TEXT-DRIVEN HUMAN MOTION SYNTHESIS USING DIFFUSION MODELS

Dissertation Phase 2 Report

Submitted by

Mr. K J ANANTHAKRISHNAN

REG NO : TKM22MEAI03

the APJ Abdul Kalam Technological University in partial fulfillment for the award of the degree of

MASTER OF TECHNOLOGY
IN
Artificial Intelligence

Under the guidance of
Dr. Sumod Sundar

Centre for Artificial Intelligence

TKM College of Engineering, Kollam

JUNE 2024

Thangal Kunju Musaliar College of Engineering
Centre for Artificial Intelligence

C E R T I F I C A T E

This is to certify that, this report titled ***TEXT-DRIVEN HUMAN MOTION SYNTHESIS USING DIFFUSION MODELS*** is a bonafide record of the **Dissertation Phase 2** work presented by **K J ANANTHAKRISHNAN (TKM22MEAI03)**, under our guidance and supervision, in partial fulfillment of the requirements for the award of the degree, **M. Tech in Artificial Intelligence** in **APJ Abdul Kalam Technological University**.

Internal Supervisor and Project Coordinator

Dr. Sumod Sundar
Associate Professor
Centre for AI
TKMCE

Head of the Department

Dr. Imthias Ahamed T P
Professor
Centre for AI
TKMCE

ACKNOWLEDGEMENT

A successful project is a fruitful culmination of efforts by many people, some directly involved and some others indirectly, by providing support and encouragement. Firstly I would like to thank the almighty for giving me the wisdom and grace for making my project a memorable one. I thank him for steering me to the shore of fulfillment under his protective wings.

I express my sincere gratitude to **Dr. T A Shahul Hameed**, Principal of T.K.M College of Engineering for allowing me to present my dissertation phase 2. I would like to thank **Dr. Imthias Ahamed T P**, Professor and Head of the Department, Centre for Artificial Intelligence, TKM College of Engineering, Kollam, for his constant support and encouragement throughout the work.

With a profound sense of gratitude, I would like to express my heartfelt thanks to my Internal Supervisor and Project Coordinator, **Dr. Sumod Sundar**, Associate Professor, Centre for Artificial Intelligence(AI), TKM College of Engineering, Kollam for his expert guidance, cooperation, and immense encouragement. I also extend my thanks to the entire faculty and staff members of the Centre for AI, TKMCE, who have encouraged me throughout this work.

I also express my thanks to my loving parents and friends, for their support and encouragement in the successful completion of this work.

K J Ananthakrishnan

Abstract

The term "text-to-motion generation" pertains to the procedure of producing sequences of human motion by using textual input. The work at hand presents considerable challenges, mostly stemming from the wide range of potential motion, the inherent sensitivity of human perception to such motion, and the inherent complexities associated in effectively articulating and characterising it. The existing generative methods for text-to-motion synthesis are characterised by either substandard quality or constrained expressiveness. To overcome this, we introduce, a diffusion model-based framework for text-driven motion generation. The model presents multiple benefits: Firstly, it employs probabilistic mapping to generate motions by refining inputs through denoising processes while introducing variations. Secondly, it features multi-level manipulation capabilities, allowing it to interpret detailed instructions regarding body parts and facilitate the synthesis of motions of various lengths in response to text prompts that change over time. The performance of model is evaluated using metrics such as FID (Fréchet Inception Distance), Diversity, and MultiModality, which measure the quality and diversity of generated motion samples.

Contents

1 Introduction **4**

2 Literature Survey **7**

3 Methodology **15**
3.1 Objectives 15
3.2 Proposed Framework 15
3.3 Techniques Used 16
3.3.1 Text Encoder 17
3.3.2 Linear Self Attention and Cross Attention 19
3.3.3 Precise Manipulation 20
3.4 Dataset Used 25
3.4.1 HUMANML3D DATASET 25
3.5 Performance Metrics 26

4 Experimental Analysis and Results **30**
4.1 Experiments on data pre-processing task 30
4.2 Environmental Setup 31
4.3 Results 31

5 Conclusion and Future Scope **42**

References **44**

List of Figures

3.1 Motion Generation Framework 15
3.2 Methodology of Multi-Level Manipulation 21
3.3 Methodology of Time Varied Signal 23
3.4 Sample data from HUMANML3D 26

4.1 Generated video showing transitions between different poses 33
4.2 Generated video showing different actions for same description 34
4.3 Generated video shows a small deviation from expected context 36
4.4 Generated video shows Multi-Level Manipulation 37
4.5 Generated video show Time Varied Signal 38

List of Tables

4.1 Quantitative evaluation on the HumanML3D test set 32

Chapter 1

Introduction

The endeavor to generate realistic 3D human motion sequences from textual descriptions sits at the cutting edge of technology, blending advancements in natural language processing and computer graphics. This pursuit holds the promise of revolutionizing interactions with virtual characters and avatars by enabling more natural and expressive animations. Despite notable progress in the field, existing solutions often fall short in terms of quality and expressiveness, struggling to capture the full nuances and dynamism present in textual inputs. Consequently, the motions generated by these methods tend to appear artificial or rigid. The primary objective of this project is to surmount these limitations by developing a system capable of interpreting textual descriptions and generating a diverse range of 3D human motions that faithfully capture the specified action types, directions, speed, timing, and styles. This advancement holds significant implications across various domains, including gaming, robotics, animations, virtual reality, and beyond.

To address the shortcomings of current methods, the project introduces a comprehensive framework for text-driven motion generation, leveraging the capabilities of diffusion models. This framework incorporates two key innovations: Probabilistic Mapping and Multi-Level Manipulation. The Probabilistic Mapping approach entails iteratively refining motion estimates through a denoising process while introducing variations at each step. This iterative refinement ensures that the generated motions exhibit both naturalness and diversity, capturing the organic variability characteristic of real human movements. By treating motion generation as a probabilistic process, this framework can explore a wide range of potential motion sequences, accommodating the inherent complexity and variability of human actions.

Probabilistic Mapping serves as a foundational component of the proposed framework, enabling the generation of realistic and diverse motion sequences. Through iterative refinement and denoising, this approach ensures that the generated motions closely adhere to the underlying structure and dynamics described in the textual inputs. Moreover, by introducing variations at each step, Probabilistic Mapping captures the inherent unpredictability and fluidity of human movements, resulting in more authentic and expressive animations.

Complementing Probabilistic Mapping, Multi-Level Manipulation enhances the precision and control over generated motions. This technique allows for fine-grained adjustments at different levels of the motion sequence, ensuring coherence and fidelity to the specified ac-

tions. Whether it involves subtle gestures or complex sequences, Multi-Level Manipulation ensures that each part of the generated motion aligns with the intended action, contributing to the overall realism and believability of the animations.

In practice, the iterative refinement process begins with a rough initial estimate of the motion sequence derived from the textual description. Each subsequent iteration introduces slight variations and refines the motion details through denoising. This process mimics the variability seen in real human movements, ensuring that the final animations are not only lifelike but also diverse. This probabilistic method allows the system to produce a broad array of authentic motion sequences from a single textual input, thereby enhancing the realism and richness of the animations.

The Multi-Level Manipulation technique offers detailed control over individual body parts, facilitating the generation of motions that can adapt to evolving textual descriptions. This capability is crucial for handling text that varies in length and detail, allowing the system to produce precise motions that align closely with the specified actions. Whether the text describes a subtle hand gesture or a complex sequence of movements, Multi-Level Manipulation ensures cohesive movement of each part of the body in accordance with the overall described action.

This technique provides fine-grained control over different levels of the human body. It enables the system to accurately depict movements specified in the text, whether they involve minor gestures or intricate, coordinated actions. Multi-Level Manipulation allows the system to dynamically adapt motions as the textual descriptions evolve, ensuring that the generated animations remain consistent and accurate. This level of control is essential for producing animations that are both realistic and expressive, as it allows for precise synchronization of body parts with the described actions.

Together, Probabilistic Mapping and Multi-Level Manipulation create a robust framework for text-driven motion generation. Probabilistic Mapping ensures diversity and naturalness in the motions, while Multi-Level Manipulation provides fine-grained control and adaptability. This combination enables the generation of highly realistic and expressive 3D human motions from textual descriptions, opening new possibilities for interactive and immersive virtual experiences.

The iterative refinement process encapsulated within Probabilistic Mapping serves as a foundational element in achieving the crucial diversity and naturalness necessary in generated motions. Initiating with an initial approximation, the system embarks on a series of iterative steps, undergoing denoising and introduction of variations with each iteration. This meticulous procedure gradually diminishes noise while integrating subtle adjustments, echoing the nuanced variability inherent in genuine human movement. Consequently, rigid and mechanical motions are effectively eliminated, paving the path for the emergence of a broad spectrum of lifelike motion sequences imbued with vitality and authenticity.

On the contrary, Multi-Level Manipulation emerges as a pivotal instrument in translating textual descriptions into seamlessly orchestrated animations. Envision a scenario where a

character is depicted performing a specific hand gesture while walking. Through Multi-Level Manipulation, every element of the body synchronizes harmoniously, ensuring that the character's actions resonate with believability and realism. This precision in governing individual body parts not only heightens the fidelity of the animations but also enhances the immersive experience for the observer.

Furthermore, as narratives evolve, so do the requisites for motion sequences. Herein lies the strength of Multi-Level Manipulation, excelling in dynamically adapting to these evolving textual descriptions. Whether it's a simple gesture or a complex dance routine, this methodology guarantees that each body part adjusts in real-time, maintaining coherence with the overarching motion. Such adaptability is paramount, not only in crafting realistic animations but also in ensuring responsiveness to the evolving narrative contexts, thus enhancing the overall engagement and immersion for the audience.

In summary, the amalgamation of Probabilistic Mapping and Multi-Level Manipulation marks a significant milestone in text-driven human motion synthesis. By leveraging the capabilities of probabilistic modeling and precise manipulation techniques, this framework not only tackles the constraints of current methodologies but also unlocks a realm of possibilities for generating lifelike animations resonating with authenticity and depth. As exploration continues in the domain of text-driven motion synthesis, this innovative approach holds the promise of shaping the future landscape of interactive storytelling, immersive gaming experiences, and beyond, ushering in an era characterized by creativity and realism in virtual realms.

Chapter 2

Literature Survey

Several studies have contributed significantly to the development of automated systems for the generation of 3D motion from textual description. These studies have employed various techniques and datasets to achieve high levels of accuracy.

TEMOS (Text-to-Motion Synthesis) is a cutting-edge system designed to generate a wide array of 3D human motions from textual descriptions [1]. By combining variational autoencoders (VAEs) with transformers, TEMOS effectively models both motion and text modalities. Unlike previous deterministic methods that produce only a single static motion per description, TEMOS generates multiple realistic motion sequences from a single textual input, addressing significant limitations in prior approaches. TEMOS utilizes a variational autoencoder (VAE) framework to encapsulate the inherent variability and complexity of human motions. The model is trained on an extensive dataset of human motions. Within the VAE, an encoder compresses motion data into a latent space, and a decoder reconstructs motion from this latent representation. The encoder learns to produce a distribution (mean and variance) of latent vectors, which represent different motion styles and variations. Additionally, a text encoder processes textual descriptions, generating distribution parameters (mean and variance) that align with the VAE's latent space. This integration ensures that textual descriptions are mapped to the same latent space as the motion data, facilitating the generation of corresponding motions. To effectively capture the sequential nature of both text and motion data, TEMOS employs transformers. The system uses symmetrical transformer-based encoders for both text and motion data, converting sequences of words or motion frames into high-dimensional embeddings that capture essential features and temporal dependencies. The motion decoder, also based on transformers, takes a latent vector from the VAE and generates a sequence of 3D human motion frames. This decoder is trained to produce realistic and coherent motion sequences that align with the latent representations derived from the text, ensuring high-quality motion synthesis.

