saturation indicates poses that occur later in time within the sequence.

or high-level planning of a person.

In contrast, intra-class variation captures the diversity that exists within a single action category. Even when performing the same action, individuals can exhibit considerable variability due to factors such as emotional state, age, cultural background, or environmental context [4–6]. For example, two persons may both perform a "running", yet differ markedly in stride length, speed, or rhythm. These variations occur at a finer granularity and tend to influence the details of execution. While intra-class variations do not influence the overall semantics, e.g., running either fast or slow is perceived as running, such intra-class variation reflects meaningful behavioural subtleties such as emotions.

While significant efforts have been denoted to for modelling human motions, the substantial diversity arising from inter-class and intra-class variations continue to pose significant challenges for conventional frameworks. For instance, traditional approaches [7, 8] typically rely on pre-defined rules and a searching strategy to utilize pre-collected motion databases, leading to limited generalization and demanding considerable manual efforts [9]. Deep generative models have therefore been introduced to learn motion distributions. For example, variational autoencoders [10]

and generative adversarial networks [11] have been widely applied to model human motions. However, these approaches often exhibit limited capacity and poor mode coverage, and thus they struggle to model inter-class and intra-class variations [12], demonstrating ambiguous patterns [13]. [14, 15] highlights the high variations with multiple modes existing in human motions to be challenging. Furthermore, [16, 17] have shown that diffusion models are learning better for multi-modal data distributions, and [18, 19] show that diffusion models have better generalization ability when compared with existing generative models.

To overcome these challenges, diffusion models have recently emerged as a powerful class of generative models and have demonstrated remarkable success in visual analysis across various domains such as generation [20, 21], understanding [22, 23], and control [24, 25]. Many existing generative approaches for human motion synthesis, including those based on variational autoencoders, generative adversarial networks, and graph neural networks that operate on skeletal joint structures, typically generate a full motion sequence through a single-step mapping from a latent representation to the output. In this context, a "step" refers to a single pass through a decoder, which may itself be composed of multiple layers, such as stacked graph attention layers or a combination of graph convolutional layers and multilayer perceptrons. Despite this internal architectural complexity, generating a sample via one invocation of the decoder is generally referred to as single-step modelling. While effective in many settings, such single-step formulations can make it challenging to capture the highly non-linear dynamics of human motion. In contrast, diffusion models adopt a multi-step generation process, in which a shared denoising network (i.e., the decoder) is applied repeatedly across a large number of timesteps to progressively refine a motion sample from noise. As a result, diffusion models obtain a sample through multiple decoder passes, whereas GANs and VAEs typically produce a sample through a single decoder pass. This fundamental difference enables diffusion models to better represent complex, multi-modal distributions and to capture highly non-linear structures in human motion [26, 27]. This thesis therefore investigates diffusion models for modelling skeletal human motion from the perspective of inter-class and intra-class variations, offering methods and insights toward

automated human motion generation.

1.1 Motivation

The motivation of this thesis arises from the fundamental challenges and practical importance of modelling skeletal human motion. Human movements exhibit substantial variability stemming from both inter-class differences between action categories and intra-class differences in how individuals execute the same action. Capturing this diversity is essential for applications ranging from animation and virtual human creation to biomechanics, robotics, and behaviour understanding. However, obtaining motion data is often costly and time-consuming, making data-driven generative modelling a highly attractive research direction.

From a modelling perspective, the intertwined nature of inter-class and intra-class variation poses a significant challenge for conventional generative frameworks. Many existing models struggle either to preserve high-level semantic distinctions or to reproduce fine-grained execution details, often resulting in over-smoothed or mode-collapsed motions. These limitations highlight a gap between the expressive complexity of human motion and the representational capacity of current methods, motivating the exploration of more expressive generative frameworks.

Diffusion models offer a promising opportunity in this respect. Their iterative denoising mechanism enables multi-step refinement of structure, providing strong mode coverage and stable training behaviour—properties that are particularly well suited to modelling diverse human motions. This thesis is therefore motivated by the potential of diffusion models to address the limitations of prior generative approaches and by the need to understand how such models can be adapted to capture the complex interplay between motion semantics and execution details across different synthesis tasks.

1.1.1 Motivation for Skeletal Human Motions

In this thesis, we adopt 3D skeletal representations to focus the modelling process on the intrinsic geometry and dynamics of human movement. Human motion can be

represented through multiple modalities, including RGB video [28–30], depth image sequences [31, 32], and 3D skeleton sequences [33–36]. While all modalities have demonstrated strong performance in recognition and synthesis tasks, skeletal data provide a compact, topologically structured representation of the body—capturing joints, bones, and their kinematic relationships—without appearance-related factors such as clothing, textures, or illumination.

This abstraction offers several advantages. RGB and depth pipelines can be sensitive to changes in viewpoint, scale, or motion speed [37, 38], and may unintentionally exploit shortcut cues from backgrounds or objects rather than the motion itself. In contrast, skeleton-based representations minimise these sources of variability and align more directly with the semantics of movement [39, 40], allowing models to concentrate on learning the structural and dynamic aspects of human motion rather than appearance. As a result, skeletons serve as an effective and robust modality for analysing and generating diverse human motions.

1.1.2 Motivation for Human Motion Generation

Human motion is complex, arising from a multifaceted interplay between cognitive goals, sensorimotor control, biomechanical constraints, and environmental context. Rather than being determined solely by intentions or external stimuli, motor execution involves the continuous integration of perceptual feedback, internal state estimation, and multi-level coordination within the motor system [41, 42]. As a behavioural and communicative medium, human motion has therefore attracted increasing attention across computer vision [35, 43], computer graphics [44, 45], multimedia [46, 47], robotics [48, 49], and human-computer interaction [50, 51]. Realistic and diverse motion is foundational for a broad range of applications, including film production, video games, AR/VR systems, human-robot interaction, and the development of digital humans [1]. Beyond creative applications such as animation and VR/AR, motions also play an important role in domains such as sports performance monitoring, clinical gait assessment, and behaviour analysis in surveillance settings. In such contexts, models of typical human motion can support downstream systems that identify unusual patterns or behaviours.

Building on the importance of human motions, the acquisition of 3D human motions is a fundamental step of 3D animation and analysis. However, obtaining such assets is usually beyond the means of the average user and is resource-intensive. Capturing real performances requires specialised equipment, e.g., motion capture systems where multiple cameras are deployed for capturing real-world motions performed by actors. Alternatively, motions can also be manually crafted, but it also demands the expertise in keyframing animation and physical simulation. In terms of these challenges, we seek to develop generative methods that facilitate the generation of smooth and diverse 3D human motions.

Generative methods learn the underlying distribution of motion sequences and enable sampling and conditional synthesis, offering broad coverage of inter- and intra-class variability for downstream systems [1, 10, 11]. While discriminative models for recognition and understanding excel at labelling or segmenting observed motion, yet they do not model the full sequence distribution and thus cannot natively produce novel samples or capture multi-modal futures [39, 40].

1.1.3 Motivation for Diffusion Models in Human Motion Generation

Generating human motion remains a challenging problem despite substantial progress in prior work [9, 46, 52]. Existing approaches often exhibit artifacts, oversmoothing, or ambiguous dynamical patterns [13], reflecting the difficulty of modelling the rich variability inherent in human movement. This challenge arises from the multi-modal nature of both inter-class and intra-class variations [2, 14].

Many traditional deep generative models struggle with this requirement. Although VAEs [10], GANs [11], and related architectures have achieved strong results across domains, their reliance on single-step mappings or restrictive latent priors can limit expressiveness and mode coverage [20]. These limitations make it challenging to represent the full range of human motion variability, where both high-level dynamics and low-level execution nuances need to be preserved.

Diffusion models have emerged as a compelling alternative because their iterative refinement process is particularly well aligned with the multi-scale nature of human

motion. Instead of producing a sequence in a single step, diffusion models generate motion through a progression of noise-conditioned denoising operations [53]. Each timestep updates the sample using a shared denoising network, allowing coarse motion structure to be shaped at early timesteps and fine-grained stylistic details to be added later. This multi-step formulation naturally mirrors the hierarchical structure of human movement and provides an effective mechanism for covering a broad range of inter- and intra-class motion variations.

Moreover, diffusion training is stable and does not rely on adversarial optimisation, reducing mode collapse and encouraging comprehensive distribution coverage [20, 54]. These properties make diffusion models particularly suitable for motion planning, where capturing coherent temporal evolution, respecting kinematic constraints, and representing multiple plausible future trajectories are critical requirements.

However, diffusion models also introduce practical challenges. The iterative sampling process can be computationally expensive [55], often requiring hundreds of denoising steps, which increases inference time compared with single-step generative models. This creates an inherent trade-off between motion quality and sampling speed [27], especially in real-time or interactive applications. Recent work on accelerated samplers and model distillation offers promising directions [56–58], but the quality–speed balance remains an important consideration when adopting diffusion models for human motion generation.

Overall, these factors, such as multi-step refinement capability, stable training, improved mode coverage, and alignment with the hierarchical structure of human motion, motivate the exploration of diffusion models in this thesis, while acknowledging the computational costs that accompany their advantages.

1.2 Problem Definitions

Human motion exhibits substantial variability arising from differences across action categories, referred to as *inter-class variation*, and differences in how a given action is executed, referred to as *intra-class variation*. These two forms of variation

are inherently entangled in real-world motion data and jointly determine the richness, diversity, and ambiguity of human movement. Formally characterising and exploiting these variations is a central problem in human motion modelling.

Let a motion sequence be denoted as $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X}$ is the space of all valid human motions. Each motion $\mathbf{x} \in \mathbb{R}^{T \times J \times D}$ is expressed as a sequence of skeletons where T is the length of a sequence, J the number of joints, and D the dimensions of features containing positions and/or rotations. Each motion $\mathbf{x}$ is associated with a high-level semantic annotation $c \in \mathcal{C}$, representing the inter-class variations such as action category or interaction type, and a latent execution variable $s \in \mathcal{S}$, representing intra-class variations such as style, tempo, coordination, or expressiveness. Under this formulation, a motion sequence is assumed to be generated from an underlying joint distribution

$$p(\mathbf{x}, c, s) = p(\mathbf{x} \mid c, s)\, p(s)\, p(c), \tag{1.1}$$

where c captures inter-class variation and s captures intra-class variation. The conditional distribution $p(\mathbf{x} \mid c, s)$ defines how semantic content and execution characteristics jointly determine the realised motion.

A core challenge in human motion modelling is to learn a generative model $p_\theta(\mathbf{x})$ or $p_\theta(\mathbf{x} \mid c)$ that can faithfully represent and manipulate both types of variation. This challenge can be decomposed into two complementary problem settings.

Problem I: Direct Modelling of Inter-Class and Intra-Class Variations. The objective in this setting is to learn a unified latent representation $(\hat{c}, \hat{s})$ from data such that inter-class and intra-class variations are explicitly disentangled. Given a dataset $\mathcal{D} = \{\mathbf{x}_i, c_i, s_i\}_{i=1}^{N}$, the goal is to learn a mapping

$$\mathbf{x} \longleftrightarrow (\mathbf{z}_c, \mathbf{z}_s), \tag{1.2}$$

where $\mathbf{z}_c \in \mathcal{Z}_c$ encodes inter-class semantics and $\mathbf{z}_s \in \mathcal{Z}_s$ encodes intra-class execution characteristics. The generative process is then defined as

$$\mathbf{x} = G(\mathbf{z}_c, \mathbf{z}_s), \tag{1.3}$$

with the requirement that $\mathbf{z}_c$ and $\mathbf{z}_s$ are statistically disentangled and independently controllable. The main difficulty lies in learning such a factorised latent space in the absence of paired samples that explicitly annotate intra-class variation, while ensuring that variations in $\mathbf{z}_s$ do not induce semantic drift in $\mathbf{z}_c$.

Problem II: Exploiting Intra-Class Variations under Inter-Class Constraints. In this setting, inter-class semantics are treated as fixed constraints, and intra-class variation is exploited as a controllable degree of freedom to achieve task-specific objectives. Given a conditioning variable c and an initial motion $\mathbf{x}_0 \sim p(\mathbf{x} \mid c)$, the goal is to generate a modified motion $\tilde{\mathbf{x}}$ such that

$$\tilde{\mathbf{x}} \sim p(\mathbf{x} \mid c), \quad \tilde{\mathbf{x}} \neq \mathbf{x}_0, \tag{1.4}$$

while satisfying additional constraints $\mathcal{K}(\tilde{\mathbf{x}})$, such as adversarial objectives and coordination requirements. The generated motion need to preserve inter-class semantics,

$$\mathbf{C}(\tilde{\mathbf{x}}) = \mathbf{C}(\mathbf{x}_0), \tag{1.5}$$

where $\mathbf{C}(\cdot)$ denotes a semantic classifier or semantic consistency measure, while allowing variations within the intra-class space.

The principal challenge in this formulation is to generate meaningful intra-class variations that satisfy task constraints without violating semantic identity. Unlike direct modelling approaches, which aim to explicitly disentangle latent factors, this setting requires implicit control mechanisms that operate within the intra-class manifold induced by fixed inter-class semantics.

1.3 Research Aims and Objectives

Grounded in the problem formulation presented in Section 1.2, this thesis aims to systematically investigate how inter-class and intra-class variations in human motion can be modelled, disentangled, and exploited within a unified generative framework. The objectives of this research are defined as follows.

Objective 1: Modelling Inter-Class and Intra-Class Variations Directly. The first objective is to study whether a single generative model can simultaneously represent both inter-class semantics and intra-class executive variations of human motion. This objective is evaluated by assessing the model's ability to generate diverse motions across different action categories while preserving meaningful variability within each category. Success is measured through both quantitative metrics like FID and qualitative evaluation like user study.

Objective 2: Disentanglement and Exploitation of Motion Variations. The second objective is to examine whether intra-class variations can be exploited as a degree of freedom within the constraint of inter-class semantics. This objective is assessed by testing whether changes in intra-class factors alter execution details without inducing semantic drift and artifacts that perceivable by humans. This objective is evaluated by assessing the model's ability to generate imperceptible variations against classifiers for recognizing inter-class semantics and humans. Success is measured through both quantitative metrics like recognition accuracy of classifiers and qualitative evaluation like user study.

Objective 3: Disentanglement and Exploitation of Interaction Variations. The third objective is to further investigate how intra-class variation can be exploited as a degree of freedom when interactions are considered. This objective is evaluated through tasks that require interaction semantic preservation under additional constraints such as multi-character social distance. Performance is measured by semantic consistency, perceptual plausibility, and task-specific criteria such as coordination quality. Success is measured through both quantitative metrics like

smoothness and qualitative evaluation like user study.

1.4 Contributions

The main contributions of this thesis are summarized as follows:

- We propose a diffusion-based method to directly model inter-class and intra-class variations of human motions, facilitating end-to-end styled motion generation. It is the first end-to-end framework to generate styled human motions facilitated by diffusion models. Local and global guidances have been proposed in a multi-task architecture for modelling the diverse distribution of human motions ..... Chapter 4

- We propose a diffusion-based method to exploit intra-class variations of human motions without altering inter-class semantics, facilitating imperceptible adversarial motion generation. The stochastic diffusion latent space has been leveraged to exploit intra-class variations and the distributional knowledge learned from pre-trained diffusion models promote the consistency of inter-class features. ..... Chapter 5.

- We propose a diffusion-based method to exploit intra-class variations in the context of human interactions without changing interaction semantics, facilitating scalable multi-character interaction generation. The temporal intra-class variations are explicitly modelled by a transition planning policy network and reinforcement learning is utilized to avoid the dependence on large-scale multi-character dataset. ..... Chapter 6

1.5 Publications

The research related to this thesis has been previously published in the following peer-reviewed publications that have been grouped by chapters:

- Chapter 3:

- **Ziyi Chang**, George A. Koulieris, Hyung Jin Chang, and Hubert P. H. Shum. "On the Design Fundamentals of Diffusion Models: A Survey." In *Pattern Recognition*, pp. 111934, Elsevier, 2025.

- Chapter 4:

  - **Ziyi Chang**, Edmund J. C. Findlay, Haozheng Zhang and Hubert P. H. Shum, "Unifying Human Motion Synthesis and Style Transfer with Denoising Diffusion Probabilistic Models," in *GRAPP '23: Proceedings of the 2023 International Conference on Computer Graphics Theory and Applications*, pp. 64-74, Lisbon, Portugal, SciTePress, Feb 2023.
  - Edmund J. C. Findlay, Haozheng Zhang, **Ziyi Chang** and Hubert P. H. Shum, "Denoising Diffusion Probabilistic Models for Styled Walking Synthesis," in *MIG '22*: Proceedings of *the 2022 ACM SIGGRAPH Conference on Motion, Interaction and Games*, Guanajuato, Mexico, ACM, 2022.

- Chapter 5:

  - **Ziyi Chang**, Kanglei Zhou, Xiaohui Liang, Hubert P. H. Shum, "Quality-Preserving Imperceptible Adversarial Attack on Skeleton-based Human Action Recognition," under revision of *IEEE Transactions on Circuits and Systems for Video Technology*.
  - Zhengzhi Lu, He Wang, **Ziyi Chang**, Guoan Yang and Hubert P. H. Shum, "Hard No-Box Adversarial Attack on Skeleton-Based Human Action Recognition with Skeleton-Motion-Informed Gradient," in *ICCV '23*: Proceedings of *the 2023 IEEE/CVF International Conference on Computer Vision*, pp. 4574-4583, Paris, France, IEEE/CVF, Oct 2023.

- Chapter 6:

  - **Ziyi Chang**, He Wang, George A. Koulieris, and Hubert P. H. Shum. "Large-Scale Multi-Character Interaction Synthesis," in *SIGGRAPH'25: Proceedings of the 2025 ACM SIGGRAPH*, Vancouver, Canada, ACM, Aug 2025. `https://doi.org/10.1145/3721238.3730750`

In addition to the listed publications above, there are other peer-reviewed publications that have not been included in this thesis:

- **Ziyi Chang**, George Alex Koulieris and Hubert P. H. Shum, "3D Reconstruction of Sculptures from Single Images via Unsupervised Domain Adaptation on Implicit Models," in *VRST '22*: Proceedings of *the 2022 ACM Symposium on Virtual Reality Software and Technology*, pp. 1-10, Tsukuba, Japan, ACM, Nov 2022.

- Xiaotang Zhang, **Ziyi Chang**, Qianhui Men and Hubert P. H. Shum, "Motion In-Betweening for two densely interacting Characters," In *SIGGRAPH Asia'25: Proceedings of the 2025 ACM SIGGRAPH Aisa*, ACM, 2025.

- Xiaotang Zhang, **Ziyi Chang**, Qianhui Men and Hubert P. H. Shum, "Real-time and Controllable Reactive Motion Synthesis via Intention Guidance," In *Computer Graphics Forum*, vol. 44, no. 6, pp. e70222, Wiley, 2025.

- Xiaotang Zhang, **Ziyi Chang**, Qianhui Men and Hubert P. H. Shum, "Physics-Based Motion Tracking of Contact-Rich Interacting Characters," In *Computer Graphics Forum*, Wiley, 2026.

1.6 Thesis Structure

This thesis is organized to systematically explore and present obtained advances in modelling skeleton-based human motions through the integration of diffusion models. Chapters are structured to guide the reader through motivations, literature review, methodologies, and findings of the research in a coherent and logical manner.

Chapter 1: This introductory chapter sets the stage by discussing the motivations behind the thesis. It highlights the limitations of traditional human motion modelling methods, the incapability of previous generative models in modelling inter-class and intra-class variations, and the necessity for integrating diffusion models to cover the significant diversity of human motions. This chapter outlines the primary aim of this thesis and divides into in three objectives where three representa-

tive tasks are considered, namely the generation of stylized motions, the generation of adversarial motions, and the generation of multi-character interactions.

Chapter 2: This literature review chapter evaluates existing methodologies in two perspectives. First, previous studies on human motion modelling are reviewed in terms of modelling single character and interactions. Then, previous studies on inter-class and intra-class variations on human motions are reviewed. It comprehensively examines previous works and establishes a foundation for the novelty of this thesis.

Chapter 3: This chapter introduces preliminaries of diffusion models, which have been served as the main methodology throughout this thesis. It provides fundamental knowledge of diffusion models for readers who are not familiar with this generative model.

Chapter 4: This chapter introduces a denoising diffusion probabilistic model for the generation of stylized human motion. As diffusion models have a high capacity brought by the injection of stochasticity, the intra-class style features is modelled jointly with the inter-class content features in the same diffusion latent space. This results in an integrated, end-to-end trained pipeline that facilitates the generation of stylized motion and advances our understanding of the joint content-style coupled latent space. By designing a multi-task architecture that strategically generates aspects of human motions for local guidance and adversarial and physical regulations for global guidance, both style and content features in human motions are captured and reused for the generation of stylized human motions.

Chapter 5: This chapter introduces a diffusion model for the generation of adversarial human motions. Modelling and understanding the inter-class action semantics and the intra-class detailed execution facilitate to generate adversarial motions to reveal the adversarial robustness of human action recognition. The introduced diffusion model generates the adversarial motions from the stochastic diffusion latent space and the distributional knowledge captured by the diffusion model.

Chapter 6: This chapter introduces a conditional diffusion model with reinforcement learning as a framework for the generation of multi-character interactions without any multi-character interaction dataset. The framework comprising a coordinatable multi-character interaction space for interaction synthesis to account for

inter-class features and a transition planning network for coordination to account for intra-class features under the context of interactions. The two component advances the modelling and understanding of interaction variations, facilitating the generation of realistic, dynamic interactions among multiple characters.

Chapter 7: The concluding chapter summarizes advancements in the design of diffusion models for the generation of stylized human motions, the development of the generation of adversarial human motions, and the generation of multi-character interactions, through diffusion models. Additionally, it outlines potential directions for future research, focusing on multi-character interaction modelling, data efficiency with prior knowledge, adversarially robust human motion modelling, and human perception-aligned evaluation metrics.

In summary, this structure is designed to offer a clear understanding of the research conducted, providing the reader with insights into how inter-class and intra-class variations of human motions are modelled through the integration of diffusion models.

CHAPTER 2

Literature Review

2.1 Human Motion Modelling

This chapter presents an overview of character animation, i.e., human motion modelling, and is organised into two complementary categories: single-character motion modelling (Section 2.1.1) and multi-character interaction modelling (Section 2.1.2). The former focuses on synthesising the motion of an individual character, whereas the latter addresses more complex scenarios involving interactions, such as synchronisation between multiple characters.

2.1.1 Single-Character Motion Modelling

The field of single-character motion modelling is concerned with generating realistic movement sequences for an individual character. This problem presents distinct challenges arising from the high diversity of actions and the intricate, subtle dynamics of human behaviour [1].

Traditional Rule-based Approaches

Single-character motion modelling has attracted considerable interest well before the advent of deep learning. Early approaches predominantly rely on rule-based frameworks coupled with dedicated, pre-collected motion databases. These methods typically offer strong interpretability and integrate expert domain knowledge to capture the temporal dynamics of single-character motion. Motion graphs [59, 60] represent a foundational paradigm among traditional rule-based approaches. In this framework, each pose, corresponding to a single frame of motion, is modelled as a node in a graph, while feasible temporal concatenations between poses are represented as edges. The construction of a motion graph requires a pre-collected database of single-character motion sequences, with edges commonly defined based on pose similarity and local continuity to ensure smooth transitions between consecutive frames [61]. Novel motion sequences are then synthesised by traversing the graph [62], for example via random walks or constraint-guided path searches, and concatenating the poses associated with the visited nodes.

A notable extension of motion graphs is motion matching [63], which was originally introduced as a greedy approximation to overcome limitations of traditional motion graph methods [64, 65]. Motion graphs often incur substantial complexity in graph construction and maintenance, particularly when the underlying database requires updates. By contrast, motion matching avoids explicitly constructing a complex graph from large, unstructured motion capture datasets. Instead, it relies on a pre-defined feature representation and efficient nearest-neighbour search to select pose transitions, thereby achieving higher computational efficiency [66].

Despite their ability to generate smooth transitions for single-character motion, traditional rule-based approaches depend heavily on pre-collected motion databases and manually designed features [67]. Consequently, they often struggle to scale to large and diverse datasets. Moreover, the quality and diversity of the generated motions are fundamentally constrained by the coverage and fidelity of the available database. As these methods primarily recombine existing motion segments, they are inherently incapable of synthesising genuinely unseen poses or motion patterns. Owing to these limitations, the field has progressively shifted towards deep

learning-based approaches, which leverage neural networks to learn complex motion representations directly from real-world human motion data.

Deep Learning–based Approaches

While traditional rule-based approaches offer interpretability and incorporate domain knowledge into their frameworks, their limited capacity to handle uncertainty and complex human behaviours has motivated a shift towards deep learning-based methods. Early deep learning approaches predominantly rely on recurrent neural networks (RNNs), formulating the task as a problem of future motion prediction. A key advantage of this formulation is that temporal dependencies are explicitly embedded within the model architecture. For instance, [68] employs Restricted Boltzmann Machines to explicitly model frame-to-frame transitions. However, the dependencies between motion frames are highly dynamic and complex. To better capture such temporal relationships, [69] proposes a hierarchical RNN trained on a large and diverse motion dataset, while [70] leverages Long Short-Term Memory (LSTM) networks to model longer-term temporal dependencies. Beyond kinematic motion features, [52] explores the phase space of human motion and proposes learning temporal dependencies in terms of motion phases.

Despite these efforts, RNN-based methods for modelling temporal dependencies in single-character motion often struggle to synthesise long motion sequences [71], with generated motions gradually converging to a static pose. This issue arises because such models tend to average over multiple plausible future poses when predicting the next state. Furthermore, these approaches typically require initialisation with several frames from an existing motion sequence and subsequently extrapolate future frames, implicitly assuming the availability of a partial motion sequence provided by the user.

To alleviate motion degradation and relax the assumption of motion availability, generative models such as autoencoders (AEs), variational autoencoders (VAEs), and generative adversarial networks (GANs) have been extensively explored. Representative work by [9] employs autoencoders to learn a motion manifold for generating realistic single-character motions, inspiring subsequent research on deep learning-

based motion manifolds. To explicitly model uncertainty, [46] combines Gated Recurrent Units (GRUs) with VAEs for motion sequence generation. In parallel, [72] proposes a sequence-level conditional VAE that enables non-autoregressive motion generation.

Another prominent line of research is based on generative adversarial networks. The introduction of GANs into single-character motion modelling enables the use of an adversarially trained discriminator to regularise the learned motion manifold [73, 74]. ActFormer [75] incorporates transformer architectures into a GAN framework for action-conditioned 3D human motion generation, while GANimator [76] proposes a hierarchical architecture that applies discriminators sequentially to better regulate the motion manifold.

Although VAE-based and GAN-based frameworks provide effective regularisation of motion manifolds and account for the inherent uncertainty in single-character motion, they often require careful tuning to avoid issues such as mode collapse or overly smooth (blurred) synthesis [20].

To further address mode collapse and blurred synthesis, diffusion-based methods have recently emerged as powerful alternatives. Diffusion models are characterised by their iterative, multi-step transformation from a noise distribution to a data distribution [20], which promotes broader mode coverage when modelling complex data. Within single-character motion modelling, two representative diffusion-based works have laid important foundations. MDM [12] is the first to introduce diffusion models for single-character motion generation, employing transformers to jointly model temporal dependencies across frames. MLD [77] further improves generation fidelity and efficiency by applying diffusion processes in a learned latent space.

Despite these advances, the capability of diffusion models to effectively capture both inter-class and intra-class variations in single-character motion remains limited. This thesis therefore addresses these gaps through two proposed frameworks:

- Chapter 4 introduces a unified pipeline for styled motion synthesis that directly models both inter-class and intra-class variations, without relying on initialised motion sequences.

- Chapter 5 presents an adversarial motion generation approach that exploits

intra-class variations under inter-class constraints to achieve adversarial objectives.

While substantial progress has been made in modelling single-character motion, real-world applications typically involve multiple characters and thus require modelling interactions between them. This motivates a further examination of approaches beyond single-character motion modelling, which is discussed in Section 2.1.2.

2.1.2 Multi-Character Interaction Modelling

The field of multi-character interaction modelling has a long history that predates the widespread adoption of deep learning. It is concerned with generating realistic interactions among multiple characters and poses distinct challenges arising from highly non-linear dynamics, self-organising collective behaviours, and multi-agent synchronisation [15].

Traditional Rule-based Approaches

Prior to the popularity of deep learning, early methods for multi-character interaction modelling predominantly rely on rule-based frameworks and pre-collected motion databases. These approaches typically employ explicitly defined rules, which confer a degree of interpretability through their manual design. Multi-character interaction modelling focuses on capturing the interdependencies between characters. Building upon the motion graph paradigm discussed in Section 2.1.1, interaction graphs [78, 79] have been proposed as an extension of traditional motion graphs. Rather than modelling a single character, interaction graphs represent multi-character poses at each frame as nodes in a graph. In addition, interaction graphs incorporate relative information between characters [7], such as relative position, relative velocity, and relative orientation, to define rules for connecting nodes and enabling plausible interaction transitions. To further improve the construction of graph edges, [80] introduces collaborative and adversarial objectives to guide the synthesis of interactive motions.

Although traditional rule-based approaches can generate smooth transitions for multi-character interactions, they face several inherent challenges. First, they rely on the availability of large-scale interaction datasets, which are significantly more difficult and in some cases impractical to collect as the number of interacting characters increases. Second, these methods struggle to scale to larger groups of characters, as they depend on manually designed features to evaluate interactions, while the complexity of inter-character dynamics grows exponentially with the number of agents, as noted by [81]. These limitations have motivated a transition towards deep learning-based approaches for multi-character interaction modelling.

Deep Learning-based Approaches

Early deep learning approaches to multi-character interaction modelling primarily rely on recurrent neural networks (RNNs) and typically formulate the problem as one of future motion prediction. At this stage, the central objective is to learn interaction semantics from observed motion sequences. One line of work models interdependencies between characters in an iterative manner. For example, [82] proposes an LSTM-based framework that iteratively captures the interdependence between two interacting characters. Another line of research follows a feature fusion strategy, where interaction features extracted from individual characters are combined for prediction. For instance, [83] employs sequential GRUs and pooling operations to encode interaction cues from observations and subsequently fuses the resulting representations for interaction prediction.

Despite progress in applying RNNs to human interaction modelling, early deep learning approaches often suffer from interaction degradation, whereby long-term predictions exhibit noticeable drift and severe interpenetration between characters. Moreover, these RNN-based methods typically require initialisation with several frames of interaction data, implicitly assuming that users have access to partial interaction sequences. This requirement can be particularly restrictive in scenarios involving multiple characters.

To address interaction quality degradation and relax the assumption of interaction availability, generative models such as variational autoencoders (VAEs) and gen-

erative adversarial networks (GANs) have been widely adopted for multi-character interaction modelling. VAE-based methods are particularly attractive due to their structured uncertainty modelling capabilities [15], with research efforts largely focused on designing effective encoders for improved feature extraction. For example, MUGL [84] employs global encoders to capture interactive dynamics alongside local encoders for individual character motion, while DSAG [85] introduces self-attention mechanisms to better model subtle motion details and temporal processing modules to handle high intra-class variance.

Another prominent line of research explores GAN-based frameworks, in which adversarial training is used to further regularise the latent representations of interactions. These methods primarily emphasise discriminator design to enhance interaction realism. For instance, [86] presents an early GAN-based approach that incorporates an attentive, part-aware generator for individual motion modelling and a dual-stream discriminator to improve interaction quality. Similarly, [87] proposes a hierarchical architecture for both the generator and discriminator to capture interaction semantics at multiple levels.

Although VAE-based and GAN-based frameworks provide regularisation of the interaction manifold and account for uncertainty in multi-character interactions, they often require careful tuning to avoid issues such as mode collapse or overly smooth synthesis. In practice, generated interactions may still suffer from insufficient synchronisation between characters.

