

TEXT-DRIVEN MULTI-HUMAN MOTION GENERATION

PABLO RUIZ PONCE

Thesis supervisor: SERGIO ESCALERA GUERRERO ((ENG)Universitat de Barcelona)

Thesis co-supervisor: GERMAN BARQUERO

Degree: Master's Degree in Artificial Intelligence

Master's thesis

School of Engineering
Universitat Rovira i Virgili (URV)

Faculty of Mathematics
Universitat de Barcelona (UB)

Barcelona School of Informatics (FIB)
Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

23/01/2024

Text-Driven Multi-Human Motion Generation

Pablo Ruiz Ponce

Supervisor: **Sergio Escalera Guerrero**
Universitat de Barcelona

Co-Supervisor: **Germán Barquero**
Universitat de Barcelona

Advisor: **José García Rodríguez**
Universidad de Alicante

Advisor: **Cristina Palmero Cantariño**
Universitat de Barcelona

Acknowledgements

Me gustaría agradecer a Susana su apoyo incondicional; sin ella no lo hubiese conseguido. A mi familia, por apoyarme desde la distancia y siempre hacer el esfuerzo de intentar entender qué narices hago. A Sergio, Germán y Cristina, por el apoyo, el conocimiento y las innumerables horas que me habéis dedicado. A Jose, por todo el apoyo y la implicación en este proyecto, pese a no ser su obligación. Y por último, a todos los amigos que me han hecho un poco más ligeros todos estos meses de trabajo.

Abstract

The task of generating human-to-human interactions presents a significant challenge, primarily due to the intricate dynamics involved in these interactions. The complexity of learning these dynamics is compounded by the vast array of possible combinations found in human motion generation. Moreover, a key aspect of generation involves conditioning the output, often through natural language, which, while increasing the complexity, simultaneously makes the approach more accessible.

In this thesis, we introduce a novel Diffusion Model incorporating a Transformer-based architecture. This model is conditioned using textual descriptions of both the motion interactions and the individual motions within these interactions. By focusing on the individual components of the interaction, our method achieves more precise conditioning in the generation of these specific motions. Concurrently, the textual descriptions of the overall interaction enable our model to effectively capture the interplay between individual motions. Our approach has been rigorously evaluated using the InterHuman dataset, demonstrating an enhancement over the results achieved by preceding methodologies.

Additionally, this thesis contributes to the field through the development of a new Motion-to-Text methodology, the implementation of an innovative multi-weight sampling technique, and the utilization of Large Language Models to augment textual descriptions from motion datasets.

Contents

List of Figures

List of Tables

1 Introduction

This chapter offers a comprehensive overview of the Master's Thesis. Section 1.1 introduces the impetus behind this research, detailing the motivations and prior work that have shaped this study. Section 1.2 delineates our unique contributions to the field of multi-human motion generation. Finally, Section 1.3 provides a structured outline of the thesis, guiding the reader through the subsequent chapters.

1.1 Motivation

The past year has witnessed unprecedented advancements in the field of generative Artificial Intelligence (AI). A unique convergence of research innovation and technological capabilities has catalyzed a rapid expansion in this domain. Numerous models, ranging in the hundreds or possibly thousands, have been developed to produce diverse outputs, both unconditionally and conditionally. Public attention has predominantly focused on image [1, 2, 3, 4] and text [5, 6, 7, 8] generation, highlighted by the release of notable tools like DALL-E[9] and ChatGPT[10]. Beyond these, there have been substantial strides in other areas, including the generation of audio [11, 12, 13, 14, 15], video [16, 17, 18, 19], and motion, marking significant yet less publicized advancements.

The ability to generate realistic motion[1] is a critical objective in contemporary society. The advent of Virtual Reality (VR)/Augmented Reality (AR) technologies and the emergence of the metaverse have made the generation of realistic human motion increasingly important. Additionally, this capability holds significant potential for other industries, including video games and animation, where it can substantially enhance user experience and realism.

Moreover, the application of motion data extends to a variety of tasks, including action forecasting, anomaly detection, robotics, and autonomous driving. However, acquiring motion data presents significant challenges compared to other data types. While Motion Capture (MoCap) suits and similar technologies exist, they are often prohibitively expensive and not universally applicable. Particularly for very young or elderly individuals, as well as those with disabilities or illnesses, these data acquisition methods can be overly intrusive or even unfeasible. Consequently, the ability to generate realistic motions under specific conditions is invaluable. It enables the creation of synthetic datasets, which would otherwise be difficult and costly to obtain using real human subjects.

[1] Throughout this thesis, the term 'motion' specifically refers to human motion.

In line with advancements in other domains, the field of motion generation has undergone significant evolution in recent years. This progress can be attributed to an increase in available data and the adoption of cutting-edge architectural paradigms, notably Transformers and Diffusion Models. These developments have substantially elevated the quality and diversity of generated motions. A more detailed exploration of this topic will be presented in Chapter 2. Presently, the field boasts models capable of producing high-quality, single human motions. These models can operate without explicit conditions, or under specified conditions, such as predefined classes and natural language descriptors.

However, the same level of advancement has not been observed in the realm of motion interactions. In dynamic real-world scenarios, it is crucial to have models capable of generating interactive motions. Rarely do humans operate in complete isolation; interactions are a fundamental aspect of human motion. These interactions can be categorized into human-human and human-environment types. This thesis, however, will primarily concentrate on human-human interactions. The complexity and diversity inherent in human interactions pose significant challenges in creating and annotating datasets for such motions. This scarcity of relevant data hampers the development of models adept at generating diverse and realistic human interactions.

The primary motivation of this thesis is to address the disparity between the generation of individual motions and interactive motion generation. Our focus is particularly on natural text conditioning, a task we recognize as both challenging and pivotal. Such an approach democratizes motion generation, enabling individuals without technical expertise to create motions simply through descriptive language. This paradigm shift is exemplified by the possibility of animating a movie using its script as the sole input for the model. While generative models have been prevalent for some time, their integration with language models has catalyzed widespread adoption.

The task at hand presents numerous challenges, particularly in the specific case of motion, the complexity lies in the fact that a single textual description of an interaction can correspond to multiple diverse motion representations. Our central hypothesis posits that by integrating more detailed information about individual motions, we can achieve greater precision and realism in the generation of human interactions.

1.2 Contributions

This thesis makes several significant contributions across various aspects of motion generation. The foremost contribution, and the central focus of this work, is the novel application of individual motion information to enhance the generation of human motion interactions. By employing detailed textual descriptions of the motions performed by individuals during an interaction, we have achieved an improvement in the overall quality of generated interactive motions. To our knowledge, this is the first instance of utilizing such fine-grained individual motion descriptions to augment the quality of motion interaction generation.

In addition to the primary contribution, this thesis also introduces an innovative Motion-to-Text (M2T) architecture. Finally, this thesis contributes to the enhancement of existing Text-to-Interaction (T2I) datasets through the utilization of pre-trained Large Language Models (LLMs). By augmenting these datasets with a broader array of textual descriptions for the same motion, we have significantly increased the variability in the textual descriptors available for characterizing interactions. This enrichment of datasets is instrumental in advancing the diversity and accuracy of motion generation models. [2]

1.3 Outline

The thesis is organized into six additional chapters beyond this introduction. Chapter 2 lays the foundational background and reviews related work essential for a comprehensive understanding of this thesis. In Chapter 3, we detail the primary datasets employed in our research. Chapter 4 methodically outlines the methodologies and innovative approaches devised for this thesis. The outcomes derived from implementing these methodologies are presented in Chapter 5. Finally, Chapters 6 and 7 engage in a critical discussion of the findings and propose avenues for future research, drawing conclusions that encapsulate the essence of this study.

[2] All the code from this thesis can be found in: https://github.com/pabloruizp/TFM

2 Background & Related Work

Prior to delving into our contributions, it is imperative to establish a comprehensive understanding of key concepts and review the extensive body of literature in our field. A clear grasp of the current state of the art in our research area is essential. This foundational knowledge enables us to recognize the significant work already accomplished and informs our approach to proposing novel methods and contributions. In this chapter, we will meticulously analyze the primary facets of our research and present the most relevant and influential studies, setting the stage for a deeper appreciation of our contributions.

In Section 2.1, a comprehensive exposition on the theoretical foundations of Diffusion Models, along with an explanation of the mechanisms underlying conditional generation, will be presented. Subsequently, Section 2.2 is dedicated to an in-depth analysis of the Transformer architecture, focusing on its operational principles and the potential for integration within diffusion models. Finally, Section 2.3 will explore the fundamental concepts of human motion generation, encompassing a thorough review of all related work pertinent to our thesis.

2.1 Diffusion Models

Generative AI currently stands as one of the most dynamic and rapidly evolving fields in technology. The domain is teeming with research focused on using deep learning models to generate a wide array of data types. Over the years, various methodologies have emerged, including Flow-based [20], Variational Auto Encoders (VAEs) [21], Generative Adversarial Networks (GANs) [22]. Presently, in the evolving landscape of this field Diffusion Models are recognized as the predominant approach.

Diffusion Models [23, 24] draw inspiration from thermodynamics. These models operate through two symmetrical processes, as depicted in Figure 2.1. The first process is the 'noising' step, wherein Gaussian noise is incrementally added to the original input over multiple stages, eventually transforming the input into pure Gaussian noise. Conversely, the second process, termed 'denoising', involves a Neural Network trained to systematically remove the noise at each stage, ultimately restoring the original input. The fundamental principle of Diffusion Models is that, given a sufficient number of steps, a model can be trained to convert random noise into coherent, meaningful output.

The diffusion model approach has gained widespread popularity across various domains, demon-

strating superior performance over other methodologies in many cases [25]. However, a notable limitation of this approach is its time-intensive nature. The requirement for consecutive steps means that both the training and inference phases of these models are generally more time-consuming compared to other solutions. This drawback has spurred considerable research [26, 27], focusing on enhancing the efficiency of these models. Despite these efforts, optimizing the time efficiency of Diffusion Models remains an active and challenging area of research.

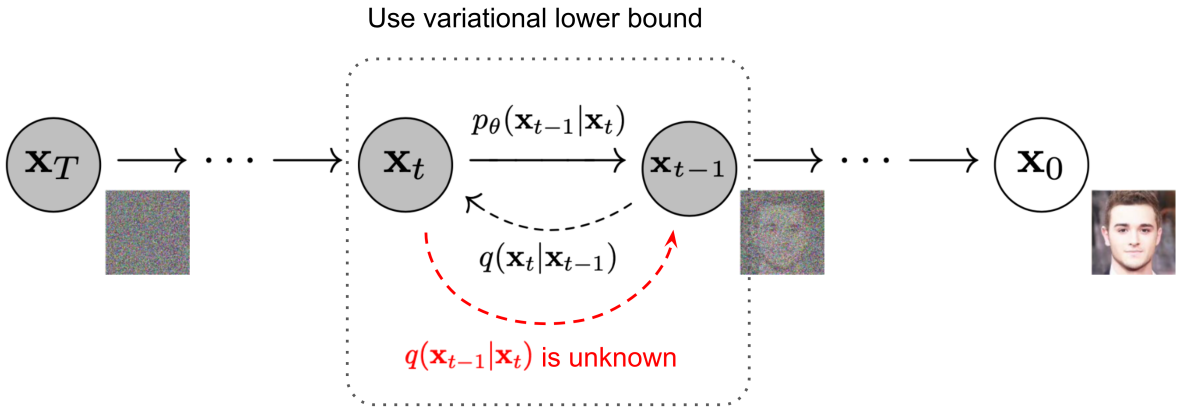


Figure 2.1: Diffusion Model process (Source: [24])

2.1.1 Forward Diffusion Process

The forward diffusion process, also known as the 'noising' process, is characterized by the gradual addition of noise to data across multiple steps. This process begins with the original data distribution and progressively transforms it into a simpler, well-known distribution, typically Gaussian, through the incremental addition of noise. Mathematically, this process is structured as a Markov chain, where each successive step involves the application of a modest quantity of Gaussian noise to the data. The formulation of this process at each timestep t is expressed as:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1 - \alpha_t}\epsilon \tag{2.1}$$

In the previous formula, the variable x_t symbolizes the data at step t, while x_{t-1} denotes the data from the preceding step. Central to this process are the coefficients α_t, collectively referred to as the variance schedule. These coefficients are pivotal in determining the noise level introduced at each step. The variance schedule is strategically selected to facilitate a seamless transition from the original data distribution to a Gaussian distribution throughout the diffusion process.

The calibration of α_t is critical: if set excessively high, the resulting noise level might exceed the model's capability to effectively denoise the data. Conversely, if α_t is set too low, the process may require an impractically large number of steps and an extended duration to achieve denoising. It is this careful tuning of α_t that governs the balance between noise addition and the feasibility of subsequent denoising.

The noise term ϵ, integral to this process, is independently sampled at each step from a standard Gaussian distribution, denoted as $\mathcal{N}(0, I)$. This stochastic component ensures the randomness necessary for the gradual transformation of the data distribution towards the target Gaussian distribution.

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon_t, \quad \text{where} \quad \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s \tag{2.2}$$

The complete forward diffusion process can be articulated more formally with the previous equation. In this refined representation, the term $\bar{\alpha}_t$ is introduced, signifying the cumulative product of the variance schedule coefficients up to the time step t.