TEMOS offers several advantages over previous methods. It can generate multiple diverse motions from the same textual description due to the probabilistic nature of the VAE, which samples from the latent space to create varied outputs. TEMOS shows significant improvements over state-of-the-art methods, especially in generating expressive SMPL (Skinned Multi-Person Linear) body motions, which are crucial for lifelike animations. However, TEMOS faces challenges with complex or ambiguous descriptions, indicating a need

for improved preprocessing or refinement steps. Additionally, scalability issues arise with very long motion sequences due to the quadratic memory cost associated with transformers, limiting the model's applicability for extended motion generation tasks. TEMOS has demonstrated superior performance on the KIT Motion-Language benchmark, surpassing existing text-conditioned motion generation methods. It generates more diverse and realistic human motions from textual descriptions, highlighting its superiority in both the quality and variability of outputs. This robust performance underscores TEMOS's potential to set new standards in the field of text-to-motion synthesis. In summary, TEMOS represents a significant advancement in text-to-motion generation, offering a robust and versatile framework capable of producing diverse and realistic 3D human motions from textual descriptions.

In another work, researchers have developed an innovative method for text-driven human motion generation, addressing challenges prevalent in aligning motion sequences accurately with textual prompts [2]. The proposed methodology for generating 3D human motions from textual descriptions introduces a novel architecture comprising two key modules: the Linguistics-Structure Assisted Module (LSAM) and the Context-Aware Progressive Reasoning (CAPR) Module. Collaboratively functioning within a diffusion model framework, these modules aim to produce human motion sequences that align well with the provided textual descriptions. LSAM enhances text encoding by leveraging dependency parsing relationships within the textual input, capturing richer semantic information by understanding the grammatical structure of sentences and identifying relationships between words. This facilitates effective information exchange and aggregation, ensuring that both local and global semantic linguistic features are comprehensively utilized during the encoding process.

On the other hand, the CAPR module focuses on learning semantic features through a multi-step inference process, akin to the progressive comprehension of text by humans. Employing graph neural networks, CAPR implements a multi-step progressive inference strategy, enabling the extraction of deeper semantic insights from the text. This iterative reasoning process refines the system's understanding of the textual descriptions, thereby enhancing the alignment of the generated motion sequences. The overall framework for generating human motion sequences is grounded in a diffusion model, which incrementally refines motion estimates through a denoising process. The integration of LSAM and CAPR ensures that the generated motions not only align semantically with the text but also exhibit natural and diverse characteristics reflective of real human movements.

The methodology's effectiveness lies in LSAM's ability to capture comprehensive linguistic features and CAPR's progressive reasoning strategy, which mimics human comprehension processes. Together, these modules enable the system to achieve competitive performance on benchmark datasets like HumanML3D and KIT, producing motion sequences that closely align with the provided textual descriptions. However, the methodology's performance heavily depends on the LSAM and CAPR modules. Without these components, the quality of the generated motion sequences may degrade significantly. Additionally, careful tuning of the CAPR module's hyper-parameter is necessary to strike a balance between text information fusion and motion feature preservation.

Experimental results have shown that the proposed Fg-T2M method outperforms pre-

vious text-to-motion generation methods on benchmark datasets, with visually confirmed alignment between the generated motion sequences and the text descriptions. User studies further validate the methodology's superiority, demonstrating better performance in metrics such as R-Precision and FID. In summary, the integration of LSAM and CAPR within a diffusion model framework presents a robust approach to generating 3D human motions from textual descriptions. Despite challenges related to module reliance and parameter tuning, this methodology represents a significant advancement in text-driven motion generation, with potential applications in animation, virtual reality, and human-computer interaction.

A unified motion-language framework called MotionGPT, which consists of a motion tokenizer and a motion-aware language model [3]. The motion tokenizer is based on the Vector Quantized Variational Autoencoders (VQ-VAE) architecture and is used to convert raw motion data into discrete motion tokens. The paper uses a motion-aware language model called T5 as the backbone model for MotionGPT. T5 is a pre-trained large language model that has been exposed to language data and represented within a text vocabulary. MotionGPT employs a 220M pre-trained Flan-T5-Base model as its backbone and fine-tunes it through the pre-training. Gesture Action Classification dataset with 1050 training samples and 259 test samples, collected from youtube conversation videos, labeled with subject, gesture, and emotion . 120 dataset for evaluating cross-dataset generalization performance, containing 12 selected action classes. MotionGPT's methodology represents a groundbreaking approach to generating human motion sequences from textual descriptions, employing a fusion of a motion tokenizer and a transformer-based architecture. At its core, MotionGPT utilizes a motion tokenizer rooted in the Vector Quantized Variational Autoencoders (VQ-VAE) architecture, which assumes a crucial role in preprocessing. By converting raw motion data into discrete motion tokens, MotionGPT streamlines the processing and manipulation of motion sequences, setting the stage for aligning motion generation with textual inputs.

The foundation of MotionGPT's architecture lies in a transformer-based model, custom-tailored for conditioned generation tasks. Renowned for its proficiency in handling sequential data, transformers effectively map input sequences to output sequences. Harnessing this architecture, MotionGPT excels in generating human motion sequences guided by textual instructions. The transformer-based approach empowers MotionGPT to discern intricate dependencies and patterns within textual descriptions and motion data, enabling the synthesis of coherent and realistic human motion sequences closely aligned with the provided textual inputs. A distinguishing feature of MotionGPT is its versatility and effectiveness across a broad spectrum of motion-related tasks. The model demonstrates cutting-edge performance in various endeavors, including text-driven motion generation, motion captioning, and motion prediction. This versatility underscores MotionGPT's relevance across diverse domains requiring human motion synthesis, spanning entertainment, gaming, robotics, and virtual reality. Moreover, MotionGPT exhibits competitive prowess in text-to-motion tasks, accurately transcribing textual descriptions into human motion sequences with fidelity and precision.

Extensive evaluations through empirical studies validate MotionGPT's proficiency in tackling diverse motion tasks within a unified framework. Consistently delivering competitive performance across all evaluated tasks, MotionGPT underscores its efficacy in generating

human motion sequences from textual inputs. Particularly notable is MotionGPT's ability to surpass recent methods in accurately describing human motion sequences based solely on textual descriptions. This superiority highlights MotionGPT's potential to redefine the landscape of text-driven motion synthesis and advance the state of the art in the field. In conclusion, MotionGPT presents a robust and adaptable methodology for generating human motion sequences from textual descriptions. Through the integration of a motion tokenizer and a transformer-based architecture, MotionGPT achieves exceptional effectiveness and competitive performance across various motion tasks. Its seamless integration of textual instructions with motion generation holds significant promise for addressing real-world challenges in human motion synthesis and propelling advancements in artificial intelligence.

Motion Diffusion Model (MDM), is a classifier-free diffusion-based generative model for human motion [4]. MDM consists of a transformer model takes feature from motion generation literature and predicts the sample, rather than the noise, in each diffusion step. Transformer-Encoder Backbone: MDM features a transformer-encoder backbone, which contributes to its lightweight nature and accurate performance in conditioned generation tasks. MDM utilizes a transformer-based architecture, which contributes to its efficiency and accuracy in conditioned generation tasks. KIT Motion-Language dataset is used here. Diffusion models can be challenging to control due to their many-to-many nature, making it harder to precisely manipulate the generated motion. MDM, or the Model for Diffusion-based Motion generation, introduces an innovative method for human motion synthesis distinguished by its classifier-free, diffusion-based generative model. Departing from conventional approaches reliant on classifiers, MDM employs a transformer-based architecture, directly predicting samples at each diffusion step. This unique strategy enables MDM to effectively leverage geometric losses such as foot contact loss, significantly enhancing the quality and realism of the resulting motion sequences.

At the core of MDM's architecture lies a transformer-encoder backbone, pivotal in prioritizing signal prediction over noise in each diffusion step. This emphasis on signal prediction not only contributes to a lightweight model, easing training and control, but also ensures state-of-the-art performance across prominent benchmarks for text-to-motion and action-to-motion tasks. This design principle bolsters the model's efficiency during both training and inference, fostering robust performance across diverse settings and tasks. A notable advantage of MDM is its efficient resource utilization for training and inference, rendering it more accessible for real-world applications where computational resources may be constrained. Despite its lightweight nature, MDM maintains competitiveness and achieves state-of-the-art results across various benchmarks for human motion generation tasks. Additionally, MDM offers conditioning flexibility and editing capabilities, enabling users to tailor motion synthesis to specific requirements.

However, akin to other diffusion models, MDM encounters challenges in precisely controlling and manipulating generated motion due to its inherently many-to-many nature. Diffusion models inherently yield diverse outputs for a given input, posing difficulties in steering the model toward accurately producing desired motion sequences. Furthermore, while MDM addresses resource demands through lightweight resource usage, diffusion models, in general, remain resource-intensive, necessitating significant computational resources

for training and inference. Empirical evaluations demonstrate MDM's superiority in conditioned generation tasks, exhibiting robust performance across varied architectures and settings. Its efficacy in generating natural and expressive human motion is evidenced by its state-of-the-art performance on leading benchmarks for text-to-motion and action-to-motion tasks. These accomplishments underscore MDM's potential as a versatile and efficient solution for human motion synthesis across diverse domains.

In summary, MDM's methodology and architecture leverage transformer-based designs and direct sample prediction to achieve efficient training, state-of-the-art results, and conditioning flexibility in motion generation tasks. Despite challenges associated with diffusion models, MDM's lightweight approach, coupled with its competitive performance, positions it as a promising model for advancing the field of human motion synthesis.

In this work, they introduced ReMoDiffuse, a retrieval-augmented motion diffusion model aimed at improving the quality and diversity of generated motion sequences efficiently [5]. ReMoDiffuse comprises retrieval and refinement stages, incorporating a Hybrid Retrieval technique and a Semantics-Modulated Transformer for generating semantic-consistent motion sequences. The model surpasses existing pipelines in motion generation, demonstrating enhanced generalizability across both common and uncommon prompts. Through qualitative comparisons with previous approaches, ReMoDiffuse exhibits superior performance in accurately conveying text descriptions involving action and path information. Its utilization of extra knowledge retrieved from samples, alongside a learnable Condition Mixture strategy during inference, contributes to the generation of high-quality and description-consistent motion sequences.

ReMoDiffuse, known as the Retrieval-Augmented Motion Diffusion Model, introduces an innovative method for synthesizing 3D human motion by integrating retrieval-augmented techniques to refine motion generation fidelity based on provided prompts. At the heart of ReMoDiffuse lies a retrieval stage, designated to acquire informative samples guiding subsequent motion generation. This retrieval mechanism significantly contributes additional knowledge crucial for producing high-fidelity motion sequences coherent with the input prompts.

Comprising two primary stages, ReMoDiffuse initiates with the retrieval stage, where it implements a Hybrid Retrieval technique. This approach meticulously evaluates both semantic and kinematic similarities to select samples, ensuring that the retrieved data align not only semantically but also kinematically with the given prompts. Consequently, in the refinement stage, ReMoDiffuse employs a Semantics-Modulated Transformer to further enhance motion sequences' coherence. This refining process meticulously aligns generated motions with the semantic context provided by the prompts, resulting in more accurate and realistic outputs.

During the inference phase, ReMoDiffuse adopts a learnable Condition Mixture strategy to produce motion sequences that are both high-fidelity and description-consistent. This adaptive approach empowers the model to dynamically adjust the motion generation process in response to the provided prompts, ensuring that the generated sequences accurately

capture the semantic and kinematic essence of the input descriptions. The efficacy of ReMoDiffuse is demonstrated through qualitative comparisons with prior works, showcasing its proficiency in accurately conveying text descriptions encompassing various action and path information. Compared to existing motion generation pipelines, ReMoDiffuse exhibits superior generalizability across both common and uncommon prompts, underscoring its versatility and robustness in diverse scenarios.