More recently, diffusion models have emerged as a dominant paradigm in this area due to their superior mode coverage when modelling complex data distributions. Two main research paradigms have gained prominence in this direction. The first focuses on the construction of large-scale datasets with detailed annotations, while their architectures are often based on cross-attention mechanisms to explicitly learn interdependencies between characters. For example, InterGen [36] introduces a large interaction dataset and utilises cross-attention modules to capture inter-character dependencies. Building on this effort, Inter-X [88] presents another large-scale interaction dataset with rich annotations, including textual descriptions, daily interaction categories, relationship and personality labels, and interaction order information.

The second paradigm emphasises architectural innovations to better model interdependencies between characters. For instance, [89] proposes a fine-tuning strategy that treats interactions as control signals for individual characters and learns an additional control layer to explicitly account for inter-character dependencies.

Despite these advances, existing methods primarily focus on modelling inter-class variations, namely interaction semantics, while largely neglecting intra-class characteristics. This thesis addresses this gap through a diffusion-based framework:

- Chapter 6 presents a diffusion-based method integrated with reinforcement learning to account for temporal intra-class variations, while preserving the underlying interaction semantics.

2.2 Modelling Intra- and Inter-Class Variations

Human motion exhibits substantial variability arising from both differences between action categories and variations in how a given action is performed. These two sources of variability are commonly referred to as inter-class and intra-class variation [2, 14, 85], respectively, and together they characterise the diversity and ambiguity inherent in human motion data. A central challenge in motion modelling is therefore not only to capture each type of variation in isolation, but also to represent their interplay in a coherent manner.

Existing research addresses this challenge from two complementary perspectives. Some approaches aim to explicitly model inter-class and intra-class variations as distinct components within the motion representation, as discussed in Section 2.2.1. Other methods treat variation as a task-driven degree of freedom, exploiting intra-class variability to fulfil task-specific objectives such as adversarial robustness or social coordination, as discussed in Section 2.2.2. In the following, we review both perspectives and examine how they contribute to a broader understanding of motion variability.

2.2.1 Direct Modelling Inter- and Intra-Class Variations

Direct modelling approaches explicitly distinguish between high-level motion semantics and low-level execution characteristics. Within this paradigm, inter-class variation is typically associated with action categories or motion content, whereas intra-class variation corresponds to stylistic or execution-level attributes such as tempo, amplitude, expressiveness, or emotional tone [90–92].

Traditional Rule-based Approaches

The direct modelling of inter-class and intra-class variations has been a long-standing research problem that predates the advent of deep learning. Traditional rule-based style transfer techniques typically rely on representative style features extracted from large motion databases, together with statistical models, to integrate stylistic attributes into given motions. A variety of feature extraction methods have been employed to capture intra-class variations, such as Laban Movement Analysis [93] and Iterative Motion Warping [94]. To better handle unstructured and heterogeneous motion data, [95] retrieves nearest-neighbour motion examples from a database based on extracted intra-class patterns. However, modelling intra-class variations in isolation, without explicitly accounting for inter-class variations, often leads to drift in action semantics.

To address this issue, several statistical frameworks have been proposed to jointly model both types of variation, such as Linear Time-Invariant models [94] and Hidden Markov Models [96]. In addition, the spectral domain has been explored as an alternative representation to achieve improved temporal integration between motions [97]. For instance, [98] represents style in terms of spectral intensity and performs style transfer by minimising differences in spectral intensity across heterogeneous motions in the database.

Although traditional rule-based approaches offer conceptual simplicity and computational efficiency, they are fundamentally constrained by manually crafted style features and exhibit limited generalisation to unseen motions. Their deterministic structures and restricted expressiveness have consequently motivated a shift towards deep learning-based approaches, which can learn complex patterns of style

and motion content and achieve more coherent integration of inter- and intra-class variations.

Deep Learning-based Approaches

While traditional approaches provide interpretable feature definitions and statistics-based integration frameworks, their limited ability to handle uncertainty and complex motion dynamics has motivated a shift towards deep learning-based methods. Early deep learning approaches largely inherit the analytical paradigms of rule-based techniques, leveraging the hidden states of recurrent neural networks to represent and integrate inter-class and intra-class variations. For instance, the Gram matrix, a statistics-based representation that captures feature co-occurrence [99], has been adapted for motion style transfer. Specifically, [9] minimises differences between Gram matrices computed from hidden network activations to incorporate stylistic variations into motion sequences. However, Gram matrix-based representations require computing pairwise correlations between feature channels, leading to computational and memory costs that scale quadratically with feature dimensionality. When applied to skeletal motion sequences [100, 101], this cost becomes more pronounced, as correlations must be computed across multiple joints and temporal dimensions, resulting in substantial computational overhead in practice. Moreover, although these early deep learning approaches adapt statistics-based methods to neural network settings, they generally rely on paired motion samples that jointly account for inter-class and intra-class variations, which are often unavailable in real-world scenarios.

To address the problem of paired samples for inter-class and intra-class variations, learning-based representations of inter-class and intra-class variations have dominated this direction and have been commonly supervised with the cycle consistency loss. Considering the lack of paired motion samples, cycle consistency, originally proposed by [102], has been employed as an indispensable part of the learning process to disentangle intra-class and inter-class features. [103] is the first work to leverage cycle consistency loss as well as a discriminator loss to capture inter-class and intra-class variations via two independent encoders for unpaired sce-

narios. Following the work of [103], subsequent studies widely adopt the architecture and strategies. Some studies further improve finer-grained intra-class variation modelling. [104,105] focuses on body parts and extract intra-class and inter-class features with respect to different parts while [106] proposes a dual-flow feature fusion architecture to better integrate intra-class and inter-class features. Others focus on improving computational efficiency [107, 108], proposing online intra-class and inter-class modelling and integration frameworks. Recently, diffusion models have been introduced for a unified framework of modelling and integration [109–111], which demonstrates superior performance in modelling and integrating inter-class and intra-class variations.

To overcome the reliance on paired samples, learning-based representations of inter-class and intra-class variations have become the dominant approach, frequently supervised using cycle consistency losses. Given the scarcity of paired motion data, cycle consistency, which was originally introduced by [102] in image domain, has been adopted as a key mechanism for learning disentangled inter-class and intra-class features. The work of [103] is among the first to employ cycle consistency loss in conjunction with adversarial objectives, using two independent encoders to capture inter-class and intra-class variations in unpaired settings. Building upon this framework, subsequent studies have widely adopted similar architectures and training strategies. Several works focus on finer-grained modelling of intra-class variations. For example, [104, 105] decompose motions into body parts and extract inter-class and intra-class features at the part level, while [106] proposes a dual-flow feature fusion architecture to more effectively integrate these features. Other studies prioritise computational efficiency [107, 108], proposing online frameworks for modelling and integrating inter-class and intra-class variations. More recently, diffusion models have been introduced as unified frameworks for both modelling and integration [109–111], demonstrating superior performance in capturing and combining inter-class and intra-class variations.

Despite these advances on unpaired scenarios, existing methods still require given motion sequences to initialise the modelling of inter-class and intra-class variations, which may not be feasible in practical applications [72]. This thesis addresses this

gap through a novel framework:

- Chapter 4 introduces a unified pipeline for styled motion synthesis that directly models inter-class and intra-class variations without reliance on initialised motion sequences.

2.2.2 Task-driven Exploitation of Intra-Class Variations

While direct modelling approaches seek to represent inter-class and intra-class variations as explicit components of motion, many practical problems instead exploit intra-class variability to achieve task-specific objectives [112]. In such settings, variation is not an end in itself, but a mechanism through which external constraints can be satisfied without compromising high-level motion semantics [113]. This perspective motivates a complementary class of methods that treat intra-class variation as a functional degree of freedom. Specifically, these approaches manipulate intra-class variation to meet external constraints while preserving the semantic identity of the motion. This thesis focuses on two representative applications of this paradigm, namely adversarial motion generation and multi-character interaction generation.

Exploitation under Adversarial Constraints

In adversarial motion generation, the objective is to produce motion sequences that remain perceptually consistent to human observers while causing a trained recognition system to produce incorrect predictions. Achieving this requires preserving inter-class semantics, such that the motion remains recognisable to humans as the original action, while introducing subtle and targeted modifications to execution-level details within the natural intra-class variability of the action class [114, 115]. These perturbations operate within the intra-class space and are constrained by inter-class semantics, exploiting the flexibility of motion execution to deceive classifiers without introducing perceptible artefacts. In this context, intra-class variation functions as a strategic resource rather than an explicitly modelled factor.

Existing approaches for generating adversarial intra-class variations can be broadly categorised into white-box and black-box methods, distinguished by the level of ac-

cess to the target classifier. White-box approaches assume full knowledge of the target model, including access to gradients, enabling direct generation of adversarial intra-class perturbations. For example, [114] introduces perceptual constraints and generates intra-class variations using gradients derived from the target classifier. Subsequent studies further incorporate motion-specific constraints to improve realism. For instance, [116] generates adversarial intra-class variations by modifying only skeletal bone lengths. Moving beyond heuristic constraint design, CIASA [117] employs pre-trained generative models, such as GANs, as generative constraints to regularise intra-class variations within given action classes. However, all of these methods rely on full access to the victim model, which may be unrealistic in practical settings [118].

The second family of approaches, namely black-box methods, operate under the assumption of partial or no knowledge of the target classifier. These methods can be further divided into query-based and transfer-based strategies. Query-based black-box methods leverage a large number of queries to the target classifier and then approximate intra-class variations for each action class based on the returned outputs. By avoiding the need for full model access, this line of work focuses on improving approximation or search strategies for intra-class variations using query–response pairs. For example, BASAR [119, 120] proposes manifold-guided search strategies to generate adversarial intra-class variations, while FGDA-GS [121] estimates gradients by approximating gradient signs to generate adversarial intra-class variations. Despite their avoidance of full model knowledge, query-based methods typically require a large number of queries, which may be impractical in real-world scenarios.

In contrast, transfer-based black-box methods generate intra-class variations using a surrogate classifier, with the expectation that the resulting adversarial motions will transfer to the target classifier [122, 123]. A common practice is to adapt existing white-box methods, such as SMART [114] and CIASA [117], for transfer-based settings. However, these approaches often suffer from limited transferability and sensitivity to the choice of surrogate models [124]. To address this issue, TARSA [125] introduces a Bayesian optimisation strategy to better characterise intra-class variations learned by surrogate classifiers and improve transferability.

Despite recent progress in transfer-based black-box methods, existing constraints are often heuristic [125] or dataset-specific [124], making them difficult to design or tune and ultimately limiting the effectiveness of intra-class variation exploitation. This thesis addresses this gap through a diffusion-based framework:

- Chapter 5 presents an adversarial motion generation method that exploits intra-class variations under inter-class semantic constraints to achieve adversarial objectives.

Exploitation under Interaction Constraints

In interaction generation, the objective is to synthesise motion sequences that reproduce realistic interaction scenarios. Achieving this requires not only modelling inter-class features, namely distinct interaction semantics [36,88], but also accommodating self-organising behaviours such as joining, leaving, or forming new interaction configurations, which correspond to intra-class features [126]. These self-organising behaviours operate within the intra-class space and are constrained by inter-class semantics, exploiting the inherent flexibility of interaction dynamics to enhance realism [127]. This setting highlights intra-class variation as a crucial component for achieving realistic interactive behaviours.

Early approaches to interaction generation have been developed well before the widespread adoption of deep learning. These methods rely on predefined composition rules and large motion databases, specifying when, where, and which motion sequences should be spatially and/or temporally concatenated according to heuristic criteria. For example, [128] partitions the surrounding space of a character into honeycomb-shaped regions and searches a motion database to populate each region with appropriate characters, thereby modelling spatial intra-class variations. Rather than explicitly partitioning space, [126] proposes concatenating motion clips based on entry and exit poses, using a matching score to identify suitable transitions. Building on this idea, [8] further improves the concatenation process by ranking candidate transitions using a motion graph. Although these rule-based approaches offer conceptual simplicity, they incur high computational costs due to exhaustive database searches. More importantly, given the complexity of multi-character in-

teractions, suitable motion clips may not exist in the collected database, leading to degraded interaction quality.

More recently, this research direction has regained attention due to the expressive capacity of diffusion models and the flexibility afforded by their training-free guidance mechanisms. Training-free guidance adjusts intermediate samples during the multi-step diffusion process using off-the-shelf differentiable constraints, without requiring additional training [129]. Autoregressive generation over characters has been proposed to incrementally introduce new interacting agents [130], where a pre-trained diffusion model preserves inter-class interaction semantics (e.g., fighting), while differentiable constraints, such as collision avoidance, facilitate intra-class variations such as tactical diversity. Subsequent works have followed this paradigm by proposing alternative constraints or optimisation strategies. For instance, [131] introduces topology-aware collision avoidance constraints, while [132] proposes optimising the noise variables of diffusion models during generation.

Despite recent advances enabled by diffusion models, existing methods primarily focus on generating larger interaction groups without altering interaction semantics, that is, exploiting spatial intra-class variations under fixed interaction constraints. Temporal intra-class variations, however, remain largely unexplored. This thesis addresses this gap through a diffusion–reinforcement learning framework:

- Chapter 6 presents a diffusion-based method combined with reinforcement learning to model temporal intra-class variations while preserving inter-class interaction features.

2.3 Datasets

Over recent decades, researchers have developed multiple datasets for human interaction motion generation research. This section first reviews representative datasets for human motion generation in Section 2.3.1, followed by an overview of representative datasets for human interaction modelling in Section 2.3.2.

2.3.1 Human Motion Datasets

Human motion datasets primarily focus on capturing large collections of single-person motions that reflect both inter-class and intra-class variations. Broadly, existing datasets follow two paradigms for representing such variations: label-guided and language-guided. Label-guided datasets use discrete action categories to represent inter-class variation, with multiple motion samples within each category capturing intra-class variation. In contrast, language-guided datasets employ natural language descriptions to encode inter-class semantics, while multiple motion sequences associated with the same description represent intra-class variation. Table 2.1 summarises representative datasets commonly used in human motion modelling.

Table 2.1: Representative human motion datasets. This table summarises key statistics and characteristics of widely used human motion datasets. *Subjects* denotes the number of individuals involved in the dataset. *Sequences* refers to the number of motion clips. *Frames* indicates the total number of frames capturing 3D human motion. *Length* represents the cumulative duration of the motion data in hours. '−' indicates unavailable information.

Dataset	**Year**	**Subjects**	**Sequences**	**Frames**	**Length**
Motion-X [133]	2023	-	81.1K	15.6M	144.2h
100STYLE [134]	2022	-	-	4.1M	18.75h
HumanML3D [35]	2022	344	14.6K	-	28.5h
NTU [34]	2019	106	114.4K	-	74h
KIT [135]	2016	111	3911	-	10.3h
Xia [95]	2015	-	-	79.8K	-

Label-guided datasets are widely adopted, as categorical labels provide a straightforward partition of daily activities. NTU [34] extends the earlier NTU RGB+D dataset [33] by introducing 60 additional action classes and 57600 new RGB+D video samples. The resulting dataset contains 120 action categories spanning daily activities and health-related motions. 100STYLE [134] is a large-scale dataset in which performers execute the same action under different stylistic variations. Its action categories are treated as content labels, while style categories represent intra-class variation. Xia [95] is a smaller dataset that provides both content and style labels. Unlike 100STYLE [134], which defines style in terms of specific motion patterns, Xia characterises style through attributes such as emotion and age, reflecting properties of real-world human behaviour. While label-guided datasets offer an in-

tuitive way to describe inter-class and intra-class variations, their categorical nature often provides only coarse-grained semantic distinctions.

To enable more fine-grained representations, language-guided datasets have emerged as a prominent trend in human motion modelling. Rather than partitioning motions into discrete categories, natural language descriptions provide flexible and expressive semantic annotations. Similarities between textual descriptions form a continuous semantic space that more faithfully captures inter-class variation, while multiple motion sequences associated with the same description reflect intra-class diversity. Motion-X [133] is a recent large-scale dataset that includes not only body motion and coarse textual descriptions, but also annotations of facial expressions, hand gestures, and fine-grained pose details, facilitating richer modelling of both inter-class and intra-class variations. HumanML3D [35] is constructed by combining the HumanAct12 [46] and AMASS [136] datasets, and provides three distinct textual descriptions for each motion sequence. It covers a wide range of activities, including daily actions, sports, acrobatics, and artistic movements. KIT [135] is a paired dataset comprising motion capture data and corresponding language annotations, where motion sequences are recorded using optical marker-based systems and annotated with natural language descriptions.

2.3.2 Human Interaction Datasets

As the difficulty of acquiring multi-character interaction data increases exponentially with the number of participants, the development of human interaction datasets has largely been driven by two main paradigms: camera-based capture and simulation-based generation. Table 2.2 summarizes these representative datasets.

Early camera-based datasets primarily represent human interactions using visual sensing systems, capturing joint coordinates or body keypoints across sequential frames. These pioneering efforts typically rely on single RGB-D cameras or multi-camera setups to record multi-person interactions. Representative datasets include 3DPW [141], MuPoTS-3D [140], and NTU [34]. Although these datasets have been widely adopted in skeletal action recognition, they have seen more limited use in interaction modelling due to noise and reconstruction inaccuracies. MuPoTS-3D [140],

Table 2.2: Representative human interaction datasets. This table summarises key statistics and characteristics of widely used human interaction datasets. *Subjects* denotes the number of individuals involved. *Sequences* refers to the number of motion clips. *Frames* indicates the total number of frames capturing 3D human motion. *Length* represents the cumulative duration of the motion data in hours. '–' indicates unavailable information.

Dataset	Year	Subjects	Sequences	Frames	Length
Inter-X [88]	2024	89	11,388	8.1M	-
InterHuman [36]	2024	60	7,779	107M	6.56h
ReMoCap [137]	2024	9	-	275.7K	2.04h
DD100 [138]	2024	10	100	210K	1.95h
GTA Combat [75]	2023	14	6,900	-	-
ExPI [139]	2022	4	115	30K	0.33h
NTU [34]	2019	106	20,579	-	18.6h
MuPoTS-3D [140]	2018	8	20	8K	-
3DPW [141]	2018	18	60	51K	-
JTA [142]	2018	10,800	512	461K	4.27h

for instance, focuses on daily interactions in real-world environments where heavy occlusions frequently occur from camera viewpoints. NTU [34], on the other hand, provides over 20000 interaction sequences and has become a standard benchmark in the field of interaction recognition. To reduce data noise, some datasets, such as 3DPW [141], further employ parametric body models, including SMPL [143] and SMPL-X [144]. Despite these efforts, camera-based interaction data often remain noisy, which limits their applicability in recent high-fidelity interaction modelling research.

More recent camera-based datasets have significantly improved in both scale and quality due to advances in professional motion capture systems. State-of-the-art capture pipelines now integrate RGB-D cameras with additional sensors in controlled laboratory environments, enabling more accurate and temporally consistent motion recordings. Consequently, recent human interaction datasets typically involve larger subject pools, longer sequences, and richer social interactions. InterHuman [36] and Inter-X [88] are two representative examples that have been widely adopted due to their high motion quality and interaction diversity. In addition, several specialised datasets have emerged for certain tasks. ReMoCap [137] focuses on a specific task of reactive motion modelling and specific interaction contexts such as dance and martial arts, providing over two hours of curated interaction data. ExPI [139] concentrates

on dyadic interactions involving Lindy Hop aerial steps, while DD100 [138] contains approximately two hours of professional dance performances across various genres, including Samba and Tango. Despite their improved quality and diversity, these datasets are costly to collect in terms of both time and labour. As a result, they often involve a limited number of participants engaged in interactions, which constrains their scalability to scenarios involving larger groups of interacting characters.

In contrast, simulation-based datasets offer an alternative means of obtaining large-scale interaction data with consistent quality. By leveraging modern game engines, human interactions can be generated using rule-based systems within simulated environments. Following this strategy, datasets such as GTA Combat [75] and JTA [142] utilise the GTA-V engine to synthesise large-scale, visually consistent multi-person scenarios featuring diverse poses and interaction patterns. While simulation-based datasets typically provide cleaner motion data than camera-based approaches, they remain constrained by the predefined rules and behaviours of the underlying simulators, often resulting in limited behavioural diversity and somewhat robotic interaction dynamics.

2.4 Evaluations

Appropriate evaluation metrics are essential for comparing different methods and driving progress in the field. However, evaluating synthesised human motion and interactions remains a non-trivial problem [145,146], due to the one-to-many nature of motion generation, the inherent subjectivity of human judgement, and the high-level semantics conveyed by conditional signals. While this challenge remains open, this section reviews commonly used evaluation metrics from multiple perspectives and discusses their respective strengths and limitations. A summary of these metrics is provided in Table 2.3.

2.4.1 Fidelity-related Metrics

Fidelity-related metrics aim to assess the overall quality of generated motion in terms of realism, smoothness, and physical coherence. This section reviews fidelity

Table 2.3: Representative evaluation metrics for human motion and interaction generation.

Category	Sub-category	Metrics
Fidelity	Comparison with Ground Truth	MPJPE [147], NDMS [148], NPSS [149]
	Naturalness	FID [150], FMD [151], MMD [152]
	Physical Plausibility	Foot Skating Ratio [35], Penetration [153]
Variation	Inter-class	Diversity [46]
	Intra-class	Multimodality [35]
Alignment	Text-conditioned	R-Precision [35], Multimodal Distance [35]
	Label-conditioned	Recognition Accuracy [72], Penetration [153]
User Study	User Study	Preference [154], Rating [137, 155]

evaluation from three complementary aspects: comparison with ground-truth motions, which assesses generated motion using distance-based metrics; naturalness, which evaluates dataset-level statistical similarity between generated motions and real motion data; and physical plausibility, which measures adherence to real-world physical constraints and interaction dynamics.

Comparison with Ground-Truth

Comparing generated motion sequences with ground-truth data represents the most direct and intuitive approach to quality assessment. Accordingly, a range of distance-based metrics has been proposed for this purpose.

Distance-based metrics quantify discrepancies between generated motions and corresponding ground-truth sequences. Among these, Mean Per-Joint Position Error (MPJPE) [147] is the most widely adopted. MPJPE measures the average Euclidean distance between corresponding joints in predicted and reference poses, with lower values indicating higher fidelity. Normalised Directional Motion Similarity (NDMS) [148] evaluates the alignment of motion directions and the relative magnitudes of movement between generated and real motions. Beyond MPJPE and NDMS, several complementary metrics have been proposed. For example, Normalised Power Spectrum Similarity (NPSS) [149] assesses long-term motion synthesis quality in the frequency domain. Collectively, these metrics evaluate differences in spatial position, motion direction, and spectral characteristics, providing a multifaceted assessment of motion accuracy across different dimensions.

However, for a given conditional input, the ground truth represents only one of many plausible motion outcomes, with numerous alternative solutions potentially

being equally valid. As a result, evaluation strategies that rely solely on ground-truth comparisons may fail to capture the full range of acceptable motion variability and thus offer limited coverage in assessing generative motion models.

Naturalness

Naturalness in human motion and interaction generation evaluates how lifelike and realistic the generated motions appear, typically by comparing dataset-level statistical and perceptual properties with those of real motion data.

Fréchet Inception Distance (FID) [150] and its motion-adapted variant, Fréchet Motion Distance (FMD) [151], have been widely adopted across the literature [36,75, 85,89,137,138,154,156–158] to quantify the divergence between feature distributions of generated and ground-truth motions. These metrics rely on deep feature representations extracted from pretrained classification models to measure distributional similarity. In addition to FID and FMD, other distribution-based metrics have also been employed. Maximum Mean Discrepancy (MMD) [152] evaluates naturalness by comparing generated and real motion distributions either on a per-timestep basis (MMD-A) or across entire motion sequences (MMD-S), the latter obtained by flattening sequences into vector representations.

Despite their intuitive formulation, these metrics face several limitations. In particular, their assessment of naturalness is highly dependent on the assumed motion distribution and on the quality and suitability of the pretrained feature extractors, which may not fully capture all aspects of perceptual motion quality.

Physical Plausibility

Physical plausibility evaluates the extent to which generated motions adhere to realistic physical constraints and natural interaction dynamics. A commonly used metric in this category is the Foot Skating Ratio [35], which quantifies unintended sliding of the feet during motion and thus reflects the stability and proper grounding of generated movements. Another widely adopted measure is penetration [153], which captures the degree of interpenetration between bodies, typically reported either as an average over all frames or as the maximum penetration observed in a

single frame.

Although these metrics provide quantitative measures of physical plausibility, they rely on several manually defined hyperparameters, such as height and velocity thresholds for determining foot–ground contact. The choice of these parameters can significantly influence the evaluation outcome and may limit the robustness and comparability of results across different studies.

2.4.2 Variation

Inter-Class Variation

In evaluating the diversity of generated motions, *Diversity* [46] is a key metric that captures variation across the overall set of generated samples. Diversity measures the range and distinctness of motions within the generated distribution, reflecting the model's ability to produce varied outputs across different action classes or conditions. It is computed as:

$$\text{Diversity} = \frac{1}{S_d} \sum_{i=1}^{S_d} |v_i - v_i'|_2, \tag{2.1}$$

where S_d denotes the number of samples used for evaluation, v_i and v_i' are deep feature vectors of the i-th samples drawn from two randomly selected subsets of generated motions, and $\| \cdot \|_2$ represents the Euclidean (L2) norm.

Inter-Class Variation

In contrast, *Multimodality* [46] evaluates variation within a single action category or text prompt, reflecting a model's capacity to generate multiple plausible motions for the same conditional input. Given a set of motions spanning C action types, two subsets of equal size S_l are randomly sampled for each action type c. Multimodality is defined as:

$$\text{Multimodality} = \frac{1}{C \times S_l} \sum_{c=1}^{C} \sum_{i=1}^{S_l} \|v_{c,i} - v_{c,i}'\|_2, \tag{2.2}$$

where $v_{c,i}$ and $v_{c,i}'$ denote the deep feature vectors of the i-th samples from the two randomly selected subsets corresponding to the c-th action type.

In summary, diversity captures inter-class variation by ensuring a broad range of

generated motions across different actions, while multimodality captures intra-class variation by assessing the ability to produce multiple valid realisations for a given input. Together, these metrics characterise the richness and flexibility of generative motion models, which are particularly important for dynamic and interactive applications.

2.4.3 Alignment

Ensuring coherence in condition-driven human motion and interaction generation is essential for producing contextually accurate and semantically consistent results. This section reviews commonly used metrics for evaluating alignment between generated motions and their conditioning signals, focusing on both label-based and text-based conditions.

Label-Conditioned

In label-to-motion generation tasks, alignment is assessed by measuring the consistency between generated motions and their corresponding labels. When labels represent action categories, many approaches employ recognition accuracy [46, 72, 159] as an evaluation metric. This metric relies on a pretrained action recognition model to determine whether generated motions are correctly classified into their intended action categories.

Beyond the use of external recognition models, consistency-based metrics have also been proposed as self-contained evaluation measures. In particular, content consistency (CC) and style consistency (SC) are adopted when multiple types of labels are involved [92, 105]. These metrics are inspired by the cycle consistency principle introduced in CycleGAN [102], where motions are iteratively regenerated using labels extracted from the previous iteration's outputs. Both accuracy-based and consistency-based metrics provide quantitative assessments of the alignment between generated motions and their conditioning labels, offering complementary perspectives on label-to-motion coherence.

Text-Conditioned

Text-conditioned alignment metrics evaluate the degree to which generated motions correspond to provided natural language descriptions. R-Precision [35] measures retrieval accuracy by ranking Euclidean distances between motion and text feature embeddings within a dataset containing both matched and mismatched descriptions. It reports precision at top-1, top-2, and top-3 ranks, indicating how frequently the correct description appears among the closest matches. Multimodal Distance [35] computes the average Euclidean distance between feature representations of generated motions and those of their corresponding textual descriptions, providing a direct measure of cross-modal alignment in the shared feature space.

2.4.4 User Study

User studies, or subjective evaluations, constitute a crucial component in assessing generated human motion, as they can reveal aspects of motion quality that are not adequately captured by objective metrics alone [160]. First, human observers are highly sensitive to subtle artefacts in biological motion, such as jitter and foot skating [161,162]. Second, existing objective metrics are often unable to capture nuanced cultural and perceptual factors, including aesthetics and emotional impact [163,164].

User studies can be broadly categorised into two main types. The first involves eliciting user preferences [154] by comparing generated motions against baseline methods or ground-truth sequences, enabling researchers to identify which motions are favoured in terms of overall appeal and effectiveness. The second type requires participants to rate generated motions along specific dimensions, such as motion quality [137], reaction plausibility [137], and realism [155]. These ratings provide a more fine-grained evaluation of particular attributes, offering insights into both the strengths and limitations of the generated motions.

Preference

Many studies employ preference-based user evaluations through pairwise comparisons between generated results and baseline methods or ground-truth motions [165].

In such studies, participants observe pairs of motion sequences and respond to questions such as: "Which motion better corresponds to the textual description?", "Which motion more accurately reflects the specified style label?", "Which motion is more likely to have been modified by an adversarial attack?", or "Which interaction appears more realistic?". Researchers then compute a win rate for the proposed method relative to the baselines. Preference-based user studies thus provide a direct and intuitive comparison between competing approaches.

Rating

Another widely used form of user evaluation involves asking participants to assign explicit scores to generated motions. Typically, volunteers are shown multiple motion sequences and asked to rate each one on a numerical scale (e.g., from 1 to 5 [166]). Beyond overall quality, some studies further request separate ratings for task-specific attributes such as diversity, consistency, or realism [145], enabling a more detailed assessment of model performance across different criteria.

CHAPTER 3

Preliminaries of Diffusion Models

This chapter introduces research topics that are foundational to this thesis. Portions of this chapter have been published in the following peer-reviewed work:

- **Ziyi Chang**, George A. Koulieris, Hyung Jin Chang, and Hubert P. H. Shum. "On the Design Fundamentals of Diffusion Models: A Survey." *Pattern Recognition*, Article 111934, Elsevier, 2025.

Diffusion models are a class of deep generative models composed of three functional components: the forward process, the reverse process, and the sampling process. The generic diffusion framework [167] learns a data distribution through the interplay between the forward and reverse processes, while the sampling process is used to generate novel data that follow the learned distribution. Together, these three components enable diffusion models to effectively model and sample from complex data distributions [168].

Specifically, the forward process progressively perturbs training data by injecting noise, whereas the reverse process learns to remove this perturbation by training a neural network to invert the noising procedure. After optimisation, the sampling process leverages the learned reverse dynamics to generate new samples that conform to the target data distribution.

3.1 The Forward Process

The forward process describes a diffusion mechanism in which an observed data sample x_0 is gradually transformed into a sequence of latent variables $\{x_t\}_{t=1}^{T}$ as the timestep index t increases, as shown in Fig. 3.1. At each step of this process, the transition distribution $p(x_t \mid x_{t-1})$ introduces a stochastic perturbation by injecting a small amount of random noise ϵ_t. Repeated application of these transitions results in the cumulative accumulation of noise, such that the intermediate variable x_t becomes increasingly dominated by randomness as the process evolves. After a sufficient number of diffusion steps, the original data distribution $p(x_0)$ is mapped to a terminal distribution $p(x_T)$ that is simple and analytically tractable, and is typically well approximated by a standard noise distribution. Because the forward diffusion mechanism consists solely of predefined noise injection operations, it does not involve any learnable parameters. Consequently, the forward process can be formally characterised as a fixed Markov chain composed of successive forward transitions:

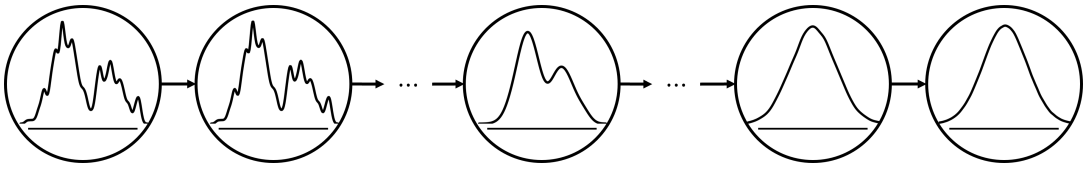

Figure 3.1: The forward process perturbs the original data distribution by gradually adding noise to training samples through a sequence of distribution transitions over multiple timesteps. Each timestep in the chain is denoted by a circle.