2.1.2 Reverse Diffusion Process

Algorithm 1 Training

1: **repeat**
2: $x_0 \sim q(x_0)$
3: $t \sim \text{Uniform}\{1, \ldots, T\}$
4: $\epsilon \sim \mathcal{N}(0, I)$
5: Take gradient descent step on:
$\nabla_\theta \|\epsilon - \epsilon_\theta(\sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t}\epsilon)\|^2$
6: **until** converged

Algorithm 2 Sampling

1: $x_T \sim \mathcal{N}(0, I)$
2: **for** $t = T, \ldots, 1$ **do**
3: $z \sim \mathcal{N}(0, I)$ if $t > 1$, else $z = 0$
4: $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t)) + \sigma_t z$
5: **end for**
6: **return** x_0

Figure 2.2: The training and sampling algorithms in DDPM (Source: [24])

The reverse diffusion process, also termed the 'denoising' process, is a cornerstone of diffusion models, enabling the generation of high-quality data from a noise distribution. Characterized by its iterative nature, the process involves a sequence of steps, each methodically reducing the noise level and gradually steering the data from an initial random noise distribution toward its original distribution.

The procedure initiates with a sample x_T drawn from a Gaussian noise distribution. Subsequently, a series of reverse diffusion steps are applied iteratively to generate $x_{T-1}, x_{T-2}, \ldots, x_0$, where x_0 represents the final, generated sample. The reverse diffusion step at each iteration is governed by the following formula:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) \tag{2.3}$$

In this equation, $\epsilon_\theta(x_t, t)$ denotes the noise model, predicted by a neural network parameterized by θ. This model is tasked with estimating the noise introduced during the forward diffusion process at each step t. The neural network is specifically trained to predict the noise ϵ that was incorporated in each step of the forward diffusion, thereby facilitating its effective reversal during the generative phase. The training objective of this network centers on minimizing the discrepancy between the predicted noise and the actual noise introduced during the forward diffusion

process. This careful alignment ensures that the reverse diffusion process can accurately and effectively revert the data from a state of randomness to its original, structured form. Through this mechanism, Diffusion Models are capable of producing outputs that are not only more varied and complex but also of a higher quality compared to earlier methods, all the while maintaining precise conditioning capabilities. However, the primary limitation of this approach, as compared to other methods, is the increased duration required for both training and sampling.

2.1.3 Conditioned Generation

Conditioned diffusion models represent an advanced form of diffusion models, specifically designed to generate data that conforms to predefined conditions or constraints. In contrast to traditional diffusion models, conditioned diffusion models integrate auxiliary information to guide throughout the generation process. This attribute is particularly advantageous in applications where the generated output must fulfil certain specifications, such as text-to-motion synthesis or generating motions corresponding to specific classes. There are two primary methodologies for incorporating conditioning into diffusion models: classifier-guided and classifier-free methods.

Classifier-Guided Conditioning

Classifier-guided diffusion [25] is a technique that infuses class-specific information into the diffusion model's process. This is achieved by integrating a classifier, denoted as $\nabla_x \log f_\theta(y|x_t)$, which is designed to predict class information from noisy input data x_t. The gradients derived from this classifier are instrumental in steering the diffusion process towards generating data corresponding to a specific class, represented by y. To facilitate this, the score function for the joint distribution $q(x_t, y)$ is formulated as a composite of the individual score functions for x_t and y:

$$\begin{aligned} \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t, y) &= \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log q(y|\mathbf{x}_t) \\ &\approx -\frac{1}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) + \nabla_{\mathbf{x}_t} \log f_\phi(y|\mathbf{x}_t) \\ &= -\frac{1}{\sqrt{1-\bar{\alpha}_t}} (\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) - \sqrt{1-\bar{\alpha}_t} \nabla_{\mathbf{x}_t} \log f_\phi(y|\mathbf{x}_t)) \end{aligned} \tag{2.4}$$

This approach leads to the development of a classifier-guided predictor $\tilde{\epsilon}_\theta$. It is constructed by modifying the original noise predictor to include a term that integrates class-specific guidance:

$$\tilde{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) - \sqrt{1-\alpha_t} \nabla_{x_t} \log f_\theta(y|x_t) \tag{2.5}$$

Additionally, a weighting factor w can be introduced to modulate the influence of the classifier guidance. This allows for fine-tuning the balance between the original noise predictor and the class-specific gradient information:

$$\tilde{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) - \sqrt{1-\alpha_t} w \nabla_{x_t} \log f_\theta(y|x_t) \tag{2.6}$$

This classifier-guided approach enhances the model's ability to generate data that is not only high in quality but also tailored to specific class characteristics, making it a powerful tool in guided data generation and targeted synthesis within diffusion models.

Classifier-Free Conditioning

Classifier-free conditioning [28] offers an alternative approach to diffusion models, obviating the need for a separate classifier. In this method, the diffusion model is directly trained to generate data conditioned on specific attributes, utilizing a modified training regimen. This process involves intermittently omitting part of the conditioning information during training with a certain probability. As a result, the model acquires the ability to generate outputs both with and without specific conditions, thereby negating the requirement for an independent classifier. The reverse diffusion process in classifier-free conditioning can be formalized as follows:

$$\begin{aligned}
\nabla_{\mathbf{x}_t} \log p(y|\mathbf{x}_t) &= \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|y) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) \\
&= -\frac{1}{\sqrt{1-\bar{\alpha}_t}} \Big(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \Big) \\
\bar{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t, y) &= \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) - \sqrt{1-\bar{\alpha}_t}\, w \nabla_{\mathbf{x}_t} \log p(y|\mathbf{x}_t) \\
&= \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) + w \big(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \big) \\
&= (w+1)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) - w\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)
\end{aligned} \tag{2.7}$$

This series of equations delineates the mathematical underpinnings of the classifier-free conditioning process. The model leverages the differential between conditioned and unconditioned noise models, represented by $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y)$ and $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$, respectively. The parameter w serves as a weighting factor that determines the relative influence of the conditioned versus unconditioned components in the final prediction. This formulation allows for control over the conditioning, facilitating the generation of data that aligns closely with the desired attributes while maintaining the flexibility afforded by the absence of a dedicated classifier. However, a drawback is that for each sampling instance, the model is required to concurrently generate predictions for both the unconditioned and conditioned scenarios.

2.2 Transformers

The transformer architecture [29] (Figure 2.4) has emerged as a fundamental innovation in artificial intelligence, especially within the realm of natural Natural Language Processing (NLP). Its hallmark is the adoption of self-attention mechanisms, marking a significant shift away from the recurrent layers that dominated earlier models. This architectural choice has not only revolutionized NLP but has also found extensive applications in fields such as image processing, audio synthesis, and motion analysis, showcasing its versatility.

The core of the transformer model is an encoder-decoder structure, where both the encoder and the decoder are composed of several identical layers. The encoder features N such layers, each comprising two distinct sub-layers: a multi-head self-attention mechanism and a fully connected feed-forward network. The multi-head self-attention mechanism enables the model to dynamically weigh the importance of different parts of the input data. The feed-forward network then processes this attention-modulated data. Both sub-layers in the encoder are enhanced with normalization and residual connections, facilitating stability in learning and efficient gradient flow during backpropagation.

The decoder, mirroring the encoder's architecture, also consists of N identical layers. Each layer in the decoder augments the two sub-layers found in the encoder with an additional sub-layer. This extra sub-layer performs multi-head attention over the encoder's output, allowing the decoder to concentrate on relevant parts of this output for generating the final result. As in the encoder, each sub-layer of the decoder is equipped with normalization and residual connections, ensuring consistency in the processing pipeline.

This design, centralizing around self-attention, enables transformers to adeptly handle complex and long-range dependencies in data. Unlike previous architectures that processed sequences in a step-wise manner, transformers assess entire sequences in parallel, significantly enhancing efficiency and effectiveness. This attribute is particularly beneficial in tasks that demand an understanding of context over long stretches of data, a feature that has made transformers a go-to choice for a variety of applications beyond their initial scope in NLP.

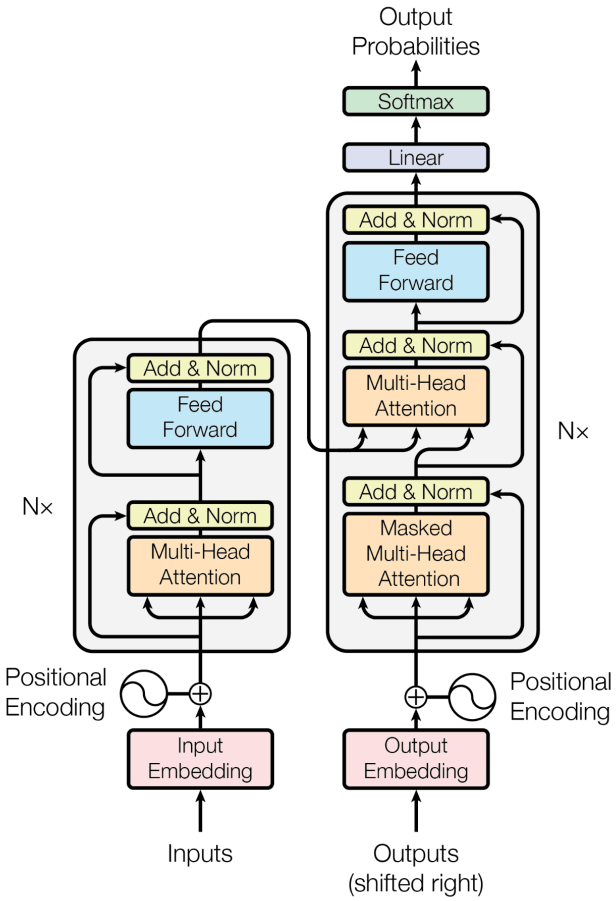


Figure 2.3: Transformer architecture (Source: [29])

2.2.1 Self-Attention Mechanism in Transformers

The self-attention mechanism is a defining feature of the transformer model, enabling it to effectively weigh the relative significance of different elements (such as words in a sentence) within

a sequence. This mechanism operates on the principle of mapping a query and a corresponding set of key-value pairs to an output, where all these elements are represented as vectors. The output is computed as a weighted sum of the values. The weights themselves are determined by a compatibility function that assesses how well each key correlates with the query.

Scaled Dot-Product Attention

A critical component of self-attention in transformers is the scaled dot-product attention. This attention function is formulated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.8}$$

In this equation, Q, K, and V represent matrices composed of the query, key, and value vectors, respectively. The term d_k denotes the dimensionality of the key vectors. The scaling factor $\sqrt{d_k}$ is crucial as it prevents the dot products from growing too large in magnitude, which could lead to vanishing gradients during training. By applying the softmax function to the scaled dot products, the model generates a distribution of weights across the values.

Multi-Head Attention

To enhance its capability to capture information from different representation subspaces and positions, the transformer employs a multi-head attention mechanism. This approach involves parallel attention layers, each producing distinct attention outputs, which are then concatenated and linearly transformed. The multi-head attention is defined as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2.9}$$

In this context, each head_i is an individual attention operation, computed as $\text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$. The matrices W_i^Q, W_i^K, and W_i^V are parameter matrices used for linearly projecting the queries, keys, and values in each head. The matrix W^O is used for the final linear transformation of the concatenated attention outputs. This structure allows the model to concurrently process the information in different ways, fostering a more comprehensive understanding of the relationships within the data.

2.2.2 Positional Encoding

Transformers, by design, do not incorporate recurrence or convolution mechanisms. This architectural choice, while beneficial in many aspects, does not inherently account for the order or position of elements in a sequence. To address this, positional encodings are integrated into the input embeddings, providing the model with necessary information about the sequential order of elements, such as words in a sentence.

The key requirement for positional encodings is that they must be of the same dimensionality, d_{model}, as the input embeddings. This dimensional alignment allows for the direct summation of positional encodings with the embeddings, thereby embedding position information into the

input representation. A widely adopted approach for positional encoding employs sinusoidal functions, defined as follows:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{2.10}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{2.11}$$

In these equations, *pos* represents the position of an element in the sequence, and i denotes the dimension. The introduction of positional encoding is a crucial step in ensuring that transformers, despite lacking recurrence or convolution, can effectively understand and utilize the order of elements within sequences, a vital aspect in tasks such as language understanding and generation.

2.2.3 Transformers in Diffusion Models

Diffusion models have gained prominence as a formidable category of generative models, demonstrating their aptitude in producing high-quality outputs across various domains, including image and audio generation. In image generation, a prevalent technique involves employing a U-Net [30] architecture for the denoising process. Yet, when it comes to generating other types of data, like time series, this architecture proves to be sub-optimal. The integration of transformer architectures into diffusion models represents a significant advancement, melding the strengths of both methodologies to achieve enhanced generative capabilities.

Transformers, renowned for their self-attention mechanisms, excel in capturing long-range dependencies and intricate patterns within data. When applied to diffusion models, these capabilities enable more effective generation of complex time-series data, such as motion sequences. The self-attention mechanism of transformers is particularly adept at processing sequences, making them well-suited for tasks that require understanding and generating coherent, temporally consistent data.