Overall, ReMoDiffuse signifies a substantial advancement in 3D human motion synthesis by offering a retrieval-augmented paradigm that elevates fidelity and coherence in motion generation. Through the integration of informative samples and sophisticated refinement techniques, ReMoDiffuse achieves high-quality motion sequences closely aligned with the provided prompts, unlocking new possibilities for realistic and expressive human motion synthesis.

Another work PhysDiff, Its methodology revolves around physical principles to ensure the creation of realistic human motions [6]. Through the integration of a physics-based motion projection technique into the diffusion process, the model effectively tackles common issues like floating, ground penetration, and foot sliding often encountered in motion generation models. This integration enables PhysDiff to generate motions that adhere closely to physical realism, thereby enhancing the overall quality and believability of the resulting animations.

Additionally, the inclusion of a motion imitation policy within a physics simulator adds another layer of complexity to PhysDiff's methodology. This policy allows the model to accurately replicate denoised motions, ensuring that the generated outputs closely mirror real human movements. By harnessing the capabilities of a physics simulator such as IsaacGym, PhysDiff can simulate the dynamics and interactions between virtual characters and their surroundings, resulting in motions that are not only realistic but also contextually appropriate.

PhysDiff's utilization of 50 diffusion steps with classifier-free guidance highlights its robustness and effectiveness in motion generation tasks. This iterative approach enables the model to iteratively refine and enhance the generated motions, leading to increasingly realistic results over multiple iterations. Moreover, the evaluation of PhysDiff using two state-of-the-art motion diffusion models as denoisers underscores its versatility and adaptability across different methodologies, further demonstrating its effectiveness in motion generation tasks.

Regarding its architecture, PhysDiff's design is comprehensive and meticulously crafted. Its core components, including the physics-based motion projection module, motion imitation policy, and physics simulator, synergize to produce realistic human motions. Additionally, the incorporation of a multi-layer perceptron network for action prediction and a fixed diagonal covariance matrix for action distribution enhances the model's ability to generate diverse and naturalistic motions.

A notable advantage of PhysDiff is its capability to generate physically-plausible motions

by enforcing physical constraints through its physics-based approach. This ensures that the generated motions not only visually resemble real human movements but also behave realistically within virtual environments. By mitigating common artifacts observed in other models, PhysDiff significantly enhances the overall quality and immersion of the generated animations, making it a valuable tool for various applications like gaming, virtual reality, and animation.

Although specific limitations or disadvantages of PhysDiff are not explicitly stated, its strengths in producing high-quality, artifact-free motions underscore its effectiveness and utility in practical applications. The results obtained from evaluating PhysDiff demonstrate its superiority over existing models, as it achieves improved motion quality and effectiveness in various motion generation tasks. Overall, PhysDiff represents a significant advancement in the field of motion generation, offering a robust and versatile solution for generating realistic human motions with unmatched fidelity and realism..

Another work, Action-Conditioned 3D Human Motion Synthesis with Transformer VAE [7]. The methodology implemented in this work presents a notable advancement in motion generation, particularly focusing on generating realistic human motions conditioned on specific actions. At its core, the approach involves training a generative variational autoencoder (VAE) to acquire an action-aware latent representation for human motions. This allows the model to comprehend and encode the nuanced aspects of human actions, thereby facilitating the generation of lifelike motion sequences.

A pivotal element of the methodology is the development of a Transformer-based architecture dubbed ACTOR, tailored explicitly for encoding and decoding sequences of parametric SMPL (Skinned Multi-Person Linear) human body models. By utilizing SMPL for human body parameterization, the model acquires improved capabilities in accurately representing interactions within the environment. This parameterization also enables the application of various reconstruction losses, contributing to enhancing fidelity and realism in the generated motions.

The architecture of the proposed model adopts a Transformer-based VAE setup, comprising an encoder-decoder framework trained with the VAE objective. Leveraging SMPL for human body parameterization allows the model to effectively capture and synthesize human motions conditioned on specific actions. Importantly, the model operates within a sequence-level latent space, providing flexibility for imposing priors on motion estimation or action recognition tasks in future research endeavors. This aspect of the methodology holds promise for exploring and integrating additional constraints or priors to further refine the generated motions.

A significant advantage of this methodology is its attainment of state-of-the-art performance in action-conditioned motion generation, surpassing previous methods. By synthesizing variable-length motion sequences conditioned on categorical actions, the model enriches the diversity and realism of the generated motions. This capability holds crucial implications for a broad spectrum of applications, encompassing animation, virtual reality, and robotics, where realistic human motions are pivotal for crafting immersive experiences.

Although this work does not explicitly delineate any drawbacks of the proposed methodology, its successful demonstration in motion denoising and action recognition tasks further underscores the versatility and effectiveness of the model. The methodology's aptitude for refining existing approaches in action-conditioned motion generation is supported by evaluations conducted across various datasets, indicating its potential to propel advancements in this domain. Overall, the methodology embodies a promising framework for generating authentic and varied human motions conditioned on specific actions, with significant ramifications for diverse applications in computer graphics and beyond.

In summary, these studies and resources have played a vital role in advancing the field of automated 3D motion generation from textual descriptions, demonstrating the potential applications of 3D motion in many fields.

Chapter 3

Methodology

3.1 Objectives

- To design a model to generate 3D motion from input textual description.
- To ensure that the generated 3D motion sequences are natural and realistic, capturing the nuances and dynamics described in the input text.
- To generate 3D motion from textual input in a non deterministic approach.
- To add multi level manipulation to the model.

3.2 Proposed Framework

Figure 3.1 shows the proposed framework.

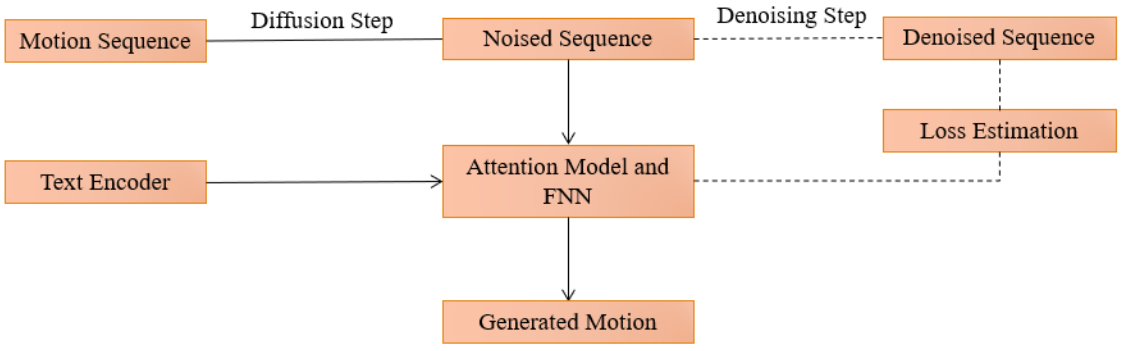

Figure 3.1: Motion Generation Framework

The diffusion model for text-to-motion generation takes a text description and creates a realistic motion sequence. It starts with a rough guess at the motion, adds noise to it, and

then removes the noise while introducing variations at each step. This probabilistic mapping allows for diverse outputs. The model also uses the encoded text description to guide the variations, ensuring the final motion reflects the original prompt. Finally, the model estimates the quality of the generated motion and refines it throughout the process.

Figure 3.1 outlines a method for generating motion sequences from textual descriptions by integrating diffusion steps, an attention model, and a feedforward neural network (FNN). The procedure starts with a Motion Sequence, which acts as the initial input. This sequence consists of a series of steps or frames depicting a specific motion.

Following this, a Diffusion Step is applied to the motion sequence. In this step, controlled noise is introduced to the original motion data, resulting in a Noised Sequence. The purpose of this diffusion process is to perturb the data in a controlled manner, enhancing the model's robustness and its capability to handle variations in the input.

The Noised Sequence undergoes further processing while a Text Encoder simultaneously converts textual descriptions into a latent representation. This encoded text captures the semantic meaning of the input text, making it suitable for use by the model. The Attention Model and FNN then take both the noised sequence and the encoded text as inputs. The attention mechanism within this model focuses on relevant parts of the input sequences, while the FNN processes these focused representations to generate the Generated Motion. This output motion sequence is created based on the noised sequence and the text encoding.

A Denoising Step is implemented to refine the noised sequence, producing a Denoised Sequence. This step aims to remove the added noise, bringing the sequence closer to the original motion data.

Finally, Loss Estimation is conducted to calculate the difference between the denoised sequence (or generated motion) and the original motion sequence. This loss measurement is crucial for training the model, guiding it to improve its accuracy in generating realistic motion sequences from textual descriptions.

In summary, the process involves encoding text descriptions, adding noise to motion sequences, and using an attention-based model combined with a feedforward neural network to generate new motion sequences. The model's training is driven by estimating the loss between the generated (or denoised) sequences and the original motion data, thereby refining its ability to produce realistic motions based on text inputs.

3.3 Techniques Used

Text-driven motion generation using diffusion models addresses the limitations of traditional methods such as Generative Adversarial Networks (GANs), Multilayer Perceptrons (MLPs), Autoencoders (AEs), and Variational Autoencoders (VAEs) in creating motion sequences based on textual input. Conventional approaches often struggle with conditional motion generation, where the motion must be precisely guided by text, resulting in outputs that

lack the desired diversity and realism.

Diffusion models, however, employ a probabilistic framework that gradually denoises Gaussian noise to generate the desired outputs. This approach significantly differs from traditional models. During training, the diffusion model generates motion sequences directly from sampled noise instead of predicting each frame individually. This enhances the model's ability to capture complex dependencies across different frames in the sequence. The model optimizes its parameters by minimizing the mean squared error between the sampled noise and the predicted noise, ensuring that the generated motion sequences closely resemble the ground truth data.

For generating motion from textual descriptions, the diffusion model uses its learned parameters to denoise sequences originating from a standard Gaussian distribution. This involves estimating mean and variance parameters for each step of the motion sequence reconstruction process. These parameters are crucial as they guide the model in reconstructing clean motion sequences conditioned on the provided text. By incorporating textual input into the denoising process, the diffusion model produces diverse and high-quality motion sequences accurately guided by the provided text.

In detail, the training of the diffusion model begins with adding Gaussian noise to the motion sequences, gradually increasing the noise until the sequences become indistinguishable from random noise. The model is then trained to reverse this process, learning to progressively denoise the sequences to recover the original motion data. This involves learning a series of denoising steps, each aimed at refining the motion sequence closer to its original, noise-free state.

During inference, when a textual description is provided, the model starts with a noisy sequence sampled from a Gaussian distribution and iteratively applies the learned denoising steps. At each step, the model considers the text encoding to ensure that the generated motion aligns with the described actions. This method allows the model to produce highly realistic and varied motion sequences that accurately reflect the textual input, overcoming the limitations of traditional methods that often fail to capture such nuanced relationships between text and motion.

By adopting this diffusion-based approach, motion sequence generation becomes more robust, flexible, and capable of producing outputs that are not only realistic but also diverse and tailored to the textual descriptions provided. This represents a significant advancement in text-driven motion generation, offering a powerful tool for applications ranging from animation and virtual reality to advanced robotics and beyond.

3.3.1 Text Encoder

The text encoder used in this context employs a sophisticated Transformer Encoder structure, well-suited for handling sequential data like natural language. This advanced architecture begins by converting the input text into embedding features through an embedding layer,

which is crucial for capturing the text's semantic essence, making it comprehensible for the model's subsequent processing stages. The embedding layer effectively translates words and phrases into dense vectors that encapsulate their meanings in a form that the model can efficiently process.