3.2 The Reverse Process

The purpose of the reverse process is to learn a model that can reconstruct clean data from progressively corrupted observations, as illustrated in Fig. 3.2. This is achieved by training a denoising neural network to operate across consecutive diffusion steps. Instead of following the forward diffusion trajectory, the model evolves in reverse time, moving from the final timestep T towards 0. At each step, the transformation from x_t to x_{t-1} is characterised by a reverse-time conditional distribution $p_\theta(x_{t-1} \mid x_t)$, whose parameters are learned through optimisation of the

network weights θ.

From a probabilistic perspective, the reverse diffusion procedure can be described as a Markov chain composed of reverse conditional transitions:

$$p_\theta(x_{0:T}) := p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t), \tag{3.1}$$

where θ denotes the parameters of the denoising model and $p_\theta(x_{t-1} \mid x_t)$ specifies the learned transition kernel at each timestep. In most practical settings, these reverse transitions are parameterised using Gaussian distributions:

$$p_\theta(x_{t-1} \mid x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \tag{3.2}$$

with the mean $\mu_\theta(x_t, t)$ and covariance $\Sigma_\theta(x_t, t)$ predicted directly by the neural network.

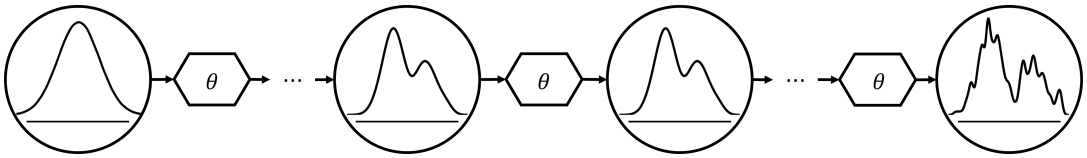

Figure 3.2: Overview of the reverse diffusion procedure, in which a neural network θ is trained to remove noise introduced by the forward process.

Learning the parameters of the denoising network is formulated as the minimisation of a variational upper bound on the negative log-likelihood of the data:

$$\begin{aligned} L = \mathbb{E}\Bigg[& D_{KL}(p(x_T \mid x_0) \,\|\, p(x_T)) \\ & + \sum_{t \geq 1} D_{KL}(p(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t)) \\ & - \log p_\theta(x_0 \mid x_1) \Bigg]. \end{aligned} \tag{3.3}$$

Here, $D_{KL}(\cdot\|\cdot)$ represents the Kullback–Leibler divergence, which serves as a measure of dissimilarity between probability distributions. As the number of diffusion steps T increases, the first term becomes negligible because the terminal distribution converges towards a standard Gaussian prior. The reconstruction term is

typically approximated using Monte Carlo sampling techniques, whereas the dominant contribution during training arises from the denoising-related divergence terms. Consequently, minimising L drives the model distribution $p_\theta(x_0)$ to align with the true data distribution $p(x_0)$.

3.3 The Sampling Process

The generation of new data samples relies on the denoising network after optimisation, denoted by θ^*, as illustrated in Fig. 3.3. Rather than following the forward diffusion direction, this procedure operates in reverse by repeatedly invoking the learned model to recover clean data from noise. The process begins by drawing an initial latent variable x_T from the predefined terminal distribution $p(x_T)$. Subsequently, the trained network is applied at each timestep to perform stochastic transitions governed by $p_{\theta^*}(x_{t-1} \mid x_t)$, gradually refining the sample.

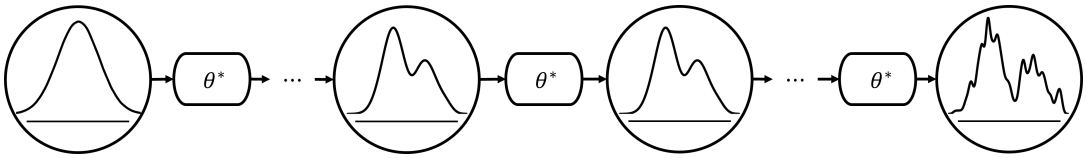

Figure 3.3: The sampling stage relies on the trained denoising network θ^* and follows the learned reverse-time transitions.

After traversing the entire sequence of transitions from timestep T down to 0, the procedure outputs a synthetic data instance $\hat{x}_0$. This sample is drawn from the model-induced distribution $p_{\theta^*}(x_0)$, which is designed to approximate the underlying data distribution $p(x_0)$. The overall sampling mechanism can be formally described as the following Markov process:

$$p_{\theta^*}\left(x_{0:T}\right) := p\left(x_T\right) \prod_{t=1}^{T} p_{\theta^*}\left(x_{t-1} \mid x_t\right), \tag{3.4}$$

where θ^* corresponds to the final learned parameters of the denoising network, $p(x_T)$ defines the noise prior at the terminal step, and $p_{\theta^*}(x_{t-1} \mid x_t)$ specifies the transition kernel employed during sample generation.

CHAPTER 4

Diffusion Models for Styled Motion Generation

Portions of this chapter have been published in the peer-reviewed publication:

- **Ziyi Chang**, Edmund J. C. Findlay, Haozheng Zhang and Hubert P. H. Shum, "Unifying Human Motion Synthesis and Style Transfer with Denoising Diffusion Probabilistic Models," in GRAPP '23: Proceedings of the 2023 International Conference on Computer Graphics Theory and Applications, pp. 64-74, Lisbon, Portugal, SciTePress, Feb 2023.

- Edmund J. C. Findlay, Haozheng Zhang, **Ziyi Chang** and Hubert P. H. Shum, "Denoising Diffusion Probabilistic Models for Styled Walking Synthesis," in *MIG '22*: Proceedings of *the 2022 ACM SIGGRAPH Conference on Motion, Interaction and Games*, Guanajuato, Mexico, ACM, 2022.

In the task of styled motion generation, content, i.e., action semantics/dynamics, varies across classes and accounts for inter-class variations, while style, i.e., low-level execution, accounts for intra-class variations. This task demands both inter-class and intra-class modelling [103–105]. Previous methods require either specifying the content alone or specifying both the content and the style. Therefore, they require two stages to generate styled motions due to the lack of modelling inter-

class and intra-class variations. This chapter presents the first end-to-end diffusion-based styled motion generation framework to demonstrate the superiority of diffusion models in modelling inter-class and intra-class variations.

4.1 Introduction

Synthesising realistic human motion has long been recognised as a fundamental yet difficult problem in the fields of computer graphics and computer vision. In contrast to traditional motion capture techniques, motion generation offers a cost-effective means of producing large-scale motion data without the need for specialised acquisition equipment. Moreover, artificially generated motions support a broad spectrum of downstream applications, including character animation and interactive game development.

Despite these advantages, achieving realism in generated human motion remains highly challenging. One key difficulty arises from the intrinsic diversity of human movement, which spans both semantic content and stylistic variation. Human motions encompass a wide range of activities, such as walking and running, and even within a single activity category, substantial variability exists. For instance, walking motions may differ markedly in style, ranging from confident, exaggerated strides to subdued or fatigued gait patterns.

Although substantial progress has been achieved in recent years through a variety of deep learning–based methods, existing work in this area largely treats motion synthesis and motion style manipulation as independent problems. Research on motion synthesis primarily aims to generate a diverse range of motion contents [14, 52], whereas style transfer methods are mainly concerned with altering stylistic attributes of motion sequences [103, 169]. While these two types of methods can be applied in succession, such a decoupled strategy often limits the quality of the resulting motions. More coherent and realistic motion generation can be expected when both content and style are jointly modelled within a shared representation.

A central difficulty in realistic human motion generation lies in capturing both inter-class behaviours, corresponding to motion content, and intra-class variations,

which reflect stylistic differences, within a unified latent space. Content defines the high-level semantic differences between action classes, corresponding to inter-class variation, whereas style captures execution-level differences within the same action class, corresponding to intra-class variation. Accurately modelling both types of variation requires neural networks to learn a common latent representation capable of encoding this combined structure. However, the resulting latent distribution is substantially more complex than those modelling only a single factor, since both motion content and style exhibit significant diversity. This increased complexity places higher demands on the expressive capacity of generative models. Otherwise, the generated motions are typically suffer from issues such as mode collapse [170], restrictive prior assumptions [171], or reliance on highly specialised network architectures [172], leading to unfaithful expression of styles and contents [46, 105].

We introduce a diffusion-based framework that jointly addresses human motion synthesis and motion style transfer within a unified model. Owing to their strong mode coverage capability facilitated by iterative stochastic refinement across multiple diffusion steps [17], diffusion models are particularly well suited for learning a shared representation of motion content and style. In particular, the denoising diffusion probabilistic model (DDPM) constitutes a class of high-capacity generative models, whose expressive power stems from the deliberate injection of stochasticity during training. This modelling paradigm is conceptually motivated by non-equilibrium thermodynamics. Building upon this foundation, we develop a multi-task DDPM architecture tailored for realistic human motion generation. The proposed design explicitly captures multiple complementary aspects of motion, including joint rotations, global translational trajectories, foot–ground contact patterns, and physical constraints. Relative to the original DDPM formulation [53], the proposed multi-task extension substantially enhances the model's ability to characterise the structured nature of human motion data. Beyond standard noise prediction, our architecture incorporates additional task-specific networks that estimate diverse motion-related attributes. To further improve global realism and consistency, adversarial training is employed to jointly regulate the outputs of the different tasks and encourage mutual coherence. In addition, explicit physical constraints are inte-

grated to promote physically plausible motion over long temporal horizons.

The effectiveness of the proposed approach is validated through quantitative and qualitative evaluations conducted on the dataset introduced by [95]. Ablation studies are further performed to assess the contribution of individual design components. For quantitative assessment, we adopt the Fréchet Inception Distance (FID) [150] to measure the discrepancy between real and generated motion distributions. Experimental results demonstrate that our method consistently achieves the lowest FID scores, indicating superior generative performance. Qualitative visualisations of synthesised motions are also provided to illustrate the high perceptual quality of the generated results, while additional ablation experiments confirm the effectiveness of the proposed multi-task architecture.

The work in [173] represents an early effort to model styled human motions via diffusion model, focusing on walking motions. However, the approach in [211] only utilizes a discriminator to generate styled motions and fails for other actions like jumping and running. In contrast, the approach developed in this chapter builds upon these early insights while addressing their limitations by adopting a multi-task framework that enables richer distribution modelling and significantly improves motion quality across a broader range of motion contents and styles. Comprehensive experimental comparisons further demonstrate the advantages of our approach over [173]. The main contributions of this chapter are summarised as follows:

- We present a single-stage pipeline unifying human motion synthesis and style transfer for high-quality motion creation. The source code is open on `https://github.com/mrzzy2021/StyledMotionSynthesis`

- To effectively represent the coupled representation of both inter-class motion contents and intra-class motion styles in a common latent space, we propose a denoising diffusion probabilistic model solution that has a large learning capacity for modelling the diverse data structure.

- To generate high-quality results, we propose a multi-task network architecture that leverages both local guidance, including joint angles, movement trajectories and supporting foot patterns, and global guidance, including physical and

adversarial regulations.

4.2 Problem Formulation

Realistic human motion generation remains a challenging research problem due to the complex structure of motion data. Effective modelling of human motion requires a neural network to simultaneously capture variations across different motion categories (inter-class content) and stylistic differences within the same category (intra-class style). Together, these two factors define a joint distribution that characterises the underlying motion manifold in a latent space. Owing to the substantial diversity present in both motion content and style, learning this joint distribution necessitates generative models with high expressive capacity. As a result, existing approaches have commonly addressed content generation and style manipulation as separate modelling tasks.

From an integrated perspective, we introduce a new problem setting termed *styled motion synthesis*, along with an associated framework, as illustrated in Fig. 4.1. Prior work has largely focused on developing neural models for either motion synthesis or motion style transfer in isolation. The absence of a unified formulation often leads to inconsistencies when multiple networks are combined in practice, which in turn degrades the quality of the generated motions. In contrast, the proposed problem formulation explicitly unifies motion synthesis and style transfer within a single learning objective, enabling end-to-end motion generation.

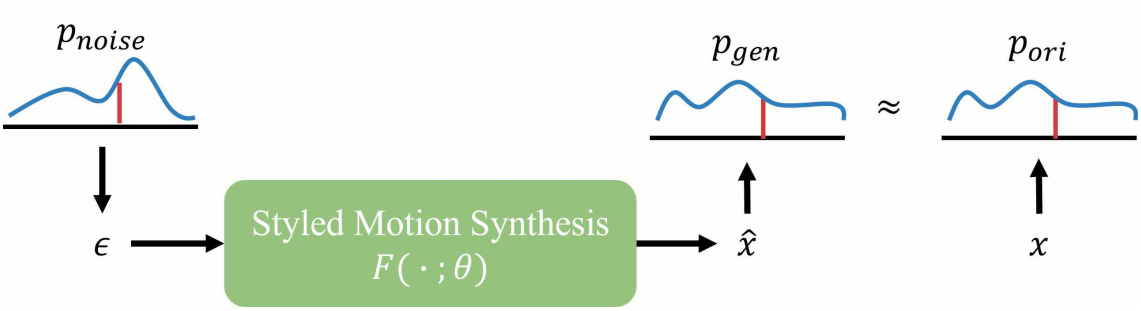

Figure 4.1: Our proposed end-to-end framework for the styled motion synthesis task.

To formally define styled motion synthesis, we introduce the following notation.

Let x denote a sequence of ground-truth motion data over all frames, and let $\hat{x}$ represent the corresponding generated motion sequence. The probability distributions of the real and generated motions are denoted by p_{ori} and p_{gen}, respectively. Given a noise variable ϵ sampled from a predefined noise distribution p_{noise}, a neural generator $F(\cdot)$ produces a motion sequence $\hat{x}$ that aims to match the distribution of real motions. The styled motion synthesis problem can therefore be expressed as

$$\hat{x} = F(\epsilon; \theta) \sim p_{gen} \approx p_{ori}, \tag{4.1}$$

where $\epsilon \sim p_{noise}$ and θ denotes the set of trainable parameters of the network $F(\cdot)$. A motion sequence is assumed to be generated from an underlying joint distribution

$$\hat{x} \sim p(x \mid c, s) \times p(s) \times p(c), \tag{4.2}$$

where c captures inter-class variation, i.e., content, and s captures intra-class variation, i.e., style. The conditional distribution $p(x \mid c, s)$ defines how semantic content and execution characteristics jointly determine the realised motion.

4.3 Method Overview

The generation of realistic human motion is fundamentally complicated by the need to represent coupled motion content and style within a shared latent space. Human motion inherently exhibits variations across motion categories (inter-class content) as well as stylistic differences within the same category (intra-class style) [103]. Many existing approaches lack sufficient modelling capacity to capture this coupled structure, and therefore treat motion synthesis and style transfer as two independent tasks. Such a decoupled strategy restricts exploration of the joint distribution and often leads to sub-optimal motion quality.

To address these limitations, we propose a styled motion synthesis framework based on denoising diffusion probabilistic models (DDPM), as illustrated in Fig. 4.2. DDPMs belong to a class of diffusion-based generative models that progressively add and remove Gaussian noise, enabling stochastic learning dynamics. This stochastic

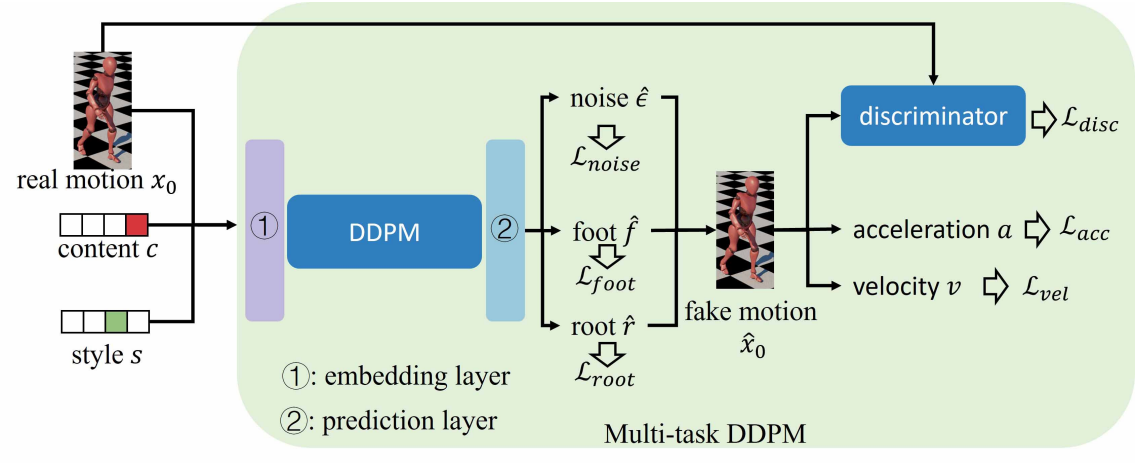

Figure 4.2: An overview of our proposed framework.

formulation substantially increases the expressive capacity of the model, allowing it to better capture the diversity of complex data distributions. Diffusion-based generative models provide an alternative framework for modelling complex and multi-modal data distributions [17]. Their iterative sampling process has been widely discussed as a mechanism for improving mode coverage [174], which is potentially beneficial for motion generation tasks that require both semantic consistency and stylistic diversity. Our proposed framework adopts a unified, end-to-end formulation with enhanced learning capacity, offering clear advantages over conventional two-stage pipelines [103, 105]. By jointly optimising motion content and style within a single model, the framework enables more effective exploration of the underlying motion manifold, resulting in improved generation performance.

The training strategy of the proposed framework follows a multi-task learning paradigm. The model is optimised to simultaneously predict multiple motion-related attributes, including joint configurations, foot contact patterns, global motion trajectories, and physically grounded constraints. In addition, an adversarial discriminator is incorporated to further regularise the learning process and encourage realism across the generated motion sequences.

Overall, the proposed end-to-end pipeline unifies human motion synthesis and motion style transfer within a single diffusion-based framework. By leveraging the high modelling capacity of DDPMs and a multi-task architecture with auxiliary

supervision, the method effectively captures the rich variability of human motion and produces high-quality stylised motion results.

4.4 Denoising Diffusion Probabilistic Models

We adopt denoising diffusion probabilistic models (DDPMs) as the core generative framework, motivated by their strong capacity to represent highly diverse data distributions. Realistic human motion exhibits substantial variability arising from both inter-class motion categories and intra-class stylistic differences. Capturing such variability places considerable demands on the expressive power of generative models. In contrast, many existing approaches, such as [103], are constrained to modelling a single component of motion due to limited network capacity, which often results in reduced diversity in the generated outputs.

In comparison with these methods, DDPMs [53, 167] offer enhanced modelling capability through the explicit introduction of stochasticity into the learning process. The generation procedure is formulated as a stochastic dynamical system in which probability distributions are gradually transformed via controlled noise perturbation and denoising steps. This stochastic formulation enables tractable manipulation of complex distributions and significantly broadens exploration of the latent space, thereby allowing the model to better capture diverse motion patterns.

During training, DDPMs progressively corrupt the input data by injecting Gaussian noise and subsequently learn to reverse this corruption process. As shown in Eq. 4.3, a Gaussian noise variable $\epsilon \sim \mathcal{N}(0, I)$ is added to the original data $x_0 = x$ over t diffusion steps, followed by a learned denoising transition. Model optimisation is performed by minimising the discrepancy between the injected noise and the noise predicted by the neural network. The overall training objective is summarised as:

$$\begin{cases} p_t(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\sigma}_t} x_0, (1 - \bar{\sigma}_t) I)\,, \\ p_t(x_{t-1} \mid x_t) = \mathcal{N}\Big(x_{t-1}; \frac{1}{\sqrt{1-\sigma_t}}\Big(x_t - \frac{\sigma_t}{\sqrt{1-\bar{\sigma}_t}}\hat{\epsilon}\Big), \sigma_t I\Big)\,, \\ \mathcal{L} = \mathbb{E}_{t, x_0, \epsilon}\big\|\epsilon - \hat{\epsilon}\big\|_2^2, \end{cases} \tag{4.3}$$

where $\sigma_t \in (0,1)$ defines the noise schedule, $\bar{\sigma}_t = \prod_{i=1}^{t}(1-\sigma_i)$ denotes the cumulative noise coefficient, $\hat{\epsilon}$ represents the network's estimate of the injected noise, and $\mathcal{L}$ corresponds to the DDPM training loss introduced in [53]. After training, motion generation is performed by sampling an initial latent variable from an isotropic Gaussian distribution $x_T \sim \mathcal{N}(0, I)$ and iteratively applying the learned reverse transitions over T diffusion steps. Each reverse update is given by

$$x_{t-1} = \frac{1}{\sqrt{1-\sigma_t}}\left(x_t - \frac{\sigma_t}{\sqrt{1-\bar{\sigma}_t}}\hat{\epsilon}\right) + \sigma_t z, \tag{4.4}$$

where $z \sim \mathcal{N}(0, I)$ is an independently sampled Gaussian noise variable. This formulation follows the denoising diffusion framework discussed in Chapter 3 and situates DDPM within the broader class of generative models reviewed in Chapter 2, where diffusion models are highlighted for their stable training and strong distributional coverage compared to earlier approaches.

4.5 Multi-task DDPM for Styled Motion Synthesis

4.5.1 Local Guidance

To achieve high-quality generated results, we propose to leverage a multi-task DDPM architecture for styled human motion synthesis. Preliminary quantitative experiments indicate that directly applying a standard DDPM to motion synthesis often leads to suboptimal results, as it does not explicitly account for certain inherent aspects of human motion, such as foot contact patterns (see Table 4.2 in Section 4.7).

The proposed multi-task formulation extends the standard DDPM training objective by incorporating explicit supervision over multiple local motion attributes. In contrast to the original DDPM, which focuses solely on minimising the discrepancy between predicted noise and ground-truth noise, our design simultaneously optimises predictions related to joint configurations, global motion trajectories, and foot–ground contact patterns. These additional objectives introduce complementary

sources of guidance that are directly relevant to the structure of human motion. By jointly learning these motion-specific components, the diffusion model is encouraged to capture a more structured and meaningful latent representation.

In this chapter, global movements refer specifically to the trajectory of the root joint, including its global translation and orientation over time. The root trajectory includes the gloab positions and rotations, thereby capturing the global spatio-temporal structure of an action.

Foot contact patterns serve as a concrete articulation of the root trajectory by constraining how global body motion is physically grounded through contacts with the ground. This articulation provides necessary contextual cues that link high-level global motion to perceptually plausible local execution.

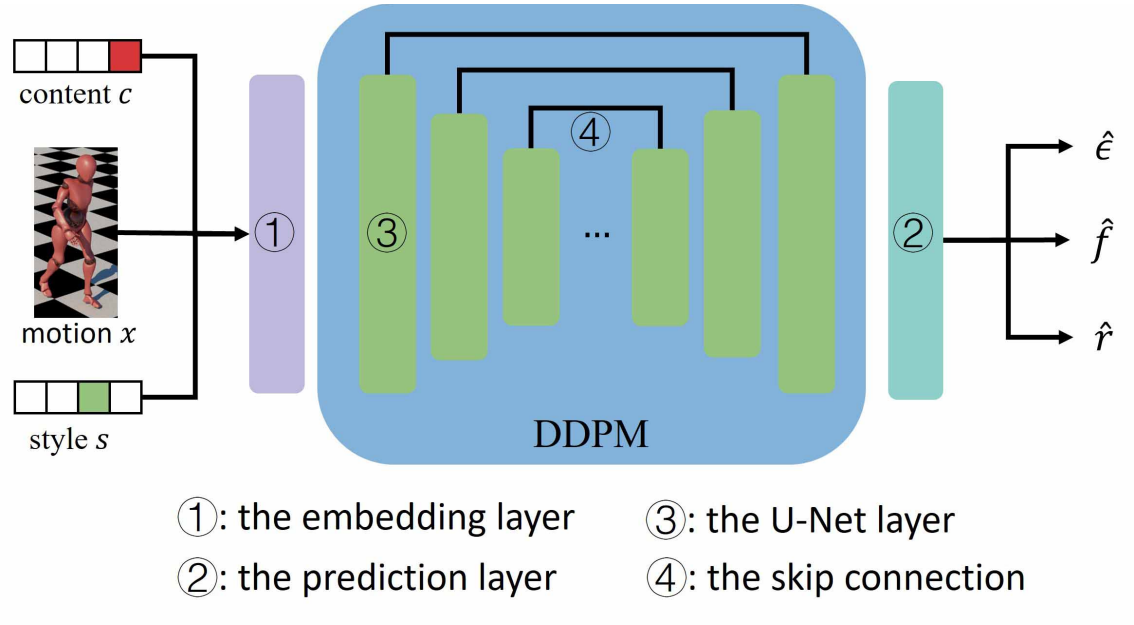

Figure 4.3: Our proposed multi-task DDPM pipeline for styled motion synthesis.

As illustrated in Fig. 4.3, the proposed multi-task DDPM is conditioned on three inputs: a motion sequence $x_0 = x$, a content label c, and a style label s. Motion sequences are represented in terms of joint angles. Prior to being fed into the diffusion model, both content and style one-hot labels are mapped to continuous representations via learnable embedding layers. The diffusion backbone adopts a U-Net architecture [175] augmented with attention mechanisms and skip connections, enabling effective modelling of the joint manifold formed by motion data x, inter-

class content categories c, and intra-class style attributes s.

Given these inputs, the multi-task DDPM simultaneously predicts the injected noise $\hat{\epsilon}$, global root motion $\hat{r}$, and foot–ground contact indicators $\hat{f}$. Foot contact patterns are represented using binary variables that specify whether each foot is in contact with the ground. Following standard practice in diffusion-based generative modelling, the network is trained to estimate the noise term $\hat{\epsilon}$ rather than directly regressing the latent joint configurations $\hat{x}$. To supervise the three output branches, the following loss functions are defined:

$$\mathcal{L}_{noise} = \mathop{\mathbb{E}}_{x_0,t,s,c} ||\epsilon - \hat{\epsilon}||_2^2, \tag{4.5}$$

$$\mathcal{L}_{foot} = \mathop{\mathbb{E}}_{x_0,t,s,c} ||f - \hat{f}||_2^2, \tag{4.6}$$

$$\mathcal{L}_{root} = \mathop{\mathbb{E}}_{x_0,t,s,c} ||r - \hat{r}||_2^2. \tag{4.7}$$

4.5.2 Global Guidance

In addition to providing supervision for individual motion attributes, we further introduce global guidance to improve coherence across all predicted components. To this end, physical constraints and a discriminator are incorporated to jointly regulate the generation process. These global regularisation mechanisms encourage consistency among the locally guided predictions and improve the overall realism of the synthesised motions.

For our global guidance, we derive from the arbitrary query property (Eq. 4.8) and propose a reconstruction formulation (Eq. 4.9). The forward diffusion mechanism described in the first term of Eq. 4.3 incrementally perturbs the original input x_0 with Gaussian noise. An important consequence of this formulation is that the noisy state x_t can be analytically accessed at any diffusion timestep t, which is given by

$$x_t = \sqrt{\bar{\sigma}_t} x_0 + \sqrt{1 - \bar{\sigma}_t} \epsilon_t. \tag{4.8}$$

This property enables direct transitions from the clean motion x_0 to any intermediate state x_t, as illustrated by the arrow connecting x_0 and x_t in Fig. 4.4.

Since the denoising network is trained to predict the injected noise term ϵ_t, it

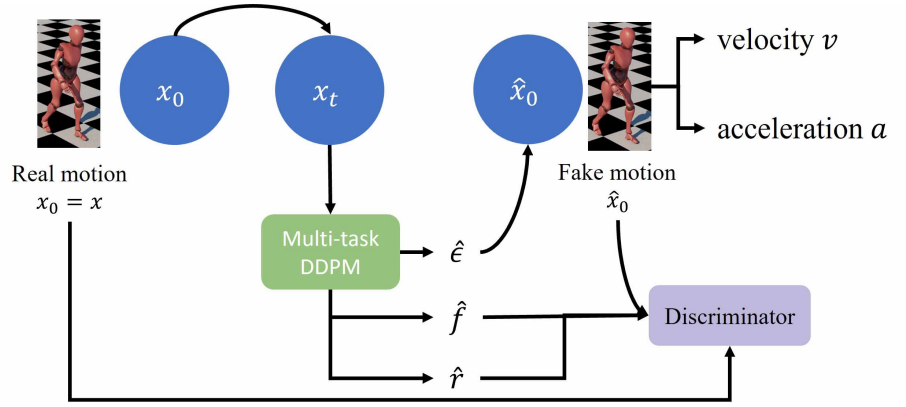

Figure 4.4: Our proposed multi-task conditional DDPM pipeline for styled motion synthesis.

becomes possible to recover an estimate of the original motion in a reverse manner. Given the predicted noise $\hat{\epsilon}$, the reconstruction of the clean motion $\hat{x}_0$ can be computed as

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\sigma}_t}\hat{\epsilon}}{\sqrt{\bar{\sigma}_t}}, \tag{4.9}$$

which corresponds to the mapping from $\hat{\epsilon}_t$ to $\hat{x}_0$ in Fig. 4.4. Through this reconstruction formulation, an estimate of the underlying motion sequence can be obtained at any diffusion timestep t.

Based on the reconstructed motion $\hat{x}$, we introduce physical regularisation terms to encourage physically plausible behaviour. These constraints are motivated by the physical properties of real human motion and aim to promote smooth and globally coherent motion sequences. In natural human movement, joint configurations between successive frames exhibit continuity, and joint velocities do not change abruptly due to physical inertia. To reflect these properties, we impose regularisation on both the velocity and acceleration of joints in the reconstructed motion. Let j_t denote the joint rotation at timestep t. The following loss terms are introduced to penalise excessive temporal variation:

$$\mathcal{L}_{vel} = \mathop{\mathbb{E}}_{t,j\in\hat{x}} ||j_t - j_{t-1}||_2^2, \tag{4.10}$$

$$\mathcal{L}_{acc} = \mathop{\mathbb{E}}_{t,j\in\hat{x}} ||j_t - 2j_{t-1} + j_{t-2}||_2^2. \tag{4.11}$$

These regularisation terms explicitly constrain joint-level velocities and accelera-

tions, thereby encouraging smooth and physically consistent motion generation.

We note that the acceleration term is computed over a minimal temporal window of three consecutive frames, which corresponds to a single-step finite-difference approximation of second-order temporal variation. While longer temporal windows or multi-frame smoothing can further suppress high-frequency fluctuations, we intentionally adopt this formulation for two reasons. First, diffusion-based motion generation already enforces temporal consistency across multiple noise scales through iterative denoising. As a result, the acceleration loss serves as a regularization rather than the sole mechanism for temporal smoothing. Second, extending acceleration constraints over longer temporal windows would introduce additional hyper-parameters and increase optimisation complexity, which may over-constrain motion dynamics and reduce responsiveness to transitions. Under this design, the single-step acceleration loss provides a lightweight and stable constraint that discourages sudden changes. More explicit multi-frame or trajectory-level smoothness constraints are therefore left as a potential direction for future work.

Beyond the incorporation of physical regularisation, we further introduce an adversarial mechanism to enforce consistency across the multiple prediction tasks. Specifically, a discriminator is employed to provide global guidance that promotes harmonious integration of the locally predicted motion components. Since joint angles $\hat{x}$, global root motions $\hat{r}$, and foot contact patterns $\hat{f}$ are estimated independently during local guidance, the role of the discriminator is to assess whether these components collectively form a temporally smooth and coherent motion sequence. Adversarial learning is adopted by training the discriminator to distinguish between real motion clips and synthesised ones. Concretely, the discriminator receives joint angles, root trajectories, and foot contact information as inputs, and outputs a score indicating whether the motion clip corresponds to ground-truth data or generated data. The adversarial objective is formulated as:

$$\mathcal{L}_{disc} = ||D(x_0, r, f) - 1||_2^2 + ||D(\hat{x}_0, \hat{r}, \hat{f}) - 0||_2^2. \tag{4.12}$$

The proposed multi-task DDPM is trained by jointly optimising all task-specific

objectives. The complete loss function is defined as:

$$\begin{aligned}\mathcal{L} =& \lambda_1 \mathcal{L}_{noise} + \lambda_2 \mathcal{L}_{foot} + \lambda_3 \mathcal{L}_{root} \\ &+ \lambda_4 \mathcal{L}_{disc} + \lambda_5 \mathcal{L}_{vel} + \lambda_6 \mathcal{L}_{acc},\end{aligned} \tag{4.13}$$

where weighting coefficients are set to $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$ and $\lambda_5 = \lambda_6 = 0.01$.