However, this integration is not without its challenges, particularly concerning scaling and computational efficiency. Transformers, especially when tasked with handling large-scale datasets or high-dimensional data, are characterized by their substantial computational demands. Characterized by an extensive array of parameters, transformers give rise to exceptionally large models, which demand considerable training time and memory allocation. This aspect can pose significant challenges in practical applications, where computational resources are a limiting factor.

The fusion of transformers with diffusion models thus opens up new possibilities in generative modelling, offering the potential to create even more sophisticated and high-fidelity generative outputs, albeit with considerations for the associated computational complexities.

2.3 Human Motion Generation

Human Motion Generation refers to the creation of synthetic human movements that closely mimic those performed by actual individuals. This field has experienced significant advance-

ments alongside the general progress in generative AI over recent years. However, unlike other areas of generative AI, such as image and text generation, the process of acquiring motion data presents unique challenges in terms of both complexity and cost. Capturing motion data necessitates a sophisticated and costly MoCap setup, along with skilled actors to perform the desired motions. As delineated in Chapter 3, the availability of datasets for motion data is relatively limited when contrasted with datasets in other modalities. Additionally, the field lacks a standardized data format. The representation and format of motion data vary widely depending on the methods used for data capture and the preferences of individual researchers. This diversity in data formats poses considerable hurdles in research, as certain methodological improvements may not be universally applicable due to these format disparities.

The spectrum of methodologies employed for motion generation reflects the broader diversity of approaches applicable to time-series data and generative AI [31]. Techniques range from 1D Convolutional Neural Netwoks (CNNs) to various adaptations of Recurrent Neural Netwoks (RNNs), and even some instances of simple Multi-Layer Perceptrons (MLPs). However, as discussed earlier in Section 2.2, Transformers have emerged as the most effective a well-performant architecture for time-series data. In terms of generative models, while VAEs and GANs are commonly used, the introduction of Diffusion Models has brought advantages that currently position them as the preferred choice. This shift towards Transformers and Diffusion Models highlights the evolving landscape of methodologies in the realm of Human Motion Generation.

2.3.1 Text-to-Motion

Having established an understanding of motion generation and its diverse methodologies, we now turn our attention to generating human motion under specific conditions. Among various conditioning inputs like audio, classes, or text, this thesis focuses on natural text conditioning.

Utilizing text as a conditioning mechanism is advantageous due to its capacity to convey detailed descriptions of specific motions. Text allows for the specification of movements in different body parts, at varying velocities, and within diverse contexts or emotional states. In contrast to other conditioning means such as categorical classes or parameters, natural language provides an intuitive and user-friendly approach for this generative task. Recent advancements with Large Language Model (LLM) have underscored the potency of text as a versatile tool across various applications.

A review of recent literature reveals significant progress in this domain over the past two years, with a plethora of methodologies being explored. This diversity in approaches suggests that the field is actively evolving, with no single method yet established as definitively superior.

In MotionCLIP [32], the authors adopt a method of encoding motions into a latent space that is aligned with the CLIP [33] model's latent space. This alignment enables the use of CLIP embeddings, derived from text or images, to generate motions. The TEMOS [34] model, on the other hand, employs a Transformer VAEs for motion generation from text. It integrates a

Text encoder powered by DistilBERT with a Motion encoder, creating a latent representation for the VAEs decoder that facilitates motion generation based on text conditions. TM2T [35], T2M-GPT [36], and MotionGPT [37] share a similar approach, where motion is tokenized into discrete codes from a learned codebook, and a Transformer architecture is used to convert text tokens into motion tokens in an autoregressive manner.

Presently, the most promising and effective methodology appears to be the use of Diffusion Models. FLAME [38] and MotionDiffusion [39] employ traditional diffusion processes with a Transformer as a denoiser, achieving impressive results. A notable challenge in applying diffusion models to motion is the selection of appropriate loss functions. MDM [40] attempts to address this by having the denoiser predict the initial data point x_0 at each step, rather than the noise. This strategy allows for the application of loss functions that are better suited to motion generation. Chapter 4 will delve deeper into this approach and the associated loss functions. Since the introduction of MDM, many new methodologies have been inspired by its concept, focusing on training more realistic models. Additional innovations include incorporating physical constraints [41] into the diffusion process, enhancing the realism of generated motions. Despite their efficacy, a notable limitation of diffusion models is the significant time required for training and inference. Overall, diffusion models have revolutionized the ability to generate complex, highly detailed outputs, and facilitate precise conditioning, albeit at the expense of training and sampling duration.

2.3.2 Motion to Text

In contrast, the M2T task focuses on converting motion into descriptive text. This is a crucial endeavor because manually annotating motions with text is an exceptionally time-consuming, costly, and challenging process. Although this field is less explored compared to its counterpart, it presents a distinct set of challenges and opportunities.

As delineated in Chapter 3, there are numerous motion datasets available. However, a scarcity of these datasets includes detailed descriptions of the included motions. This shortage hinders the development of a robust Language Model capable of accurately converting motion to text. The diverse formats in which motion data is represented further complicate the task, making the amalgamation of multiple datasets for a generalized model challenging.

The literature does present some notable attempts to address M2T conversion. A singular study [42] focuses on generating text from motion using a GANs framework, where the generator creates textual descriptions from motion, and the discriminator differentiates between genuine and artificial descriptions. The generator employs RNNs, and the discriminator utilizes 1D CNNs. Other studies, while not directly aimed at M2T conversion, contribute indirectly through the development of Text-to-Motion (T2M) models. For instance, in [43], the authors translate motions into a discrete token set using vector quantization, then train dual networks for converting between motion and text tokens, akin to a translation task. Similarly, [37] also employs vector quantization to tokenize motions, but distinctively creates a unified vocabulary that merges text

and motion tokens. This approach enables fine-tuning an existing LLM to process motion tokens, resulting in a versatile model capable of handling text, motion, or their combination as input and output.

Furthermore, the M2T task can be recontextualized into more research-intensive domains, such as image or video captioning. By transforming motions into images or videos, the task aligns with the more active fields of image or video captioning [44, 45, 46], which offer richer datasets. However, this thesis reveals a lack of specific research focusing on motion in this redefined context.

2.3.3 Text to Interaction

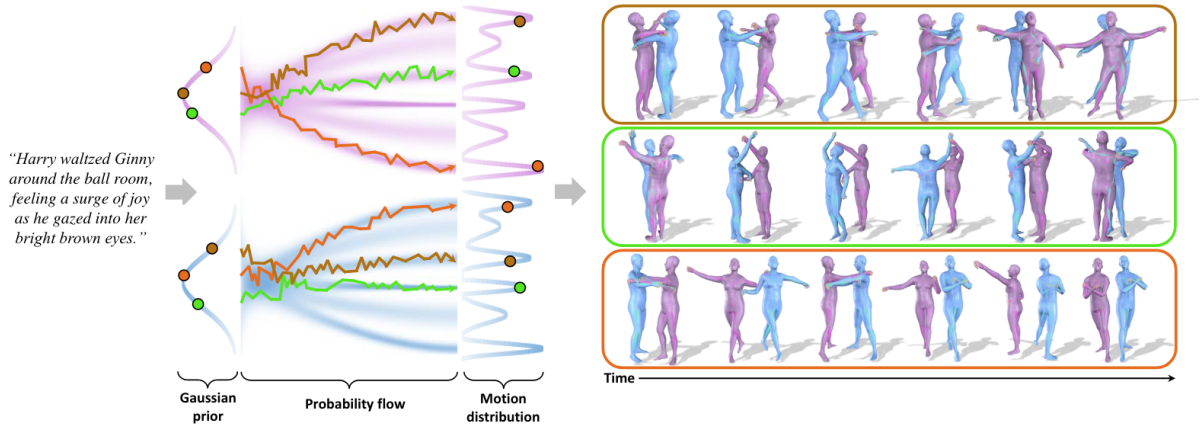


Figure 2.4: Diffusion Text-to-Interaction Pipeline (Source: [47])

Having established a foundational understanding of motion generation, T2M, and M2T, we now transition to the focal theme of this thesis: T2I. As implied by its name, T2I pertains to the generation of human motion that encompasses some form of interaction. This thesis explores multi-human interactions, a phenomenon prevalent in various environments due to the inherently social nature of humans. Our motions are often influenced by those we interact with. Leveraging natural language to generate these motion interactions not only affords greater control but also simplifies the utilization of such models.

The introductory chapter highlighted the nascent stage of research in motion interaction generation, particularly when conditioned on textual descriptions. A significant bottleneck, as we will explore in the forthcoming chapter, is the scarcity of annotated datasets capturing these interactions. Furthermore, the expensive nature of data collection, the intricate dynamics of motion interaction, and the extensive range of possible combinations of individual motions with corresponding valid textual descriptions constitute additional challenges in this area.

In [48], the authors, building on their work in MDM, adapt their individual diffusion model for T2M to facilitate multi-human interactions. They introduce ComMDM, a construct compris-

ing a single-layer transformer integrated into specific layers of the denoisers in two frozen MDM models. This transformer processes the activations from the two models and adjusts them to foster interaction. In [49], a similar concept is employed with two distinct models that, unlike the frozen models in ComMDM, share parameters. Interaction modelling is achieved through a cross-attention module that connects these parameter-sharing models, an architecture particularly suited for asymmetric interactions involving an actor and a receiver. Lastly, [50] contributes the most extensive annotated dataset of human interactions to date, proposing an architecture akin to those previously described. Section 3.3.2 will provide a more comprehensive discussion on this dataset.

Despite some existing proposals, this field remains largely uncharted. The scarcity of annotated datasets has been a key impediment to research in this area. Although new proposals and methods are emerging, there is still considerable work to be done. Numerous architectures await testing and development, a variety of loss functions to be utilized and innovated, extensive data-enhancing and preprocessing techniques to be explored, and various forms of specialized conditioning to be integrated. In the following chapters, we will present some of our proposals aimed at progressing in this field.

3 Data

Throughout this thesis, we have repeatedly emphasized the paramount importance of possessing extensive, high-quality data for the perfection of generative artificial intelligence models. This requirement becomes particularly challenging in the context of motion data, which is notably more difficult to acquire in comparison to other modalities like text and images. Additionally, the representation of motion data varies significantly in format.

In Section 3.1, we will delineate the most pertinent motion representations as documented in current literature, and describe the specific representation chosen for use in our thesis. Subsequently, Section 3.2 is devoted to detailing existing datasets that facilitate single-human motion generation, with a particular focus on those datasets that have been integral to our research. Lastly, Section 3.3 will introduce datasets related to multi-human motion, highlighting those that have been pivotal in the development of our final approach.

3.1 Motion Representation

Prior to delving into the specifics of various datasets and those we have utilized, it is crucial to grasp the nature of the data these datasets comprise. Essentially, human motion can be conceptualized as a sequence of human poses arrayed along a temporal axis, analogous to how a video is perceived as a series of image frames in time. Nonetheless, the representation of a human pose entails considerable complexities and can be approached in several distinct manners. The three most prominent methods are:

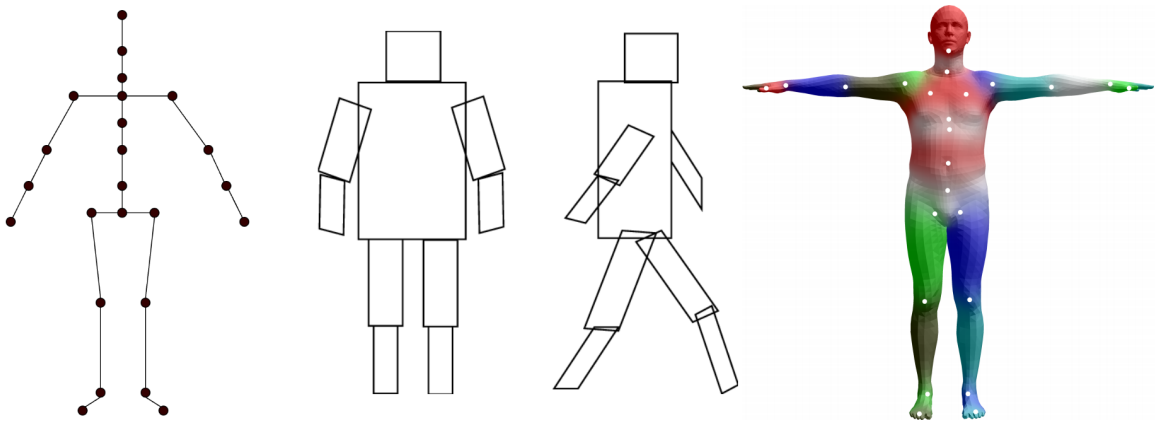
- **Kinematic:** Also known as the skeleton-based model, this approach depicts a human pose through a set of joints arranged in a tree or graph-like structure. Typically, the pelvis is the root of the skeleton, with other joints represented as relative positions/angles. A significant limitation of this method is its inability to convey the body's shape, which may lead to the generation of unrealistic motions, particularly in cases of varying human body shapes.

- **Planar:** This model was developed to address the shortcomings related to representing the human shape in the kinematic model. Here, a human pose is depicted using boxes (in 2D) or cubes (in 3D) to represent different body parts.