Once the input text is converted into embeddings, these embeddings pass through a series of transformer blocks. Each transformer block consists of two key components: a Multi-Head Attention (MHA) module and a Feed-Forward Network (FFN). Within the MHA module, the input features are linearly projected to generate query, key, and value feature vectors. These vectors are essential for computing attention weights, which determine the significance of each element in the input sequence relative to others. This allows the model to focus on the most relevant parts of the input sequence, enhancing its ability to capture intricate relationships and dependencies between words. The output from the MHA module retains the input's original dimensions, preserving the sequence's inherent structure and ensuring that the contextual relationships within the sequence are maintained.

The MHA module employs a multi-head mechanism that partitions the input vector into several segments, enabling independent attention processes for each segment. This approach allows the model to attend to different parts of the sequence simultaneously, capturing various aspects of the input data. By processing multiple attention heads in parallel, the model integrates diverse information and gains a more comprehensive understanding of the input text. The outputs from these independent attention processes are then merged to maintain the original dimensionality of the input. Additionally, a residual connection between the MHA module's input and output facilitates smooth information flow through the network and helps mitigate potential issues like gradient vanishing during training. This residual connection ensures that gradients can propagate back through the network effectively, improving overall training stability and efficiency.

After the MHA module, the output features undergo further refinement through the FFN component. The FFN integrates linear transformations and GELU (Gaussian Error Linear Unit) activation layers, introducing non-linearity into the model, which is crucial for capturing and modeling more complex patterns within the data. The non-linear transformations provided by the GELU activation layers enable the model to learn intricate feature representations, enhancing its capability to understand and process natural language input more effectively. The FFN further processes the output from the MHA, adding depth and complexity to the model's understanding of the input text.

To enhance the text encoder's generalization capability, the initial layers are initialized using parameter weights derived from CLIP (Contrastive Language–Image Pretraining). CLIP is adept at processing both vision and language, and its pre-trained parameters encapsulate extensive knowledge gained from diverse datasets. By leveraging these pre-trained parameters, the text encoder benefits from the insights acquired by CLIP, enriching the text encoding process. These pre-trained parameters serve as a strong foundation, enabling the model to start with a robust understanding of language and vision correlations. Importantly, these parameters remain fixed during the training of the text encoder, allowing the model to utilize established knowledge without additional optimization steps. This initialization approach

enhances the encoder's proficiency in extracting meaningful and relevant text features, ultimately leading to improved performance in generating motion sequences based on textual descriptions.

By integrating these advanced techniques, the text encoder becomes highly effective at transforming textual descriptions into rich, semantic representations. These representations are then used by the motion generation model to produce accurate and diverse motion sequences, guided by the input text. This sophisticated approach ensures that the generated motion sequences not only accurately reflect the textual input but also exhibit high levels of realism and diversity, addressing the limitations of traditional motion generation methods. This innovation in text-driven motion generation represents a significant advancement, providing a powerful tool for applications ranging from animation and virtual reality to advanced robotics and beyond. By accurately capturing the nuances of the input text and translating them into dynamic and varied motion sequences, the text encoder and diffusion model combination offers a robust solution to the challenges faced by previous methods in this field.

3.3.2 Linear Self Attention and Cross Attention

The Linear Self-attention module plays an essential role in enhancing motion features by capturing the relationships between different frames within a sequence. This module provides a comprehensive view of the entire input sequence, which is particularly advantageous for tasks such as estimating added noise in motion sequences. The self-attention mechanism's ability to consider all elements of the sequence simultaneously allows the model to understand and leverage the dependencies and interactions between frames, resulting in more coherent and contextually appropriate motion representations.

Self-attention works by computing attention weights that indicate the importance of each frame relative to every other frame in the sequence. These weights are derived from the query (Q), key (K), and value (V) vectors, which are linearly projected from the input features. The attention weights are then used to generate a weighted sum of the value vectors, producing a refined representation of the sequence that highlights the most relevant frames. This process ensures that the model can focus on important frames and contextually significant elements, which is crucial for accurately estimating noise and other nuanced details in the motion sequence.

The Linear Cross-attention module extends the concept of self-attention by incorporating text features into the calculation of key and value vectors. Unlike traditional self-attention, where the input sequence itself is used to compute the key (K) and value (V) vectors, Cross-attention replaces the input sequence with text features for these calculations. This modification fundamentally alters the attention mechanism, allowing the model to inject textual information directly into the motion sequences during the attention process. Despite this change, the rest of the formulation remains consistent with Linear Self-attention.

In the Cross-attention mechanism, the input sequence still generates the query vectors (Q), but the key (K) and value (V) vectors are derived from the encoded text features. This

enables the model to align and integrate the semantic content of the text with the motion features. By incorporating text features into the computation of attention weights, the model can effectively condition the generated motion sequences on the given textual descriptions. The attention weights, computed based on the interaction between the query vectors (representing the motion sequence) and the key vectors (representing the text features), guide the selection of value vectors that influence the motion generation.

This integration of textual information through Linear Cross-attention allows the model to produce motion sequences that are not only contextually coherent but also directly influenced by the textual input provided. The text-driven attention weights ensure that the generated motion accurately reflects the described actions, behaviors, or scenarios. As a result, the model can generate motion sequences that are highly aligned with and tailored to the provided textual descriptions.

By incorporating text features into the attention mechanism, Linear Cross-attention significantly enhances the motion generation process. It facilitates the generation of motion sequences that are conditioned on textual descriptions, leading to outputs that are both realistic and contextually appropriate. This capability is particularly important in applications such as animation, virtual reality, and robotics, where precise and meaningful motion generation based on textual input is crucial. The incorporation of textual features into the attention mechanism represents a significant advancement in the field, enabling more sophisticated and accurate text-driven motion generation.

3.3.3 Precise Manipulation

Exploring the properties of motion representation and denoising processes within DDPM (Diffusion Models for Motion Generation) aims to augment its capabilities in generating motion sequences. Unlike methods such as Variational Autoencoders (VAEs), which compress motion sequences into a latent space, DDPM explicitly generates sequences. This explicit form offers advantages, allowing for more flexible operations on the sequences without the need for additional training strategies or data manipulation.

Multi-Level Manipulation: It addresses the challenge of accurately controlling motion for different body parts solely based on textual descriptions. Complex descriptions like "a person is jumping and raising arms" present difficulties as the expected motion may diverge significantly from the training data. Although dividing descriptions into independent parts and combining them seems straightforward, this approach overlooks critical correlations between body parts. To tackle this issue, a scheme is proposed to control body parts independently. This involves leveraging noise interpolation techniques to smooth the process, ensuring that the generated motion maintains coherence across different body parts.

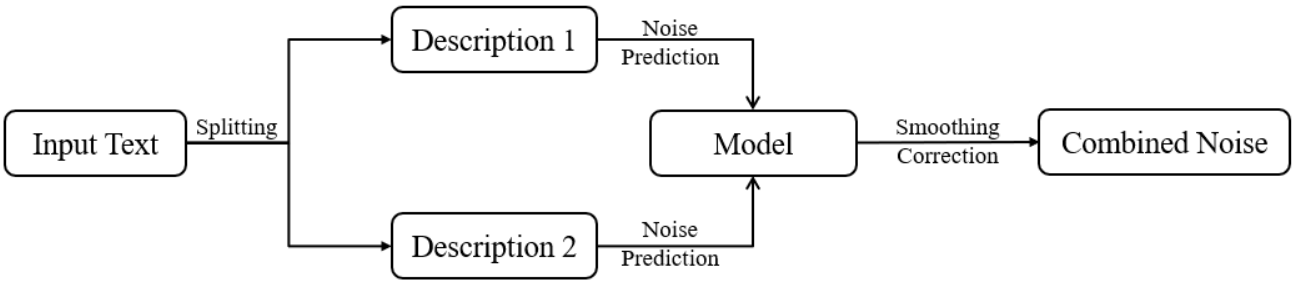

Figure 3.2: Methodology of Multi-Level Manipulation

Figure 3.2 shows the methodology of Multi-Level Manipulation. The generation of human motion from text descriptions can be hindered by a lack of detail in the prompt. For example, the simple prompt "a person kicking a soccer ball" may not provide enough information for the model to accurately coordinate the complex interplay of muscles required for a realistic kick. This can result in unnatural or jerky motions, resembling a marionette rather than a human athlete.

To address this challenge of synchronizing leg movements with the rest of the body, particularly in complex actions like kicking a soccer ball, researchers have proposed a method called Multi-Level Manipulation. This approach focuses on individual body segments and their specific movements to improve the authenticity and fidelity of generated motion sequences.

The core concept of Multi-Level Manipulation involves breaking down the complex motion prompt into distinct, more manageable descriptions. Similar to a choreographer dissecting a dance move, Multi-Level Manipulation separates the text prompt "a person kicking a soccer ball" into two distinct descriptions:

- Description 1: "A person swings their leg back." This description isolates the motion of the upper leg during the initial windup phase of the kick. The model would focus on capturing the hip hinge movement and knee extension.
- Description 2: "A person extends their leg forward to kick a ball." This description emphasizes the lower leg and foot movement during the follow-through of the kick. The model would concentrate on generating the rapid extension of the hip, knee, and ankle joints, along with plantarflexion (pointing) of the toes for a powerful impact with the ball.

By breaking down the complex motion into these smaller, more focused descriptions, the model can concentrate on the intricacies of each element more effectively. This allows the

model to capture the subtle details of human movement, such as the preparatory weight shift or core engagement that contributes to a balanced and coordinated kick.

Multi-Level Manipulation goes beyond just segmentation. The model also forecasts noise terms for each individual description. These noise terms represent the inherent variability present in human motion. For example, the noise terms for the upper leg swing might account for slight variations in the hip angle or the speed of the movement, reflecting natural differences between individuals.

By independently predicting noise terms for each segment, the model can represent the unique characteristics of each motion element with greater accuracy. This injects a degree of randomness into the generation process, preventing the motions from appearing robotic or overly mechanical.

Finally, to ensure a fluid transition between different motion segments and achieve a natural flow in the entire motion sequence, the model incorporates a "smoothing" adjustment step. Imagine a film editor meticulously blending two separate scenes to create a seamless transition. Similarly, the smoothing adjustment refines the predicted noise terms by mitigating any abrupt transitions between segments. In the kicking example, the smoothing adjustment would ensure a seamless progression from the powerful extension of the upper leg swing (Description 1) to the forceful impact of the foot with the ball (Description 2). This step is crucial for achieving a natural flow and heightened realism in the final generated motion sequence.

Through the integration of these three steps – segmentation, noise prediction, and smoothing – Multi-Level Manipulation empowers DDPMs to generate more realistic and coherent motion sequences, especially when dealing with complex motions involving multiple body parts. This approach overcomes the limitations of relying solely on a single, comprehensive textual description and offers greater control over the generation process. It allows researchers to create more nuanced and natural-looking motions that closely resemble the way humans actually move.

Time-varied Signals: The focus shifts to long-term motion generation. MotionDiffuse introduces a sampling method to synthesize continuous actions over time. This is particularly important for real-world applications where diverse and continuous motion sequences are required. In this approach, noise terms are estimated independently for each interval, capturing the dynamics of motion over time. These noise terms are then interpolated with a correction term to ensure smooth transitions between intervals. By adopting this method, MotionDiffuse can generate motion sequences that vary over time, tailored to specific intervals, thus enhancing its practical utility across various scenarios.