The weighting parameter λ is used to balance the contribution of the auxiliary loss terms relative to the primary diffusion objective. In practice, these loss components operate on different quantities and scales, and an explicit weighting is required to prevent any single term from dominating the optimisation.

The value of λ is selected empirically based on stability and convergence behaviour observed during training. Specifically, λ is chosen such that the auxiliary losses provide a meaningful regularising signal while remaining comparable in magnitude to the primary loss throughout training. This selection reflects a trade-off between enforcing auxiliary constraints (e.g., physical consistency or execution-level structure) and preserving the model's capacity to learn the overall motion distribution.

4.6 Experimental Setup

Our approach is trained and evaluated using a publicly available human motion dataset [95]. For quantitative comparison, we compare our method against the original DDPM framework [53] as well as a recent DDPM-based motion generation approach [173]. Qualitative evaluation is conducted by visualising synthesised motion sequences exhibiting diverse motion contents and stylistic variations. Furthermore, to examine the contribution of individual components within the proposed multi-task architecture, ablation studies are performed.

All models are trained on a single NVIDIA RTX 3080 Ti GPU using 32-bit floating-point precision. Training converges within approximately one day. During inference, the generation of a single motion sequence requires around 20 seconds. Detailed hyperparameter configurations used throughout the experiments are sum-

marised in Table 4.1.

Table 4.1: Hyperparameters for training diffusion models.

Learning Rate	0.0002
Discriminator Learning Rate	0.0001
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	$1e^{-8}$
Batch Size	128
number of timesteps T	1000
EMA decay rate m	0.9999

The dataset [95] consists of six distinct motion content categories, with each category annotated with eight different motion styles. As part of the post-processing stage, a Gaussian smoothing filter is applied to the generated motion sequences, followed by inverse kinematics to ensure kinematic consistency and visual plausibility.

4.7 Quantitative Comparison

Quantitative evaluation is conducted using the Fréchet Inception Distance (FID) [150]. This metric measures the discrepancy between the probability distribution of motion sequences synthesised by the diffusion model and that of real motion samples from the dataset. To compute FID, feature representations are extracted for both generated and real motion sequences using a neural network, and each set of features is approximated by a multivariate Gaussian distribution. The resulting distributions are denoted as $\mathcal{N}(\mu_g, \Sigma_g)$ for generated motions and $\mathcal{N}(\mu_r, \Sigma_r)$ for real motions, respectively.

$$FID = ||\mu_r - \mu_g||_2^2 + tr(\Sigma_g + \Sigma_r - 2(\Sigma_r\Sigma_g)^{\frac{1}{2}}). \tag{4.14}$$

Direct comparison with existing methods is non-trivial, as the problem formulation considered here differs from those addressed by prior work. Adapting other methods is non-trival because the inputs, network architectures, objectives, and the

task formulation are all different. In the meanwhile, this chapter therefore focuses on analysing how diffusion models can be adapted to this new setting and the state-of-art work [173] has already been compared.

We conduct a quantitative comparison between the proposed method, an existing diffusion-based motion generation approach [173], and the original DDPM baseline [53], all evaluated on the dataset introduced by [95]. The Fréchet Inception Distance (FID) is computed using the same number of generated motion samples for each method, with features extracted by the same pre-trained classifier. The quantitative results are summarised in Table 4.2.

Table 4.2: Quantitative comparison of methods for styled motion generation.

Model	FID (↓)
[173]	158.47
[53]	198.67
Ours	**56.73**

As reported in Table 4.2, the proposed approach achieves a substantially lower FID score than the compared methods, indicating a closer match between the distribution of generated motions and that of the ground-truth dataset. The comparatively poor performance of the original DDPM baseline [53] can be largely attributed to the absence of motion-specific constraints, which limits its ability to capture structured human motion characteristics. Although the method of [173] incorporates adversarial learning, its performance remains limited due to degraded motion quality in non-walking categories, a consequence of its relatively simple network design. Overall, these results demonstrate that the proposed multi-task DDPM framework delivers superior generative performance across diverse motion contents and styles.

4.8 Qualitative evaluations

In addition to quantitative metrics, we present qualitative results to further assess the effectiveness of the proposed method. Fig. 4.5 illustrates examples of motions generated with different content categories, demonstrating that the proposed approach is capable of producing diverse motion types corresponding to distinct motion contents.

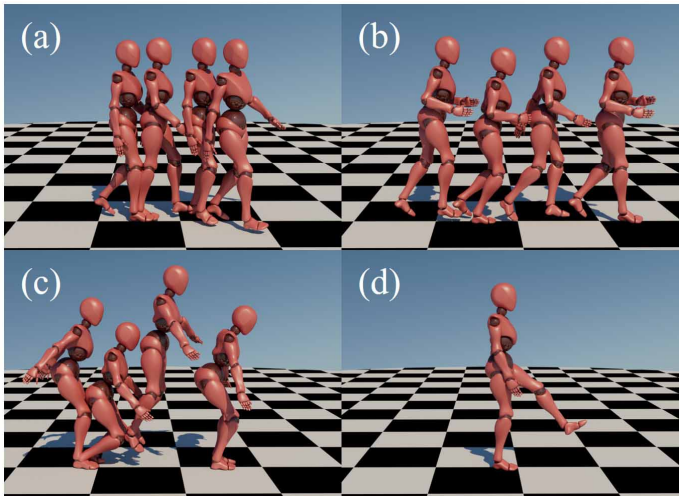

Figure 4.5: Generated motions with different contents. (a) is walking. (b) is running. (c) is jumping. (d) is kicking.

Fig. 4.6 presents several examples of stylised walking motions generated by our model. These samples exhibit noticeable stylistic variations, such as old walking patterns, while preserving coherent walking dynamics. Owing to the stochastic nature of the diffusion-based generation process, the model is able to capture a wide range of stylistic behaviours. In addition to walking, stylised motions of other content categories, including running, are shown in Fig. 4.7, further highlighting the generality of the proposed framework across different motion types.

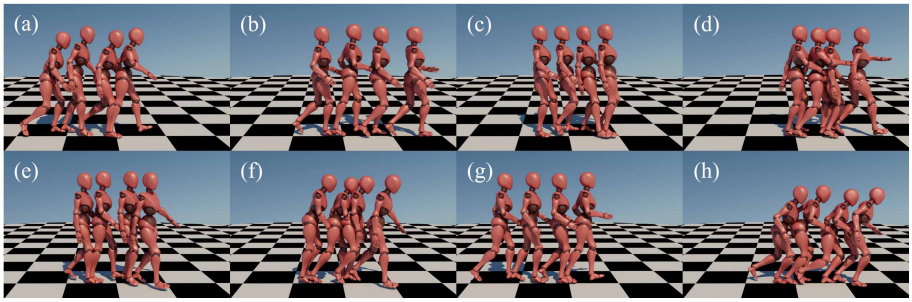

Figure 4.6: Walking motion with styles. (a) is angry walking. (b) is sexy walking. (c) is proud walking. (d) is strutting walking. (e) is neutral walking. (f) is depressed walking. (g) is childlike walking. (h) is old walking.

In addition to quantitative evaluations, we conducted a user study to assess the perceptual quality of the generated motions. A total of 71 participants from diverse geographical regions worldwide took part in the study. Participants were shown motion sequences generated by different methods and were asked to rate them with

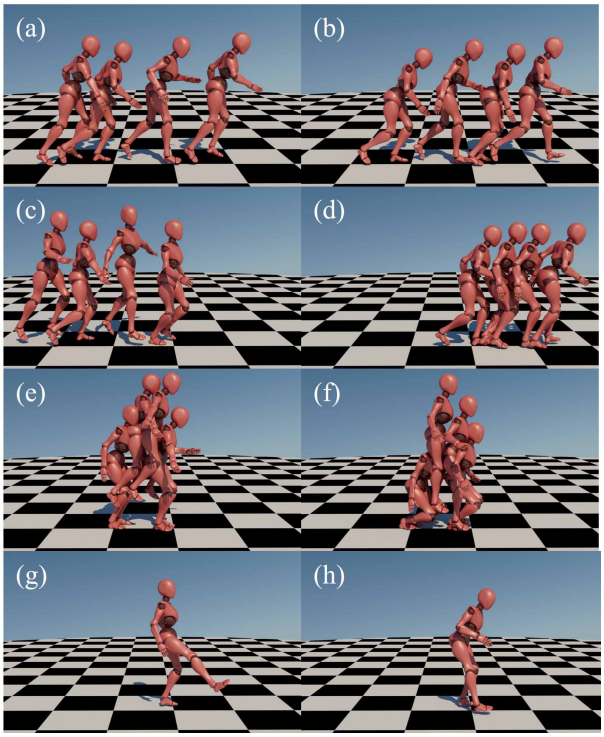

Figure 4.7: We also provide generated styled motions with other contents. (a) is angry running. (b) is depressed running. (c) is strutting running. (d) is old running. (e) is sexy jumping. (f) is proud jumping. (g) is angry kicking. (h) is old kicking.

respect to consistency with the given content and style labels, as well as overall visual quality.

To reduce potential bias, motion clips were presented in a random order without revealing the underlying generation method. All participants provided their ratings independently, and higher ratings indicate better perceived motion quality. The aggregated results were analysed to complement the quantitative metrics reported earlier. This user study offers a human-centred evaluation of motion quality, which is difficult to fully capture using numerical measures alone. Fig. 4.8 demonstrates the superiority of our method against other methods with a higher score value.

4.9 Ablation Study

An ablation study is conducted to investigate the contribution of individual components within the proposed multi-task architecture. The results presented in Table 4.3 show that the full model achieves the lowest FID score among all ablated variants,

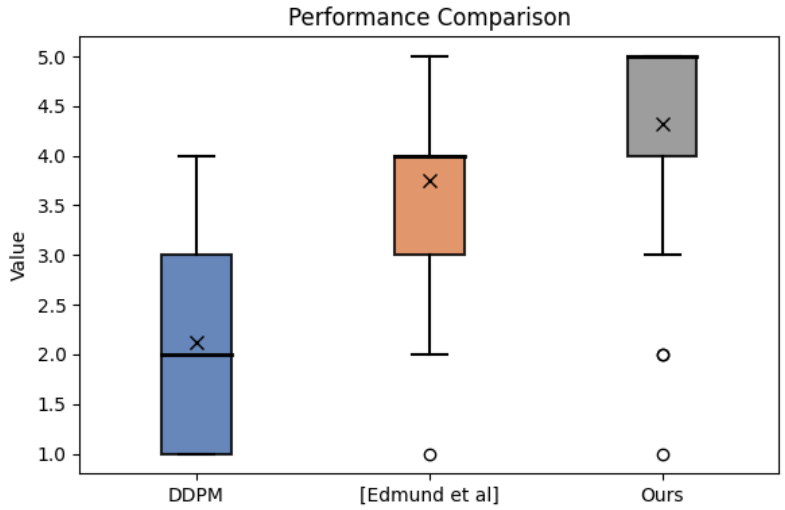

Figure 4.8: User study results presented as box plots of participant ratings across different methods. Higher ratings indicate better perceived motion quality.

confirming the effectiveness of the proposed multi-task DDPM design. Removing the discriminator leads to a noticeable degradation in performance, indicating its importance in coordinating the independently predicted components produced by local guidance. The adversarial objective encourages these components to be combined into temporally smooth and coherent motion sequences. A significant drop in performance is also observed when the root motion prediction is excluded. Global movement trajectories encoded by the root play a crucial role in distinguishing inter-class motion content, suggesting that realistic human motion generation strongly depends on accurate trajectory modelling. Furthermore, eliminating the physical regularisation terms, including velocity and acceleration losses, adversely affects performance. These losses capture stylistic variations within the same motion category, as differences in joint dynamics are primarily reflected through velocity and acceleration patterns. Finally, removing the foot contact loss also results in inferior performance. Estimating supporting foot patterns provides essential cues for maintaining temporal smoothness and physical plausibility in generated motions.

Table 4.3: Ablation study on multi-task architecture.

Model	FID (↓)
w/o foot loss	74.68
w/o root loss	118.20
w/o physical loss	139.44
w/o discriminator	106.29
Ours - full	**56.73**

4.10 Summary

Styled motion synthesis represents a fundamental yet challenging problem with wide-ranging applications across animation, simulation, and interactive media. In this work, we present an end-to-end framework that unifies human motion synthesis and motion style transfer within a single pipeline. Instead of treating these two tasks independently, our approach jointly models inter-class motion content and intra-class stylistic variation within a shared latent space. This unified formulation enables more effective exploration of the coupled content–style distribution and yields improved motion generation quality compared with conventional two-stage pipelines.

An important research direction concerns improving control over limb endpoints, such as hands and feet, remains an open challenge. As illustrated in Fig. 4.9, these regions often exhibit subtle yet complex dynamics that are difficult for diffusion models to capture accurately. While our approach incorporates foot contact estimation and global trajectory modelling, future work may benefit from incorporating stronger physics-based constraints or alternative guidance mechanisms to further improve the realism and controllability of limb-end motions.

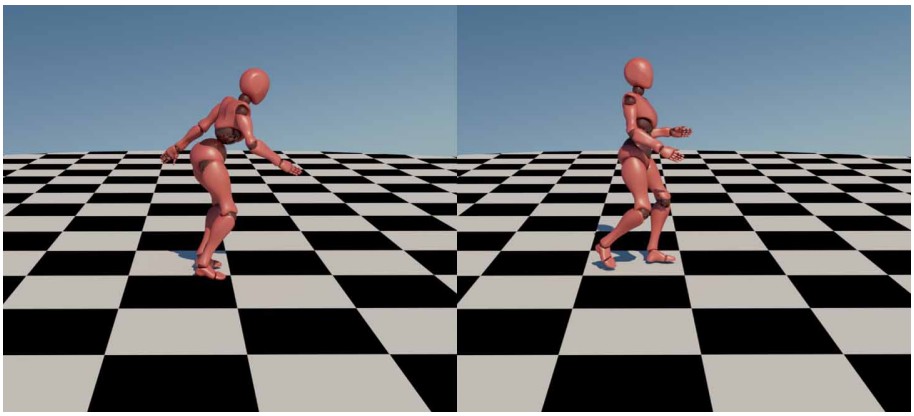

Figure 4.9: Punching is a challenging case.

CHAPTER 5

Diffusion Models for Adversarial Motion Generation

Portions of this chapter have previously been published or will be submitted to the following peer-reviewed publications:

- **Ziyi Chang**, Kanglei Zhou, Xiaohui Liang, Hubert P. H. Shum, "Quality-Advocating Imperceptible Adversarial Attack on Skeleton-based Human Action Recognition," under revision of IEEE Transactions on Circuits and Systems for Video Technology.
- Zhengzhi Lu, He Wang, **Ziyi Chang**, Guoan Yang and Hubert P. H. Shum, "Hard No-Box Adversarial Attack on Skeleton-Based Human Action Recognition with Skeleton-Motion-Informed Gradient," in *ICCV '23*: Proceedings of *the 2023 IEEE/CVF International Conference on Computer Vision*, pp. 4574-4583, Paris, France, IEEE/CVF, Oct 2023.

The goal of adversarial motion generation is to produce motions that are perceptually indistinguishable to human observers while causing a target model to make incorrect predictions. To preserve perceptual consistency, intra-class variations need to be exploited without altering inter-class semantic features. Existing methods typically rely on heuristic rules to define and generate intra-class variations, which often

fail to disentangle intra-class variations from inter-class semantics and may consequently undermine the latter. This chapter introduces a diffusion-based framework for adversarial motion generation that facilitate imperceptibility.

5.1 Introduction

Adversarial attacks on skeletal human action recognition (S-HAR) have attracted increasing attention due to concerns over the robustness of systems deployed in safety-critical and human-centric applications, such as healthcare, action assessment, and surveillance [39]. In this context, adversarial motion generation aims to mislead S-HAR models while preserving perceptual realism, thereby exposing vulnerabilities and improving system robustness [176, 177] of daily applications. For example, assistive robotics involved in interactions with vulnerable users [178, 179] and social robots that rely on human motion understanding to detect hazardous behaviours and prevent accidents [180, 181] need to be robust enough for safety reasons.

From a perceptual perspective, successful adversarial motion generation requires exploiting *intra-class motion variations* while strictly preserving *inter-class semantic features*. That is, adversarial perturbations should remain within the natural variation of a given action class, such that human observers still perceive the motion as belonging to the same action category. However, existing attack methods [114, 117] typically rely on noise-like perturbations applied directly to skeletal inputs, which often violate this constraint. As a result, intra-class variations become entangled with inter-class semantics, leading to noticeable degradation in motion quality and reduced imperceptibility [161, 162].

This limitation is frequently masked by dataset noise [33, 34, 182] and the limited expressiveness of early classifiers [183–185]. Moreover, existing evaluation metrics based on paired pre- and post-attack comparisons fail to faithfully assess post-attack motion quality, as skeletal motions are sparsely distributed and local neighborhoods are not guaranteed to be smooth or physically plausible [25]. With recent advances in high-quality skeletal motion capture and reconstruction [186–188], such perceptual artifacts have become increasingly evident.

We attribute the degradation of adversarial motion quality in prior work to two fundamental issues. First, most attacks optimise paired losses on individual samples, leading to a substantial gap between empirical risk and true risk [189, 190]. Second, gradient-based optimisation is driven by classifier decision boundaries that extend beyond the true motion manifold, pushing perturbations outside the space of natural intra-class variations and resulting in unsmooth, noise-like motions [191].

To address these issues, we propose a diffusion-based adversarial attack framework that explicitly constrains adversarial optimisation within the data manifold. By leveraging a diffusion model to learn the underlying motion distribution, our method generates adversarial motions by exploiting intra-class variations under the constraint of preserving inter-class semantics. This formulation reduces the gap between empirical and true risks and avoids noise-like perturbations. Furthermore, to faithfully evaluate adversarial motion quality, we introduce a new smoothness metric inspired by [192, 193].

Extensive experiments demonstrate that the proposed method achieves the better quality when attacking four state-of-the-art S-HAR classifiers on both a high-quality dataset (100STYLE [134]) and a widely used benchmark (HDM05 [194]). User studies further confirm that our adversarial motions are least perceptible to humans. We also conduct ablation studies on diffusion model configurations and generative design choices. Code is available at `https://github.com/mrzzy2021/QualityPreservingAttack`. Our contributions are summarised as follows:

- We identify a previously overlooked limitation in S-HAR adversarial attacks, where existing methods fail to disentangle intra-class motion variations from inter-class semantics, resulting in perceptible adversarial motions.
- We propose a diffusion-based, distribution-driven adversarial attack framework that exploits intra-class variations under the constraint of preserving inter-class semantic features, enabling imperceptible adversarial motion generation.
- We introduce a smoothness metric to faithfully assess adversarial motion quality and reveal vulnerabilities that existing evaluation metrics fail to capture.

5.2 Distribution-based S-HAR Attack Method

This section presents a distribution-based adversarial attack framework against skeleton-based human action recognition (S-HAR) systems. We first analyse the origin of noise-like perturbations in existing attack methods and motivate the incorporation of data distributions into adversarial optimisation via a generative diffusion latent (Section 5.2.1). We then introduce an attack strategy that leverages this latent to achieve better adversarial motion generation in Section 5.2.2.

5.2.1 The Diffusion Latent for Intra-Class Variation

The Risk Gap of Previous Optimization The noise-like perturbations observed in prior S-HAR attacks primarily arise from a mismatch between empirical risk and true risk. Empirical risk is defined over a finite set of observations, whereas true risk is the expectation over the underlying data distribution [189]. In adversarial attacks, optimisation is typically performed on a single motion instance, resulting in a severe risk gap [190]. This mismatch drives optimisation toward spurious directions, producing noise-like perturbations that degrade post-attack motion quality [191].

Constructing Generative Diffusion Latent To reduce the risk gap and the degradation of motion quality, we introduce a distributional latent derived from generative diffusion models. Diffusion models provide semantically structured latent spaces that capture motion distributions across different noise levels [195], making them well suited for constraining adversarial optimisation within natural intra-class variations. Compared with discriminative models [196] or VAEs [72], diffusion models offer a richer hierarchical representation of motion distributions.

Instead of optimising adversarial perturbations directly in the sparse original data space where smooth motions are unevenly distributed [25], we perform optimisation in the stochastic latent space of diffusion models, as illustrated in Fig. 5.1. This latent space is approximately smooth [197] and enables adversarial modification while preserving inter-class semantics. By mapping a pre-attack motion to a

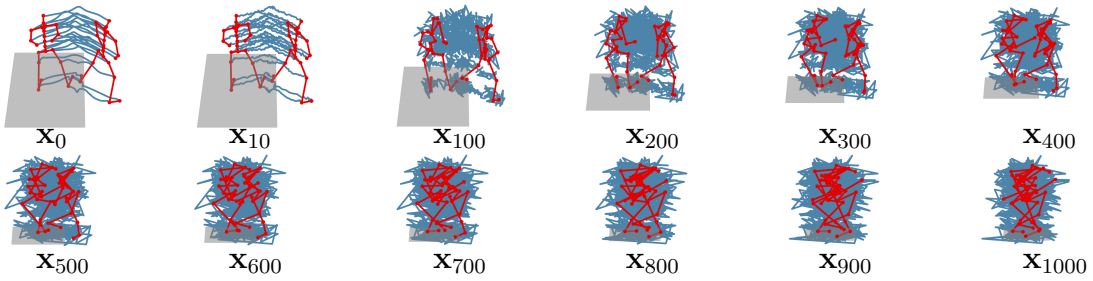

Figure 5.1: The visualization of diffusion latents at different timesteps. As shown, the earlier timesteps maintain more low-level details, the later timesteps focus on high-level structures until latents become pure noise.

latent distribution, the empirical risk is effectively evaluated over infinitely many samples drawn from the distribution, rather than a single instance.

Specifically, we construct a latent proxy consisting of a posterior mean and a stochastic regularisation term. The forward diffusion process maps a pre-attack motion $\mathbf{x}_0$ to a distribution:

$$p(\mathbf{x}_{t+1}|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_{t+1}}\mathbf{x}_0, (1-\alpha_{t+1})\mathbf{I}), \tag{5.1}$$

where $\mathbf{x}_{t+1}$ is a stochastic latent sample. We then consider a single reverse diffusion step to obtain a manipulable latent:

$$p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{x}_0) = \mathcal{N}(\boldsymbol{\mu}, \sigma_t\mathbf{I}), \tag{5.2}$$

where the posterior variance is fixed by the noise schedule and the posterior mean $\boldsymbol{\mu}_t$ encodes the denoising direction. Using Tweedie's formula [198], the posterior mean can be expressed as:

$$\boldsymbol{\mu}_t = \gamma_t\mathbf{x}_0 + \lambda_t\epsilon_t + \delta_t\nabla \log p_\theta(\mathbf{x}_{t+1}, \mathbf{y}), \tag{5.3}$$

where $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$. The first term preserves inter-class semantic content, while the distributional term aligns the optimisation with the learned motion manifold. Unlike prior attacks that impose point-wise constraints in data space, we regularise the adversarial process over the latent distribution $\mathbf{x}_t \sim p(\mathbf{x}_t|\mathbf{x}_0)$, explicitly constraining

intra-class variation.

The final latent representation used for adversarial optimisation is defined as

$$\boldsymbol{\kappa}_t = \boldsymbol{\mu}_t + \mathbf{x}_t. \tag{5.4}$$

Discussion The proposed distributional latent is derived from the data distribution rather than an individual motion sample. Compared with deterministic latents [55], stochastic diffusion latents provide higher expressive capacity [195] and guide optimisation toward high-density regions of the motion manifold. By operating on a single denoising step, the attack remains efficient while sampling different timesteps enables the exploitation of diverse intra-class variations. This formulation allows adversarial optimisation to remain within the natural motion distribution, achieving imperceptibility of adversarial motions.

Despite their effectiveness, latent-based approaches also potentially introduce several limitations. A primary drawback is the potential loss of fine-grained motion details when motion data are encoded into a latent space. This mapping can lead to over-smoothing or reduced diversity, particularly for subtle execution-level variations. In addition, the learned latent representation may be difficult to interpret, making it challenging to explicitly control or diagnose specific motion attributes. Latent-based models are also sensitive to the choice of latent dimensionality and/or regularisation strength, which can affect both training stability and generation quality. Furthermore, errors introduced in the latent space can propagate through the decoding process, potentially amplifying artifacts in the generated motions.

5.2.2 The Attack Strategy for Diffusion-Driven Attack

Overview of Strategy To leverage the diffusion latent space for adversarial attacks, we propose perturbing the source motion observation by randomly sampling different class labels rather than relying on the gradient of a specific classifier [114, 120, 125], which has been commonly used in this field. This is not only because obtaining the gradient of a specific classifier is challenging in real-world applications [125], but also because the gradient of a classifier may not be reliable. As

classifiers focus on the label distribution rather than data distribution, their gradient may point to out-of-distribution regions. While following this gradient achieves the shortest trajectory of deceiving classifiers, it leads to a decline in post-attack motion quality and undermines the imperceptibility. As a result, we design our method to only rely on the classifier's decisions to determine when to stop and drive the motion towards adversarial samples by sampling adversarial labels for the conditional diffusion model.

Optimization Objective Specifically, we denote the stochastic distributional latents of the source motion observation and the adversarial motion as $\boldsymbol{\kappa}_t^{\text{src}}$ and $\boldsymbol{\kappa}_t^{\text{adv}}$. The source motion and the adversarial motion are mapped to the latent space by the diffusion forward process and then the desired posterior mean are obtained within a single timestep denoising conditioned on the ground truth label $\mathbf{y}^{\text{src}}$ and the randomly sampled the adversarial label $\mathbf{y}^{\text{adv}}$ from the set of all possible labels excluding the ground truth label, respectively. The regularizer is obtained through the forward process. We define our objective function as follows:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\kappa}_t} :=& 0.5 \times \mathbb{E}_{t,\epsilon_t}\left[\|\boldsymbol{\kappa}_t^{\text{adv}} - \boldsymbol{\kappa}_t^{\text{src}}\|_2^2\right], \\ =& 0.5 \times \mathbb{E}_{t,\epsilon_t}\| \underbrace{\boldsymbol{\mu}_t^{\text{adv}} - \boldsymbol{\mu}_t^{\text{src}}}_{\text{latent}} + \underbrace{(\mathbf{x}_t^{\text{adv}} - \mathbf{x}_t^{\text{src}})}_{\text{regularization}} \|_2^2\end{aligned} \tag{5.5}$$

where $\boldsymbol{\kappa}_t^{\text{src}}$ and $\boldsymbol{\kappa}_t^{\text{adv}}$ are our defined latents and serve as proxies for adversarial attack. The first term, $\boldsymbol{\kappa}_0^{\text{adv}}$, represents the direction towards being adversarial, while the second term, $\boldsymbol{\kappa}_0^{\text{adv}}$, represents the direction of maintaining the representative information within the original class. Our optimization objective is an expectation over the diffusion timesteps t and the stochastic distributional latent, indicated by the randomly sampled noise ϵ in the forward process where information is encoded into the latent space of the diffusion model.

By minimizing the defined object, our optimization aligns the stochastic distributional latents of the source and the adversarial motions. Specifically, we calculate

Algorithm 1 Diffusion-based Adversarial Motion Attack on S-HAR

Require: Diffusion model θ, a classifier φ, a motion $\mathbf{x}_0^{\text{src}}$ with label $\mathbf{y} \in \mathbf{Y}$, maximum iteration I, diffusion timesteps T

$\mathbf{x}_0^{\text{adv}} \leftarrow \mathbf{x}_0^{\text{src}}$
$i \leftarrow 0$
$\mathbf{y}^{\text{adv}} \sim \mathbf{Y}_{/\ \mathbf{y}^{\text{src}}}$ ▷ Randomly sample an adversarial label
while $i \leq I$ *and* $\mathbf{y}^{\text{pred}} \neq \mathbf{y}^{\text{src}}$ **do**
    $t \sim [1, T]$
    $\kappa_t^{\text{src}} = \mu_t(\mathbf{x}_0^{\text{src}}, \mathbf{y}^{\text{src}}; \theta) + \mathbf{x}_t^{\text{src}}$
    $\kappa_t^{\text{adv}} = \mu_t(\mathbf{x}_0^{\text{adv}}, \mathbf{y}^{\text{adv}}; \theta) + \mathbf{x}_t^{\text{adv}}$
    $\text{grad} = \kappa_t^{\text{adv}} - \kappa_t^{\text{src}}$
    $\mathbf{x}^{\text{adv}} = \mathbf{x}^{\text{adv}} + \text{grad}$
    $\mathbf{y}^{\text{pred}} = \arg\max p_\varphi(\mathbf{y}|\mathbf{x}^{\text{adv}})$ ▷ Get classifier decision
end while

the gradient of $\mathcal{L}_{\kappa_t}$ with respect to $\boldsymbol{\kappa}_t^{\text{adv}}$ and obtain our adversarial gradient:

$$\begin{aligned} \text{grad} :=& \nabla \mathcal{L}_{\kappa_t} \\ =& \mathbb{E}_{t,\epsilon_t} \left[\kappa_t^{\text{adv}} - \kappa_t^{\text{src}} \right], \end{aligned} \tag{5.6}$$

on which we rely to iteratively update the $\mathbf{x}_0^{\text{adv}}$. The Eq. 5.6 facilitates the adversarial effect. Instead of directly requiring the paired pre-attack and post-attack motions to be close with each other in data space, our optimization aligns the latent distribution of pre-attack and post-attack motions. Constraining the two distributional latents $\boldsymbol{\kappa}_t^{\text{adv}}$ and $\boldsymbol{\kappa}_t^{\text{src}}$ promotes that the latent distributions of $\mathbf{x}_0^{\text{adv}}$ and $\mathbf{x}_0^{\text{src}}$ remain closely aligned rather than merely examine the pre-attack and the post-attack motions and rely on the gradients from an external classifier. Our optimization represents a single timestep examination, which is different from a multi-timestep generative process conditioned on $\mathbf{y}^{\text{adv}}$. By optimizing over the expectation of randomly sampled timesteps, the trajectory defined by the posteriors is expected to be closely aligned with each other. Consequently, our attack strategy facilitates the generation of $\mathbf{x}_0^{\text{adv}}$ that corresponds with $\mathbf{y}^{\text{adv}}$ to deceive the target model, while simultaneously reducing the influence on the quality of $\mathbf{x}_0^{\text{src}}$. Detailed adversarial attack methodology is provided in Algorithm 1 and Fig. 5.2.