- **Volumetric:** For a more realistic and complete representation of human shapes, the volumetric human model is employed. This involves representing a human with a volumetric

mesh that can adapt to specific body shapes. Commonly, parametric models like SMPL [51] are used to define these shapes. However, applying this representation directly in human motion generation poses significant challenges due to the high dimensional nature of the data.

For our thesis, we have chosen to work with the Kinematic model. This decision is rooted in its widespread application across most solutions in human motion generation. Moreover, all the datasets engaged in our research utilize various renditions of a kinematic model. The kinematic model not only provides an intuitive and detailed framework for understanding human poses but also maintains a level of generality conducive to efficient computational handling. Despite some limitations, it allows for potential conversion between different models if necessary.



a) Kinematic b) Planar c) Volumetric

Figure 3.1: Different Human Pose Representations

3.2 Single Human Motion Datasets

Having established an understanding of motion representation, we now turn to examining the datasets available for use in various components of our thesis, focusing initially on single human motion datasets. In Table 3.1, we have selectively included only those single-human datasets that are accompanied by textual descriptions, as these are particularly relevant to our research objectives. Despite the relatively limited number of these datasets, they encompass a substantial collection of motions and accompanying annotations.

3.2.1 AMASS

AMASS [52] constitutes a substantial collection of human motion data. It amalgamates diverse optical marker-based motion capture datasets into a unified framework and parameterization,

proving immensely beneficial for animation, visualization, and training data for deep learning models. Notably, while this dataset employs a volumetric approach for storage, the motions are articulated through a kinematic representation of these parametric body shapes. Specifically, it utilizes the SMPL version, SMPL-H [53], encompassing 52 joints, with 22 in the body and 30 in the hands. This dataset has been pivotal in our research, particularly for the MotionCLIP [32] methodology that forms the foundation of our motion-to-text approach, as detailed in Section 4.1. The AMASS dataset itself does not provide any textual description. While AMASS itself lacks textual descriptions, the BABEL dataset [54] complements it by providing language labels for each frame. These labels annotate over 40 hours of motion from AMASS, covering 250 distinct actions. In MotionCLIP, the authors concatenate these language labels from a motion sequence to create textual descriptions. However, these generated texts are notably brief, and concise and lack syntactic coherence.

3.2.2 HumanML3D

HumanML3D [55] emerges as a comprehensive 3D human motion-language dataset, born from the integration of HumanAct12 [56] and AMASS. It encompasses a wide array of human activities ranging from everyday actions (like 'walking', 'jumping') to sports ('swimming', 'playing golf'), acrobatics ('cartwheel'), and artistic movements ('dancing'). A pose $\mathbf{p}$ in HumanML3D is defined by a tuple $\left(\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y, \mathbf{j}^p, \mathbf{j}^v, \mathbf{j}^r, \mathbf{c}^f\right)$. Here, $\dot{r}^a \in \mathbb{R}$ represents the root angular velocity along the Y-axis; $(\dot{r}^x, \dot{r}^z \in \mathbb{R})$ are the root's linear velocities on the XZ-plane; $r^y \in \mathbb{R}$ denotes root height; $\mathbf{j}^p \in \mathbb{R}^{3j}, \mathbf{j}^v \in \mathbb{R}^{3j}$, and $\mathbf{j}^r \in \mathbb{R}^{6j}$ describe the local joint positions, velocities, and rotations in root space, respectively, with j being the number of joints; $\mathbf{c}^f \in \mathbb{R}^4$ consists of binary features highlighting foot ground contacts. This dataset employs a $6D$ continuous rotation representation, which, though redundant, has been found to outperform less redundant representations like those originally used in AMASS. The authors have enriched the existing textual annotations from both source datasets, providing more comprehensive descriptions with a richer vocabulary. Consequently, HumanML3D stands as the largest annotated single-motion dataset for human motions and a primary benchmark for text-to-motion tasks.

3.3 Multi Human Motion Datasets

After establishing a comprehensive understanding of single-human motion datasets, our focus shifts to multi-human datasets. In contrast to single-human datasets, the multi-human category is characterized by a relatively scarce availability of data, particularly in terms of annotated content. A notable development in this field is the release of InterHuman in September 2023, which marked a significant enhancement in available data resources.

A critical aspect distinguishing multi-human datasets from their single-human counterparts is the importance of the initial positioning of subjects. While single-human datasets can uniformly position the individual at the origin of coordinates, multi-human interactions necessitate distinct initial positions and orientations for each subject. This consideration is vital for accurately visualizing and understanding the interactions. The spatial dynamics and relative positioning

of multiple humans are essential for representing complex interactions, thereby adding another layer of complexity to multi-human motion dataset analysis.

3.3.1 3D Poses in the Wild

The 3D Poses in the Wild (3DPW) [57] dataset is a pioneering collection in the realm of outdoor human motion datasets, featuring accurate 3D poses for evaluation in natural settings. Prior to 3DPW, outdoor datasets were limited by their confinement to small recording volumes. This dataset sets itself apart by incorporating video footage captured using a mobile phone camera in motion, a first of its kind. As observed in Table 3.1, while the volume of data in 3DPW is not extensive, it was the sole annotated dataset available when we commenced our thesis work.

The dataset was annotated for the paper PriorMDM [48]. The data format aligns with that of HumanML3D, augmented by the inclusion of the initial root position and rotation for each individual involved in an interaction.

3.3.2 InterHuman

InterHuman [47] stands as a comprehensive and large-scale 3D human interactive motion dataset, unparalleled in its extensive coverage of 3D motions involving two interacting individuals, each paired with natural language annotations. It represents the most expansive collection of 3D human-to-human interaction data currently available. The dataset is segmented into two broad categories: daily motion, which includes routine interactions between two people such as passing objects, greeting, and communicating; and professional motions, encompassing specific human-to-human interactions in contexts like Taekwondo, Latin dance, and boxing.

The dataset adopts a unique non-canonicalization approach for representing multi-person interactive motions. Unlike conventional methods that transform joint positions and velocities to the root frame, InterHuman retains these elements in the world frame. This approach facilitates direct access to global translation and rotation data from the root position and inverse kinematics (IK). The representation is expressed as $x^i = \left[\mathbf{j}_g^p, \mathbf{j}_g^v, \mathbf{j}^r, \mathbf{c}^f\right]$, where x^i, the i-th motion state, encompasses global joint positions $\mathbf{j}_g^p \in \mathbb{R}^{3N_j}$, velocities $\mathbf{j}_g^v \in \mathbb{R}^{3N_j}$ in the world frame, $6D$ representation of local rotations $\mathbf{j}^r \in \mathbb{R}^{6N_j}$ in the root frame, and binary foot-ground contact features $\mathbf{c}^f \in \mathbb{R}^4$.

With its public release in September 2023, InterHuman has significantly transformed the text-to-interaction research paradigm. Although it is still in the early stages of adoption by the research community, it has already been a foundational element in our methodologies, particularly in the approaches described in Sections 4.2 and 4.3.

Examining the data presented in Table 3.1, a significant disparity is evident when comparing the largest datasets in single-human text-to-motion. The complexity of interaction surpasses that of individual motion, attributable to the intricate dynamics involved in interactions and the myriad combinations of single motions. Therefore, despite the considerable advancements made

by the InterHuman dataset, the need for more data in this domain remains pressing.

Dataset	Descriptions	Interactive	Motions	Vocab.	Descriptions	Duration
KIT[58]	✓	-	3911	1623	6278	11.23 h
HumanML3D[55]	✓	-	14616	5371	44970	28.59 h
NTU RGB+D[59]	-	✓	739	-	-	0.47 h
UMPM[60]	-	✓	36	-	-	2.22 h
You2Me[61]	-	✓	42	-	-	1.4 h
3DPW[57]	✓	✓	60	135	56	0.47 h
InterHuman [47]	✓	✓	**6022**	**5656**	**16756**	6.56 h

Table 3.1: Dataset Comparison

4 Methodology

Having delineated essential preliminary concepts, surveyed pertinent related work in the field we aim to contribute to, and presented both the available data and our selected datasets, we are now poised to delve into the exposition of the various contributions and methods proposed in this thesis.

This chapter is dedicated to a thorough theoretical explanation of our proposed methodologies. It aims to furnish a comprehensive understanding of the intricate workings of these systems and their implementation details. Following this, Chapter 5 will shift the focus towards evaluating these methods, discussing the results obtained, and analyzing their implications and effectiveness within the broader context of our research objectives.

Our objective is to generate high-quality motion interactions using textual descriptions as a conditioning element. Moreover, we aim to enhance the generation process by employing individual descriptions to partially condition the individual motions within the interaction. In Section 4.1, we will introduce a model capable of generating textual descriptions from individual motions. Section 4.2 will detail a methodology for augmenting existing textual descriptions within an interaction dataset. Finally, Section 4.3 will present our proposed approach for creating motion interactions from textual descriptions.

4.1 Motion-to-Text

This segment of my master's thesis focuses on the development of a M2T model. Our hypothesis posits that the use of individual descriptions of motions performed by participants in an interaction, as a partial conditioning factor, will enhance the generation of interactive motions. A challenge arises with the InterHuman dataset, which lacks textual descriptions for individual motions. Consequently, we decided to develop our own M2T model to serve as a baseline.

As discussed in Section 2.3.2, prevailing methodologies in this domain typically treat the problem as a translation task, converting motion tokens into text tokens. However, this approach might lead to overfitting specific to the dataset employed due to the limited diversity in the textual descriptions. Instead, our method extends the MotionCLIP framework [32], illustrated in Figure 4.1. As previously detailed, MotionCLIP employs an encoder-decoder architecture capable of encoding motion into CLIP [33] embeddings, which then enables the decoder to transform these embeddings back into motion representations. Building on this concept, we propose incorpo-

rating a text-decoder, denoted as $\mathcal{D}_{\text{text}}$, capable of translating encoded motion embeddings into textual descriptions.

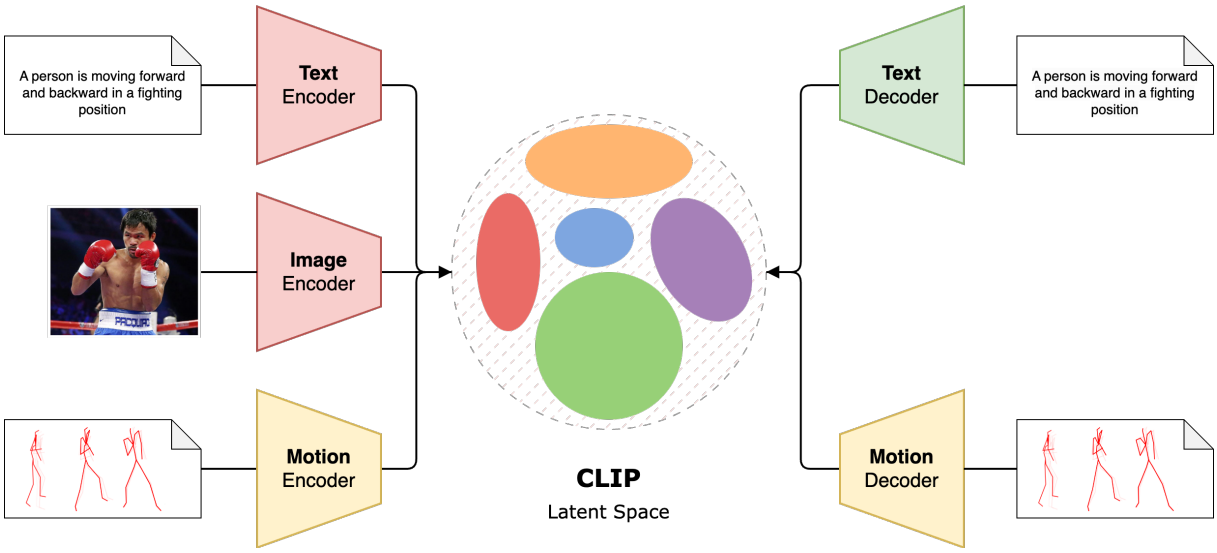


Figure 4.1: Motion-to-Text diagram. **red:** CLIP [33], **yellow:** MotionCLIP [32], **green:** Ours

The $\mathcal{D}_{\text{text}}$ was implemented as a 4-layer Transformer Decoder. This design choice is predicated on the hypothesis that leveraging CLIP embeddings would allow our model to achieve superior generalization, particularly in interpreting rare or out-of-distribution motion scenarios. In this implementation, the components from CLIP and MotionCLIP were frozen, allowing the focus to be solely on refining the text-generation aspect of the model.

This approach was initially trained using the AMASS dataset, chosen for its use in the foundational MotionCLIP research. However, as detailed in Section 3.2.1, the textual descriptions in the AMASS dataset are notably limited, resulting in the generation of descriptions that are often lacking in detail and syntactic coherence. To address this, we altered our architecture to support the HumanML3D format, leading to the retraining of both the motion Encoder-Decoder from the MotionCLIP framework and our model's Text Decoder. This modification was driven by the objective of training a model capable of generating more comprehensive and detailed textual descriptions of motion data, thereby enhancing the model's ability to provide extended explanations and improving its overall effectiveness in interpreting complex motion scenarios.

4.1.1 Loss Function

In the development of our proposed model, we have elected to employ the Negative Log-Likelihood (NLL) as our primary loss function. This choice aligns with standard practices in the field, particularly for training autoregressive models of this nature.