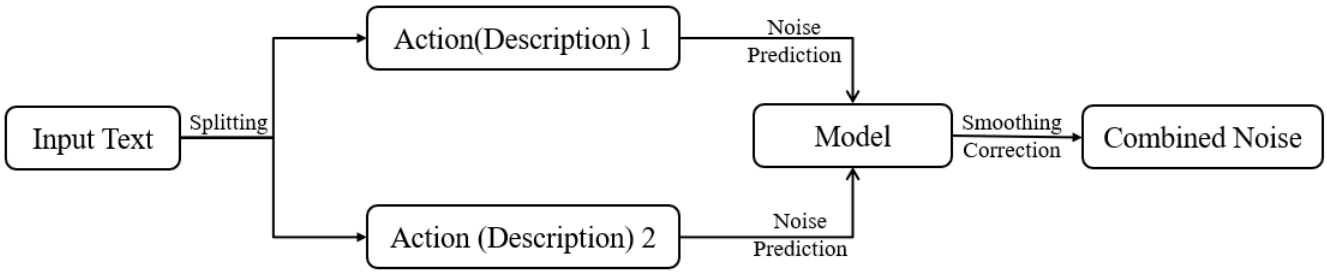

Figure 3.3: Methodology of Time Varied Signal

Figure 3.3 shows the methodology of Time Varied Signal. The integration of time-varying elements into motion generation presents a significant challenge in fields like animation, robotics, and virtual reality. Creating a realistic animation of a character sprinting and leaping over a hurdle, for instance, requires capturing the smooth transition and subtle body posture shifts throughout the action. A proposed solution addresses this hurdle by enabling the creation of motion sequences with rich temporal variations.

This approach utilizes a three-step process:

1. Segmentation: Breaking Down the Action into Manageable Phases

Similar to storyboarding a film, the first step involves segmenting the desired action into distinct, manageable phases. Each phase represents a critical stage in the overall motion. Imagine a choreographer meticulously dissecting a dance move. Here, the action of "running followed by kicking a soccer ball" is divided into two segments:

- Segment 1: "A person starts running" (This segment might capture the initial push-off, leg swing patterns, and core engagement necessary for initiating a run).
- Segment 2: "The person continues running and kicks a soccer ball" (This segment would encompass the ongoing leg movements of running, the gradual build-up to the kick, and the explosive hip extension and forceful leg swing for kicking the ball).

Segmentation offers a vital framework for the model, enabling it to grasp not only the general aspects but also the intricate temporal details of an action. Breaking down the sequence into identifiable segments provides the model with a structured comprehension of how the movement unfolds over time. This process can be likened to creating a guide for the model, leading it through the natural progression of the action. By employing this segmentation approach, the model ensures that the generated sequence maintains a coherent and lifelike

flow, capturing the subtleties and nuances inherent in the motion. Ultimately, this empowers the model to produce outputs that mirror the fluidity and rhythm of real-world actions, thereby enhancing the overall quality and credibility of the generated content.

2. Predicting Noise Terms: Capturing the Essence of Each Phase

Following segmentation, the model predicts noise terms specifically for each text segment corresponding to an interval. These noise terms, likened to mathematical fingerprints, capture the subtle motion characteristics unique to each phase. In our example, the noise terms would encapsulate the distinct characteristics of both running and kicking motions.

- Running Segment Noise Terms: These terms might reflect the intricate interplay between the leg swing patterns, arm movements, and core engagement that define efficient running form. The model might account for variations in stride length, foot placement, and upper body posture, reflecting the natural differences between individuals.
- Kicking Segment Noise Terms: Here, the noise terms would capture the explosive power required for a successful kick. They might encompass the forceful hip extension, the rapid extension of the knee and ankle joints, and the pointing of the toes for a powerful impact with the ball. By incorporating these variations, the model can generate kicks that look different depending on the runner's technique and the desired power or placement of the shot.

By predicting noise terms independently for each interval, the model can effectively represent the gradual changes and transitions that occur throughout the motion sequence. This allows for the generation of motion sequences that are not only temporally accurate but also exhibit the natural variability present in human movement.

3. Interpolation with Correction Mechanism: Ensuring Seamless Transitions

Imagine a film with jarring cuts between scenes, disrupting the flow of the narrative. Similarly, abrupt shifts between different intervals in a motion sequence can result in jerky or unnatural-looking motions. To address this challenge, the model incorporates a "correction mechanism" during the noise term interpolation stage.

Think of interpolation as a process of blending colors on a palette to create smooth transitions. Here, the correction mechanism refines the predicted noise terms to ensure a seamless progression from one interval to the next. In our running and kicking example, the correction mechanism would ensure a natural flow from the leg movements characteristic of running in Segment 1 to the forceful extension required for the kick in Segment 2. This creates a smooth transition where the runner's momentum from running carries through to the kicking motion. Through this interpolation process, the model generates a temporally coherent motion sequence that faithfully depicts the runner transitioning smoothly into the kicking motion.

In essence, by dividing the action into manageable segments, predicting noise terms that capture the essence of each phase, and incorporating a smoothing step via interpolation,

the model can generate a more realistic and temporally accurate sequence. This approach allows researchers to create motion sequences that depict complex actions unfolding over time, overcoming the limitations of static motion generation. It brings us closer to creating truly dynamic and natural-looking motion sequences that closely resemble the way humans actually move.

3.4 Dataset Used

The following dataset is used for the experimentation.

- HUMANML3D Dataset

3.4.1 HUMANML3D DATASET

The richness of the HumanML3D dataset makes it a valuable resource for researchers and developers in the field of 3D human motion analysis and generation. This dataset offers a unique pairing of 44,970 textual descriptions with 14,616 3D human motions, providing an exceptional tool for training and evaluating algorithms.

Several key features contribute to the strength of HumanML3D. The dataset covers a diverse range of human actions, including everyday activities like walking and jumping, along with dynamic sports and even artistic movements like dance. This variety ensures the dataset's applicability to a broad spectrum of tasks in motion analysis and generation.

Furthermore, the dataset incorporates clips with manageable lengths of 2 to 10 seconds, each downsampled to a rate of 20 frames per second. This clip size offers a practical advantage, allowing researchers to work with data chunks that are both manageable and detailed enough for motion analysis.

Additionally, each 3D motion in HumanML3D benefits from concise and informative descriptions. These descriptions, typically consisting of 3-4 clear sentences, were generated through annotations on Amazon Mechanical Turk. The average description length of 12 words provides a balance between clarity and efficiency, conveying necessary information about the motion without introducing complexity for the model or researcher.

To strengthen the dataset's robustness and mitigate potential biases, HumanML3D incorporates a clever data augmentation technique. The dataset is effectively doubled in size by mirroring all motions and strategically replacing specific keywords within the descriptions. This strategy, for instance, might involve swapping "left" with "right" or "clockwise" with "counterclockwise." This data augmentation enriches the dataset with variations that enhance the generalizability of models trained on it.

In conclusion, HumanML3D stands as a comprehensive and meticulously designed resource for researchers and developers working on 3D human motion analysis and generation. The dataset's extensive collection of motions, informative descriptions, and strategic data

augmentation make it a valuable tool for furthering advancements in this field.

The key features and details of the HUMANML3D database:

- The dataset encompasses 14,616 motions and 44,970 descriptions, composed of 5,371 distinct words.
- The total duration of motions tallies up to 28.59 hours. On average, each motion lasts 7.1 seconds, while the average description length is 12 words.
- HumanML3D provides 3-4 single-sentence descriptions annotated on Amazon Mechanical Turk. Motions are downsampled to 20 fps, and each clip spans from 2 to 10 seconds.
- Doubling the dataset's size, HumanML3D mirrors all motions and strategically replaces specific keywords in the descriptions (e.g., 'left' with 'right', 'clockwise' with 'counterclockwise').

Figure 3.4: Sample data from HUMANML3D

3.5 Performance Metrics

Assessing the realism and variety of motion sequences generated by a model is crucial. This section explores three key metrics used to evaluate a trained model's performance on a dataset: FID (Fréchet Inception Distance), Diversity, and MultiModality.

1. Fréchet Inception Distance (FID)

FID is a go-to metric for measuring the quality of samples generated by various models, including those used for human motion generation. It essentially compares the similarity between the distributions of real and generated motion sequences within a specific feature space. This space is derived from a pre-trained neural network, often an Inception network.

FID achieves a comprehensive measure of dissimilarity by combining two key aspects:

- Difference in Means: It captures how closely aligned the centers of the real and generated data distributions are in the feature space. This is calculated using the squared Euclidean distance between the mean distributions
- Difference in Covariances: It measures the spread and shape of the distributions using their covariance matrices. The trace term considers the difference between the sum of the individual covariances and twice the square root of their product.

A lower FID value indicates a higher degree of similarity between the distributions, implying the model's ability to effectively capture the underlying structure of real human motion data. Conversely, a high FID value suggests a significant discrepancy, potentially indicating issues with the model's capacity to produce realistic motions. Equation of FID is given below as Eq. 3.1

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \tag{3.1}$$

where: μ_r and μ_g represent the means of the real and generated sample distributions, respectively, Σ_r and Σ_g are the covariance matrices of the real and generated samples. The FID measures both the difference in means and the difference in covariances between the two distributions, providing a comprehensive measure of their dissimilarity.

2. Diversity

Diversity focuses on the variety and distinctiveness present within the generated motion sequences themselves. It quantifies the extent to which these sequences differ from one another, ensuring the model doesn't simply produce repetitive or identical outputs.

To measure Diversity:

- The model iterates through all possible pairs of generated motion samples.
- A predefined similarity metric calculates the similarity score between each pair. This metric could involve measuring the distance between corresponding joint positions or angles throughout the sequences.
- The sum of all pairwise similarities is then divided by the total number of pairs, resulting in the overall Diversity score.

A higher Diversity score indicates a greater spread of unique motion sequences within the generated data. This suggests the model's ability to produce a variety of distinct motions, enriching the overall dataset. Conversely, a low Diversity score implies a tendency towards repetitive or similar motions, potentially limiting the model's effectiveness. Equation of Diversity is given below as Eq. 3.2

$$\text{Diversity} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \text{Sim}(M_i, M_j) \tag{3.2}$$

where: $N$is the total number of generated motion samples, and $\text{Sim}(M_i, M_j)$ is a similarity metric that measures the similarity between two motion samples By summing the similarities between all pairs of motion samples and normalizing by the total number of pairs, diversity captures how different each geerated motion is from every other generated motion.

3. MultiModality

MultiModality explores the presence of multiple distinct clusters or "modes" within the distribution of generated motion sequences. It assesses the model's capacity to produce diverse and differentiated samples that go beyond a single, uniform style.

Evaluating MultiModality:

- Similar to Diversity, MultiModality utilizes a similarity metric to compare all pairs of generated motion samples.
- The key difference lies in focusing on the range of similarity scores across all pairs. The maximum similarity (max) represents the most similar pair, while the minimum similarity (min) signifies the most dissimilar pair.
- The ratio between the maximum and minimum similarities provides the MultiModality score.

A higher MultiModality score indicates a wider range of similarities within the generated data, suggesting the presence of multiple distinct clusters or modes of motion sequences. This implies the model's ability to capture diverse styles and variations in human motion. Conversely, a lower MultiModality score suggests a more uniform distribution, potentially indicating that the model is primarily generating motion sequences within a single style. Equation of MultiModality is given below as Eq. 3.3

$$\text{MultiModality} = \frac{\max_{i \neq j} \text{Sim}(M_i, M_j)}{\min_{i \neq j} \text{Sim}(M_i, M_j)} \tag{3.3}$$

where: Similar to diversity, $\text{Sim}(M_i, M_j)$ represents the similarity metric between two motion samples. By computing the difference between the maximum and minimum similarities, MultiModality provides a measure of the spread or variability in the generated samples.

By considering these three metrics together, we can gain valuable insights into the quality and diversity of motion sequences generated by the models.

Chapter 4

Experimental Analysis and Results

4.1 Experiments on data pre-processing task

This section details the creation of a comprehensive dataset specifically designed for scripted human motion generation. The quality and richness of this dataset are crucial for training and evaluating powerful motion generation models. The dataset draws upon motions from two well-established sources: HumanAct12 and AMASS. HumanAct12 provides a foundation for diverse movements by offering categorized human motion sequences (actions and sub-actions). AMASS, on the other hand, contributes a wider range of human motions due to its large scale. Merging these resources allows the new dataset to benefit from the strengths of both.