Relationship with Previous Optimization We further demonstrate that our proposed method implicitly integrates previous approaches while additionally offer

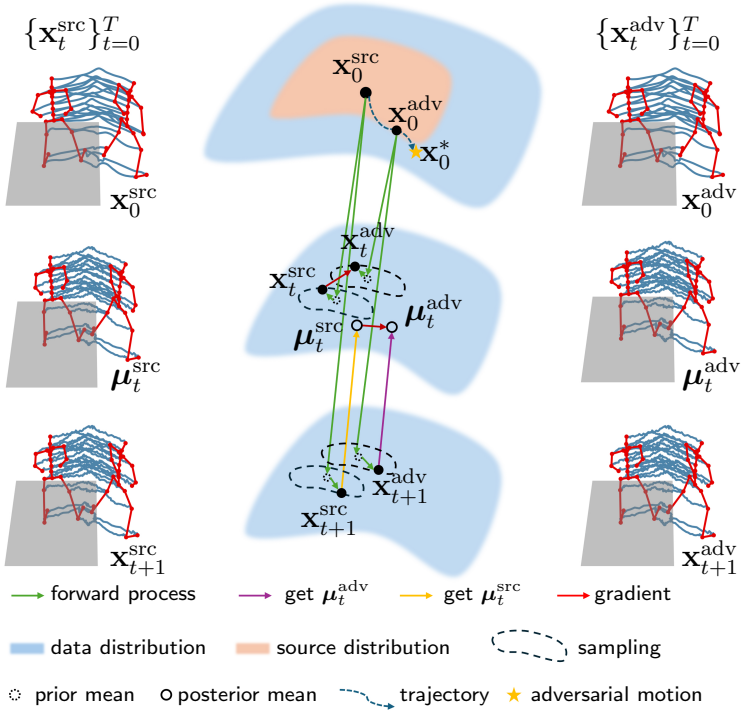

Figure 5.2: The illustration of attack strategy. We illustrate an intermediate calculation at the timestep t during the optimization of achieving the final adversarial motion $\mathbf{x}_0^*$.

the advantage of distributional prior knowledge provided by a pre-trained diffusion model to minimize the risk gap. The Eq. 5.6 is equivalently represented using the input motion and the learned distribution, from which the following detailed formulation is derived:

$$\begin{aligned}
\text{grad} &:= \mathbb{E}_{t,\epsilon_t}\left[\kappa_t^{\text{adv}} - \kappa_t^{\text{src}}\right] \\
&= \mathbb{E}_{t,\epsilon_t}\Bigg[\psi(t)\Big(\underbrace{\mathbf{x}_0^{\text{adv}} - \mathbf{x}_0^{\text{src}}}_{\substack{\text{Distance Function} \\ \text{in Original Space}}}\Big) \\
&\quad + \chi(t)\Bigg(\underbrace{\left(\nabla \log p(\mathbf{x}_t^{\text{adv}}) - \nabla \log p(\mathbf{x}_t^{\text{src}})\right)}_{\text{Distribution Constraint}} \\
&\quad + \Big(\underbrace{\nabla \log p(\mathbf{y}^{\text{adv}}|\mathbf{x}_t^{\text{adv}})}_{\substack{\text{Adversarial} \\ \text{Gradient}}} - \underbrace{\nabla \log p(\mathbf{y}^{\text{src}}|\mathbf{x}_t^{\text{src}})}_{\substack{\text{Representiveness} \\ \text{of the Given Input}}}\Big)\Bigg)\Bigg],
\end{aligned} \tag{5.7}$$

where we decompose the probability with Bayes' theorem. The four terms serve distinct yet complementary purposes in our objective function. The first term demonstrates a similar measurement to previous attack methods, ensuring that the

adversarial sample does not deviate excessively from the input sample. However, our ability of reducing the influence on quality, which distinguish our method from previous methods, is ensured through the remaining terms. Unlike traditional adversarial optimization, which considers only an individual motion pair as shown in the first term, our perturbation strategy provides the modification gradient based on data distributions that are composed of infinite motions. The second term measures the distribution density and ensures that the adversarial motions remain not only neighborhood but also high density area. The third term considers the adversarial gradient while the fourth term involves the representativeness of the observed sample as a consideration for the range of pre-attack motion neighborhood by evaluating how representative the given input is with respective to the label. These considerations leverage the distributional prior knowledge and enables better quality under adversarial modifications.

5.3 Perception Aligned Smoothness Metric

In this section, we examine previous metrics and formulate the proposed metric to faithfully measure the quality of generated adversarial motions in terms of smoothness.

Unfaithfulness of Existing Metrics The evaluation of existing metrics fails to accurately measure quality. Previous metrics rely on paired comparisons between the pre-attack and post-attack motions to assess quality, especially smoothness [114]. However, remaining within the neighborhood of a pre-attack motion does not ensure motion quality comparable to clean motions because smooth and plausible motions are sparsely distributed [25]. Consequently, these metrics cannot reliably determine whether adversarial motions are sufficiently smooth due to their misalignment with human perception.

Human Perception Aligned Metric Human motion exhibits characteristic temporal regularities. As a result, real-world movements tend to be smooth and continuous over time, without abrupt changes in joint trajectories. Motivated by these

observations, we introduce a smoothness-based metric to assess the temporal regularity of generated adversarial motions from a kinematic perspective. The proposed metric operates directly on joint trajectories and captures high-order temporal variations that are indicative of abrupt or irregular motion.

Specifically, we define the following smoothness measure:

$$\text{smoothness} = \frac{1}{T}\int_0^T \left\| \frac{\mathrm{d}^4 J(t)}{\mathrm{d}t^4} \right\| \mathrm{d}t, \tag{5.8}$$

where T denotes the temporal length of the motion sequence and $J(t)$ represents the joint positions at time t. Higher values correspond to stronger high-frequency temporal variations, indicating reduced smoothness in the motion.

5.4 Experiments

In this section, we begin by outlining the experimental settings. Subsequently, we quantitatively and qualitatively analyze the performance of our proposed method and existing adversarial techniques. Finally, we perform a user study to empirically evaluate the imperceptibility of post-attack motions and also ablation studies to validate our configurations.

5.4.1 Experimental Settings

Datasets To evaluate our imperceptible adversarial attacks, we select the 100STYLE [134] dataset due to its noise-free and inherently high-quality characteristics. The 100STYLE dataset is collected using a motion capture system and comprises 100 classes of different styles. We represent the skeletons using Cartesian coordinates for 23 joints. This dataset is pre-processed by segmenting long sequences into several segments according to the valid periods provided by [134]. Additionally, we employ the HDM05 [194] dataset to evaluate our method on a smaller scale dataset. Although HDM05 is commonly used for action recognition tasks, it has slightly lower quality compared to 100STYLE. We adhere to the pre-processing procedure outlined in [114]. The HDM05 dataset includes 65 classes of different human actions,

and the hip joint is fixed to the origin after pre-processing.

Evaluated Models Given the significant advancements in the field of human action recognition, we adopt the latest S-HAR models as victim classifiers to effectively evaluate the performance of adversarial attack methods against advanced classifiers. Specifically, we select the Style classifier [199], STTFormer [200], Skateformer [201], and FR-Head [202] as victim models. These models encompass both transformer-based and graph-based architectures. We utilize their publicly available codebases to train their models.

Evaluation Metrics Human motions are governed by physical and biomechanical constraints, thereby necessitating the assessment of visual motion quality in terms of smoothness and plausibility [1]. To evaluate the motion quality in adversarial attacks, we introduce the smoothness metric, as discussed in Section 5.3, which is grounded in the characteristics of smooth human movements. Additionally, we report the Frechet Inception Distance (FID) and Maximum Mean Discrepancy (MMD) based on acceleration to measure the smoothness by assessing the distributional similarity between pre-attack and post-attack motions. As for plausibility, we report the foot skating ratio and bone length variations between frames. Following [35], foot skating is quantified by the consistency between foot velocity and foot height. A high foot skating ratio indicates significant violation in terms of physical constraints. Bone length variation [203] measures the consistency of bone lengths across frames. Higher deviations indicating distortions in the skeleton structure. Finally, we report the success rate of the adversarial attacks to evaluate the threatfulness.

Attacking Methods We compare our method with the state-of-the-art (SOTA) S-HAR attack technique, i.e. SMART [114], as well as other adversarial attack methods including I-FGSM [204], MI-FGSM [122], and MIG [205]. To ensure a fair comparison, we execute 2000 iterations for each attack method, allowing all methods to explore a broader solution space in their pursuit of effective adversarial motions. Since our method requires a pre-trained diffusion model, we adopt the diffusion model proposed by [206] and follow the prescribed settings to pre-train it

on the two datasets.

5.4.2 Adversarial Motion Quality Evaluation

We qualitatively and quantitatively evaluate the performance of adversarial attack in terms of deceitfulness, motion quality including smoothness and plausibility, and human imperceptibility. Deceitfulness measures the effectiveness of an adversarial method. In addition to deceitfulness, motion quality assesses whether the adversarial motions are plausible and smooth, and serves as an indicator for imperceptibility. Beyond analytical measurements, a user study is essential for evaluating human imperceptibility, as this property fundamentally concerns human perception rather than purely numerical differences. While quantitative metrics can characterise statistical or kinematic deviations between motions, they cannot fully capture whether such differences are perceptible to human observers. Imperceptibility in the context of adversarial attack refers to the extent to which humans are unable to distinguish between pre-attack and post-attack motions when they are presented together. A user study therefore provides critical validation by directly assessing perceptual indistinguishability, complementing analytical evaluations and ensuring that the generated adversarial motions remain visually and perceptually plausible. For simplicity, we present only representative results in this chapter.

Deceitfulness The success rates presented in Table 5.1 indicate that our method is the most deceitful compared with other methods on the 100STYLE dataset. Our method consistently achieves an average success rate of 100%. Our attack effectively modifies nearly every input to become adversarial without relying on the gradient of victim model within a limited number of iterations. In contrast, other methods may struggle to generate adversarial motions within iteration constraints. This demonstrates that the stochastic latent features convey comprehensive information [207] about the underlying dynamics via different timesteps while the information cannot be easily represented in the original motion space.

Table 5.1: Generated Adversarial Motion Quality Comparison on 100STYLE dataset. *FID* is the Fréchet inception distance. *MMD* is the maximum mean discrepancy. *Phys. Nat.* stands for physiciological smoothness. *FS* means the ratio between the frames with foot sliding and total frames. *Bone Variation* calculates the differences of bone lengths in two consecutive frames.

Victim	Method	Success Rate ↑	FID ↓	MMD ↓	Phys. Nat. ↓	FS ↓	Bone Variation ↓
Style [199]	I-FGSM	81.72%	194.95	0.053	129.42	0.147	10.96
	MI-FGSM	82.08%	195.22	0.053	131.33	0.147	11.16
	MIG	70.61%	238.23	0.071	264.46	0.234	23.22
	SMART	42.65%	242.12	0.072	97.73	0.119	4.00
	Ours	**100%**	**18.91**	**0.011**	**16.05**	**0.075**	**2.94**
STTFormer [200]	I-FGSM	80.29%	162.55	0.043	190.98	0.196	19.28
	MI-FGSM	80.29%	162.78	0.043	191.10	0.197	19.29
	MIG	72.76%	190.75	0.050	280.56	0.241	27.63
	SMART	29.39%	195.06	0.053	103.56	0.167	3.92
	Ours	**100%**	**20.28**	**0.013**	**16.42**	**0.077**	**3.09**
SkateFormer [201]	I-FGSM	68.46%	79.63	0.019	43.90	0.053	4.53
	MI-FGSM	68.82%	79.06	0.019	44.70	0.053	4.61
	MIG	60.22%	102.77	0.026	70.72	0.069	7.30
	SMART	31.90%	125.46	0.030	29.32	**0.047**	**1.60**
	Ours	**100%**	**20.16**	**0.014**	**16.33**	0.079	2.99
FR-Head [202]	I-FGSM	86.38%	226.80	0.068	312.90	0.276	25.65
	MI-FGSM	86.74%	226.67	0.068	316.77	0.280	26.08
	MIG	78.14%	242.85	0.075	489.37	0.396	38.12
	SMART	32.97%	262.75	0.100	214.63	0.301	6.33
	Ours	**100%**	**19.13**	**0.011**	**16.75**	**0.084**	**3.23**
Average	I-FGSM	79.21%	165.98	0.046	169.3	0.168	15.11
	MI-FGSM	79.48%	165.93	0.046	170.98	0.169	15.29
	MIG	69.34%	193.65	0.056	276.28	0.235	24.07
	SMART	34.23%	206.35	0.064	111.31	0.159	3.96
	Ours	**100%**	**19.62**	**0.012**	**16.39**	**0.079**	**3.06**

Motion Quality Regarding smoothness, the FID and MMD scores in Table 5.1 show that the distribution of our generated motions closely resembles that of smooth motions, whereas other methods exhibit significant distributional deviations. Our adversarial motions maintain proximity to the ground truth distribution of motion dynamics by a considerable margin. The lowest FID and MMD scores indicate that our adversarial motions are indistinguishable from those in the ground truth dataset. In other words, the post-attack motion quality are better because our method leverages the distributional knowledge introduced by the pre-trained diffusion model. By utilizing data distribution, our method minimizes the gap between empirical and true risks by assessing deviations between the modified motions and the data distribution. Conversely, other methods can only access single source motions, and disrupt the smoothness of post-attack motions, resulting in much higher FID and MMD scores.

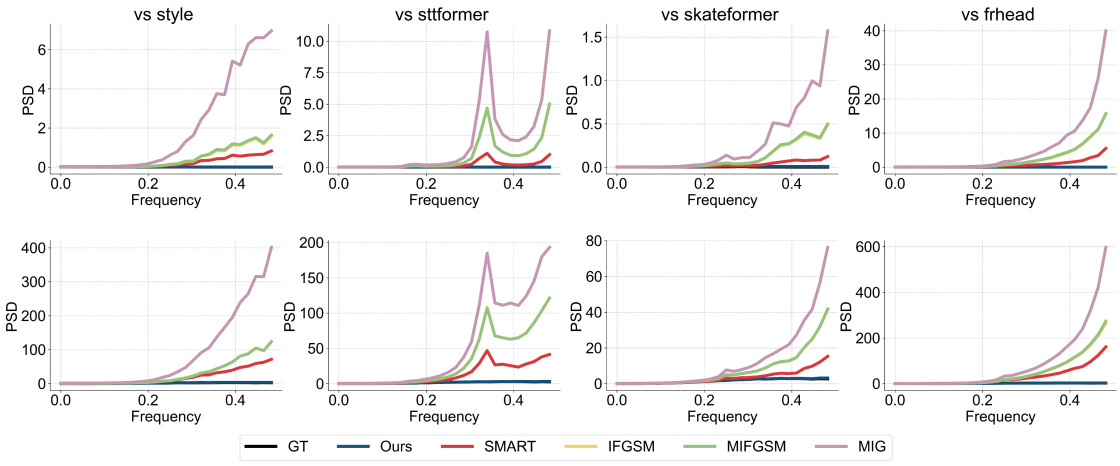

Figure 5.3: The mean power spectral density of adversarial samples found on 100STYLE (upper row) and HDM05 (lower row) against four classifiers.

Smoothness reflects whether adversarial motions conform to smooth movements from the biomechanical perspective of muscle activation. Our method achieves the best performance by a large margin, as shown in Table 5.1. This indicates that even after adversarial attacks, our generated motions retain plausible movements that real-world humans can physically perform. To further evaluate smoothness, as illustrated in Fig. 5.3, we present the mean power spectral density (PSD) of adversarial

samples. The spectral density of our post-attack motions in the high-frequency domain is significantly lower than that of other adversarial motions. Besides, our spectral density closely aligns with the ground truth curve. This suggests that our adversarial motions suffer from few noise-like perturbations because we minimize empirical risk over the data distribution rather than a single sample.

We further compare the effect of explicitly regulating motion dynamics with respect to a given sample in SMART and implicitly regulating motion dynamics through data distribution in our diffusion-based method. Fig. 5.4 displays the acceleration changes of a sample for comparison. Our method introduces modifications that are not only smaller in magnitude but also smoother and more consistent, whereas SMART does not perturb the motion coherently and consistently. These results imply that minimizing empirical risk over only a given sample leads to deviations from the smooth motion distribution and results in the post-attack quality decline.

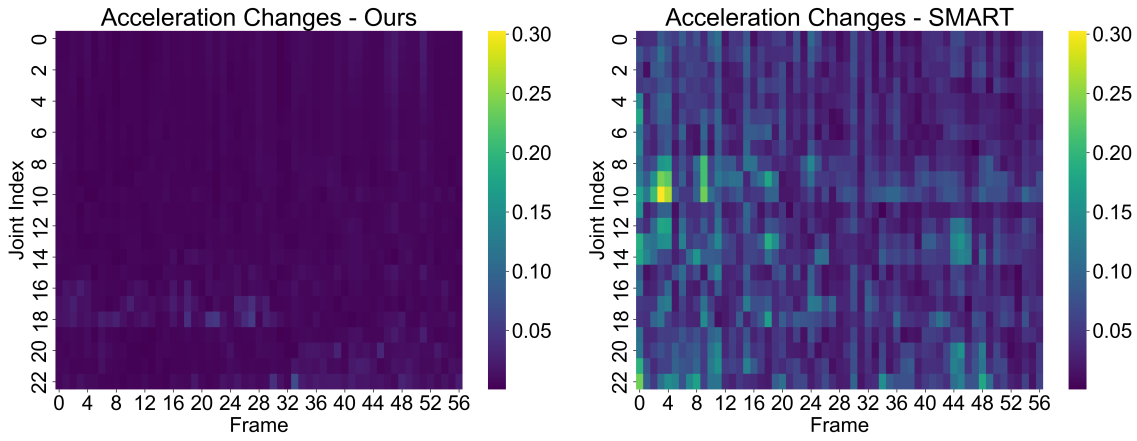

Figure 5.4: The visualization of acceleration changes.

In terms of physical plausibility, the foot sliding ratio and bone length variation in Table 5.1 demonstrate that our generated adversarial motions exhibit superior movement coherency and skeletal consistency compared to other methods. The foot skating ratio indicates the coherence of movements concerning foot contact with the ground. Our method achieves the lowest foot skating ratio, suggesting that post-attack movements are still well synchronized. This is attributed to our method's ability to minimize the risk gap through the integration of data distribution rather

than relying on a single sample. Additionally, our adversarial motions exhibit the lowest variation in bone lengths, maintaining consistent skeletal structures across frames. By leveraging access to the data distribution rather than individual samples, our method ensures that modifications preserve cross-frame skeletal integrity.

Table 5.2: Generated Adversarial Motion Quality Comparison on HDM05 dataset. *FID* is the Fréchet inception distance. *MMD* is the maximum mean discrepancy. *Phys. Nat.* stands for physiciological smoothness. *Bone Variation* calculates the differences of bone lengths in two consecutive frames.

Victim	Method	Success Rate ↑	FID ↓	MMD ↓	Phys. Nat. ↓	Bone Variation ↓
Style [199]	I-FGSM	93.91%	129.78	0.086	1226.33	126.39
	MI-FGSM	93.91%	129.79	0.086	1226.98	126.46
	MIG	92.47%	194.33	0.170	2273.09	238.34
	SMART	68.46%	142.39	0.116	1052.44	41.00
	Ours	**100%**	**22.91**	**0.012**	**187.28**	**32.15**
STTFormer [200]	I-FGSM	86.74%	162.98	0.124	1488.08	144.97
	MI-FGSM	86.74%	163.18	0.125	1487.92	144.95
	MIG	76.34%	202.24	0.188	1945.83	189.95
	SMART	75.63%	174.86	0.129	974.33	33.30
	Ours	**100%**	**21.89**	**0.011**	**185.73**	**30.31**
Skateformer [201]	I-FGSM	83.87%	36.66	0.022	521.52	54.50
	MI-FGSM	84.23%	36.83	0.022	523.93	54.83
	MIG	78.14%	57.59	0.037	692.53	78.87
	SMART	65.59%	50.84	0.028	364.26	**13.54**
	Ours	**100%**	**18.67**	**0.011**	**181.89**	28.09
FR-Head [202]	I-FGSM	96.06%	169.34	0.124	1902.39	177.91
	MI-FGSM	96.77%	169.96	0.126	1920.98	179.56
	MIG	92.47%	215.88	0.206	2757.84	243.96
	SMART	81.72%	211.88	0.173	1574.95	55.46
	Ours	**99.28%**	**20.80**	**0.013**	**182.96**	**31.07**
Average	I-FGSM	90.15%	124.69	0.089	1284.58	125.94
	MI-FGSM	90.41%	124.94	0.090	1289.95	126.45
	MIG	84.86%	167.51	0.150	1917.32	187.78
	SMART	72.85%	144.99	0.112	991.50	35.83
	Ours	**99.82%**	**21.07**	**0.012**	**184.47**	**30.41**

We also achieve the best performance on the HDM05 dataset, as shown in Table 5.2. Our method successfully fools all systems with an average success rate of 99.82%. This further validates the efficacy of our attack method even when the data quality and amount are not as high as that of the 100STYLE dataset, demonstrating its threatfulness to such systems. Moreover, the distribution of our adversarial motions closely aligns with the ground truth dataset, evidenced by the lowest FID and MMD scores. This validates that our adversarial motions are as smooth as pre-attack motions. Besides, smoothness indicates that our generated motions adhere to real-world biomechanical constraints. Since the HDM05 dataset data have been centered to the origin of the coordinates, we do not report the foot skating ratio for

this dataset. Additionally, our method achieves superior motion plausibility with the lowest bone length variation. Both the best physical plausibility demonstrate that our method achieves better quality of adversarial motions.

Human Imperceivability We visualize the trajectories of joints in adversarial motions generated by our method and previous methods in Fig. 5.5 against the four victim models. In real-world smooth movements, joint trajectories are typically smooth and stable. As shown in Fig. 5.5, our method produces the most stable and smooth trajectories for all joints in the adversarial motions. By minimizing the risk gap over the data distribution rather than a single motion, our optimization results do not introduce noise-like perturbations, whereas other methods suffer from significant gaps by optimizing over a single input. Consequently, the noise-like perturbations lead to a decline in the post-attack motion quality with observable unstable trajectories and undermines imperceptibility.

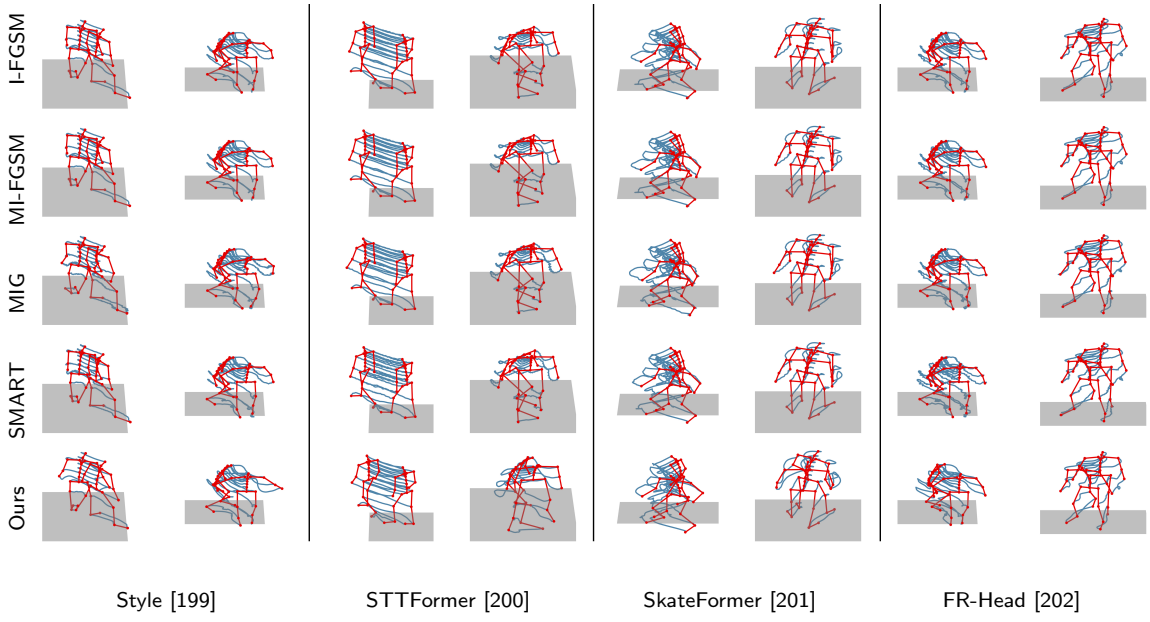

Figure 5.5: Visual comparison among the adversarial motions generated by different attack methods against victim models. We visualize the starting and the ending poses in red, the trajectories of all joints in blue, and the ground floor in grey. Our adversarial motions exhibit the most smooth and stable trajectories.

Additionally, we recruited volunteers without visual impairments from diverse backgrounds and a balanced gender representation. They participate in question-

naires to evaluate the imperceptibility of our adversarial samples. We provided participants with batches of motions from the same label, which pre-attack motions and post-attack motions generated by our method are mixed together. Participants are asked to select the motions they consider most likely to have potentially been attacked without any time constraints. As illustrated in Fig. 5.6, our method produces adversarial motions that are the least perceivable by humans.

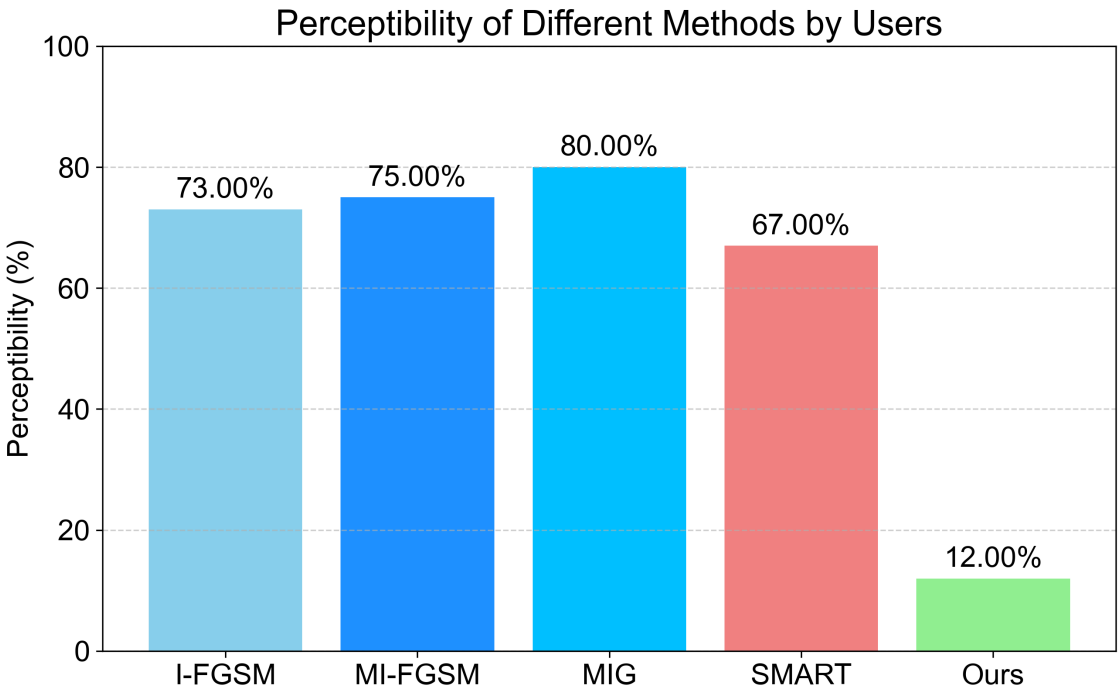

Figure 5.6: Perceptibility comparison across different methods.

5.4.3 Ablation Study

Given that our method is based on the latent space of a diffusion model, we first perform an ablation study to investigate the influence of the chosen timesteps in constructing the stochastic latent space. Secondly, we examine the impact of different latents used for adversarial attacks to determine which latent is the most suitable for the quality-advocating imperceptible attack task. Finally, we validate the choice of diffusion models by exploring alternative generative models for quality-advocating adversarial attack.

Table 5.3: The quality of adversarial motions generated via different variants and configurations. *FID* is the Fréchet inception distance. *MMD* is the maximum mean discrepancy. *Phys. Nat.* stands for physiciological smoothness. *FS* means the ratio between the frames with foot sliding and total frames. *Bone Variation* calculates the differences of bone lengths in two consecutive frames.

Variants	Config.	Success Rate ↑	FID ↓	MMD ↓	Phys. Nat. ↓	FS ↓	Bone Variation ↓
Timestep	[1, 20]	100%	14.01	0.007	13.28	0.090	3.22
	[980, 1000]	100%	26.14	0.017	30.37	0.050	1.62
Latent	$\boldsymbol{\kappa}_t$	100%	18.91	0.011	16.05	0.075	2.94
	$\hat{\mathbf{x}}_0$	100%	19.26	0.011	56.70	0.083	3.31
Architecture	Diffusion [206]	100%	18.91	0.011	16.05	0.075	2.94
	VAE [72]	96.06%	198.00	0.056	1058.08	0.083	214.75

Timestep of Stochastic Latent Space

We investigate the impact of different timestep ranges used to map motions from data space to diffusion latent space for adversarial attacks. It has been shown that the information encapsulated by the latents have smooth transitions with respect to the diffusion timesteps [208, 209], we conduct experiments using two representative timestep ranges: the earliest timestep range ($t \sim [1, 20]$) and the latest timestep range ($t \sim [980, 1000]$) for comparison.

A trade-off exists between motion smoothness and plausibility, as shown in Table 5.3. This trade-off effect arises from the different emphasis on information encoded in the latents. Generally, latents derived from earlier timesteps primarily capture low-level details, while those from later timesteps focus on high-level structural patterns [210]. Utilizing earlier timesteps to construct the stochastic distributional latents is biased toward detailed movements, with a reduced understanding of global coherency and consistency. This bias enhances smoothness by conveying motion dynamics. However, it also exacerbates foot skating and bone length variation due to the lack of global rationality in the latents. Conversely, constructing the distributional latents using later timesteps undermines smoothness, as the generated adversarial motions deviate from the smooth distribution, leading to higher FID and MMD scores.

Alternative Latents for Attack

We examine the influence of different latents used for our quality-advocating adversarial attacks. Specifically, we compare the constructed latent with the pre-

dicted pre-attack data as the latent for adversarial attack. The constructed latent $\boldsymbol{\kappa}$ evaluates the required perturbations based on a single timestep of denoising with regularization, and allows for step-by-step motion modifications. In contrast, the predicted pre-attack data $\hat{\mathbf{x}}_0$ represents modifications along the entire chain of previous timesteps with approximation. We conduct experiments to determine whether the constructed latent is more effective to advocate post-attack motion quality than leveraging the predicted pre-attack data.

As shown in Table 5.3, using the predicted pre-attack data as the latent leads to a consistent decline in all performances of adversarial attack. Although both latent choices achieve the same success rate, the quality of adversarial motions deteriorates significantly when switching from the constructed latent to the predicted data. This deterioration stems from the approximation error when calculating the required latents. The predicted $\hat{\mathbf{x}}_0$ is derived from the $\mathbf{x}_{t+1}$ by approximating all previous timesteps $\{i\}_{i=1}^{t}$ collectively, thereby ignoring information from intermediate timesteps. Conversely, the constructed latent κ_t, also derived from the $\mathbf{x}_{t+1}$, considers only the desired changes within a single timestep. This finer-grained modification enhances the quality during adversarial attacks by integrating fewer approximation errors.

Alternative Distribution Modeller

Given that our attack method is based on generative models, we explore the use of alternative generative architectures. Similarly, we utilize their latent spaces to modify motions adversarially. Specifically, we conduct experiments using either diffusion models or variational autoencoders (VAEs). We train the VAE with label conditions following the architecture and training configurations outlined in [72]. To maintain consistency with our use of the posterior mean as latents, we employ the mean of the latent distribution in the VAE. All other configurations remain consistent with those used in our diffusion-based method.

As shown in Table 5.3, the performance of adversarial attacks using a VAE model significantly declines compared to our diffusion-based method. It indicates that effective motion modification requires a comprehensive representation of under-

lying patterns. In contrast to diffusion models, VAEs utilize a single latent feature rather than a hierarchy of features. Consequently, using a single latent feature in VAEs results in ambiguity regarding motion dynamics and leads to under-expressed movement coherency and consistency.

In addition to quantitative results, we present qualitative visual comparisons in Fig. 5.7 to illustrate the effects of different ablation settings. Using latent representations from timesteps that are either too early or too late results in slight jittering in the right hand, while directly operating on $\mathbf{x}_t$ leads to a more pronounced degradation in overall motion quality. Furthermore, replacing the diffusion model with a VAE introduces severe jitter artefacts, highlighting the critical role of the diffusion latent space in producing stable and realistic motions.