Let $X = \{x_1, x_2, ..., x_n\}$ denote a sequence of tokens (such as words or characters) in a given text.

The autoregressive model is designed to estimate the probability of each token x_t conditional on the preceding tokens $x_1, x_2, ..., x_{t-1}$. Mathematically, this relationship can be expressed as:

$$P(X) = \prod_{t=1}^{n} P(x_t | x_1, x_2, ..., x_{t-1}) \tag{4.1}$$

The NLL for this model is formulated as follows:

$$\mathcal{L}_{\mathcal{D}_{\text{text}}} = -\sum_{t=1}^{n} \log P(x_t | x_1, x_2, ..., x_{t-1}) \tag{4.2}$$

In practical applications, the computation of conditional probabilities $P(x_t | x_1, x_2, ..., x_{t-1})$ is carried out by our $\mathcal{D}_{\text{text}}$ model. The primary objective in training this model is to minimize the NLL. Achieving this minimization effectively maximizes the likelihood of the observed data under the model's parameters, leading to more accurate and reliable text generation.

4.2 InterHuman Textual Descriptions Augmentation

The introduction of the InterHuman dataset significantly impacted our thesis. Transitioning from a context of limited data resources to one with a substantially larger and fully annotated dataset, our focus shifted correspondingly. An initial analysis of this dataset revealed a critical insight: despite the diversity of motions it contained, the textual descriptions associated with these interactions were notably homogenous.

Building on our prior experience with LLMs in this thesis, we identified a promising opportunity to employ these models for augmenting the diversity of the dataset's textual descriptions. Our approach involved feeding the original interaction descriptions from the dataset into the LLMs and prompting it to generate new, varied, yet plausible textual descriptions. This relatively straightforward method proved effective, enabling us to expand the dataset's textual description volume by two to threefold.

LLM Prompt

Having the description of an interaction, generate a list of other possible description variations for the same interaction. Each of the new descriptions must be significantly different from the original.
Original Interaction Description: Oppenheimer and Einstein walked side by side in the garden and had profound discussions.
- **Interaction Description Variations** -

1. Two people are walking side by side while having a conversation.
2. One person is walking while moving his head and hands, while the other is doing the same.
3. Two people are going for a walk while arguing about something.

Original Interaction Description: <interaction motion description>

For this phase, we exclusively utilized GPT3.5 from OpenAI. Although the more advanced GPT4

was available, its higher cost led us to opt for the former version. It's worth noting that during the course of this thesis, notable Open Source LLMs were released. However, our extensive experience and accumulated expertise in prompt engineering with OpenAI's models influenced our decision to continue using these, despite the availability of potentially viable alternatives.

4.3 Diffusion-based Text-to-Interaction

The culmination of our research and the central contribution of this thesis is the development of a diffusion model for generating motion interactions from textual descriptions.

The architecture of our proposed model extends beyond the framework presented in the InterGen paper [47]. It is founded on the principle that interactions between two persons exhibit a commutative property. Specifically, an interaction involving two individuals, denoted as $\{x_a, x_b\}$, is considered equivalent to $\{x_b, x_a\}$. This implies that the distribution of interaction data adheres to the following symmetry:

$$p\left(\mathbf{x}_a, \mathbf{x}_b\right) \equiv p\left(\mathbf{x}_b, \mathbf{x}_a\right) \tag{4.3}$$

Building on this concept, the authors introduce what they term as cooperative denoisers. These consist of two transformer networks operating in a Siamese configuration, sharing parameters. Each network is responsible for processing its respective noisy motion inputs, $\mathbf{x}_a^{(t)}$ and $\mathbf{x}_b^{(t)}$, and aims to produce the denoised versions, $\mathbf{x}_a$ and $\mathbf{x}_b$. This approach diverges from traditional methods in diffusion models but aligns with the latest trends in motion domain applications. A more in-depth discussion on this will be presented in Section 4.3.1. The denoising prediction process is conditioned on several factors: the diffusion timestep t, a control condition c, and the hidden states $\mathbf{h}^{(i)}$ from the counterpart network.

The process begins with embedding the noisy motion into a shared latent space, followed by positional encoding to transform it into an internal representation, termed as the initial hidden states $\mathbf{h}^{(0)}$. This representation then passes through N attention-based blocks, leading to the generation of denoised hidden states $\mathbf{h}^{(N)}$. The final step involves applying a common inverse embedding layer to produce the denoised motion output.

Each block within the architecture consists of two multi-head attention layers (denoted as $Attn$) followed by a feed-forward network (denoted as FF). The first attention layer, a self-attention layer, serves to embed the current hidden states $\mathbf{h}^{(i)}$ into a context vector $\mathbf{c}^{(i)}$. The computation of $\mathbf{c}_a^{(i)}$ is formulated as follows:

$$\mathbf{c}_a^{(i)} = \text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{C}}\right)\mathbf{V} \tag{4.4}$$

$$\mathbf{Q} = \mathbf{h}_a^{(i)}\mathbf{W}_s^Q, \mathbf{K} = \mathbf{h}_a^{(i)}\mathbf{W}_s^K, \mathbf{V} = \mathbf{h}_a^{(i)}\mathbf{W}_s^V \tag{4.5}$$

Here, C represents the number of channels in the attention layer, and $\mathbf{W}_s$ are trainable weights.

The computation for $\mathbf{c}_b^{(i)}$ follows the same process, utilizing the shared weights $\mathbf{W}_s$.

The second attention layer, the mutual attention layer, takes the key $\mathbf{K}$ and value $\mathbf{V}$ from the hidden states $\mathbf{h}^{(i)}$ of the counterpart block. This design incorporates the embedded representation of the other motion in the interaction. The process for computing the next hidden states $\mathbf{h}^{(i+1)}$ is given by:

$$\mathbf{h}_a^{(i+1)} = FF\left(\text{Attn}\left(\mathbf{Q}_a, \mathbf{K}_b, \mathbf{V}_b\right)\right) \tag{4.6}$$

$$\mathbf{h}_b^{(i+1)} = FF\left(\text{Attn}\left(\mathbf{Q}_b, \mathbf{K}_a, \mathbf{V}_a\right)\right) \tag{4.7}$$

$$\mathbf{Q}_a = \mathbf{c}_a^{(i)}\mathbf{W}_m^Q, \mathbf{K}_a = \mathbf{h}_a^{(i)}\mathbf{W}_m^K, \mathbf{V}_a = \mathbf{h}_a^{(i)}\mathbf{W}_m^V \tag{4.8}$$

$$\mathbf{Q}_b = \mathbf{c}_b^{(i)}\mathbf{W}_m^Q, \mathbf{K}_b = \mathbf{h}_b^{(i)}\mathbf{W}_m^K, \mathbf{V}_b = \mathbf{h}_b^{(i)}\mathbf{W}_m^V \tag{4.9}$$

In this case, $\mathbf{W}_m$ are the shared weights between the two branches of the model.

Furthermore, adaptive layer normalization is applied before each attention layer and the feed-forward network. This normalization is conditioned on the control condition c and the diffusion timestep t, ensuring that the model's processing is fine-tuned to these specific parameters at each step.

Our Proposal

The exploration of attention mechanisms in this context reveals that self-attention primarily focuses on the motion of the individual currently undergoing denoising. In contrast, the mutual attention layer incorporates not only the motion of the individual being denoised but also the latent representation of the other individual's motion within the interaction.

Building on this understanding, we propose the implementation of distinct conditionings for the different attention layers. Specifically, for the mutual attention layer, the textual description of the overall interaction provided by the dataset remains the conditioning element. This choice is justified by the layer's consideration of both individuals' motions in the interaction.

Conversely, for the self-attention layer, which focuses on the individual motion being denoised, a different approach is warranted. The textual description of the overall interaction lacks the specificity required for individual motion conditioning. To address this, our methodology introduces a textual description that is exclusive to the individual motion. This description is derived using a M2T method akin to those discussed in previous sections. The proposed conditioning strategy can be formally expressed as follows:

$$c = c_{\text{interaction}} \oplus c_{\text{individual-a}} \oplus c_{\text{individual-b}} \tag{4.10}$$

In this formula, $c_{\text{interaction}} \in \mathbb{R}^l$ represents the interaction description, $c_{\text{individual-a}} \in \mathbb{R}^l$ denotes the motion description of one individual in the interaction, and $c_{\text{individual-b}} \in \mathbb{R}^l$ is the textual representation of the other individual's motion. Here, l signifies the dimensionality of the CLIP latent representation of the textual description. And $\oplus$ represents a tensor concatenation. This innovative conditioning strategy ensures more precise and effective attention layer functioning, tailored to the nuances of individual and combined motions within interactions.

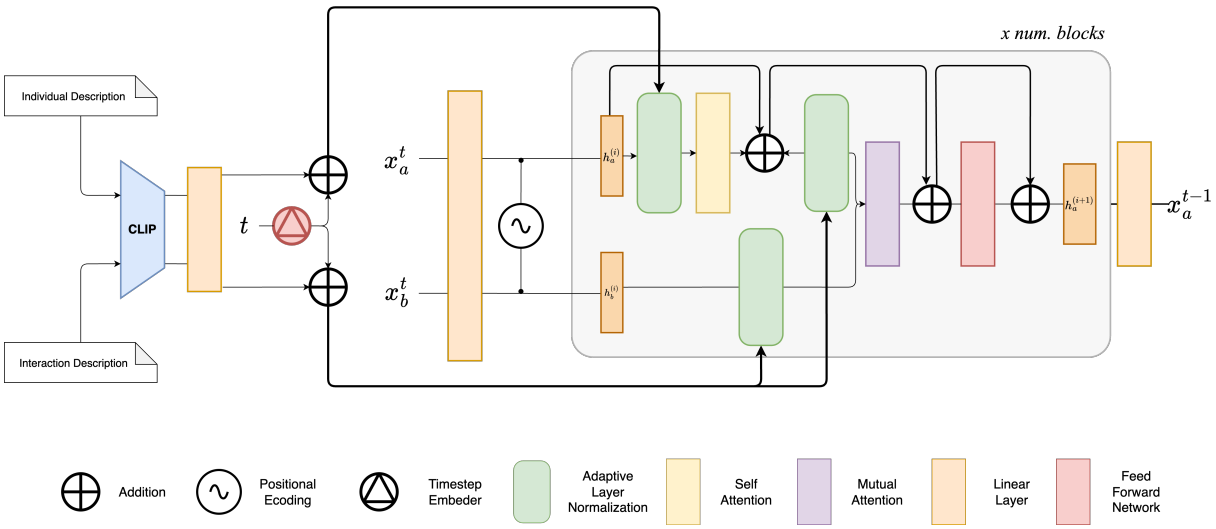


Figure 4.2: Transfomer-based denoiser architecture.

The remainder of the architecture remains consistent, with a key distinction in the adaptive layer normalization process, to the InterGen architecture. The denoiser architecture can be visualized in Figure 4.2. In this aspect, the choice of conditioning is contingent on whether the layer precedes the self-attention or the mutual attention mechanism. This differentiation in conditioning aligns with the respective focus of each attention layer, ensuring that the adaptive layer normalization is appropriately tuned to the specific requirements of the subsequent attention process, be it self-focused or mutually inclusive of both individuals' motions.

4.3.1 Loss Functions

Following the introduction of our network's architecture, this section delineates the loss functions employed for parameter optimization. As previously mentioned in Section 2.1, standard practice involves the network predicting the noise term ϵ_t. However, this convention precludes the utilization of more intricate, motion-related loss functions. To address this, our approach deviates by predicting x_0 directly [40, 62]. The primary, or 'simple', loss function is formulated as:

$$\mathcal{L}_{\text{simple}} = E_{x_0 \sim q(x_0|c), t \sim [1,T]} \left[\|x_0 - G(x_t, t, c)\|_2^2 \right] \tag{4.11}$$

Beyond the simple loss, we incorporate geometric loss functions, which are pivotal in upholding

physical accuracy and minimizing artefacts, thus fostering natural and coherent motion in generated sequences. In this thesis, we integrate two prevalent geometric losses that are instrumental in regulating foot contact and velocities.

$$\mathcal{L}_{\text{foot}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left\| \left(FK\left(\hat{x}_0^{i+1}\right) - FK\left(\hat{x}_0^{i}\right) \right) \cdot f_i \right\|_2^2 \tag{4.12}$$

$$\mathcal{L}_{\text{vel}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left\| \left(x_0^{i+1} - x_0^{i} \right) - \left(\hat{x}_0^{i+1} - \hat{x}_0^{i} \right) \right\|_2^2 \tag{4.13}$$

In these equations, $FK(\cdot)$ represents the forward kinematics function, which translates joint rotations into spatial joint positions. The term $f_i \in \{0,1\}^J$ denotes a binary foot contact mask for each frame i. This mask, pertinent exclusively to the feet, signifies ground contact and is determined based on binary ground truth data. It plays a crucial role in mitigating the foot-sliding phenomenon by neutralizing velocities when the feet are in contact with the ground.