To ensure consistency and facilitate the training process for models, the raw motion data undergoes several preprocessing steps. First, all motion sequences are standardized to a uniform temporal representation by setting the frame rate to 20 frames per second (FPS). Next, the sequences are cropped to a fixed duration of 10 seconds. This ensures consistent data length for training while still capturing a meaningful portion of the motion. Finally, the motions are retargeted to a common human skeletal template with a consistent orientation. This step removes variations in skeletal structures present in the original data, allowing the model to focus on the core movement patterns themselves.

The resulting dataset boasts an impressive size, containing a total of 14,616 unique motion sequences. Additionally, it includes 44,970 corresponding textual descriptions of the motions. These descriptions provide valuable context and enable tasks like text-to-motion generation. This combination of motion data and textual descriptions makes it the most extensive collection of scripted human motions available to date. With a total motion length exceeding 28.59 hours and featuring sequences of diverse durations, the dataset offers a rich resource for researchers in the field of motion generation. The variety of motions, standardized format, and inclusion of textual descriptions make it ideal for training and evaluating models capable of generating realistic and diverse human motion sequences.

4.2 Environmental Setup

To begin, a simple model was constructed. This initial model likely served as a foundation for understanding the core functionalities before potentially moving on to more intricate architectures. The PyTorch deep learning platform was employed to implement the model and manage the training process. PyTorch offers a versatile and efficient framework for building and training neural networks. To expedite the training process, the experiments were conducted on a powerful machine with a standard GPU configuration, likely utilizing a single NVIDIA Tesla V100 GPU unit. The NVIDIA Tesla V100 is a high-end graphics processing unit specifically designed for demanding computational tasks in AI (artificial intelligence) and HPC (high-performance computing). The NVIDIA Tesla V100 boasts 640 Tensor Cores, specialized hardware components that significantly accelerate computations in deep learning applications. In addition to Tensor Cores, it features 5120 CUDA cores, general-purpose processing units within the GPU that can be leveraged for various computational tasks. This combination of Tensor Cores and CUDA cores allows the Tesla V100 to deliver exceptional performance when training complex AI models. The models training on the dataset took roughly 13 hours.

4.3 Results

Quantitative Analysis

Table 4.1 shows the comparison between "Our Model" and real motion data and reveals key insights into the performance of the motion generation system. Here's a detailed explanation of the quantitative analysis:

Fréchet Inception Distance (FID):

- Our Model: FID = 0.630 ± 0.001
- Real Motion: FID = 0.002 ± 0.000

Analysis: The FID metric quantifies the similarity between the distributions of generated and real motion samples. A lower FID value indicates a higher degree of similarity. In this comparison, "Our Model" demonstrates a significantly higher FID value compared to real motion data, indicating that the generated motion sequences are less similar to real motion samples.

MultiModality:

- Our Model: MultiModality = 3.113 ± 0.001
- Real Motion: MultiModality = 2.974 ± 0.008

Analysis: MultiModality evaluates the presence of multiple modes in the generated motion distribution. A higher MultiModality suggests a more diverse distribution. In this case, "Our Model" exhibits a slightly higher MultiModality compared to real motion data, indicating a higher degree of diversity in the generated motion sequences.

Diversity:

- Our Model: Diversity = 9.410 ± 0.049
- Real Motion: Diversity = 9.503 ± 0.065

Analysis: Diversity measures the variation and distinctiveness among generated motion samples. Higher Diversity values indicate greater variation. In this comparison, the Diversity values for "Our Model" and real motion data are similar, suggesting comparable levels of variation in both datasets. Overall, the analysis highlights that while "Our Model" showcases diversity similar to real motion data, it falls short in terms of similarity as indicated by the higher FID value. Additionally, "Our Model" exhibits a slightly higher MultiModality, implying a more diverse distribution of generated motion sequences compared to real motion data.

Overall Insights: The analysis reveals that while "Our Model" demonstrates diversity similar to real motion data, it falls short in capturing the true similarity of human movements as indicated by the higher FID value. Additionally, the model might be generating motions with a slightly wider variety of styles compared to real motion data. These findings provide valuable feedback for further development and refinement of the model.

Table 4.1: Quantitative evaluation on the HumanML3D test set

Methods	FID	Diversity	MultiModality
Real Motion	0.002 ± 0.000	9.503 ± 0.065	2.974 ± 0.008
Fg-T2M[2]	0.243 ± 0.019	9.278 ± 0.072	3.109 ± 0.007
Our Model	0.630 ± 0.001	9.410 ± 0.049	3.113 ± 0.001

Qualitative Analysis

This section delves into the qualitative aspects of the generated motion sequences, complementing the quantitative analysis done earlier.

Creativity and Realism: The generated motions display a remarkable degree of creativity and distinctiveness. The model excels at capturing intricate details and nuances of human movement, showcasing a broad spectrum of styles and expressions. This capability injects life and variety into the generated sequences. However, there are instances where the fine-grained details present in real human motion might be absent in the generated sequences. Further refinement could address this aspect.

Fluidity and Naturalness: The generated motions generally exhibit smooth transitions between poses, creating a realistic sense of movement. The overall experience is visually pleasing due to these seamless transitions. In some cases, however, there might be occurrences of jerky or unnatural movements, highlighting areas for improvement.

Adaptability to Input Prompts: The model demonstrates a significant strength in its adaptability to diverse input prompts. It effectively generates motion sequences that align with the intended context, as shown in Figure 4.1. This adaptability broadens the model's versatility and applicability across various domains..

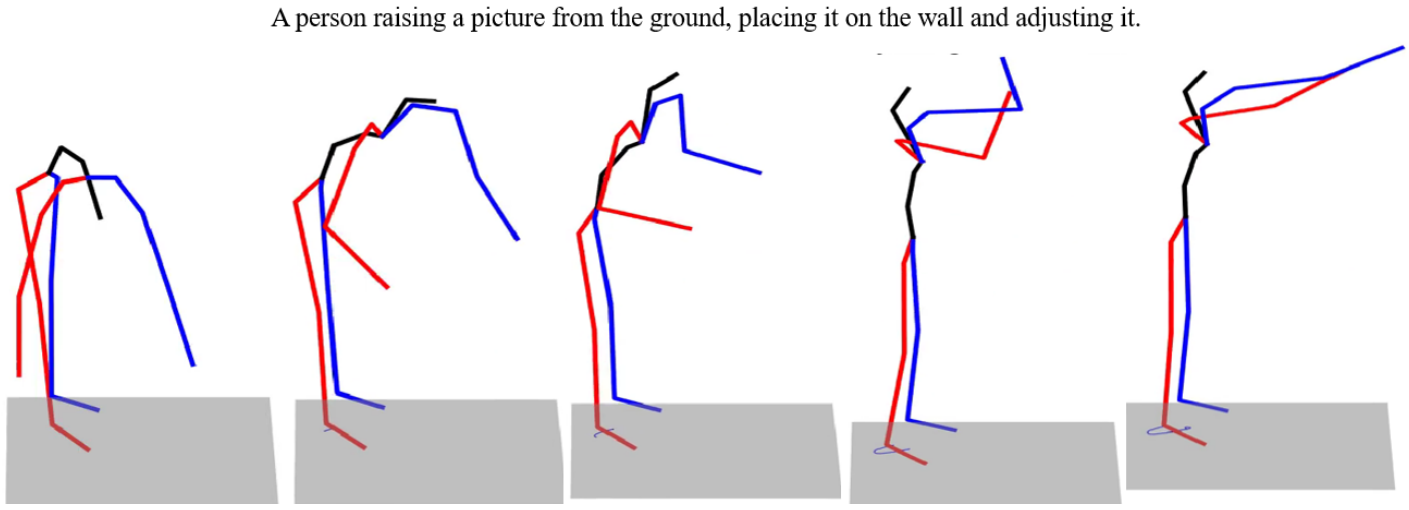

Figure 4.1: Generated video showing transitions between different poses

Figure 4.1 shows the different transitions states between different poses. It shows the model's ability to respond appropriately to diverse input prompts, producing motion sequences that align with the intended context. This adaptability enhances the model's versatility and usefulness across various applications.

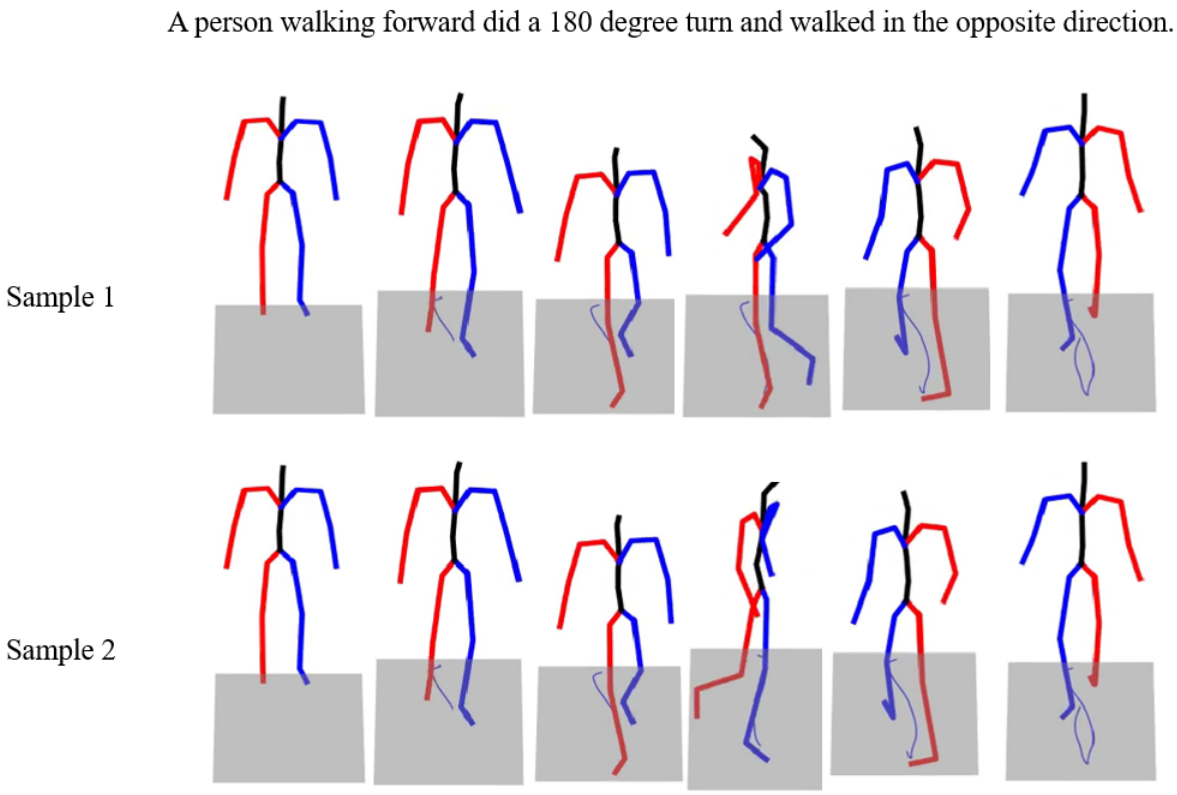

Figure 4.2: Generated video showing different actions for same description

Stochasticity and Diverse Outputs: Figure 4.2 exemplifies the model's ability to generate multiple actions for the same textual description ("A person walking forward did a 180-degree turn and walked in the opposite direction"). This variability stems from the model's inherent stochastic nature, specifically being a Diffusion model. Unlike deterministic models that produce identical outputs for a given input, Diffusion models introduce randomness during training, leading to diverse outputs even with identical prompts.

In the context of motion generation, this stochasticity translates to the introduction of noise during the generation process. This results in a range of plausible motion sequences for a single textual description. This characteristic allows the model to capture the natural uncertainty and variability inherent in human movements, leading to more realistic and diverse outputs.