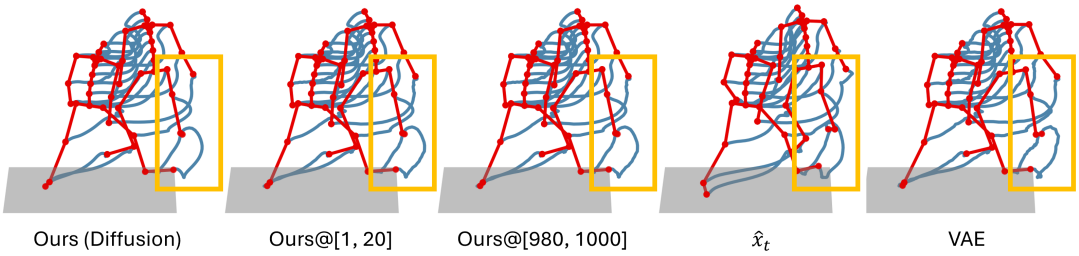

Figure 5.7: Qualitative comparison of ablation results. The yellow rectangles highlight key differences between variants, particularly in motion smoothness. Compared with *Ours*, *Ours[1,20]* and *Ours[980,1000]* exhibit jitters on the right hand. Compared with *Ours*, using $\hat{x}_t$ leads to larger trajectory changes of the right hand and the right foot. Compared with *Ours*, *VAE* leads to heavy jitters.

5.5 Summary

We propose a novel attack application that imperceptible adversarial motions are achieved with better post-attack motion quality. Additionally, we introduce a distribution-based adversarial attack method targeting skeleton-based human action recognition (S-HAR) systems by minimizing the optimization gap inherent in previous approaches. Our method integrates a generative diffusion model, wherein the posterior mean of single timestep denoising with regularization is constructed as

the proxy to fulfill our attack strategy. To faithfully assess the smoothness of adversarial motions, we develop a new metric aligned with human perception of smooth real-world human movements. We evaluate the quality of adversarial motions in terms of threatfulness, motion quality, and imperceptibility, demonstrating that our adversarial motions achieve superior performance across these metrics. The success of our proposed quality-advocating attack application and distribution-based attack method raises significant concerns regarding the robustness of action recognizers, highlighting the necessity for further enhancements in this area.

While achieving smooth adversarial motions, our approach presents opportunities for future research. As shown in our experiment, there exists a trade-off in motion quality with respect to the timesteps of the diffusion model. Future work can potentially investigate the influence of trading-off post-attack motion quality on different S-HAR systems. Our proposed smoothness focuses on the perspective of motion smoothness. Future work can potentially integrate the research field of action quality assessment [211, 212] and character animation [213] for more comprehensive motion quality measurements as well as constraints. Although this work aims to reduce the extent of quality degradation introduced by the proposed constraints, there exist an inherent trade-off between introducing adversarial perturbation and retaining motion quality.

CHAPTER 6

Diffusion Models for Multi-Character Interaction Generation

Portions of this chapter have been published in the peer-reviewed publication:

- **Ziyi Chang**, He Wang, George Koulieris, and Hubert P. H. Shum. 2025. Large-Scale Multi-Character Interaction Synthesis. In Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '25), August 10–14, 2025, Vancouver, BC, Canada. ACM, New York, NY, USA, 10 pages. `https://doi.org/10.1145/3721238.3730750`

The goal of multi-character interaction generation is to synthesize realistic interaction scenes that reproduce the real-world behaviours of interaction groups. Interactions among multiple characters exhibit not only interaction semantics (i.e., inter-class variations) but also coordination among agents (i.e., intra-class variations), which has been overlooked by previous works and hinders the realism of generated interactions. This chapter introduces the first framework for multi-character interaction generation that learns and exploits temporal intra-class variations in the form of transitions between interactions groups under the constraint of inter-class semantic features to achieve realistic multi-character interaction scenes.

6.1 Introduction

Generating large-scale multi-character interactions is challenging due to the need to model both interaction structure and temporal coordination. Existing work primarily focuses on modelling *inter-class interaction patterns*, i.e., different interactions [36,214]. While effective for synthesising isolated interactions, these approaches largely ignore *intra-class temporal variation*, namely how interactions evolve over time through partner changes and transitions.

Existing methods are limited in their ability to exploit intra-class temporal variation under the constraint of inter-class interaction features. In realistic scenarios such as social dancing, characters do not remain within fixed interaction pairs. Instead, they continuously transition between interaction partners based on spatial and temporal context [7]. These transitions represent a form of intra-class variation constrained by the underlying interaction structure. However, current interaction generation methods either restrict attention to two-character settings [36] or assume sparse [215] and passive [216] interactions in larger groups [217], thereby failing to model coordinated transitions in dense multi-character environments. Addressing this limitation requires not only generating plausible interactions, but also explicitly modelling how interactions transition and reconfigure over time in a scalable manner.

Large-scale multi-character interaction synthesis faces two challenges. The first challenge is the lack of data. Existing datasets for interactions [36] focus on two characters and do not consider coordinated interactions. Existing datasets for crowd simulation [217] do not contain dense and close interactions. Capturing such a dataset for our task would be time-consuming and labour-intensive, which becomes unmanageable as the number of characters scales up. The second challenge is to plan dense and close interactions based on spatial and temporal context for multiple characters. Scheduling suitable interactions for multiple characters is a highly correlated problem. In the temporal domain, previous coordination could heavily influence the interactions that follow, and in the spatial domain, the difficulty of planning increases with the increasing number of characters.

We propose a generative pipeline for large-scale multi-character interaction syn-

thesis that explicitly separates *interaction synthesis* from *transition planning*. Intra-class temporal variations are facilitated through a coordinatable interaction space that supports multiple characters and are modelled via a transition planning policy. In the absence of large-scale multi-character data, we decompose multi-character interactions into multiple two-character groups, each modelled using a pre-trained two-character interaction diffusion model. This decomposition allows the learned two-character interaction manifold to be generalised to multi-character settings without requiring additional data. Importantly, the grouping strategy is independent of the number of characters and therefore scales naturally to larger groups. To model temporal intra-class variation in coordinated interactions, we introduce a transition planning network that predicts high-level transition plans in the form of re-grouping decisions among characters. As the planning operates at an abstract level, it is motion-agnostic and transferable across different motion types. Training is conducted using reinforcement learning, where the coordinatable interaction space serves as the environment and the transition planner acts as the policy network. Transition smoothness and transition diversity are defined as reward signals to guide learning.

We train the proposed method using a two-character dancing subset from the InterHuman dataset [36], and evaluate its scalability by synthesising interactions involving a larger number of characters, as well as its transferability to other motion types. Transition smoothness and hip distance are used as evaluation metrics. Experimental results demonstrate that the proposed framework generalises effectively to large-scale multi-character settings and different motion domains.

The main contributions of this work are summarised as follows:

- We introduce a scalable framework for large-scale multi-character interaction synthesis by decomposing coordinated interactions into interaction synthesis and transition planning.
- We propose a data-efficient learning strategy that combines a pre-trained two-character diffusion model with a transition planning network, enabling the learning of coordinated multi-character interactions via deep reinforcement learning without requiring dedicated multi-character datasets.

6.2 Problem Formulation

Our objective is to synthesise large-scale multi-character interactions in the absence of dedicated training data. To this end, we decompose coordinated interactions into two complementary components: interaction synthesis for maintaining inter-class semantics and transition planning for modelling intra-class temporal variations. Interaction synthesis focuses on generating plausible interactions among multiple characters within a shared interaction space, whereas transition planning aims to coordinate interaction changes involving close and dense multi-character interactions.

We represent interactions using full-body poses and short motion clips. Full-body pose representations are adopted to capture social cues that are essential for realistic interaction modelling and transition planning [218]. To explicitly model temporal coordination, motions are represented as short clips containing several consecutive frames, which provide contextual information beyond single-frame poses and help infer character intent for upcoming transitions.

Specifically, we denote multi-character interactions as

$$M_{1:N}^{1:T} = \left[M_{1:N}^{1}, M_{1:N}^{2}, \cdots, M_{1:N}^{t}, \cdots, M_{1:N}^{T}\right], \tag{6.1}$$

where T is the total number of motion clips and N is the total number of characters. The t-th clip $M_{1:N}^{t} = [m_1^t, m_2^t, \cdots, m_n^t, \cdots, m_N^t]$ consists of motions for N characters where each clip contains w frames and m_n^t is the t-th clip of the n-th character. Following [36], each motion clip is represented using global joint positions, local joint rotations, and joint velocities.

6.3 Method Overview

Our pipeline is formulated as an autoregressive conditional generative model for synthesising coordinated interactions among multiple characters. The autoregressive design enables characters to plan future interaction transitions dynamically based on previously observed interactions, rather than committing to a fixed global plan in advance. As illustrated in Fig. 6.1, the proposed framework $\mathcal{F}$ consists of two tightly

coupled components: a coordinatable interaction space and a transition planning module. The interaction space is responsible for generating interactions among multiple characters, while the transition planner predicts how interaction groupings should evolve over time.

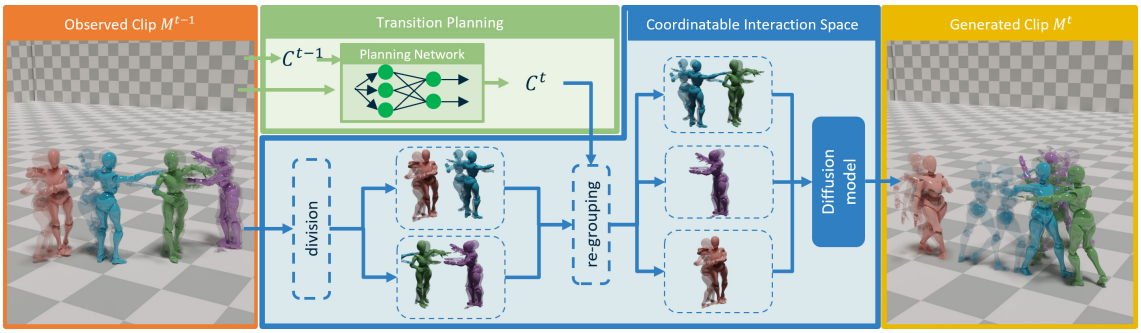

Figure 6.1: Framework overview. Our pipeline is an autoregressive conditional generative model to plan transitions and synthesize interactions for multiple characters. It has two components: The first component divides multiple characters into groups and leverages a pre-trained diffusion-based model to autoregressively generate interactions for each group. The second component predicts a transition plan based on the observed interactions and serves as the conditional signal for the interaction synthesis.

The coordinatable interaction space decomposes a multi-character interaction into multiple two-character groups. In the absence of large-scale multi-character data, this decomposition allows the interaction space to be constructed from learnable two-character interaction models. Specifically, we leverage a pre-trained two-character diffusion model to approximate the multi-character interaction manifold. Character groupings are determined by their indices, and historical motion clips are used as conditioning inputs. Interactions for each group are generated autoregressively. This formulation is independent of the total number of characters and therefore naturally scalable.

The transition planning module predicts a high-level transition plan C^t based on the observed interactions $M_{1:N}^{t-1}$. Transition plans are represented as re-grouping decisions over characters, which determine how characters are coordinated to interact in the subsequent motion clip. Specifically, our method follows an autoregressive

conditional generative formulation:

$$M_{1:N}^{t} = \mathcal{F}_{\theta}(M_{1:N}^{t-1}, \epsilon_{1:N}^{t}, C^{t}), \tag{6.2}$$

where $\mathcal{F}$ represents our autoregressive generative pipeline, $\epsilon_{1:N}^{t}$ is the sampled standard Gaussian noise for generation, $M_{1:N}^{t-1}$ is the last observed interaction clip for all characters $1, 2 \cdots, N$, $M_{1:N}^{t}$ is the next interaction clip for all characters, C^{t} is the transition plan for next motion clip $M_{1:N}^{t}$, and θ represents all trainable parameters.

6.4 Multi-Character Interaction Space

A coordinatable interaction space is essential for synthesising coherent multi-character interactions. Such a space capture realistic interaction manifolds and remain controllable for coordination. In the absence of large-scale data, we adopt a divide-and-conquer strategy to simplify the interaction space.

We approximate multi-character interactions by decomposing characters into two-character groups. In practice, complex interactions often consist of multiple smaller interaction groups [219], each of which can be modelled independently. In our implementation, we leverage a pre-trained two-character diffusion model to synthesise interactions for each group. This decomposition enables reuse of the expressive capacity of existing two-character models while allowing interaction groups to be dynamically reconfigured according to transition plans. Moreover, the grouping strategy is independent of the total number of characters, making the interaction space naturally scalable.

To generalise a two-character diffusion model to multi-character settings, we synthesise interaction groups autoregressively. Diffusion models are well suited for this purpose due to their high modelling capacity for complex interaction manifolds [27] and their ability to incorporate conditional signals without requiring explicit conditional training data [20]. When generating the next motion clip for a given two-character group, the diffusion process is conditioned on both the observed motion history of that group and the interactions already generated for other groups. As illustrated in Fig. 6.2, guidance from previously synthesised groups provides coordi-

nation cues that ensure consistency across the multi-character interaction space.

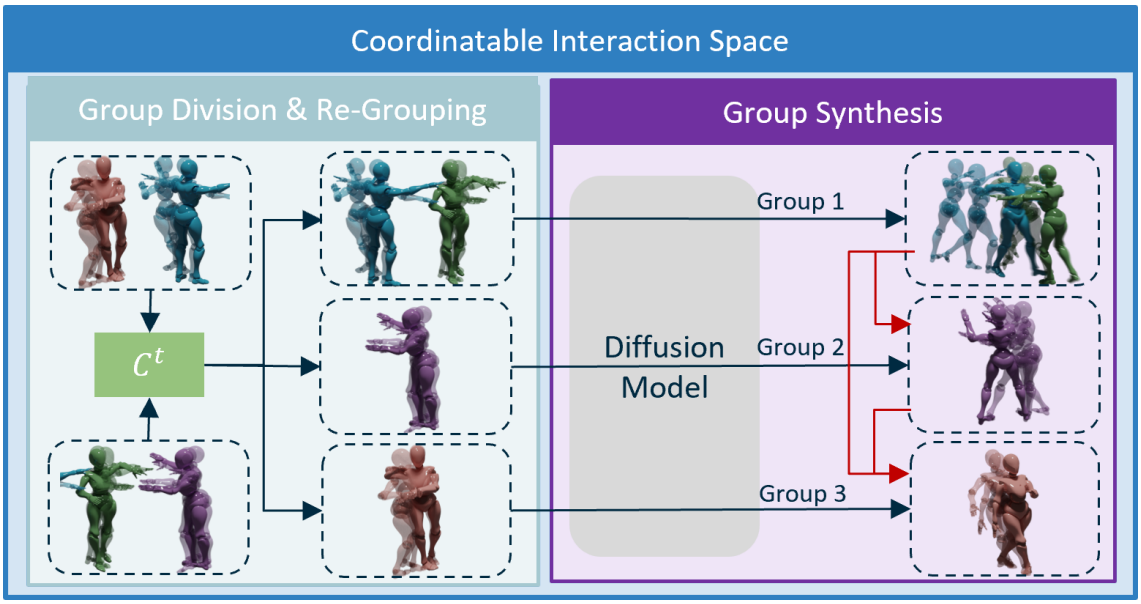

Figure 6.2: Coordinatable multi-character interaction space by group division. We divide multiple characters into groups and re-group them for potential coordination. The group synthesis generates new motions group by group. The newly generated group is conditioned on the already generated ones, which is indicated by red arrows.

Given the observed multi-character interaction clip $M_{1:N}^{t-1}$, we extend a state-of-the-art two-character interaction diffusion model [36] to generate the next interaction clip for all characters. The next clip is synthesised by autoregressively generating two-character social groups under the condition of previously generated groups. These groups collectively form the multi-character interaction clip $M_{1:N}^{t}$, which is defined as

$$M_{1:N}^{t} = \left[g(M_{i,j}^{t-1}, M') \mid (i,j) \in C^t \right], \tag{6.3}$$

where g denotes the two-character diffusion model, M' represents interaction groups already generated in the current autoregressive step, $M_{i,j}^{t-1}$ is the observed interaction clip for characters i and j, and C^t specifies the re-grouping transition plan.

To ensure coordination among interaction groups, we incorporate training-free guidance in the diffusion process to regulate spatial relationships between the newly generated group $M_{i,j}^{t}$ and previously generated groups M'. We define a distance-based constraint $d(\cdot)$ to maintain appropriate social spacing. Following Proxemics Theory [220], hip positions are used as the primary spatial cue, as they effectively

characterise interpersonal distance while avoiding high-dimensional joint constraints. The distance term is computed as

$$d(M_{i,j}^t, M') = \frac{1}{|M'|} \sum_{n'}^{|M'|} \min\left(\|p_{i,j} - p_{n'}\|_2^2 - \tau, 0\right), \tag{6.4}$$

where τ is a distance threshold and $p_{i,j}$ is the hip positions of characters i and j. In addition, temporal continuity between consecutive clips is enforced by constraining motion smoothness between the generated and observed clips:

$$d(M_{i,j}^t, M_{i,j}^{t-1}) = \sum \|acc_{i,j}\|_2^2, \tag{6.5}$$

where *acc* denotes joint-wise accelerations. The overall constraint is defined as

$$d = d(M_{i,j}^t, M') + d(M_{i,j}^t, M_{i,j}^{t-1}). \tag{6.6}$$

The complete procedure for constructing the coordinatable interaction space is summarised in Algorithm 2.

Algorithm 2 Coordinate-able Multi-Character Interaction Space

Require: Re-grouping choice (i, j) from a transition plan C^t, motion mask m for motion inpainting, observed motions $M_{1:N}^{t-1}$, other groups M', a pre-trained model $g(\cdot)$, distance function $d(\cdot)$

Ensure: a group $M_{i,j}^t$

$M_{i,j}^{t-1} \leftarrow M_{1:N}^{t-1}[(i,j)]$ ▷ Re-grouping by transition plan
$u \leftarrow U$ ▷ U is total number of diffusion timesteps
$\epsilon^t \leftarrow \mathcal{N}(0, I)$ ▷ Sample a random noise
$x^u \leftarrow \epsilon^t$
while $u \neq 0$ **do**
    $x^0 \leftarrow g(x^u, u)$ ▷ Diffusion predicts x_{start}
    $x^0 \leftarrow m \otimes M_{i,j}^{t-1} + (1 - m) \otimes x^0$ ▷ Masking for inpainting
    $x^0 \leftarrow x^0 + \nabla_x d$ ▷ Classifier guidance
    $u \leftarrow u - 1$
end while
$M_{i,j}^t \leftarrow x^0$
$M' \leftarrow M' \cup M_{i,j}^t$

6.5 Transition Planning

Beyond synthesising realistic interactions, multi-character coordination requires the ability to plan suitable interaction transitions over time. Transition planning operates at a high level by determining how characters should reconfigure their interaction groups based on the currently observed interactions. Once a transition plan is decided, the subsequent interactions are generated by the interaction space described previously.

We model transition planning through a conditional re-grouping mechanism integrated with the multi-character interaction space. While coordinating all characters jointly is theoretically possible, we restrict planning to local subsets of four characters to balance transition expressiveness and learning complexity. Considering a larger candidate set increases representational power but significantly enlarges the action space, whereas limiting choices to the nearest neighbour overly constrains transition diversity.

The planning signal is defined as high-level re-grouping decisions, represented by character indices. This abstraction avoids reasoning about low-level joint trajectories in a high-dimensional spatio-temporal space and makes the planning network motion-agnostic. Concretely, the planning network takes the motion clips of the selected characters as input and predicts a re-grouping configuration as the transition plan:

$$C^t = f_\theta(M^t_{i,j,i',j'}). \tag{6.7}$$

In the absence of ground-truth transition annotations, we formulate transition planning as a Markov Decision Process and train the planning network using deep reinforcement learning, as illustrated in Fig. 6.3. The MDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, E)$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ the action space, $\mathcal{R}$ a scalar reward function, and E the environment.

Specifically, a state $s \in \mathcal{S}$ is defined as the motion clips of the selected characters:

$$s := M_{i,j,i',j'}. \tag{6.8}$$

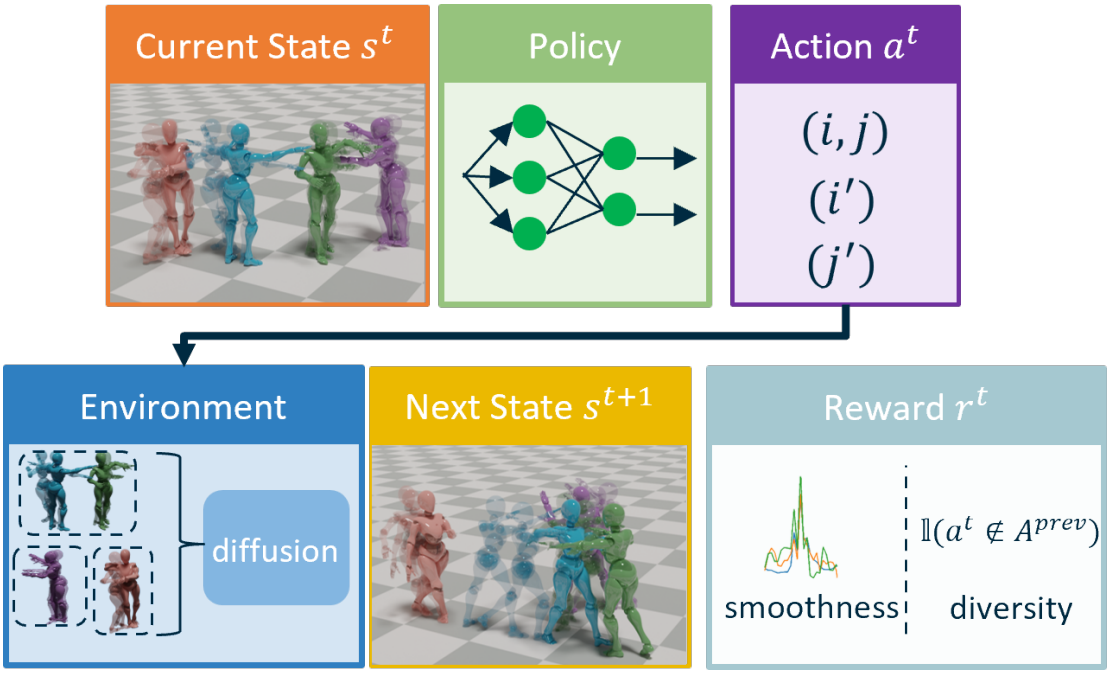

Figure 6.3: The planning network is learned as a policy network via deep reinforcement learning. The action is a transition plan that contains a high-level grouping choice.

An action $a \in \mathcal{A}$ corresponds to a transition plan,

$$a := C, \tag{6.9}$$

where C specifies the re-grouping decision among the four characters.

The previously introduced interaction space serves as the environment in our reinforcement learning formulation. The environment transition function $E : (s^t, a^t) \rightarrow s^{t+1}$ models the state evolution induced by executing action a^t in state s^t. Concretely, the interaction synthesis module defined in Section 6.4 acts as the environment model:

$$E := g(\cdot). \tag{6.10}$$

Given the current state, the next state is generated as

$$s^{t+1} = g(\epsilon^t, s^t), \tag{6.11}$$

where $\epsilon^t \sim \mathcal{N}(0, I)$. The initial state M^1 is generated by starting from an empty

state.

The transition planning network is formulated as the policy

$$\pi := f_\theta(\cdot), \tag{6.12}$$

which maps the current state to a transition action:

$$a^t = \pi(s^t). \tag{6.13}$$

We define a reward function $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ to evaluate transitions. The objective is to encourage smooth and diverse interaction transitions. Smoothness is quantified by

$$r_{\text{smooth}} = \exp\left(-\|acc^t - acc^{t+1}\|_2^2\right), \tag{6.14}$$

where acc denotes joint accelerations computed over a temporal window of ten frames. To promote exploration of different transition patterns, a diversity reward is defined as

$$r_{\text{div}} = \begin{cases} 1, & \text{if } a^t \text{ is novel,} \\ 0, & \text{otherwise.} \end{cases} \tag{6.15}$$

The final reward is given by

$$r = r_{\text{smooth}} + r_{\text{div}}. \tag{6.16}$$

6.6 Experiment Setup

We run a series of experiments to (i) investigate the effectiveness of our method by employing two metrics and (ii) validate the scalability and transferability of our method to three applications.

Quantitative metrics remain an open challenge. Essentially, without directly comparable ground truth, only self-contained metrics can be calculated. Therefore, we evaluate our method with transition smoothness (TS) [221] and hip distance (HD) for comparison. Following [221], transition smoothness calculates the change of acceleration, and we report its maximum change (peak jerk) as a metric. For

hip distance, we calculate the average distance between the hip of a character and those of other characters. This metric is designed to indicate whether characters overlapped in the same position. We report the average value over all characters and frames, which is calculated by

$$HD = \frac{2}{N(N-1)F} \sum_{f=1}^{F} \sum_{i,j \in N} \|h_i^f - h_j^f\|_2^2, \tag{6.17}$$

where N is the set of all characters, F is the number of frames, $\frac{2}{N(N-1)F}$ is an averaging term, and h_i^f and h_j^f are the hip positions for characters i and j at frame f. The TS metric has already integrated speed information by considering acceleration changes, which are crucial for physical plausibility and perceptual quality. Biomechanically, smooth acceleration results from the gradual modulation of neural signals regulating muscle force, aligning with Newton's second law. Additionally, following common practice, we measure variance as the diversity metric.

Hyper-parameters related to the reinforcement learning part have been shown in Table 6.1. DQN strategy is used to train the transition planning network.

Table 6.1: Hyperparameters of reinforcement learning strategy.

Hyperparameter	Value
batch size	16
gamma	0.98
epsilon start	0.08
epsilon end	0.01
epsilon decay	200
lr	0.0005

We employ a Deep Q-Network (DQN) to train the transition planning network, as the problem can be naturally formulated as a discrete sequential decision-making task. At each planning step, the agent selects a transition action that determines how motion segments are connected over time, with the objective of producing smooth and coherent transitions across generated sequences.

DQN is chosen due to its suitability for discrete action spaces and its ability to learn value functions that capture long-term rewards. In the context of transition planning, rewards are defined to encourage temporally smooth transitions and

penalise abrupt changes in motion dynamics. By optimising expected cumulative reward, the DQN-based planner learns policies that favour stable and consistent transition paths rather than locally optimal but perceptually jarring connections.

Alternative reinforcement learning algorithms, such as policy-gradient or actor–critic methods, were considered less suitable in this setting, as they typically require continuous action representations or introduce additional training complexity without clear benefit for the discrete transition selection problem addressed here. DQN therefore provides a balanced trade-off between modelling capacity, training stability, and implementation simplicity.

We balanced trade-off between transition representativeness and learning complexity. A character evaluating all others as potential transition partners improves representativeness but increases training complexity. Conversely, considering only the nearest character removes the need for learning but limits transition diversity.

To address the lack of ground truth data, we use re-grouping as high-level semantic control, avoiding computational burden of determining low-level joint movements in a high-dimensional spatiotemporal space. Currently, a 4-character group is formed by a greedy distance-based strategy, and future work could explore other advanced methods such as a first-person receptive field.

For large scenes, if total number of characters isn't divisible by four, we introduce an imaginary character in decision-making. If selecting it as a partner, a character behaves independently. Future improvements could include predefined actions like walking via prompt or recycling characters in [7] to enhance transitions further.

6.7 Comparison

As no existing method directly addresses large-scale multi-character interaction synthesis with transition planning, we adopt InterGen [36], a state-of-the-art interaction synthesis method, as our baseline. In addition, we implement an adapted variant, denoted as InterGen†, to enable a fair comparison. Specifically, InterGen† employs the same coordinatable interaction space as our method to support multiple characters. Since InterGen does not incorporate transition planning, transition signals are

randomly sampled and used as control inputs. All evaluation metrics are computed by averaging results over multiple generations. Quantitative results are reported in Table 6.2.

Table 6.2: Comparison with interaction synthesis models. † represents our implementation of the coordinatable interaction space in the original method. TS denotes transition smoothness and HD, the hip distance. Div denotes the diversity of the generated results.

Methods	TS ↓	HD	Div
InterGen	0.073	0.567	4.938
InterGen†	0.117	1.578	5.001
Ours	**0.071**	1.963	5.010

Our method achieves the highest transition smoothness (TS), indicating fewer visual artefacts during interaction transitions. Although InterGen attains a TS value close to ours, qualitative inspection reveals severe character overlap, leading to degraded motion quality. We attribute this behaviour to the fact that transitions in InterGen occur at very small inter-character distances, which artificially reduces the TS value. This observation is further supported by its significantly lower hip distance (HD) compared to our method, as illustrated in Fig. 6.4.

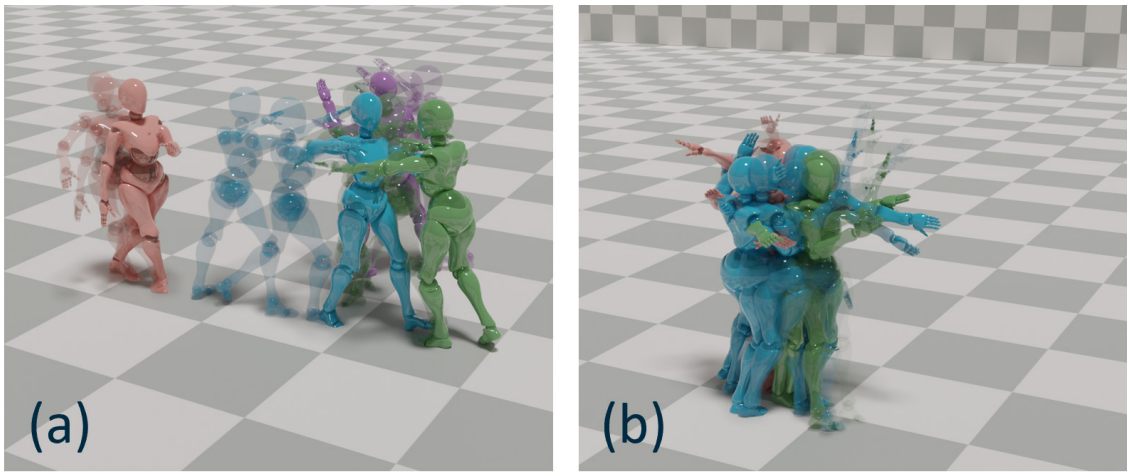

Figure 6.4: (a) An example result from our method. (b) An example from InterGen where characters heavily overlap.

Fig. 6.5 shows the density distributions of HD values for all methods. The bimodal distribution produced by our approach indicates well-separated characters

and effective transition planning. In contrast, InterGen† exhibits a unimodal distribution, reflecting the absence of explicit transition planning. InterGen demonstrates a similar distribution shape to InterGen†, but with a substantially smaller mode, confirming frequent character overlap.

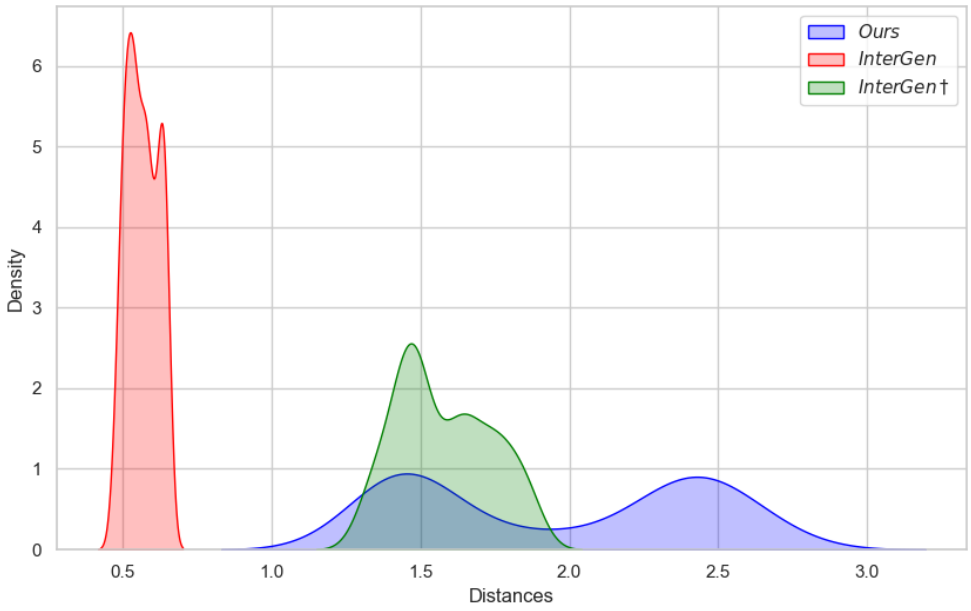

Figure 6.5: The density of hip distance for the three methods evaluated. The two modes in our hip distance density demonstrate minimal character overlap and clear transitions. InterGen† does not have the ability of transition planning, leading to an averaged distance density with a single mode. InterGen has a similar curve shape with InterGen† as both of them do not have transition planning. Its much smaller mode value indicates that characters heavily overlap.