In addition to employing the geometric losses foundational to the MDM framework, the researchers in InterGen introduce further loss functions tailored to their unique, non-canonical representation. A notable addition is the Bone Length Loss, $\mathcal{L}_{BL}$, designed to maintain skeletal consistency by constraining global joint positions. This loss implicitly captures the kinematic structure of the human body. The Bone Length Loss is mathematically expressed as:

$$\mathcal{L}_{BL} = \left\| B\left(\hat{\mathbf{x}}_0\right) - B\left(\mathbf{x}_0\right) \right\|_2^2 \tag{4.14}$$

Here, B denotes the calculation of bone lengths within a predefined kinematic tree of the human body, based on the global joint positions represented in $\mathbf{x}$. It's important to note that all previously discussed losses, including the Bone Length Loss, are applied to each individual motion within the interaction. This comprehensive approach ensures that the generated motions are not only realistic and coherent but also maintain structural integrity with respect to individual skeletal dynamics.

To adeptly manage the intricacies of spatial dynamics in multi-person interactions, the authors of InterGen introduce interactive loss functions. These consist of the Masked Joint Distance Map (DM) loss and the Relative Orientation (RO) loss. The DM loss quantifies the spatial interplay between two individuals by measuring and aligning the $N_j \times N_j$ joint distance maps with the ground truth, where N_j represents the number of joints per person. The DM loss is formulated as:

$$\mathcal{L}_{DM} = \left\| \left(M\left(\hat{\mathbf{x}}_a, \hat{\mathbf{x}}_b\right) - M\left(\mathbf{x}_a, \mathbf{x}_b\right) \right) \odot I\left(M_{xz}\left(\mathbf{x}_a, \mathbf{x}_b\right) < \bar{M} \right) \right\|_2^2 \tag{4.15}$$

In this context, M is the function generating the joint distance map for two people, derived from their global joint positions. $I(\cdot)$ is an indicator function that applies a masking mechanism based on a 2D distance threshold on the XZ-plane. This masking selectively activates the loss, primarily when the horizontal distance between the two individuals is within a specified limit. M_{xz} denotes the distance map projected onto the XZ-plane, and $\bar{M}$ is the predetermined

distance threshold. The symbol $\odot$ represents the Hadamard product. This strategic approach ensures that the DM loss is effectively utilized to reinforce spatially realistic and contextually appropriate interactions between individuals nearby.

The Relative Orientation (RO) loss, another critical component of the interactive losses, is designed to accurately estimate and align the relative orientation of two individuals with the ground truth. This loss plays a pivotal role in ensuring that the interactions between the characters are spatially and contextually realistic. The formulation of the RO loss is as follows:

$$\mathcal{L}_{RO} = \left\| O\left(IK\left(\hat{\mathbf{x}}_a\right), \mathrm{IK}\left(\hat{\mathbf{x}}_b\right)\right) - O\left(IK\left(\mathbf{x}_a\right), IK\left(\mathbf{x}_b\right)\right) \right\|_2^2 \tag{4.16}$$

Here, $IK(\cdot)$ denotes the inverse kinematics process, which computes the joint rotations from the given joint positions. O refers to the calculation of 2D relative orientation between the two people, particularly around the Y-axis, based on these rotations. These interactive losses, specifically the RO loss, play an indispensable role in constraining both the positional spatial relations and the relative frontal orientation of the two interacting individuals. This constraint is essential to mirror the natural dynamics and nuances inherent in human interactions.

The comprehensive loss function employed in our methodology is an amalgamation of various individual loss components, each targeting specific aspects of motion realism and consistency. This loss function is articulated as:

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{vel}}\, \mathcal{L}_{\text{vel}} + \lambda_{\text{foot}}\, \mathcal{L}_{\text{foot}} + \lambda_{BL}\mathcal{L}_{BL} + \lambda_{DM}\mathcal{L}_{DM} + \lambda_{RO}\mathcal{L}_{RO} \tag{4.17}$$

In this formulation, each term is weighted by a corresponding coefficient λ, signifying the relative importance of each loss component in the overall optimization process.

Building upon insights from Physdiff [41], which indicated that network-predicted denoised motion becomes increasingly implausible with larger diffusion timesteps t (or higher noise levels), the authors of InterGen developed an innovative diffusion training strategy. We adopt this approach in our work, involving the truncation of diffusion timesteps using a threshold $\bar{t}$. Regularization loss is applied only when the sampled timestep t is below this threshold. Consequently, the total loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{simple}} + \lambda_{\text{reg}}\, \mathbb{E}_t \left[I(t \leq \bar{t}) \cdot \mathcal{L}_{\text{reg}}^{(t)} \right] \tag{4.18}$$

In this equation, $I(t \leq \bar{t})$ operates as an indicator function, effectively omitting the regularization term when $t > \bar{t}$. This nuanced approach allows for a more targeted and effective training regime, focusing on the most critical aspects of motion realism while managing the challenges associated with higher noise levels in the diffusion process.

4.3.2 Multi-Weight Classifier-Free Guidance

During each timestep t in our model's operation, the clean sample $\hat{x}_0$ is predicted using the formula $\hat{x}_0 = G(x_t, t, c)$, after which it is re-noised to obtain x_{t-1}. This process aligns with

the classifier-free guidance approach, as detailed in Section 2.1.3. The core principle of this approach involves balancing the trade-off between diversity and fidelity in the model's output. This balance is achieved by interpolating, or in some cases extrapolating, between conditioned and unconditioned variants of the model's predictions, using a specific weight factor.

However, our approach introduces a nuanced layer of complexity due to the presence of different conditions. We address this by employing distinct sampling and weighting strategies for each condition. The equation representing our model's sampling process, denoted as G_s, is as follows:

$$\begin{aligned} G_s\left(x_t, t, c\right) = & G\left(x_t, t, \emptyset\right) \\ & + w_1 \cdot \left(G\left(x_t, t, c\right) - G\left(x_t, t, \emptyset\right)\right) \\ & + w_2 \cdot \left(G\left(x_t, t, c_{\text{interaction}}\right) - G\left(x_t, t, \emptyset\right)\right) \\ & + w_3 \cdot \left(G\left(x_t, t, c_{\text{individual}}\right) - G\left(x_t, t, \emptyset\right)\right) \end{aligned} \tag{4.19}$$

In this equation, $G_s(x_t, t, c)$ represents the sampling function, $G(x_t, t, \emptyset)$ is the unconditioned output of the model, and the terms $G(x_t, t, c)$, $G(x_t, t, c_{\text{interaction}})$, and $G(x_t, t, c_{\text{individual}})$ denote the model outputs conditioned on different types of context. The weights w_1, w_2, and w_3 adjust the influence of each conditioned output relative to the unconditioned baseline. This method allows for a more refined control over the generation process, ensuring that the model can effectively capture and express the nuances of both individual and interaction-specific conditions. If a weight is set to 0, it indicates that the generation using that particular conditioning is not considered in the overall prediction.

A notable limitation of this approach is the necessity to perform quadruple sampling from the denoiser, as opposed to the dual sampling required in a conventional classifier-free guidance methodology. This increase in sampling demand may impact the method's efficiency.

5 Results

Having delineated our novel contributions and methodologies in this thesis, we now turn our attention to the empirical evaluation of these proposals. While we have introduced various concepts such as T2M and M2T, our primary focus will be on the results emanating from interaction generation. This aspect constitutes the crux of the thesis, and by examining the outcomes of the experiments dedicated to interaction generation from textual descriptions, we can indirectly gauge the efficacy of the various methodologies proposed in the preceding chapters.

The subsequent sections of this document are organized to offer a thorough evaluation of our research. Section 5.1 is dedicated to delineating the metrics used for evaluating the performance of our proposed methodologies. This is followed by Section 5.2, where we will detail the implementation aspects and training parameters that underpin our experimental setup. Next, Section 5.3 is reserved for presenting and analyzing the quantitative results obtained from the various experiments conducted as part of this thesis. Finally, Section 5.4 will focus on providing a qualitative assessment of the predictions generated by our method.

5.1 Evaluation Metrics

In order to effectively evaluate and benchmark any proposed methodology, particularly in the realm of machine learning, it is crucial to have clear and relevant criteria. While loss functions play a pivotal role in the learning and parameter optimization of machine learning models, their values often lack direct interpretability. This is where metrics come into play. Metrics provide human-interpretable values that offer insights into a model's functionality. The choice of metrics depends significantly on the specific task at hand, as different metrics yield different types of information relevant to that task. Moreover, in research, these metrics are instrumental for comparing methodologies and determining the most effective approaches for specific tasks and datasets.

For the task of generating human motion from textual descriptions, we have adopted the same metrics as those used in InterGen [47], which, are the same as the proposed in HumanML3D [55]. As our goal is to surpass the current state-of-the-art (SOTA) performance on the InterHuman dataset, employing the same evaluation procedure is crucial for a valid comparison between our method and existing approaches.

Concerning the evaluation metrics for this generative task, the necessity of a feature extrac-

tor for both motions and textual descriptions is evident. We have utilized the feature extractor provided by the InterGen authors. This extractor is trained with contrastive losses, enabling it to produce aligned feature vectors between the motions and the corresponding textual descriptions.

Frechet Inception Distance (FID)

The Frechet Inception Distance (FID) is a highly regarded metric for assessing the quality of predictions produced by generative models. To compute FID, we first need to extract feature vectors from both the generated motions and the ground-truth motions. These features are used to characterize their respective statistical distributions. The FID is then calculated as the distance between these two distributions. A lower FID score is indicative of better model performance, as it suggests a closer similarity between the statistical distributions of the generated motions and the ground-truth motions. This similarity is crucial in evaluating how well the generative model replicates the expected motion characteristics.

Diversity

Diversity, in the context of evaluating generative models, refers to the variance observed across all the motions generated by the model in response to the various textual descriptions from the test set. An ideal model would exhibit a level of diversity closely matching that of the ground truth. Such alignment indicates that the motions generated by our model, based on the provided textual descriptions, are consistent with the diversity inherent in the dataset's actual motion data.

Multimodality (MModality)

Multimodality, as an evaluation metric, measures the average variance among the motions generated by a model for a given textual description. A higher Multimodality score is desirable, as it indicates that the model is capable of generating a more diverse array of motions from the same textual description while still maintaining alignment with the specified conditioning. This metric is crucial for determining the model's ability to produce varied yet contextually appropriate motions, reflecting the natural diversity found in human responses to the same set of instructions or descriptions.

R-Precision

R-Precision is a metric that evaluates how well the latent representations of generated motions align with their conditioning textual descriptions. To compute this metric, for each motion generated in the test set, its correct textual description (from the test set) is paired with 31 other randomly selected, mismatched textual descriptions. A ranking is then established based on the Euclidean distance between the feature vector of the generated motion and the feature vectors of each of these textual descriptions. The effectiveness of the model is assessed by calculating the average accuracy for the highest-ranking (Top-1), second-highest (Top-2), and third-highest (Top-3) matches in this ranking. A higher R-Precision score is indicative of better performance, signifying that the generated motions are more closely aligned with the textual descriptions intended to guide their creation.

Multimodal Distance (MM Dist)

Multimodal Distance (MM Dist) is a metric that quantifies the closeness of the generated motions to their corresponding textual descriptions in the latent space. It is calculated as the average Euclidean distance between the motion feature vector and the text feature vector for each generated motion in the test set. A lower MM Dist score indicates better performance of the model. This is because a lower distance signifies that the latent representations of the motions generated by the model and their textual descriptions are closely aligned in the latent space.

5.2 Implementation Details & Training Parameters

Our methodology is implemented with $N = 8$ blocks, each featuring a latent dimension of 1024. Additionally, each attention layer within these blocks comprises 8 heads. For text encoding, we utilize a frozen CLIP-ViT$L/14$ model [33], resulting in a total of 182M trainable parameters.

During training, we set the number of diffusion timesteps to 1,000. For sampling, the DDIM strategy [27] is applied, with 50 timesteps and with a noise reduction coefficient $\eta = 0$. We incorporate the cosine noise level schedule [63] and employ our proposed classifier-free guidance variation. In this setup, 10% of the random CLIP embeddings are set to zero during training, and a guidance coefficient of 3.5 is used during sampling.

The training of all models is conducted using the AdamW [64] optimizer. This optimizer features betas of $(0.9, 0.999)$, a weight decay of 2×10^{-5}, a maximum learning rate of 10^{-4}, and a cosine learning rate schedule with an initial 10-epoch linear warm-up period. To balance the disparity in magnitude orders among various loss terms as outlined in Eq. 4.17, we adjust them using multiple hyper-parameters. This reweighting harmonizes these terms to a similar order of magnitude, ensuring an equitable contribution from each. Specifically, we set $\lambda_{\text{vel}} = 30, \lambda_{\text{foot}} = 30, \lambda_{BL} = 10, \lambda_{DM} = 3$, and $\lambda_{RO} = 0.01$. For Eq. 4.18, $\lambda_{\text{reg}} = 1$ is maintained across all experiments. The diffusion denoisers are trained over 2,000 epochs with a batch size of 16, utilizing an Nvidia 3090 GPU. The entire training process spans over 288 hours.

5.3 Quantitative Results

With a clear understanding of the evaluation metrics in place, we are now positioned to conduct various experiments and compare our proposed methodology with existing approaches. Besides comparing our approach with InterGen, the method introduced by the creators of the InterHuman dataset, these authors also have adapted several existing methods, training and evaluating them on the InterHuman dataset, thus providing additional points of comparison.