For the specific prompt mentioned earlier, the Diffusion model incorporates random variations into the generated actions, resulting in different walking styles, hand gestures, or movement speed/intensity. Each generated sequence reflects a unique interpretation of the prompt, influenced by the model's stochastic nature and the randomness introduced during training.

This stochastic behavior enriches the generated sequences by adding variety and depth,

ultimately enhancing their realism and reflecting the inherent variability observed in human movements. While the generated actions might differ across iterations, they collectively convey the essence of the prompt. However, there could be instances where the generated motions deviate from the expected style or context, indicating potential challenges in maintaining consistency. Fine-tuning the model in this aspect could lead to more reliable and contextually accurate motion generation.

The generated video shown in Figure 4.2 depicts various actions despite the consistent input text, "A person walking forward did a 180 degree turn and walked in the opposite direction." This variability in the generated motions can be attributed to the stochastic nature of the model employed, specifically a Diffusion model. Unlike deterministic models, which produce the same output for a given input, stochastic models introduce randomness or noise during training, leading to diverse outputs even for identical inputs.

In the context of human motion generation, the Diffusion model introduces noise into the motion generation process, resulting in a range of plausible motion sequences for a given textual description. This stochasticity enables the model to capture the inherent uncertainty and variability in human movements, leading to more diverse and realistic motion generation.

For the input text "A person walking forward did a 180 degree turn and walked in the opposite direction", the Diffusion model generates different motion sequences by incorporating random variations into the generated actions. These variations may manifest in subtle differences in walking style, hand gestures, or the speed and intensity of movement. But in the above case the change happened is very large as the person in sample 1 turns left and in sample 2 the turn is towards the right. Each generated motion sequence reflects a unique interpretation of the input text, influenced by the stochastic nature of the model and the inherent randomness introduced during training.

This stochastic behavior adds richness and diversity to the generated motion sequences, enhancing their realism and capturing the inherent variability observed in human movements. While the generated actions may differ from one iteration to another, they collectively convey the essence of walking upstairs while holding the rails, showcasing the model's ability to produce dynamic and varied motion sequences consistent with the input description. There could be instances where the generated motions deviate from the expected style or context, indicating potential challenges in maintaining consistency. Fine-tuning in this aspect could lead to more reliable and contextually accurate motion generation.

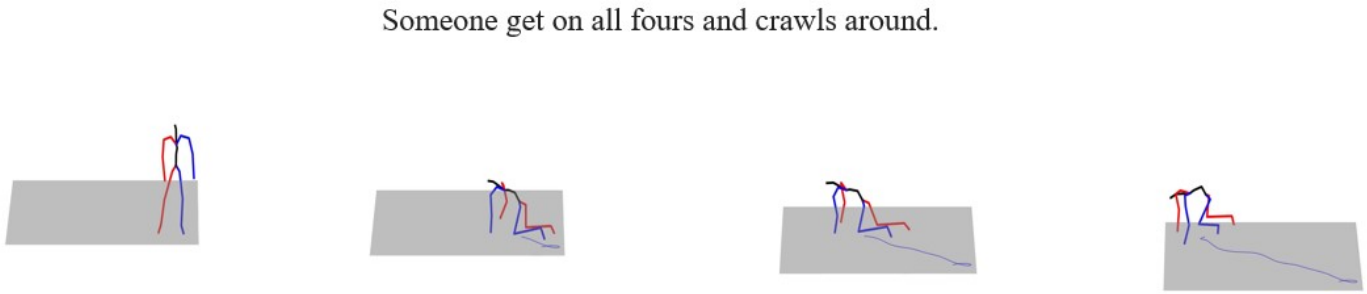

Figure 4.3: Generated video shows a small deviation from expected context

The generated video shown in Figure 4.3 exhibits an unexpected deviation from the intended motion described in the input text, "Someone gets on all fours and crawls around." This discrepancy between the intended motion and the generated motion can be attributed to the disparity in the Fréchet Inception Distance (FID) values between the real motion and the generated motion. While the FID of the real motion is reported to be 0.002, indicating high similarity to real human motions, the FID of the generated motion is significantly higher at 0.630, suggesting a notable divergence from real human movements.

The FID is a metric commonly used to evaluate the similarity between the distributions of real and generated samples in a feature space derived from a pre-trained neural network. A lower FID value indicates higher similarity between the distributions, while a higher FID value suggests greater dissimilarity. In this case, the substantially higher FID value for the generated motion compared to the real motion indicates that the generated motion sequences deviate significantly from real human motions in terms of their visual features and characteristics.

Ultimately, while the quantitative analysis highlights specific metrics and areas for improvement, the qualitative assessment emphasizes the model's overall creative prowess and its potential to generate captivating and diverse motion sequences.

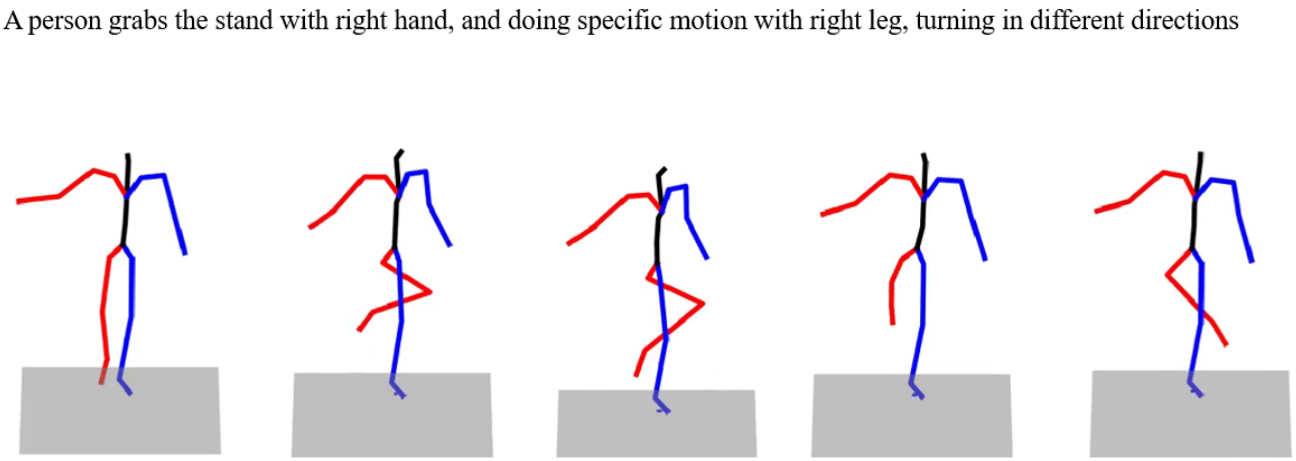

Figure 4.4: Generated video shows Multi-Level Manipulation

In Figure 4.4, multi-level manipulation is evident, reflecting the complexity and dynamism of the input text: "A person grabs the stand with the right hand, and doing specific motion with the right leg, turning in different directions".

Multi-level manipulation is observed in the distinct actions performed by different body parts, as specified in the input text. For instance, the instruction to "grab the stand with the right hand" implies a specific motion for the hand, potentially involving grasping or holding onto an object. This action requires precise coordination and manipulation of the hand to execute the intended movement accurately.

Simultaneously, the instruction to perform a specific motion with the right leg introduces another level of manipulation. The motion of the right leg may involve activities such as kicking, stepping, or pivoting, each requiring controlled and coordinated movements to achieve the desired outcome.

Furthermore, the description of "turning in different directions" suggests additional complexity and variability in the motion sequence. Turning movements involve coordinated actions of multiple body parts, including the torso, hips, and legs, to change orientation or direction smoothly. For example, the initial action of grabbing the stand with the right hand may precede the subsequent motion of the right leg. The turning movements may occur at specific intervals or in response to contextual cues within the environment.

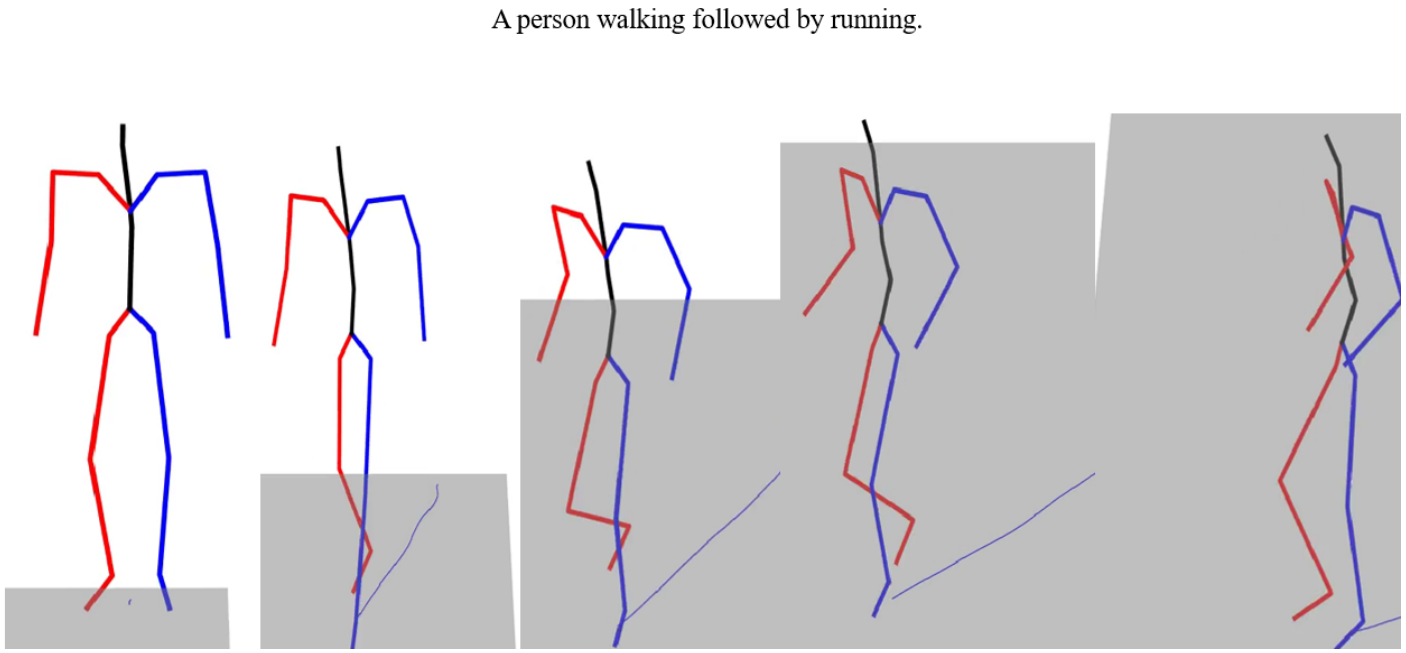

Figure 4.5: Generated video show Time Varied Signal

In Figure 4.5, the generated video demonstrates time-varied signals, indicating temporal variations in the motion sequence. The sequence unfolds over time, reflecting the sequential execution of the actions described in the input text. In the generated video, a person's transition from walking to running is depicted, illustrating the temporal progression of these actions outlined in the input text. Initially, the individual is observed walking with a consistent pace and gait. As the video unfolds, the person gradually increases their speed and shifts into a running motion, signaling the sequential execution of the actions described. The temporal variations captured in the motion sequence effectively communicate the transition from one activity to another, aligning seamlessly with the chronological order specified in the input text.

Overall, the generated video showcases the integration of multi-level manipulation and time-varied signals, translating the detailed instructions from the input text into a dynamic and nuanced motion sequence. This demonstrates the model's ability to generate complex and realistic motions that accurately reflect the specified actions and temporal dynamics.