Overall, the comparison demonstrates the effectiveness of both proposed components. Relative to InterGen†, our method benefits from explicit transition planning. Comparing InterGen and InterGen†, both of which lack planning, shows that the coordinatable interaction space alone already reduces character overlap. Together, these results validate the necessity of both the interaction space and the transition planning network in achieving high-quality multi-character coordinated interactions.

We provide an example as an illustration of the effects of the transition smoothness metric. Generally speaking, higher values indicate larger discontinuities like abrupt movements, which is indicated by the orange colour in Figure 6.6.

In terms of diversity, we attribute the slightly increased diversity to our coordinatable space and planning for synthesizing transitions that are absent in the

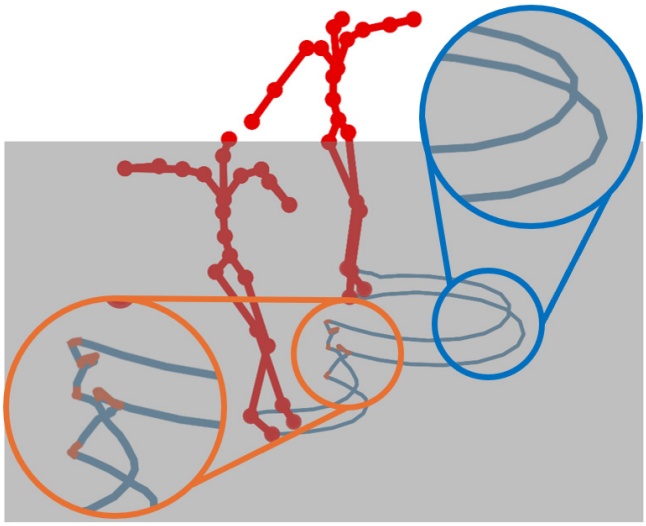

Figure 6.6: An illustrative figure for the effects of transition smoothness metric. Discontinuities in trajectories are highlighted in orange color.

original two-character data.

We further conduct a small user study where participants were asked to rank the quality of results generated by our method against InterGen and InterGen†. 94.12% prefers ours to the other two, 5.88% prefers InterGen†, and InterGen is unfavourable due to heavy character overlap.

6.8 Extended Applications

We showcase the scalability of the proposed framework and the transferability of the transition planning network through three extended applications. Although training relies only on a two-character dataset, the method scales naturally to larger numbers of characters, either by incrementally adding new characters or by generating large scenes in a single pass. Moreover, since the planning network operates on high-level re-grouping decisions, it is transferable across different motion types. We demonstrate this property by transferring a planner trained on dancing motions to boxing motions.

6.8.1 Adding New Character Synthesis

The proposed framework supports incremental character addition. In this setting, generation starts with four characters, after which additional characters are progressively introduced. Newly added characters are coordinated with existing ones through re-grouping decisions predicted by the planning network. Quantitative results for this setting are reported in Table 6.3.

6.8.2 Generating Large Scenes

We further demonstrate scalability by synthesising scenes containing a large number of characters simultaneously. Despite being trained only on two-character interactions, the framework is able to generate dense multi-character scenes in a single generation process, without incremental addition. Quantitative results are summarised in Table 6.3.

Table 6.3: Method performance on extended applications. TS denotes transition smoothness and HD, the hip distance.

Methods	TS	HD
Adding New Characters	0.026	2.450
Generating Large Scenes	0.075	3.372
Boxing	0.057	2.155

We clarify that our two-character interaction setup is a modelling choice driven by two factors: (1) Unlike prior work that loosely couples single-character motions, we focus on close and continuous interactions (CCI) with frequent, prolonged contact (e.g., dancing), which rarely involve more than two people. (2) Data availability is a major constraint. Capturing two-person CCI is already rare, and, to our knowledge, no datasets exist for multi-person CCI.

Furthermore, our model is not inherently limited to two-person interactions. To demonstrate this, we adapt our method using Multi-Track Timeline Control [222], allowing a character to be in multiple groups (e.g., A-B, A-C) with minimal modifications: diffusion model taking two interaction groups as input and further constraining distances among three individuals. Results are shown in Table 6.4:

Table 6.4: Allowing three-character division choice.

	TS	HD
Two-character division	0.071	1.963
Allowing three-character division	0.079	1.888

This confirms our method's high extensibility with minimal modifications, enabling diverse scenarios and demonstrating potential for broader multi-character interaction synthesis.

6.8.3 Other Interaction Semantics

Since the planning network predicts re-grouping choices independently of low-level motion dynamics, it can be transferred to other motion domains. We evaluate this transferability by applying a planner trained on dancing motions to boxing motions without additional adaptation. As shown in Table 6.3, transition smoothness remains stable, while hip distance increases slightly. This behaviour reflects the larger interpersonal spacing typically observed in boxing motions and confirms that the planning network generalises across motion types.

6.8.4 Ablation Studies

We conduct two ablation studies to analyse the contribution of each component that we propose and the distance threshold, which is a hyper-parameter in the scalable multi-character interaction generation.

Specifically, we first evaluate the effects of the coordinatable space and the transition planning module by progressively removing them from the full model. The results are reported in Table 6.5. Variant (a) removes both the coordinatable space and the planning module, resulting in limited transition smoothness (TS) and low hip distance (HD). This configuration represents a baseline without explicit mechanisms for coordination or transition modelling and highlights the difficulty of scaling interactions under such settings. Variant (b) introduces the coordinatable space while excluding the planning module. Compared to (a), this variant achieves a substantial increase in hip distance (HD), indicating that the coordinatable space plays

a key role in enabling suitable social distance across multiple characters. However, the transition smoothness (TS) degrades significantly, suggesting that coordination alone is insufficient to ensure smooth and coherent transitions over time.

Table 6.5: An ablation study on method scaling.

	Coordinatable Space	Planning	TS	HD
a	×	×	0.071	0.564
b	✓	×	0.202	3.422
Full	✓	✓	0.075	3.372

The full model combines both the coordinatable space and the planning module. This configuration achieves a strong balance between transition smoothness and hip distance, demonstrating that explicit transition planning is necessary to stabilise temporal evolution when scaling to multi-character scenarios. Together, these results confirm that both components are essential and complementary for achieving scalable and coherent multi-character motion generation.

We further conduct a second ablation study to examine the effect of the distance threshold used in interaction modelling. The distance threshold controls the spatial tolerance within which characters are considered to be suitable, thereby influencing both coordination flexibility and transition behaviour. The results are summarised in Table 6.6.

Table 6.6: An ablation study on the distance threshold.

Distance threshold	TS	HD
0.5	0.059	1.723
1.0	0.057	1.841
1.5	0.067	1.901
2.0	0.071	1.963
2.5	0.085	2.003
3.0	0.092	2.355

As the distance threshold increases from 0.5 to 3.0, hip distance (HD) consistently increases, indicating that larger thresholds allow greater spatial separation and more diverse interaction configurations. This reflects increased flexibility in group formation and reduced spatial constraints among characters.

However, this increased flexibility comes at the cost of transition smoothness (TS). While TS remains relatively stable for smaller thresholds, it gradually degrades as the threshold becomes larger. This suggests that excessively loose spatial constraints make it more difficult for the planning module to maintain smooth transitions, as characters may undergo larger positional adjustments between consecutive interaction states.

These results highlight an inherent trade-off between interaction diversity and transition smoothness. Moderate distance thresholds provide a balance between allowing sufficient spatial variation (HD) and preserving stable temporal transitions (TS), supporting the choice of the threshold used in the full model.

6.9 Summary

We present a framework for synthesising coordinated multi-character interactions without requiring any multi-character training data. The proposed method decomposes coordinated interactions into interaction synthesis and transition planning, realised by a coordinatable interaction space and a planning network, respectively. To address data scarcity, multi-character interactions are approximated through two-character groups using a pre-trained two-character interaction diffusion model, while the planning network provides high-level control via re-grouping decisions. Experimental results demonstrate that the proposed approach produces smoother transitions and reduces inter-character penetration compared with existing methods. We further show that the framework scales to larger numbers of characters and generalises across motion types.

Despite these advantages, the method has several limitations primarily stemming from the lack of multi-character datasets. First, the interaction space relies on a divide-and-approximate strategy that decomposes multi-character interactions into two-character groups. Second, the controllability of diffusion models remains limited in the absence of explicitly conditioned training data. Although classifier guidance and motion inpainting are employed, control accuracy could be further improved to enhance generation quality [223]. In addition, the overall performance depends on

the quality of the pre-trained two-character interaction model.

Future work includes collecting datasets containing dense, close-range interactions among multiple individuals, which would enable finer-grained interaction modelling and improved visual realism. For large-scale scenes, the current use of virtual characters to complete interaction groups could be replaced by predefined actions or character recycling strategies [7]. Moreover, while four-character groups are currently formed using a greedy distance-based heuristic, more principled grouping strategies, such as first-person receptive fields, offer promising directions for further improvement.

CHAPTER 7

Conclusions

In this thesis, we systematically explored the powerful role of diffusion models in human motion modelling in terms of inter-class variation modelling and intra-class variation modelling. This thesis explores diffusion models for three representative tasks on modelling the two variations: styled motion generation, adversarial motion generation, and multi-character interaction generation. Through the three tasks, this thesis demonstrates how diffusion models can drive realistic, diverse, and scalable capability that surpass previous methods.

Our findings explicitly demonstrate that diffusion models are critical and capable to accommodate the substantial diversity in human motions. Whether faced with combinations of content-style, action-execution, or interaction-coordination, diffusion models are consistently capable to cover the variations of human motions, facilitating a wide range of real-world applications. This thesis thus establishes diffusion models as a central strategy for advancing human motion modelling.

7.1 Achievement of Aims and Objectives

this thesis aims to systematically investigate how inter-class and intra-class variations in human motion can be modelled, disentangled, and exploited within a unified generative framework. This thesis demonstrate that diffusion models facilitate applications that require to model inter-class variations and intra-class variations by considering content-style, action-execution, and interaction-coordination. The progress made in each technical objective demonstrates how this approach meets various demands, as detailed below.

In Chapter 4, we showed the superiority of diffusion models by addressing the task of styled motion generation where inter-class and intra-class variations are directly modelled. We presented a single-stage diffusion-based pipeline to unify human motion synthesis and motion style transfer for high-quality motion creation. To effectively represent the coupled representation of both inter-class motion contents and intra-class motion styles in a common latent space, we proposed a denoising diffusion probabilistic model solution that has a large learning capacity for modelling the diverse data structure. To generate high-quality results, we proposed a multi-task network architecture that leverages both local guidance, including joint angles, movement trajectories and supporting foot patterns, and global guidance, including physical and adversarial regulations. These advancements demonstrated the utility of diffusion models for modelling human motions to cover both inter- and intra-class variations, establishing the feasibility for future styled animation systems.

In Chapter 5, we showed the superiority of diffusion models by addressing the task of adversarial motion generation where intra-class variations are exploited as a degree of freedom without altering inter-class semantics. We discover a critical yet previously overlooked vulnerability in existing attack methods where the intra-class variations are usually modelled by noise-like perturbations that are inherently perceptible to humans. To exploit intra-class variations, we proposed a diffusion-based attack method where imperceptible adversarial motions are generated through a pre-trained diffusion model. To faithfully quantify the quality of adversarial motions, we also proposed a new metric based on existing physiological analysis of real-world human motions. These advancements demonstrated the utility of diffusion models

for exploiting intra-class variations under the constraints of inter-class features to generate adversarial motions, establishing the potential for future diffusion-based adversarial attack systems.

Finally, Chapter 6 further exploits the intra-class variations in the context of interactions. We showed the superiority of diffusion models by supporting a transition planning network for modelling temporal intra-class variations while maintaining inter-class interaction semantics. We proposed a framework to synthesize large-scale multiple characters by decomposing their coordinated interactions into interaction synthesis and transition planning as inter-class and intra-class variations, respectively. Building on the decomposition, we proposed a method of combining a pre-trained two-character diffusion model and a transition planning network to learn the coordinated interactions via deep reinforcement learning without the requirement for data. These advancements demonstrate the exploitation of intra-class temporal variations facilitated by diffusion models establishes the feasibility for future multi-character interaction synthesis systems.

Furthermore, the integration of diffusion models into human motion modelling in this thesis demonstrates clear relevance to several industry domains where controllable, scalable, and reliable motion generation is required.

- **Animation and Game Production Pipelines:** In animation and game development, character motion is a core asset that is traditionally expensive to author and difficult to scale. The styled motion generation framework developed in this thesis enables end-to-end synthesis of motions with controllable execution styles, supporting rapid prototyping and content variation. Such capabilities can assist animators during pre-visualisation, reduce reliance on large motion capture libraries, and facilitate scalable character behaviour generation in interactive environments.

- **Robustness Evaluation in Safety-Critical AI Systems:** Skeleton-based motion analysis is increasingly used in applications such as surveillance, sports analytics, healthcare monitoring, and human–robot interaction. The adversarial motion generation methods proposed in this thesis provide a principled

way to assess the robustness of motion-based recognition and decision systems under subtle but structured perturbations. This is particularly relevant for responsible AI deployment, where understanding potential failure modes prior to real-world operation is critical.

- **Multi-Character Simulation and Virtual Human Platforms:** Large-scale simulations involving multiple interacting characters are central to crowd simulation, virtual training environments, and digital human platforms. The scalable multi-character interaction modelling framework presented in this thesis addresses the challenge of generating coordinated group behaviour with smooth transitions over time. By emphasising transition strength and structured interaction modelling, the proposed approach aligns with the needs of industry systems that require coherent and extensible multi-agent motion generation.

Taken together, the contributions of this thesis situate diffusion-based human motion modelling as a practical and versatile tool across multiple industries. By addressing stylisation, robustness, and scalable interaction within a unified generative framework, this work bridges methodological advances with concrete industry requirements in animation, intelligent systems, and virtual human technologies.

7.2 Future Research Directions

This thesis demonstrates that diffusion-based models provide a powerful framework for human motion modelling, particularly in terms of modelling inter-class and intra-class variations. Building on these contributions, future research should move towards a unified goal: developing motion generation systems that are *scalable*, *efficient*, *robust*, and *aligned* with human perception preference. The following research directions outline a coherent roadmap towards this objective, each addressing a critical limitation identified in the current work.

7.2.1 Learning-based Multi-Character Interactions

Human motion is inherently interactive, yet current generative models remain largely limited to single-person scenarios, with two-character interactions only recently receiving focused attention. Although several studies have begun to explore multi-character settings [130, 131, 224], most approaches primarily extend spatial layouts to include additional characters, without fully addressing the temporal coordination that emerges uniquely in polyadic interactions.

This thesis highlights that realistic multi-character behaviour is largely determined by transition dynamics and coordination over time. However, existing methods often rely on heuristic or rule-based constraints when extending from dyadic to multi-character interactions, which struggle to generalise to complex and highly non-linear group behaviours [224]. A key challenge lies in learning such coordination patterns directly from limited interaction data.

Future research could therefore focus on learning-based formulations that reduce reliance on hand-crafted constraints. In particular, few-shot or meta-learning strategies [225] offer a promising pathway for enabling diffusion models to adapt to new interaction configurations with minimal additional data. This direction would allow models to progressively scale from controlled pairwise interactions to more realistic, polyadic social scenarios.

7.2.2 Adversarially Robust Human Motion Modelling

While diffusion models have demonstrated strong performance in terms of motion diversity and quality, their robustness to adversarial perturbations remains an open challenge. Existing research on adversarial attacks in human motion has largely focused on skeleton-based action recognition [114], leaving the vulnerability of other motion-based tasks largely unexplored.

Many practical applications, such as reactive motion synthesis [137] in human–robot collaboration, rely on motion generation systems that directly influence physical actions. In such settings, adversarially perturbed motions could induce unsafe or even dangerous system responses. This highlights the need to study adversarial robustness beyond recognition tasks and towards generative and reactive motion pipelines.

Future work should therefore investigate whether adversarial attacks can be ex-

tended to a broader range of motion-based tasks such as motion prediction and whether common structural vulnerabilities exist across these tasks. Ultimately, this line of research may lead to unified adversarial and defence frameworks that improve the reliability and safety of motion-driven systems in real-world deployments.

7.2.3 Data-Efficient Motion Diffusion Models with Prior Domain Knowledge

Despite their expressive power, diffusion models typically require large-scale training data. Unlike images or videos, high-quality human motion data are expensive to capture and difficult to scale, resulting in a persistent gap between academic benchmarks and real-world requirements. Although existing datasets range from large-scale single-person motion corpora (e.g., HumanML3D [35]) to two-person interaction datasets [36, 88], data scarcity remains a fundamental bottleneck due to the difficulty of obtaining motions.

A promising research direction is to improve data efficiency by integrating domain knowledge into diffusion models. Psychological models, such as OCEAN [226] and PAD [227], can provide structured priors over affect and style, reducing the burden on data-driven learning. Similarly, simplified physical models, such as the inverted pendulum [228], can act as inductive biases to capture fundamental aspects of human balance and response to perturbations [216].

Future research should explore principled ways to incorporate such priors into diffusion architectures, enabling models that generalise better under limited data while maintaining plausibility and generalization ability.

7.2.4 Human Perception-Aligned Evaluation Metrics

The evaluation of human motion models remains a significant challenge. Many commonly used metrics are adapted from computer vision and emphasise statistical or feature-level similarity [35, 36, 229], which may not fully capture perceptual or interaction quality. This limitation becomes particularly pronounced in multi-character settings, where interaction quality depends on implicit and relational con-

straints [230].

User studies are often employed to address this gap, but they introduce practical challenges, including high cost, limited scalability, and potential participant bias. To move beyond these limitations, future research should aim to develop perception-aligned evaluation metrics that approximate human judgments while remaining computationally tractable.

One promising direction is to model human preferences directly using learned reward models [231–233] or using multi-modal data like EEG signals [234] that are potentially informed by insights from psychology and cognitive science. Such potentional solutions could provide scalable and consistent evaluation tools, supporting both model development and real-world deployment.

Bibliography

[1] W. Zhu, X. Ma, D. Ro, H. Ci, J. Zhang, J. Shi, F. Gao, Q. Tian, and Y. Wang, "Human motion generation: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2430–2449, 2023. 1, 1.1.2, 2.1.1, 5.4.1

[2] R. Zhao, H. Su, and Q. Ji, "Bayesian adversarial human motion synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6225–6234, 2020. 1, 1.1.3, 2.2

[3] M. S. Islam, K. Bakhat, M. Iqbal, R. Khan, Z. Ye, and M. M. Islam, "Representation for action recognition with motion vector termed as: Sdqio," *Expert Systems with Applications*, vol. 212, p. 118406, 2023. 1

[4] K. Chen, J. Zhang, M. Li, Z. Zheng, and H. Fan, "Clusterstyle: Modeling intra-style diversity with prototypical clustering for stylized motion generation," *arXiv preprint arXiv:2512.02453*, 2025. 1

[5] H. Fei and I. Reid, "Dynamic classifier for non-rigid human motion analysis.," in *BMVC*, pp. 1–10, 2004. 1

[6] H. Yu, J. Liu, X. Gui, M. Wong, Y. Hou, and Y.-S. Ong, "A plug-and-play multi-criteria guidance for diverse in-betweening human motion generation," *arXiv preprint arXiv:2508.01590*, 2025. 1

[7] H. P. Shum, T. Komura, M. Shiraishi, and S. Yamazaki, "Interaction patches for multi-character animation," *ACM transactions on graphics (TOG)*, vol. 27, no. 5, pp. 1–8, 2008. 1, 2.1.2, 6.1, 6.6, 6.9

[8] J. Won, K. Lee, C. O'Sullivan, J. K. Hodgins, and J. Lee, "Generating and ranking diverse multi-character interactions," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 6, pp. 1–12, 2014. 1, 2.2.2

[9] D. Holden, J. Saito, and T. Komura, "A deep learning framework for character motion synthesis and editing," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–11, 2016. 1, 1.1.3, 2.1.1, 2.2.1

[10] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013. 1, 1.1.2, 1.1.3

[11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014. 1, 1.1.2, 1.1.3

[12] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, "Human motion diffusion model," *arXiv preprint arXiv:2209.14916*, 2022. 1, 2.1.1

[13] S. Starke, P. Starke, N. He, T. Komura, and Y. Ye, "Categorical codebook matching for embodied character controllers," *ACM Trans. Graph.*, vol. 43, July 2024. 1, 1.1.3

[14] L. Mourot, L. Hoyet, F. Le Clerc, F. Schnitzler, and P. Hellier, "A survey on deep learning for skeleton-based human animation," in *Computer Graphics Forum*, vol. 41, pp. 122–157, Wiley Online Library, 2022. 1, 1.1.3, 2.2, 4.1

[15] K. Sui, A. Ghosh, I. Hwang, B. Zhou, J. Wang, and C. Guo, "A survey on human interaction motion generation," *arXiv preprint arXiv:2503.12763*, 2025. 1, 2.1.2, 2.1.2

[16] S. Dasari, O. Mees, S. Zhao, M. K. Srirama, and S. Levine, "The ingredients for robotic diffusion transformers," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 15617–15625, IEEE, 2025. 1

[17] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, 2024. 1, 4.1, 4.3

[18] P. Li, Z. Li, H. Zhang, and J. Bian, "On the generalization properties of diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 2097–2127, 2023. 1

[19] X. Li, Y. Dai, and Q. Qu, "Understanding generalizability of diffusion models requires rethinking the hidden gaussian structure," *Advances in neural information processing systems*, vol. 37, pp. 57499–57538, 2024. 1

[20] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021. 1, 1.1.3, 2.1.1, 6.4

[21] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022. 1

[22] Y. Wu, W. Ji, K. Zheng, Z. Wang, and D. Xu, "Mote: Learning motion-text diffusion model for multiple generation tasks," *arXiv preprint arXiv:2411.19786*, 2024. 1

[23] F. Daneshfar, A. Bartani, and P. Lotfi, "Image captioning by diffusion models: A survey," *Engineering Applications of Artificial Intelligence*, vol. 138, p. 109288, 2024. 1

[24] Y. Xie, V. Jampani, L. Zhong, D. Sun, and H. Jiang, "Omnicontrol: Control any joint at any time for human motion generation," in *The Twelfth International Conference on Learning Representations*, 2024. 1

[25] K. Karunratanakul, K. Preechakul, E. Aksan, T. Beeler, S. Suwajanakorn, and S. Tang, "Optimizing diffusion noise can serve as universal motion priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1334–1345, 2024. 1, 5.1, 5.2.1, 5.3

[26] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Comput. Surv.*, vol. 56, Nov. 2023. 1

[27] Z. Chang, G. A. Koulieris, H. J. Chang, and H. P. Shum, "On the design fundamentals of diffusion models: A survey," *Pattern Recognition*, p. 111934, 2025. 1, 1.1.3, 6.4

[28] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211, 2019. 1.1.1

[29] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018. 1.1.1

[30] A. Thatipelli, S. Narayan, S. Khan, R. M. Anwer, F. S. Khan, and B. Ghanem, "Spatio-temporal relation modeling for few-shot action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19958–19967, 2022. 1.1.1

[31] C. Xu, L. N. Govindarajan, Y. Zhang, and L. Cheng, "Lie-x: Depth image based articulated object pose estimation, tracking, and action recognition on lie groups," *International Journal of Computer Vision*, vol. 123, no. 3, pp. 454–478, 2017. 1.1.1

[32] S. Baek, Z. Shi, M. Kawade, and T.-K. Kim, "Kinematic-layout-aware random forests for depth-based action recognition," *arXiv preprint arXiv:1607.06972*, 2016. 1.1.1

[33] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010–1019, 2016. 1.1.1, 2, 5.1

[34] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019. 1.1.1, 2.1, 2, 2, 2.2, 5.1

[35] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3d human motions from text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5152–5161, 2022. 1.1.1, 1.1.2, 2.1, 2, 2.3, 2.4.1, 2.4.3, 5.4.1, 7.2.3, 7.2.4

[36] H. Liang, W. Zhang, W. Li, J. Yu, and L. Xu, "Intergen: Diffusion-based multi-human motion generation under complex interactions," *International Journal of Computer Vision*, pp. 1–21, 2024. 1.1.1, 2.1.2, 2.2.2, 2.2, 2, 2.4.1, 6.1, 6.2, 6.4, 6.7, 7.2.3, 7.2.4

[37] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception & psychophysics*, vol. 14, no. 2, pp. 201–211, 1973. 1.1.1

[38] C. Liu, M. Zhao, B. Ren, M. Liu, N. Sebe, *et al.*, "Spatio-temporal graph diffusion for text-driven human motion generation.," in *BMVC*, pp. 722–729, 2023. 1.1.1

[39] B. Ren, M. Liu, R. Ding, and H. Liu, "A survey on 3d skeleton-based action recognition using learning method," *Cyborg and Bionic Systems*, vol. 5, p. 0100, 2024. 1.1.1, 1.1.2, 5.1

[40] C. Wang and J. Yan, "A comprehensive survey of rgb-based and skeleton-based human action recognition," *IEEE Access*, vol. 11, pp. 53880–53898, 2023. 1.1.1, 1.1.2

[41] B. Hommel, "Toward an action-concept model of stimulus-response compatibility," *Advances in psychology*, 1997. 1.1.2

[42] S.-J. Blakemore and J. Decety, "From the perception of action to the understanding of intention," *Nature reviews neuroscience*, vol. 2, no. 8, pp. 561–567, 2001. 1.1.2

[43] J. Kim, H. Oh, S. Kim, H. Tong, and S. Lee, "A brand new dance partner: Music-conditioned pluralistic dancing controlled by multiple dance genres," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3490–3500, 2022. 1.1.2

[44] S. Alexanderson, R. Nagy, J. Beskow, and G. E. Henter, "Listen, denoise, action! audio-driven motion synthesis with diffusion models," *ACM Trans. Graphics*, vol. 42, no. 4, pp. 1–20, 2023. 1.1.2

[45] T. Ao, Z. Zhang, and L. Liu, "Gesturediffuclip: Gesture diffusion model with clip latents," *ACM Trans. Graphics*, 2023. 1.1.2

[46] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, and L. Cheng, "Action2motion: Conditioned generation of 3d human motions," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2021–2029, 2020. 1.1.2, 1.1.3, 2.1.1, 2, 2.3, 2.4.2, 2.4.2, 2.4.3, 4.1

[47] J. Gao, J. Pu, H. Zhang, Y. Shan, and W.-S. Zheng, "Pc-dance: Posture-controllable music-driven dance synthesis," in *Proceedings of the 30th ACM international conference on multimedia*, pp. 1261–1269, 2022. 1.1.2

[48] Y. Nishimura, Y. Nakamura, and H. Ishiguro, "Long-term motion generation for interactive humanoid robots using gan with convolutional network," in *Proc. ACM/IEEE Int. Conf. on Human-Robot Interaction*, pp. 375–377, 2020. 1.1.2

[49] G. Gulletta, W. Erlhagen, and E. Bicho, "Human-like arm motion generation: A review," *Robotics*, vol. 9, no. 4, p. 102, 2020. 1.1.2

[50] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström, "Analyzing input and output representations for speech-driven gesture generation," in *Proc. Int. Conf. on Intelligent Virtual Agents*, p. 97–104, 2019. 1.1.2

[51] T. Yin, L. Hoyet, M. Christie, M.-P. Cani, and J. Pettré, "The one-man-crowd: Single user generation of crowd motions using virtual reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 5, pp. 2245–2255, 2022. 1.1.2

[52] D. Holden, T. Komura, and J. Saito, "Phase-functioned neural networks for character control," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–13, 2017. 1.1.3, 2.1.1, 4.1

[53] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020. 1.1.3, 4.1, 4.4, 4.4, 4.6, 4.7, 4.2, 4.7

[54] C. Luo, "Understanding diffusion models: A unified perspective," *arXiv preprint arXiv:2208.11970*, 2022. 1.1.3

[55] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021. 1.1.3, 5.2.1

[56] Z. Ma, Y. Zhang, G. Jia, L. Zhao, Y. Ma, M. Ma, G. Liu, K. Zhang, N. Ding, J. Li, *et al.*, "Efficient diffusion models: A comprehensive survey from principles to practices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1.1.3

[57] H. Shen, J. Zhang, B. Xiong, R. Hu, S. Chen, Z. Wan, X. Wang, Y. Zhang, Z. Gong, G. Bao, *et al.*, "Efficient diffusion models: A survey," *Transactions on Machine Learning Research (TMLR)*, 2025. 1.1.3

[58] I. Yacoub and R. Dina, "Unlocking the potential of diffusion models through efficiency," 2025. 1.1.3

[59] L. Kovar, M. Gleicher, and F. Pighin, "Motion graphs," *ACM Trans. Graph.*, vol. 21, p. 473–482, July 2002. 2.1.1

[60] O. Arikan and D. A. Forsyth, "Interactive motion generation from examples," *ACM Trans. Graphics*, vol. 21, no. 3, pp. 483–490, 2002. 2.1.1

[61] J. Lee, J. Chai, P. S. Reitsma, J. K. Hodgins, and N. S. Pollard, "Interactive control of avatars animated with human motion data," in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pp. 491–500, 2002. 2.1.1

[62] P. S. Reitsma and N. S. Pollard, "Evaluating motion graphs for character animation," *ACM Transactions on Graphics (TOG)*, vol. 26, no. 4, pp. 18–es, 2007. 2.1.1

[63] M. Büttner and S. Clavet, "Motion matching and the road to next-gen animation." 2.1.1

[64] J.-R. Chen and A. Steed, "Planning plausible human animation with environment-aware motion sampling," in *International Conference on Motion in Games*, pp. 51–62, Springer, 2011. 2.1.1

[65] A. Safonova and J. K. Hodgins, "Construction and optimal search of interpolated motion graphs," *ACM Trans. Graph.*, vol. 26, p. 106–es, July 2007. 2.1.1

[66] K. Zadziuk, "Motion matching, the future of games animation...today." 2.1.1

[67] D. Holden, O. Kanoun, M. Perepichka, and T. Popa, "Learned motion matching," *ACM Transactions on Graphics (ToG)*, vol. 39, no. 4, pp. 53–1, 2020. 2.1.1

[68] G. W. Taylor, G. E. Hinton, and S. Roweis, "Modeling human motion using binary latent variables," *Advances in neural information processing systems*, vol. 19, 2006. 2.1.1

[69] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1110–1118, 2015. 2.1.1

[70] T. Komura, I. Habibie, D. Holden, J. Schwarz, and J. Yearsley, "A recurrent variational autoencoder for human motion synthesis," in *The 28th British Machine Vision Conference*, 2017. 2.1.1

[71] D. Holden, J. Saito, T. Komura, and T. Joyce, "Learning motion manifolds with convolutional autoencoders," in *SIGGRAPH Asia 2015 technical briefs*, pp. 1–4, 2015. 2.1.1

[72] M. Petrovich, M. J. Black, and G. Varol, "Action-conditioned 3d human motion synthesis with transformer vae," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10985–10995, 2021. 2.1.1, 2.2.1, 2.3, 2.4.3, 5.2.1, 5.3, 5.4.3

[73] H.-Y. Lee, X. Yang, M.-Y. Liu, T.-C. Wang, Y.-D. Lu, M.-H. Yang, and J. Kautz, "Dancing to music," *Advances in neural information processing systems*, vol. 32, 2019. 2.1.1