Owing to the challenges in replicating the evaluation metrics results reported in the InterGen paper, as evidenced in Table 5.1, we have decided to use the metrics derived from our retrained version of the InterGen model as the standard for comparison against our proposed method.For this retrained version of InterGen, we adhered to the same code and parameters as those specified

in the method's official implementation. The only alteration was in the batch size, where we used a batch size of 16 instead of the originally specified 64, a change necessitated by our hardware constraints.

Methods	R Precision ↑			FID ↓	MM Dist ↓	Diversity →	MModality ↑
	Top 1	Top 2	Top 3				
Real	$0.452^{\pm.008}$	$0.610^{\pm.009}$	$0.701^{\pm.008}$	$0.273^{\pm.007}$	$3.755^{\pm.008}$	$7.948^{\pm.064}$	-
TEMOS	$0.224^{\pm.010}$	$0.316^{\pm.013}$	$0.450^{\pm.018}$	$17.375^{\pm.043}$	$6.342^{\pm.015}$	$6.939^{\pm.071}$	$0.535^{\pm.014}$
T2M	$0.238^{\pm.012}$	$0.325 \pm .010$	$0.464^{\pm.014}$	$13.769^{\pm.072}$	$5.731^{\pm.013}$	$7.046^{\pm.022}$	$1.387^{\pm.076}$
MDM	$0.153^{\pm.012}$	$0.260^{\pm.009}$	$0.339^{\pm.012}$	$9.167^{\pm.056}$	$7.125^{\pm.018}$	$\mathbf{7.602}^{\pm.045}$	$\mathbf{2.35}^{\pm.080}$
ComMDM*	$0.067^{\pm.013}$	$0.125^{\pm.018}$	$0.184^{\pm.015}$	$38.643^{\pm.098}$	$14.211^{\pm.013}$	$3.520^{\pm.058}$	$0.217^{\pm.018}$
ComMDM	$0.223^{\pm.009}$	$0.334^{\pm.008}$	$0.466^{\pm.010}$	$7.069^{\pm.054}$	$6.212^{\pm.021}$	$7.244^{\pm.038}$	$1.822^{\pm.052}$
InterGen	$\mathbf{0.371} \pm .010$	$\mathbf{0.515}^{\pm.012}$	$\mathbf{0.624}^{\pm.010}$	$\mathbf{5.918}^{\pm.079}$	$\mathbf{5.108}^{\pm.014}$	$7.387^{\pm.029}$	$2.141^{\pm.063}$
InterGen (Re-trained)	$0.252^{\pm0.007}$	$0.382^{\pm0.005}$	$0.466^{\pm0.005}$	$10.780^{\pm0.179}$	$3.882^{\pm0.002}$	$7.808^{\pm0.027}$	$1.534^{\pm0.061}$

Table 5.1: Method comparison in InterHuman dataset

5.3.1 InterHuman Augmentation Evaluation

Following the establishment of a suitable baseline for comparison, as outlined in our methodology, we proceeded with various experimental validations. The initial aspect we aimed to scrutinize was the efficacy of our proposed augmentation technique, as detailed in Section 4.2. We postulated that enriching the model with an expanded range of textual descriptions for identical motions would lead to a more comprehensive model capable of representing a broader array of motions corresponding to diverse textual inputs.

To establish a point of reference, we initially trained the InterGen model solely with the textual descriptions available in the dataset. Subsequently, the model was trained incorporating the additional textual descriptions generated by our approach. Additionally, we evaluated a model trained exclusively with a single textual description from the original dataset. The outcomes of these experimental setups are presented in Table 5.2.

The evaluation metrics clearly indicate that the model trained with augmented textual descriptions achieves the highest score in Multimodality (MModality). As previously discussed, a higher Multimodality score implies a broader distribution of generated motions for identical textual descriptions, aligning with our expectations. However, it is noteworthy that in other metrics, the model trained with a single textual description from the original set exhibits superior performance. This can be primarily attributed to the fact that this model is more overfitted to the dataset. Given the abundance of similar motions associated with similar textual descriptions in the dataset, the loss of information resulting from limiting the model to one textual description per motion is relatively negligible.

For all other experiments, we utilized the standard InterHuman dataset. This approach was

chosen to ensure that our comparisons align as closely as possible with the results presented in the InterGen paper.

Methods	R Precision ↑			FID ↓	MM Dist ↓	Diversity →	MModality ↑
	Top 1	Top 2	Top 3				
InterHuman	$0.252^{\pm0.007}$	$0.382^{\pm0.005}$	$0.466^{\pm0.005}$	$10.780^{\pm0.179}$	$3.882^{\pm0.002}$	$7.808^{\pm0.027}$	$1.534^{\pm0.061}$
InterHuman Reduced	$\mathbf{0.337}^{\pm0.009}$	$\mathbf{0.477}^{\pm0.009}$	$\mathbf{0.564}^{\pm0.008}$	$\mathbf{6.388}^{\pm0.099}$	$\mathbf{3.839}^{\pm0.002}$	$\mathbf{7.951}^{\pm0.030}$	$1.075^{\pm0.044}$
InterHuman Plus	$0.149^{\pm0.006}$	$0.242^{\pm0.007}$	$0.308^{\pm0.010}$	$18.689^{\pm0.271}$	$3.960^{\pm0.002}$	$7.703^{\pm0.031}$	$\mathbf{1.744}^{\pm0.081}$

Table 5.2: Experiments in the InterHuman Textual Descriptions Augmentation

5.3.2 M2T Evaluation

This section delves into a series of experiments conducted to evaluate our proposed architecture aimed at enhancing T2I generation. We utilized the previously defined and self-retrained baseline as the comparative standard.

In the initial experimental phase, we juxtaposed the baseline with our novel architecture. One model was trained using individual textual descriptions derived via our M2T methodology outlined in Section 4.1. Another model employed MotionGPT, which at the time of writing this thesis, stood as the SOTA in M2T conversion. The outcomes of these experiments are detailed in Table 5.3. Interestingly, the model leveraging our method exhibited superior performance, a phenomenon we will elucidate at the end of this section.

Methods	R Precision ↑			FID ↓	MM Dist ↓	Diversity →	MModality ↑
	Top 1	Top 2	Top 3				
InterGen	$0.252^{\pm0.007}$	$0.382^{\pm0.005}$	$0.466^{\pm0.005}$	$10.780^{\pm0.179}$	$3.882^{\pm0.002}$	$7.808^{\pm0.027}$	$\mathbf{1.534}^{\pm0.061}$
Ours + M2T	$\mathbf{0.297}^{\pm0.007}$	$\mathbf{0.436}^{\pm0.009}$	$\mathbf{0.523}^{\pm0.009}$	$\mathbf{7.794}^{\pm0.164}$	$\mathbf{3.853}^{\pm0.001}$	$\mathbf{7.843}^{\pm0.024}$	$1.358^{\pm0.043}$
Ours + MGPT	$0.277^{\pm0.008}$	$0.413^{\pm0.008}$	$0.504^{\pm0.006}$	$8.010^{\pm0.121}$	$3.862^{\pm0.002}$	$7.842^{\pm0.050}$	$1.422^{\pm0.036}$

Table 5.3: M2T evaluation using different M2T models

Using our proposed method for M2T, we conducted an ablation study focusing on the different weights applied during sampling, as per the classifier-free guidance approach described in Section 4.3.2. These results are presented in Table 5.4. Overall, our proposal appears to outperform the baseline, particularly in metrics like R-precision when more weight is given to interaction. Conversely, models with a more balanced approach fare better in terms of FID.

However, it is important to acknowledge that this evaluation was not entirely fair. The InterHuman dataset lacks individual motion textual descriptions for interactions, leading us to generate these descriptions. The same methods were applied for annotating motions in both the training and test sets, resulting in our M2T models having exposure to test set motions during annotation. This scenario creates an information leakage, inadvertently providing cues to the generative model.

CFG Weights			R Precision ↑			FID ↓	MM Dist ↓	Diversity →	MModality ↑
w1	w2	w3	Top 1	Top 2	Top 3				
3.5	0	0	$0.297^{\pm0.007}$	$0.436^{\pm0.009}$	$0.523^{\pm0.009}$	$7.794^{\pm0.164}$	$3.853^{\pm0.001}$	$7.843^{\pm0.024}$	$1.358^{\pm0.043}$
3.5	0	3.5	$0.253^{\pm0.007}$	$0.376^{\pm0.007}$	$0.458^{\pm0.007}$	$9.531^{\pm0.088}$	$3.876^{\pm0.001}$	$7.487^{\pm0.045}$	$1.333^{\pm0.021}$
3.5	3.5	0	$\mathbf{0.314}^{\pm0.008}$	$\mathbf{0.460}^{\pm0.008}$	$\mathbf{0.554}^{\pm0.007}$	$8.198^{\pm0.145}$	$\mathbf{3.847}^{\pm0.003}$	$8.159^{\pm0.042}$	$1.345^{\pm0.065}$
3.5	3.5	3.5	$0.299^{\pm0.007}$	$0.442^{\pm0.011}$	$0.529^{\pm0.012}$	$\mathbf{6.963}^{\pm0.175}$	$3.852^{\pm0.003}$	$\mathbf{7.970}^{\pm0.027}$	$1.352^{\pm0.041}$
3.5	3.5	7	$0.262^{\pm0.006}$	$0.392^{\pm0.007}$	$0.475^{\pm0.006}$	$8.626^{\pm0.120}$	$3.873^{\pm0.002}$	$7.756^{\pm0.047}$	$1.386^{\pm0.048}$
3.5	7	3.5	$0.126^{\pm0.006}$	$0.201^{\pm0.005}$	$0.256^{\pm0.006}$	$18.127^{\pm0.347}$	$3.984^{\pm0.002}$	$7.721^{\pm0.037}$	$\mathbf{1.613}^{\pm0.017}$

Table 5.4: M2T evaluation using different weights

This reveals why the model utilizing our annotation method emerged as the most effective. Our model, despite being significantly less capable than MotionGPT and highly overfitted to the HumanML3D dataset, benefited from the similarity between training and test set motions and descriptions. Consequently, the individual descriptions generated by our model, despite their lack of precision, ended up conveying excessive information, skewing the evaluation results.

5.3.3 LLM Evaluation

To address the issue of information leakage and conduct an equitable evaluation, we developed a method for annotating the test set without prior exposure to the motions. This approach involved utilizing a pre-trained LLMs to generate individual motion descriptions from a broader interaction description, using a specific prompt:

LLM Prompt

Having the description of an interaction, extract individual descriptions for the motions of each individual.

Interaction Description: In an intense boxing match, one person attacks the opponent with straight punch, and then the opponent falls over.
Individual Motion 1: One person is moving and then throws a punch.
Individual Motion 2: One person falls over and stays on the ground.
—
Interaction Description: <interaction motion description>

The inherent limitation of this method is that the LLMs generates hypothetical descriptions of the individual motions, which may or may not accurately reflect the actual events. Despite this, in the absence of a more viable alternative, this was the only feasible way to ensure a fair evaluation.

Similar to the previous section, we evaluated our proposal by applying two different techniques for creating individual textual descriptions from the training set. The results of this comparison are detailed in Table 5.5. Here, the results are more logical, showing that annotations using MotionGPT are significantly more effective than those generated by our method. However, in this unbiased assessment, the baseline model outperforms our proposed approach.

Methods	R Precision ↑			FID ↓	MM Dist ↓	Diversity →	MModality ↑
	Top 1	Top 2	Top 3				
InterGen	**0.252**$^{\pm0.007}$	**0.382**$^{\pm0.005}$	**0.466**$^{\pm0.005}$	**10.780**$^{\pm0.179}$	**3.882**$^{\pm0.002}$	**7.808**$^{\pm0.027}$	**1.534**$^{\pm0.061}$
Ours + M2T	0.081$^{\pm0.004}$	0.133$^{\pm0.006}$	0.173$^{\pm0.007}$	17.650$^{\pm0.222}$	4.015$^{\pm0.002}$	7.326$^{\pm0.038}$	1.406$^{\pm0.057}$
Ours + MGPT	0.127$^{\pm0.003}$	0.204$^{\pm0.004}$	0.260$^{\pm0.005}$	14.521$^{\pm0.252}$	3.974$^{\pm0.002}$	7.477$^{\pm0.042}$	1.483$^{\pm0.038}$

Table 5.5: LLM evaluation using different M2T models

An extensive ablation study was conducted on the model using MotionGPT for individual motion descriptions. The findings of this study are presented in Table 5.6. Notably, specific weight configurations led to substantial improvements over the original results, even surpassing the performance of the InterGen model.

The most successful experiments were those that heavily favored the interaction description, in some cases completely disregarding the other weights. This suggests that the individual motion descriptions provided by the LLMs might not contribute valuable information for motion generation. In fact, being so divergent from the descriptions used during training, they likely introduce more noise than useful guidance.