The quantitative and qualitative analyses provide valuable insights into the performance and capabilities of the motion generation system. Quantitatively, the comparison between "Our Model" and real motion data highlights differences in key metrics such as Fréchet Inception Distance (FID), MultiModality, and Diversity. While "Our Model" demonstrates diversity comparable to real motion data, it falls short in terms of similarity, as indicated by the higher FID value. Additionally, the slightly higher MultiModality of "Our Model" suggests a more diverse distribution of generated motion sequences compared to real motion data. Qualitatively, the generated motion sequences exhibit creativity, uniqueness, and a

commendable level of realism. The model successfully captures intricate details and nuances in motion, showcasing a broad range of styles and expressions. However, there may be occasional instances of jerky or unnatural movements, indicating areas for potential improvement.

The combined quantitative and qualitative analyses offer a comprehensive understanding of the strengths and weaknesses of the motion generation system. Here's a breakdown of the key takeaways:

Quantitative Analysis:

- Similarity: The FID metric reveals a significant discrepancy between the generated motions and real human movements. This suggests that the model needs further refinement to achieve more realistic outputs that closely resemble actual human motion data.
- Diversity and MultiModality: The model demonstrates diversity in the generated motions, comparable to real motion data. Additionally, it exhibits a slightly higher MultiModality, indicating a potentially broader range of styles in the generated sequences. This highlights the model's capability to produce varied outputs.

Qualitative Analysis:

- Strengths:
  - Creativity and Realism: The generated motions showcase a commendable level of creativity and capture intricate details, reflecting a variety of styles and expressions. This adds life and richness to the sequences.
  - Adaptability: The model effectively adapts to diverse input prompts, generating motion sequences that align with the intended context. This versatility broadens its applicability across various domains.
  - Multi-Level Manipulation and Temporal Dynamics: The model can handle complex descriptions involving coordinated movements of different body parts (as seen in Figure 4.4) and accurately capture temporal variations within the motion sequence (as seen in Figure 4.5). This demonstrates its ability to generate intricate and realistic motions.
  - Stochasticity for Natural Movement: The inherent randomness of the Diffusion model allows for the introduction of variability in the generated motions, mimicking the natural uncertainty observed in human movements. This results in more dynamic and realistic outputs.
- Areas for Improvement:
  - Occasional Jerkiness/Unnatural Movements: In some instances, the generated motions might exhibit jerky or unnatural movements. This suggests a need for further refinement to achieve smoother and more natural-looking motions.
  - Maintaining Contextual Consistency: While the model adapts well to prompts, there can be occasional deviations from the expected style or context. Fine-tuning the model could lead to more reliable and contextually accurate outputs.

- Bridging the FID Gap: The significant difference between the FID of generated and real motion data indicates a need to improve the model's ability to produce motions that are visually indistinguishable from real human movements.

Overall Assessment: Despite the identified areas for improvement, the motion generation system possesses significant potential for generating diverse, realistic, and contextually accurate motion sequences based on textual descriptions. Continued development and fine-tuning hold promise for enhancing performance and fidelity, making it a valuable tool for various applications in the field of human motion generation.

The model's adaptability to diverse input prompts is a notable strength, as it produces motion sequences that align with the intended context. Multi-level manipulation and time-varied signals are evident in the generated videos, reflecting the complexity and dynamism of the input text. Despite the stochastic nature of the model, it effectively captures the inherent uncertainty and variability in human movements, resulting in dynamic and varied motion sequences consistent with the input description. Overall, while there are areas for refinement and improvement, the motion generation system demonstrates significant potential for generating diverse, realistic, and contextually accurate motion sequences based on textual descriptions. Continued development and fine-tuning of the model could lead to further enhancements in performance and fidelity, making it a valuable tool for various applications in human motion generation.

The motion generation system exhibits promising capabilities in generating a wide array of realistic and contextually appropriate motion sequences from textual descriptions, despite some identified areas for enhancement. Its ability to adapt to diverse input prompts stands out as a significant strength, consistently producing motion sequences that align closely with the intended context. This adaptability underscores the system's versatility and suitability for various applications across different scenarios.

An essential factor contributing to the system's effectiveness is its utilization of multi-level manipulation and time-varying signals, observable in the resulting videos. These methodologies empower the system to capture the intricacies and dynamics inherent in the input text, resulting in motion sequences characterized by nuanced movements and seamless transitions. By integrating these techniques, the system can accurately depict the subtleties described in the input descriptions.

Despite the inherently stochastic nature of the model, it adeptly captures the inherent uncertainty and variability present in human movements. This capability enables the system to generate motion sequences that exhibit dynamic and varied motions consistent with the input descriptions, thereby enhancing the realism and authenticity of the animations produced. The system's capacity to embrace uncertainty and variability contributes significantly to the depth and naturalness of the generated motion sequences.

While opportunities for refinement exist, such as improving the smoothness and fluidity of motion transitions, the motion generation system demonstrates substantial potential for further advancement. Continued refinement and optimization of the model hold the promise of significant improvements in performance and fidelity, enhancing its utility as a valuable

tool for diverse applications in human motion generation. With ongoing development efforts, the system is poised to emerge as a reliable and versatile solution for generating lifelike motion sequences from textual descriptions, driving progress in the field of human motion synthesis.

Chapter 5

Conclusion and Future Scope

In summary, the motion generation system have shown promising results in producing varied and contextually appropriate motion sequences based on textual descriptions. Both quantitative and qualitative assessments have shed light on its performance and areas where enhancements are possible.

Quantitatively, although the system exhibits diversity comparable to real motion data, there's room for improvement in terms of similarity to actual motions, as evidenced by the higher Fréchet Inception Distance (FID) value. Moreover, its capability to generate motion sequences with a higher MultiModality implies a broader distribution, demonstrating its adaptability in capturing diverse motion styles.

Qualitatively, the system demonstrates creativity and realism, notwithstanding occasional instances of unnatural movements. Its flexibility in handling different input prompts, along with the incorporation of multi-level manipulation and time-varied signals, underscores its capacity to translate detailed textual descriptions into dynamic motion sequences.

Looking forward, several possibilities for exploration and refinement present themselves:

- Improving Similarity to Real Motion: Further refining the model architecture and training approaches could enhance the system's resemblance to real motion data, potentially reducing the FID value and boosting overall realism.
- Fine-tuning for Naturalness: Addressing sporadic occurrences of jerky or unnatural movements through optimization of the motion generation process could yield smoother and more lifelike motion sequences.
- Exploring Innovative Architectures: Experimenting with novel architectures, like incorporating attention mechanisms or reinforcement learning techniques, might lead to advancements in motion generation quality and diversity.
- Enhancing User Interaction and Control: Developing interfaces or tools enabling users to interactively influence the generated motion sequences could improve the system's usability across various domains.
- Expanding the Dataset: Enlarging the diversity and size of the training dataset by incorporating motion sequences from diverse sources and styles could enrich the model's

understanding of human motion dynamics.

- Refining Evaluation Metrics: Continuously refining evaluation metrics and methodologies to capture the subtleties of generated motion sequences could offer more comprehensive insights into the system's performance.

By pursuing these avenues of research and development, the motion generation system can continue evolving, offering valuable solutions for applications in entertainment, virtual reality, robotics, and human-computer interaction. Additionally, expanding the evaluation metrics to include qualitative assessments by human evaluators could provide a more comprehensive understanding of the model's performance. Fine-tuning the model based on both quantitative metrics and human judgment will be crucial for achieving a more robust and reliable text-to-motion generation system.

The work lays a solid foundation for future research in advancing the expressiveness and quality of text-to-motion synthesis. As technology evolves, incorporating feedback loops for model improvement and exploring interdisciplinary collaborations with experts in motion analysis and perception could further enhance the model's capabilities. The journey toward achieving a seamless and lifelike text-to-motion synthesis is ongoing, and this study opens avenues for continued exploration and refinement in this exciting field.

References

[1] Petrovich, M., Black, M.J. and Varol, G., 2022, October. TEMOS: Generating diverse human motions from textual descriptions. In European Conference on Computer Vision (pp. 480-497). Cham: Springer Nature Switzerland.

[2] Wang, Y., Leng, Z., Li, F.W., Wu, S.C. and Liang, X., 2023. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 22035-22044).

[3] Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G. and Chen, T., 2023. MotionGPT: Human Motion as a Foreign Language. arXiv preprint arXiv:2306.14795.

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

[5] Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L. and Liu, Z., 2023. Remodiffuse: Retrieval-augmented motion diffusion model. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 364-373).

[6] Yuan, Y., Song, J., Iqbal, U., Vahdat, A. and Kautz, J., 2023. Physdiff: Physics-guided human motion diffusion model. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 16010-16021).

[7] Petrovich, M., Black, M.J. and Varol, G., 2021. Action-conditioned 3d human motion synthesis with transformer vae. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10985-10995).

[8] Ahn, H., Ha, T., Choi, Y., Yoo, H. and Oh, S., 2018, May. Text2action: Generative adversarial synthesis from language to action. In 2018 IEEE International Conference on Robotics and Automation (ICRA) (pp. 5915-5920). IEEE.

[9] Ahuja, C. and Morency, L.P., 2019, September. Language2pose: Natural language grounded pose forecasting. In 2019 International Conference on 3D Vision (3DV) (pp. 719-728). IEEE.

[10] Aksan, E., Kaufmann, M. and Hilliges, O., 2019. Structured prediction helps 3d human motion modelling. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7144-7153).

[11] Norman Badler. Temporal Scene Analysis: Conceptual De scriptions of Object Movements. PhD thesis, University of Toronto, 1975.

[12] Barsoum, E., Kender, J. and Liu, Z., 2018. Hp-gan: Probabilistic 3d human motion prediction via gan. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 1418-1427).

[13] Bowden, R., 2000, July. Learning statistical models of human motion. In IEEE Workshop on Human Modeling, Analysis and Synthesis, CVPR (Vol. 2000).

[14] Cai, H., Bai, C., Tai, Y.W. and Tang, C.K., 2018. Deep video generation, prediction and completion of human action sequences. In Proceedings of the European conference on computer vision (ECCV) (pp. 366-382).

[15] Cheng, X., Xu, W., Wang, T. and Chu, W., 2018. Variational semi-supervised aspect-term sentiment analysis via transformer. arXiv preprint arXiv:1810.10437.

[16] Corona, E., Pumarola, A., Alenya, G. and Moreno-Noguer, F., 2020. Context-aware human motion prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6992-7001).

[17] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[18] Duan, Y., Shi, T., Zou, Z., Lin, Y., Qian, Z., Zhang, B. and Yuan, Y., 2021. Single-shot motion completion with transformer. arXiv preprint arXiv:2103.00776.

[19] Fang, L., Zeng, T., Liu, C., Bo, L., Dong, W. and Chen, C., 2021. Transformer-based conditional variational autoencoder for controllable story generation. arXiv preprint arXiv:2101.00828.

[20] Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M. and Cheng, L., 2020, October. Action2motion: Conditioned generation of 3d human motions. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 2021-2029).

[21] Habibie, I., Holden, D., Schwarz, J., Yearsley, J. and Komura, T., 2017. A recurrent variational autoencoder for human motion synthesis. In Proceedings of the British Machine Vision Conference (BMVC).

[22] Henter, G.E., Alexanderson, S. and Beskow, J., 2020. Moglow: Probabilistic and controllable motion synthesis using normalising flows. ACM Transactions on Graphics (TOG), 39(6), pp.1-14.

[23] Holden, D., Kanoun, O., Perepichka, M. and Popa, T., 2020. Learned motion matching. ACM Transactions on Graphics (TOG), 39(4), pp.53-1.

[24] Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G. and Black, M.J., 2019. AMASS: Archive of motion capture as surface shapes. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 5442-5451).

[25] Martinez, J., Black, M.J. and Romero, J., 2017. On human motion prediction using recurrent neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2891-2900).

[26] Osman, A.A., Bolkart, T. and Black, M.J., 2020. Star: Sparse trained articulated human body regressor. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16 (pp. 598-613). Springer International Publishing.