[74] B. Li, Y. Zhao, S. Zhelun, and L. Sheng, "Danceformer: Music conditioned 3d dance generation with parametric motion transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 1272–1279, 2022. 2.1.1

[75] L. Xu, Z. Song, D. Wang, J. Su, Z. Fang, C. Ding, W. Gan, Y. Yan, X. Jin, X. Yang, *et al.*, "Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2228–2238, 2023. 2.1.1, 2.2, 2, 2.4.1

[76] P. Li, K. Aberman, Z. Zhang, R. Hanocka, and O. Sorkine-Hornung, "Ganimator: Neural motion synthesis from a single sequence," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, p. 138, 2022. 2.1.1

[77] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, and G. Yu, "Executing your commands via motion diffusion in latent space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18000–18010, 2023. 2.1.1

[78] K. H. Lee, M. G. Choi, and J. Lee, "Motion patches: building blocks for virtual environments annotated with motion data," in *ACM SIGGRAPH 2006 Papers*, pp. 898–906, 2006. 2.1.2

[79] K. Chen, Z. Tan, J. Lei, S.-H. Zhang, Y.-C. Guo, W. Zhang, and S.-M. Hu, "Choreomaster: Choreography-oriented music-driven dance synthesis," *ACM Trans. Graphics*, vol. 40, jul 2021. 2.1.2

[80] H. P. Shum, T. Komura, and S. Yamazaki, "Simulating multiple character interactions with collaborative and adversarial goals," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 5, pp. 741–752, 2010. 2.1.2

[81] X. Zhang, Z. Chang, Q. Men, and H. P. Shum, "Motion in-betweening for densely interacting characters," in *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025. 2.1.2

[82] J. N. Kundu, H. Buckchash, P. Mandikal, R. M. V, A. Jamkhandi, and V. B. RADHAKRISHNAN, "Cross-conditioned recurrent networks for long-term synthesis of inter-person human motion interactions," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 2.1.2

[83] V. Adeli, E. Adeli, I. Reid, J. C. Niebles, and H. Rezatofighi, "Socially and contextually aware human motion and pose forecasting," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6033–6040, 2020. 2.1.2

[84] S. Maheshwari, D. Gupta, and R. K. Sarvadevabhatla, "Mugl: Large scale multi person conditional action generation with locomotion," 2021. 2.1.2

[85] D. Gupta, S. Maheshwari, S. S. Kalakonda, M. Vaidyula, and R. K. Sarvadevabhatla, "Dsag: A scalable deep framework for action-conditioned multi-actor full body motion synthesis," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4300–4308, 2023. 2.1.2, 2.2, 2.4.1

[86] Q. Men, H. P. H. Shum, E. S. L. Ho, and H. Leung, "Gan-based reactive motion synthesis with class-aware discriminators for human-human interaction," 2021. 2.1.2

[87] A. Goel, Q. Men, and E. S. L. Ho, "Interaction mix and match: Synthesizing close interaction using conditional hierarchical gan with multi-hot class embedding," 2022. 2.1.2

[88] L. Xu, X. Lv, Y. Yan, X. Jin, S. Wu, C. Xu, Y. Liu, Y. Zhou, F. Rao, X. Sheng, *et al.*, "Inter-x: Towards versatile human-human interaction analysis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22260–22271, 2024. 2.1.2, 2.2.2, 2.2, 2, 7.2.3

[89] Y. Shafir, G. Tevet, R. Kapon, and A. H. Bermano, "Human motion diffusion as a generative prior," in *The Twelfth International Conference on Learning Representations*, 2024. 2.1.2, 2.4.1

[90] R. McDonnell, S. Jorg, J. McHugh, F. Newell, and C. O'Sullivan, "Evaluating the emotional content of human motions on real and virtual characters," in *Proceedings of the 5th symposium on Applied perception in graphics and visualization*, pp. 67–74, 2008. 2.2.1

[91] S. M. A. Akber, S. N. Kazmi, S. M. Mohsin, and A. Szczesna, "Deep learning-based motion style transfer tools, techniques and future challenges," *Sensors*, vol. 23, no. 5, p. 2597, 2023. 2.2.1

[92] chuan guo, Y. Mu, X. Zuo, P. Dai, Y. Yan, J. Lu, and L. Cheng, "Generative human motion stylization in latent space," in *The Twelfth International Conference on Learning Representations*, 2024. 2.2.1, 2.4.3

[93] A. Aristidou, Q. Zeng, E. Stavrakis, K. Yin, D. Cohen-Or, Y. Chrysanthou, and B. Chen, "Emotion control of unstructured dance movements," in *Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation*, pp. 1–10, 2017. 2.2.1

[94] E. Hsu, K. Pulli, and J. Popović, "Style translation for human motion," in *ACM SIGGRAPH 2005 Papers*, pp. 1082–1089, 2005. 2.2.1

[95] S. Xia, C. Wang, J. Chai, and J. Hodgins, "Realtime style transfer for unlabeled heterogeneous human motion," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, pp. 1–10, 2015. 2.2.1, 2.1, 2, 4.1, 4.6, 4.6, 4.7

[96] M. Brand and A. Hertzmann, "Style machines," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 183–192, 2000. 2.2.1

[97] M. Unuma, K. Anjyo, and R. Takeuchi, "Fourier principles for emotion-based human figure animation," in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pp. 91–96, 1995. 2.2.1

[98] M. E. Yumer and N. J. Mitra, "Spectral style transfer for human motion between independent actions," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–8, 2016. 2.2.1

[99] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423, 2016. 2.2.1

[100] D. Holden, I. Habibie, I. Kusajima, and T. Komura, "Fast neural style transfer for motion data," *IEEE computer graphics and applications*, vol. 37, no. 4, pp. 42–49, 2017. 2.2.1

[101] H. Du, E. Herrmann, J. Sprenger, K. Fischer, and P. Slusallek, "Stylistic locomotion modeling and synthesis using variational generative models," in *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pp. 1–10, 2019. 2.2.1

[102] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017. 2.2.1, 2.4.3

[103] K. Aberman, Y. Weng, D. Lischinski, D. Cohen-Or, and B. Chen, "Unpaired motion style transfer from video to animation," *ACM Transactions On Graphics (TOG)*, vol. 39, no. 4, pp. 64–1, 2020. 2.2.1, 4, 4.1, 4.3, 4.3, 4.4

[104] D.-K. Jang, S. Park, and S.-H. Lee, "Motion puzzle: Arbitrary motion style transfer by body part," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 3, pp. 1–16, 2022. 2.2.1, 4

[105] B. Kim, J. Kim, H. J. Chang, and J. Y. Choi, "Most: Motion style transformer between diverse action contents," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1705–1714, 2024. 2.2.1, 2.4.3, 4, 4.1, 4.3

[106] W. Song, X. Jin, S. Li, C. Chen, A. Hao, and X. Hou, "Finestyle: Semantic-aware fine-grained motion style transfer with dual interactive-flow fusion," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 11, pp. 4361–4371, 2023. 2.2.1

[107] T. Tao, X. Zhan, Z. Chen, and M. van de Panne, "Style-erd: Responsive and coherent online motion style transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6593–6603, 2022. 2.2.1

[108] D.-K. Jang, Y. Ye, J. Won, and S.-H. Lee, "Mocha: Real-time motion characterization via context matching," in *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023. 2.2.1

[109] L. Zhong, Y. Xie, V. Jampani, D. Sun, and H. Jiang, "Smoodi: Stylized motion diffusion model," in *European Conference on Computer Vision*, pp. 405–421, Springer, 2024. 2.2.1

[110] W. Song, X. Jin, S. Li, C. Chen, A. Hao, X. Hou, N. Li, and H. Qin, "Arbitrary motion style transfer with multi-condition motion latent diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 821–830, 2024. 2.2.1

[111] S. Raab, I. Gat, N. Sala, G. Tevet, R. Shalev-Arkushin, O. Fried, A. H. Bermano, and D. Cohen-Or, "Monkey see, monkey do: Harnessing self-attention in motion diffusion for zero-shot motion transfer," in *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–13, 2024. 2.2.1

[112] R. Villegas, J. Yang, D. Ceylan, and H. Lee, "Neural kinematic networks for unsupervised motion retargetting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8639–8648, 2018. 2.2.2

[113] X. B. Peng, P. Abbeel, S. Levine, and M. Van de Panne, "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills," *ACM Transactions On Graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018. 2.2.2

[114] H. Wang, F. He, Z. Peng, T. Shao, Y.-L. Yang, K. Zhou, and D. Hogg, "Understanding the robustness of skeleton-based action recognition under adversarial attack," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14656–14665, 2021. 2.2.2, 5.1, 5.2.2, 5.3, 5.4.1, 5.4.1, 7.2.2

[115] H. Wang, Y. Diao, Z. Tan, and G. Guo, "Defending black-box skeleton-based human activity classifiers," *arXiv preprint arxiv.2203.04713*, 2022. 2.2.2

[116] N. Tanaka, H. Kera, and K. Kawamoto, "Adversarial bone length attack on action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 2335–2343, 2022. 2.2.2

[117] J. Liu, N. Akhtar, and A. Mian, "Adversarial attack on skeleton-based human action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 4, pp. 1609–1622, 2020. 2.2.2, 5.1

[118] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, "Simple black-box adversarial attacks," in *International conference on machine learning*, pp. 2484–2493, PMLR, 2019. 2.2.2

[119] Y. Diao, T. Shao, Y.-L. Yang, K. Zhou, and H. Wang, "Basar: Black-box attack on skeletal action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7597–7607, 2021. 2.2.2

[120] Y. Diao, H. Wang, T. Shao, Y. Yang, K. Zhou, D. Hogg, and M. Wang, "Understanding the vulnerability of skeleton-based human activity recognition via black-box attack," *Pattern Recognition*, vol. 153, p. 110564, 2024. 2.2.2, 5.2.2

[121] Z. Kang, H. Xia, R. Zhang, S. Jiang, X. Shi, and Z. Zhang, "Fgda-gs: Fast guided decision attack based on gradient signs for skeletal action recognition," *Computers & Security*, vol. 135, p. 103522, 2023. 2.2.2

[122] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018. 2.2.2, 5.4.1

[123] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," *arXiv preprint arXiv:1611.02770*, 2016. 2.2.2

[124] Z. Lu, H. Wang, Z. Chang, G. Yang, and H. P. Shum, "Hard no-box adversarial attack on skeleton-based human action recognition with skeleton-motion-informed gradient," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4597–4606, 2023. 2.2.2

[125] Y. Diao, B. Wu, R. Zhang, A. Liu, X. Hao, X. Wei, M. Wang, and H. Wang, "TASAR: Transfer-based attack on skeletal action recognition," in *The Thirteenth International Conference on Learning Representations*, 2025. 2.2.2, 5.2.2

[126] K. Hyun, M. Kim, Y. Hwang, and J. Lee, "Tiling motion patches," *IEEE transactions on visualization and computer graphics*, vol. 19, no. 11, pp. 1923–1934, 2013. 2.2.2

[127] I. Bae, J. Lee, and H.-G. Jeon, "Continuous locomotive crowd behavior generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22416–22431, 2025. 2.2.2

[128] W. Wang, M. Liu, and X. Wang, "Multi-characters interaction based on honeycomb," in *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*, vol. 1, pp. 65–69, IEEE, 2012. 2.2.2

[129] H. Ye, H. Lin, J. Han, M. Xu, S. Liu, Y. Liang, J. Ma, J. Y. Zou, and S. Ermon, "Tfg: Unified training-free guidance for diffusion models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 22370–22417, 2024. 2.2.2

[130] K. Fan, J. Tang, W. Cao, R. Yi, M. Li, J. Gong, J. Zhang, Y. Wang, C. Wang, and L. Ma, "Freemotion: A unified framework for number-free text-to-motion

synthesis," in *European Conference on Computer Vision*, pp. 93–109, Springer, 2024. 2.2.2, 7.2.1

[131] W. Xu, S. Fan, P. Henderson, and E. S. Ho, "Multi-person interaction generation from two-person motion priors," in *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025. 2.2.2, 7.2.1

[132] S. Ota, Q. Yu, K. Fujiwara, S. Ikehata, and I. Sato, "Pino: Person-interaction noise optimization for long-duration and customizable motion generation of arbitrary-sized groups," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10676–10685, 2025. 2.2.2

[133] J. Lin, A. Zeng, S. Lu, Y. Cai, R. Zhang, H. Wang, and L. Zhang, "Motion-x: A large-scale 3d expressive whole-body human motion dataset," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2.1, 2

[134] I. Mason, S. Starke, and T. Komura, "Real-time style modelling of human locomotion via feature-wise transformations and local motion phases," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 5, may 2022. 2.1, 2, 5.1, 5.4.1

[135] M. Plappert, C. Mandery, and T. Asfour, "The kit motion-language dataset," *Big Data*, vol. 4, no. 4, pp. 236–252, 2016. 2.1, 2

[136] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "AMASS: Archive of motion capture as surface shapes," in *Proceedings of the IEEE international conference on computer vision*, pp. 5441–5450, Oct. 2019. 2

[137] A. Ghosh, R. Dabral, V. Golyanik, C. Theobalt, and P. Slusallek, "Remos: 3d motion-conditioned reaction synthesis for two-person interactions," in *European Conference on Computer Vision*, pp. 418–437, Springer, 2024. 2.2, 2, 2.3, 2.4.1, 2.4.4, 7.2.2

[138] L. Siyao, T. Gu, Z. Yang, Z. Lin, Z. Liu, H. Ding, L. Yang, and C. C. Loy, "Duolando: Follower GPT with off-policy reinforcement learning for dance accompaniment," in *The Twelfth International Conference on Learning Representations*, 2024. 2.2, 2, 2.4.1

[139] W. Guo, X. Bie, X. Alameda-Pineda, and F. Moreno-Noguer, "Multi-person extreme motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13053–13064, 2022. 2.2, 2

[140] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt, "Single-shot multi-person 3d pose estimation from monocular rgb," in *2018 international conference on 3D vision (3DV)*, pp. 120–130, IEEE, 2018. 2, 2.2

[141] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *European Conference on Computer Vision (ECCV)*, sep 2018. 2, 2.2

[142] M. Fabbri, F. Lanzi, S. Calderara, A. Palazzi, R. Vezzani, and R. Cucchiara, "Learning to detect and track visible and occluded body joints in a virtual world," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 430–446, 2018. 2.2, 2

[143] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," 2019. 2

[144] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, *SMPL: A Skinned Multi-Person Linear Model*. New York, NY, USA: Association for Computing Machinery, 1 ed., 2023. 2

[145] J. Voas, Y. Wang, Q. Huang, and R. Mooney, "What is the best automated metric for text to motion generation?," in *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023. 2.4, 2.4.4

[146] A. Ismail-Fawaz, M. Devanne, S. Berretti, J. Weber, and G. Forestier, "Establishing a unified evaluation framework for human motion generation: A comparative analysis of metrics," *Computer Vision and Image Understanding*, vol. 254, p. 104337, 2025. 2.4

[147] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013. 2.3, 2.4.1

[148] J. Tanke, C. Zaveri, and J. Gall, "Intention-based long-term human motion anticipation," in *International Conference on 3D Vision*, pp. 596–605, IEEE, 2021. 2.3, 2.4.1

[149] A. Gopalakrishnan, A. Mali, D. Kifer, L. Giles, and A. G. Ororbia, "A neural temporal model for human motion prediction," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12116–12125, 2019. 2.3, 2.4.1

[150] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017. 2.3, 2.4.1, 4.1, 4.7

[151] A. Maiorca, Y. Yoon, and T. Dutoit, "Evaluating the quality of a synthesized motion with the fréchet motion distance," in *ACM SIGGRAPH 2022 Posters*, pp. 1–2, 2022. 2.3, 2.4.1

[152] S. Maheshwari, D. Gupta, and R. K. Sarvadevabhatla, "Mugl: Large scale multi person conditional action generation with locomotion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 257–265, January 2022. 2.3, 2.4.1

[153] N. Ugrinovic, B. Pan, G. Pavlakos, D. Paschalidou, B. Shen, J. Sanchez-Riera, F. Moreno-Noguer, and L. Guibas, "Multiphys: Multi-person physics-aware 3d motion estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2331–2340, 2024. 2.3, 2.4.1

[154] Z. Wang, J. Wang, Y. Li, D. Lin, and B. Dai, "Intercontrol: Zero-shot human interaction generation by controlling every joint," *Advances in Neural Information Processing Systems*, vol. 37, pp. 105397–105424, 2024. 2.3, 2.4.1, 2.4.4

[155] J. Wang, H. Xu, M. Narasimhan, and X. Wang, "Multi-person 3d motion prediction with multi-range transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6036–6049, 2021. 2.3, 2.4.4

[156] L. Xu, Y. Zhou, Y. Yan, X. Jin, W. Zhu, F. Rao, X. Yang, and W. Zeng, "Regennet: Towards human action-reaction synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1759–1769, 2024. 2.4.1

[157] M. Tanaka and K. Fujiwara, "Role-aware interaction generation from textual description," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15999–16009, 2023. 2.4.1

[158] A. Goel, Q. Men, and E. S. Ho, "Interaction mix and match: Synthesizing close interaction using conditional hierarchical gan with multi-hot class embedding," in *Computer Graphics Forum*, vol. 41, pp. 327–338, Wiley Online Library, 2022. 2.4.1

[159] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or, "Motionclip: Exposing human motion generation to clip space," in *European Conference on Computer Vision*, pp. 358–374, Springer, 2022. 2.4.3

[160] R. Rekik, S. Wuhrer, L. Hoyet, K. Zibrek, and A.-H. Olivier, "Quality assessment of 3d human animation: Subjective and objective evaluation," *arXiv preprint arXiv:2505.23301*, 2025. 2.4.4

[161] N. F. Troje, "Decomposing biological motion: A framework for analysis and synthesis of human gait patterns," *Journal of vision*, vol. 2, no. 5, pp. 2–2, 2002. 2.4.4, 5.1

[162] S. Shimada and K. Oki, "Modulation of motor area activity during observation of unnatural body movements," *Brain and cognition*, vol. 80, no. 1, pp. 1–6, 2012. 2.4.4, 5.1

[163] S. Saint-Auret, F. Multon, R. Gaugne, L. Hoyet, R. Kulpa, and V. Gouranton, "How do people perceive changes in physical bounce model for virtual racket interactions?," in *ACM Symposium on Applied Perception 2025*, pp. 1–10, 2025. 2.4.4

[164] H. Wang, W. Zhu, L. Miao, Y. Xu, F. Gao, Q. Tian, and Y. Wang, "Aligning human motion generation with human perceptions," in *The Thirteenth International Conference on Learning Representations*, 2025. 2.4.4

[165] J. Sheng, M. Lin, A. Zhao, K. Pruvost, Y.-H. Wen, Y. Li, G. Huang, and Y.-J. Liu, "Exploring text-to-motion generation with human preference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1888–1899, 2024. 2.4.4

[166] A. T. Jebb, V. Ng, and L. Tay, "A review of key likert scale development advances: 1995–2019," *Frontiers in psychology*, vol. 12, p. 637547, 2021. 2.4.4

[167] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*, pp. 2256–2265, PMLR, 2015. 3, 4.4

[168] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 26565–26577, 2022. 3

[169] Z. Ye, H. Wu, and J. Jia, "Human motion modeling with deep learning: A survey," *AI Open*, 2021. 4.1

[170] Y. Dong, A. Aristidou, A. Shamir, M. Mahler, and E. Jain, "Adult2child: Motion style transfer using cyclegans," in *Proceedings of the 13th ACM SIGGRAPH Conference on Motion, Interaction and Games*, MIG '20, (New York, NY, USA), Association for Computing Machinery, 2020. 4.1

[171] H. Y. Ling, F. Zinno, G. Cheng, and M. Van De Panne, "Character controllers using motion vaes," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 40–1, 2020. 4.1

[172] G. E. Henter, S. Alexanderson, and J. Beskow, "Moglow: Probabilistic and controllable motion synthesis using normalising flows," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–14, 2020. 4.1

[173] E. J. Findlay, H. Zhang, Z. Chang, and H. P. Shum, "Denoising diffusion probabilistic models for styled walking synthesis," *arXiv preprint arXiv:2209.14828*, 2022. 4.1, 4.6, 4.7, 4.2, 4.7

[174] D. McAllester, "On the mathematics of diffusion models," *arXiv preprint arXiv:2301.11108*, 2023. 4.3

[175] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical*

image computing and computer-assisted intervention, pp. 234–241, Springer, 2015. 4.5.1

[176] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21* (Z.-H. Zhou, ed.), pp. 4312–4321, International Joint Conferences on Artificial Intelligence Organization, 8 2021. Survey Track. 5.1

[177] B. Chander, C. John, L. Warrier, and K. Gopalakrishnan, "Toward trustworthy artificial intelligence (tai) in the context of explainability and robustness," *ACM Computing Surveys*, 2024. 5.1

[178] S. Kotsovolis and Y. Demiris, "Garment diffusion models for robot-assisted dressing," *IEEE Robotics and Automation Letters*, 2024. 5.1

[179] M. Bronars, S. Cheng, and D. Xu, "Legibility diffuser: Offline imitation for intent expressive motion," *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 10161–10168, 2024. 5.1

[180] A. Martin-Ozimek, I. Jayarathne, S. L. Mon, and J. Y. Chew, "Diffusion-based imitation learning for social pose generation," in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 1488–1492, IEEE, 2025. 5.1

[181] S. Samavi, A. Lem, F. Sato, S. Chen, Q. Gu, K. Yano, A. P. Schoellig, and F. Shkurti, "Sicnav-diffusion: Safe and interactive crowd navigation with diffusion trajectory predictions," *IEEE Robotics and Automation Letters*, 2025. 5.1

[182] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017. 5.1

[183] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018. 5.1

[184] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *Advances in neural information processing systems*, vol. 32, 2019. 5.1

[185] A. Li, Y. Wang, Y. Guo, and Y. Wang, "Adversarial examples are not real features," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 5.1

[186] L. Kovács, B. M. Bódis, and C. Benedek, "Lidpose: Real-time 3d human pose estimation in sparse lidar point clouds with non-repetitive circular scanning pattern," *Sensors*, vol. 24, no. 11, p. 3427, 2024. 5.1

[187] J. Xu, Y. Guo, and Y. Peng, "Finepose: Fine-grained prompt-driven 3d human pose estimation via diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 561–570, 2024. 5.1

[188] S. Shin, J. Kim, E. Halilaj, and M. J. Black, "Wham: Reconstructing world-grounded humans with accurate 3d motion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2070–2080, 2024. 5.1

[189] V. Vapnik, "Principles of risk minimization for learning theory," *Advances in neural information processing systems*, vol. 4, 1991. 5.1, 5.2.1

[190] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*, vol. 31. Springer Science & Business Media, 2013. 5.1, 5.2.1

[191] G. Claeskens and N. L. Hjort, "Model selection and model averaging," *Cambridge books*, 2008. 5.1, 5.2.1

[192] Y. P. Ivanenko, R. E. Poppele, and F. Lacquaniti, "Five basic muscle activation patterns account for muscle activity during human locomotion," *The Journal of physiology*, vol. 556, no. 1, pp. 267–282, 2004. 5.1

[193] Y. Ueyama, "Costs of position, velocity, and force requirements in optimal control induce triphasic muscle activation during reaching movement," *Scientific Reports*, vol. 11, no. 1, p. 16815, 2021. 5.1

[194] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation mocap database hdm05," Tech. Rep. CG-2007-2, Universität Bonn, June 2007. 5.1, 5.4.1

[195] C. H. Wu and F. De la Torre, "A latent space of stochastic diffusion models for zero-shot image editing and guidance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7378–7387, 2023. 5.2.1, 5.2.1

[196] P. Jaini, K. Clark, and R. Geirhos, "Intriguing properties of generative classifiers," in *The Twelfth International Conference on Learning Representations*, 2024. 5.2.1

[197] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, "Diffusion autoencoders: Toward a meaningful and decodable representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10619–10629, 2022. 5.2.1

[198] K. Kim and J. C. Ye, "Noise2score: tweedie's approach to self-supervised image denoising without clean images," *Advances in Neural Information Processing Systems*, vol. 34, pp. 864–874, 2021. 5.2.1

[199] L. Zhong, Y. Xie, V. Jampani, D. Sun, and H. Jiang, "Smoodi: Stylized motion diffusion model," in *European Conference on Computer Vision*, pp. 405–421, Springer, 2025. 5.4.1, 5.1, 5.2, 5.4.2

[200] H. Qiu, B. Hou, B. Ren, and X. Zhang, "Spatio-temporal tuples transformer for skeleton-based action recognition," *arXiv preprint arXiv:2201.02849*, 2022. 5.4.1, 5.1, 5.2, 5.4.2

[201] J. Do and M. Kim, "Skateformer: Skeletal-temporal transformer for human action recognition," in *European Conference on Computer Vision*, Springer, 2025. 5.4.1, 5.1, 5.2, 5.4.2

[202] H. Zhou, Q. Liu, and Y. Wang, "Learning discriminative representations for skeleton based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10608–10617, 2023. 5.4.1, 5.1, 5.2, 5.4.2

[203] C. Duan, Z. Zhang, X. Liu, Y. Dang, and J. Yin, "Physics-constrained attack against convolution-based human motion prediction," *Neurocomputing*, vol. 575, p. 127272, 2024. 5.4.1

[204] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, pp. 99–112, Chapman and Hall/CRC, 2018. 5.4.1

[205] W. Ma, Y. Li, X. Jia, and W. Xu, "Transferable adversarial attack for both vision transformers and convolutional networks via momentum integrated gradients," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4630–4639, 2023. 5.4.1

[206] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano, "Human motion diffusion model," in *The Eleventh International Conference on Learning Representations*, 2023. 5.4.1, 5.3

[207] Y.-H. Park, M. Kwon, J. Choi, J. Jo, and Y. Uh, "Understanding the latent space of diffusion models through the lens of riemannian geometry," *Advances in Neural Information Processing Systems*, vol. 36, pp. 24129–24142, 2023. 5.4.2

[208] A. Sclocchi, A. Favero, and M. Wyart, "A phase transition in diffusion models reveals the hierarchical nature of data," *arXiv preprint arXiv:2402.16991*, 2024. 5.4.3

[209] Y. Huang, J. Wang, Y. Shi, B. Tang, X. Qi, and L. Zhang, "Dreamtime: An improved optimization strategy for diffusion-guided 3d generation," in *The Twelfth International Conference on Learning Representations*, 2023. 5.4.3

[210] M. Kwon, J. Jeong, and Y. Uh, "Diffusion models already have a semantic latent space," in *The Eleventh International Conference on Learning Representations*, 2023. 5.4.3

[211] K. Zhou, L. Wang, X. Zhang, H. P. H. Shum, F. W. B. Li, J. Li, and X. Liang, "Magr: Manifold-aligned graph regularization for continual action quality assessment," in *Proceedings of the 2024 European Conference on Computer Vision*, ECCV '24, Springer, 2024. 5.5

[212] K. Zhou, R. Cai, Y. Ma, Q. Tan, X. Wang, J. Li, H. P. Shum, F. W. Li, S. Jin, and X. Liang, "A video-based augmented reality system for human-in-the-loop muscle strength assessment of juvenile dermatomyositis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 5, pp. 2456–2466, 2023. 5.5

[213] H. Wang, E. S. Ho, H. P. Shum, and Z. Zhu, "Spatio-temporal manifold learning for human motions via long-horizon modeling," *IEEE transactions on visualization and computer graphics*, vol. 27, no. 1, pp. 216–227, 2019. 5.5

[214] S. Starke, I. Mason, and T. Komura, "Deepphase: Periodic autoencoders for learning motion phase manifolds," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–13, 2022. 6.1

[215] P. Charalambous, J. Pettre, V. Vassiliades, Y. Chrysanthou, and N. Pelechano, "Greil-crowds: Crowd simulation with deep reinforcement learning and examples," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–15, 2023. 6.1

[216] J. Yue, B. Li, J. Pettré, A. Seyfried, and H. Wang, "Human motion prediction under unexpected perturbation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1501–1511, 2024. 6.1, 7.2.3

[217] J. Zhong, D. Li, Z. Huang, C. Lu, and W. Cai, "Data-driven crowd modeling techniques: A survey," *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, vol. 32, no. 1, pp. 1–33, 2022. 6.1

[218] S. M. Fiore, T. J. Wiltshire, E. J. Lobato, F. G. Jentsch, W. H. Huang, and B. Axelrod, "Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior," *Frontiers in psychology*, vol. 4, p. 859, 2013. 6.2

[219] J. Jeong, D. Park, and K.-J. Yoon, "Multi-agent long-term 3d human pose forecasting via interaction-aware trajectory conditioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1617–1628, 2024. 6.4

[220] J. Rios-Martinez, A. Spalanzani, and C. Laugier, "From proxemics theory to socially-aware navigation: A survey," *International Journal of Social Robotics*, vol. 7, pp. 137–153, 2015. 6.4

[221] G. Barquero, S. Escalera, and C. Palmero, "Seamless human motion composition with blended positional encodings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 6.6

[222] M. Petrovich, O. Litany, U. Iqbal, M. J. Black, G. Varol, X. Bin Peng, and D. Rempe, "Multi-track timeline control for text-driven 3d human motion generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1911–1921, 2024. 6.8.2

[223] K. Karunratanakul, K. Preechakul, S. Suwajanakorn, and S. Tang, "Guided motion diffusion for controllable human motion synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2151–2162, October 2023. 6.9

[224] Z. Chang, H. Wang, G. Koulieris, and H. P. Shum, "Large-scale multi-character interaction synthesis," in *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–10, 2025. 7.2.1

[225] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020. 7.2.1

[226] J. S. Wiggins, *The five-factor model of personality: Theoretical perspectives*. Guilford Press, 1996. 7.2.3

[227] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current psychology*, vol. 14, no. 4, pp. 261–292, 1996. 7.2.3

[228] J. Hwang, J. Kim, I. H. Suh, and T. Kwon, "Real-time locomotion controller using an inverted-pendulum-based abstract model," in *Computer Graphics Forum*, vol. 37, pp. 287–296, Wiley Online Library, 2018. 7.2.3

[229] A. Maiorca, H. Bohy, Y. Yoon, and T. Dutoit, "Objective evaluation metric for motion generative models: Validating fréchet motion distance on foot skating and over-smoothing artifacts.," in *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pp. 1–11, 2023. 7.2.4

[230] X. Zhang, Z. Chang, Q. Men, and H. P. Shum, "Real-time and controllable reactive motion synthesis via intention guidance," in *Computer Graphics Forum*, p. e70222, Wiley Online Library, 2025. 7.2.4

[231] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li, "Human preference score: Better aligning text-to-image models with human preference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2096–2105, 2023. 7.2.4

[232] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, and H. Li, "Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis," *arXiv preprint arXiv:2306.09341*, 2023. 7.2.4

[233] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "Imagereward: Learning and evaluating human preferences for text-to-image

generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 15903–15935, 2023. 7.2.4

[234] Y. Bai, X. Wang, Y.-P. Cao, Y. Ge, C. Yuan, and Y. Shan, "Dreamdiffusion: High-quality eeg-to-image generation with temporal masked signal modeling and clip alignment," in *European Conference on Computer Vision*, pp. 472–488, Springer, 2024. 7.2.4

APPENDIX A

Hardware Acknowledgements

In addition to the individuals acknowledged for their contributions to this thesis, we also recognize the essential hardware support that made this research possible.

We extend our sincere gratitude to Durham University's NVIDIA CUDA Center (NCC) GPU system (https://nccadmin.webspace.durham.ac.uk), whose computational resources were instrumental in conducting the experiments presented in this work. The NCC cluster, established through Durham University's strategic investment funds and managed by the Department of Computer Science, provided a high-performance computing environment that enabled the efficient processing of large-scale datasets and the execution of complex deep learning models. This infrastructure was vital for the rigorous testing and refinement of the methodologies developed in this thesis, and we are grateful for the access to such advanced resources, which have been critical to the success of this research.