CFG Weights			R Precision ↑			FID ↓	MM Dist ↓	Diversity →	MModality ↑
w1	w2	w3	Top 1	Top 2	Top 3				
3.5	0	0	0.127$^{\pm0.003}$	0.204$^{\pm0.004}$	0.260$^{\pm0.005}$	14.521$^{\pm0.252}$	3.974$^{\pm0.002}$	7.477$^{\pm0.042}$	1.483$^{\pm0.038}$
0	3.5	0	0.263$^{\pm0.005}$	0.405$^{\pm0.007}$	0.490$^{\pm0.006}$	10.538$^{\pm0.220}$	3.877$^{\pm0.003}$	8.169$^{\pm0.045}$	1.506$^{\pm0.046}$
0	5	0	0.276$^{\pm0.006}$	0.409$^{\pm0.009}$	0.501$^{\pm0.009}$	10.633$^{\pm0.144}$	3.870$^{\pm0.002}$	8.315$^{\pm0.035}$	1.489$^{\pm0.044}$
0	7	0	**0.280**$^{\pm0.008}$	**0.423**$^{\pm0.004}$	**0.514**$^{\pm0.005}$	11.893$^{\pm0.238}$	**3.869**$^{\pm0.002}$	8.301$^{\pm0.045}$	1.435$^{\pm0.068}$
2	3.5	0	0.253$^{\pm0.007}$	0.385$^{\pm0.008}$	0.469$^{\pm0.008}$	**8.940**$^{\pm0.187}$	3.882$^{\pm0.003}$	7.992$^{\pm0.028}$	1.463$^{\pm0.042}$
2	5	0	0.267$^{\pm0.008}$	0.399$^{\pm0.006}$	0.486$^{\pm0.005}$	9.492$^{\pm0.213}$	3.875$^{\pm0.002}$	8.112$^{\pm0.035}$	1.425$^{\pm0.063}$
2	7	0	0.264$^{\pm0.005}$	0.404$^{\pm0.005}$	0.495$^{\pm0.006}$	11.133$^{\pm0.208}$	3.877$^{\pm0.002}$	8.060$^{\pm0.043}$	1.454$^{\pm0.045}$
3.5	2.5	0	0.199$^{\pm0.006}$	0.312$^{\pm0.004}$	0.388$^{\pm0.005}$	10.088$^{\pm0.147}$	3.911$^{\pm0.002}$	7.776$^{\pm0.041}$	1.444$^{\pm0.095}$
3.5	3.5	0	0.221$^{\pm0.009}$	0.341$^{\pm0.006}$	0.419$^{\pm0.008}$	9.907$^{\pm0.225}$	3.899$^{\pm0.002}$	7.817$^{\pm0.052}$	1.497$^{\pm0.049}$
3.5	3.5	3.5	0.124$^{\pm0.004}$	0.200$^{\pm0.005}$	0.255$^{\pm0.006}$	15.553$^{\pm0.224}$	3.977$^{\pm0.002}$	7.544$^{\pm0.051}$	1.524$^{\pm0.017}$
3.5	5	0	0.228$^{\pm0.005}$	0.350$^{\pm0.008}$	0.438$^{\pm0.008}$	10.353$^{\pm0.144}$	3.894$^{\pm0.002}$	**7.943**$^{\pm0.038}$	1.427$^{\pm0.089}$
3.5	7	0	0.242$^{\pm0.008}$	0.371$^{\pm0.009}$	0.450$^{\pm0.007}$	12.161$^{\pm0.242}$	3.892$^{\pm0.002}$	8.011$^{\pm0.048}$	1.484$^{\pm0.050}$
3.5	7	3.5	0.164$^{\pm0.006}$	0.253$^{\pm0.008}$	0.321$^{\pm0.006}$	15.739$^{\pm0.307}$	3.947$^{\pm0.003}$	7.761$^{\pm0.037}$	1.517$^{\pm0.067}$
3.5	10	0	0.216$^{\pm0.006}$	0.335$^{\pm0.005}$	0.411$^{\pm0.007}$	18.032$^{\pm0.310}$	3.932$^{\pm0.002}$	8.055$^{\pm0.043}$	**1.597**$^{\pm0.071}$
7	3.5	0	0.160$^{\pm0.006}$	0.252$^{\pm0.008}$	0.321$^{\pm0.007}$	15.315$^{\pm0.218}$	3.950$^{\pm0.003}$	7.734$^{\pm0.029}$	1.478$^{\pm0.068}$

Table 5.6: LLM evaluation using different weights

5.3.4 Best Results

Following extensive experimentation, we have successfully enhanced all relevant metrics beyond the prior SOTA performance on the InterHuman dataset. The most effective strategy within our proposal involved training the model using individual descriptions but shifting the focus entirely to the interaction description during evaluation. The comprehensive details of these optimally performing experiments are presented in Table 5.7.

CFG Weights			R Precision ↑			FID ↓	MM Dist ↓	Diversity →	MModality ↑
w1	w2	w3	Top 1	Top 2	Top 3				
3.5	-	-	$0.252^{\pm0.007}$	$0.382^{\pm0.005}$	$0.466^{\pm0.005}$	$10.780^{\pm0.179}$	$3.882^{\pm0.002}$	$7.808^{\pm0.027}$	$1.534^{\pm0.061}$
0	7	0	$\mathbf{0.280}^{\pm0.008}$	$\mathbf{0.423}^{\pm0.004}$	$\mathbf{0.514}^{\pm0.005}$	$11.893^{\pm0.238}$	$\mathbf{3.869}^{\pm0.002}$	$8.301^{\pm0.045}$	$1.435^{\pm0.068}$
2	3.5	0	$0.253^{\pm0.007}$	$0.385^{\pm0.008}$	$0.469^{\pm0.008}$	$\mathbf{8.940}^{\pm0.187}$	$3.882^{\pm0.003}$	$7.992^{\pm0.028}$	$1.463^{\pm0.042}$
3.5	5	0	$0.228^{\pm0.005}$	$0.350^{\pm0.008}$	$0.438^{\pm0.008}$	$10.353^{\pm0.144}$	$3.894^{\pm0.002}$	$\mathbf{7.943}^{\pm0.038}$	$1.427^{\pm0.089}$
3.5	10	0	$0.216^{\pm0.006}$	$0.335^{\pm0.005}$	$0.411^{\pm0.007}$	$18.032^{\pm0.310}$	$3.932^{\pm0.002}$	$8.055^{\pm0.043}$	$\mathbf{1.597}^{\pm0.071}$

Table 5.7: Best models trained with LLM Evaluation. The first row correspond to the base InterGen model re-trained

5.4 Qualitative Results

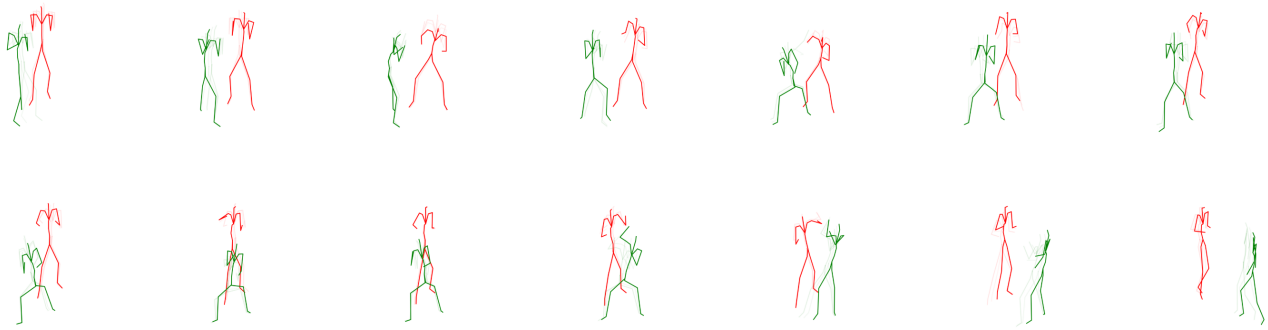


Figure 5.1: **Interaction:** Two persons are in an intense boxing match., **Individual:** One person is continuously punching.,**Individual:** One person is defending and counterattacking.

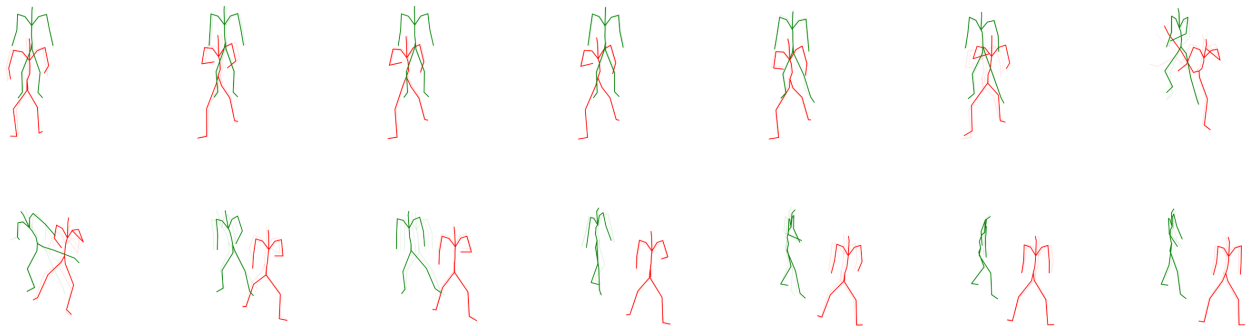


Figure 5.2: **Interaction:** Two humans are practicing taekwondo moves. one person executes strikes and the other person performs evasive maneuvers., **Individual:** One person is executing strikes with precise movements.,**Individual:** The other person is swiftly moving and performing evasive maneuvers to avoid the strikes.

In conclusion, after detailing the various experiments conducted and their respective results, we turn our attention to the visual assessment of the model's inferencing capabilities using the most

effective approach. The ensuing figures demonstrate not only the quantitative quality but also the ability of our approach to generate precise and realistic interactions from textual descriptions. The visualizations here utilize basic kinematic skeletons; however, the format of the predicted motions is sufficiently versatile to enable their application to any mesh or volumetric body.

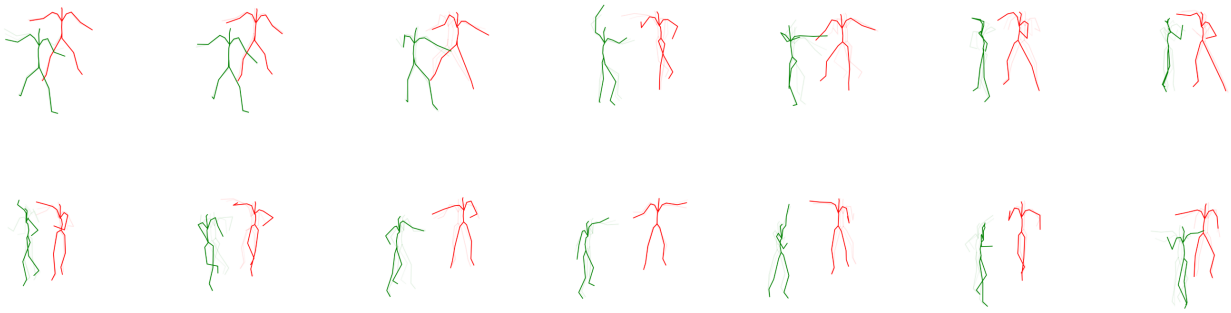


Figure 5.3: **Interaction:** These two are approaching each other, chests swaying with every step. they move closer, leaning towards each other and twirling around, then break into an impromptu dance., **Individual:** One person is approaching the other, with their chest swaying as they walk. They lean towards the other person and twirl around. ,**Individual:** One person is approaching the other, with their chest swaying as they walk. They break into an impromptu dance.

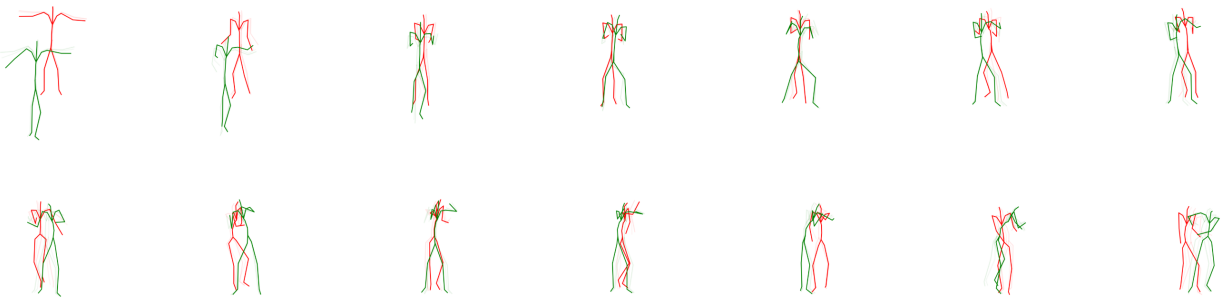


Figure 5.4: **Interaction:** The two guys rush towards one another and embrace tightly., **Individual:** One person sprints towards the other person.,**Individual:** One person sprints towards the other person and wraps their arms around them tightly.

6 Discussion

While this master's thesis has yielded promising results, it is important to acknowledge that certain aspects did not unfold as anticipated, and there is substantial scope for further research and development. This chapter aims to provide a balanced critique of all the work undertaken, highlighting both the successes and the limitations. In Section 6.1, we will delve into the role and impact of individual descriptions on the outcomes of our thesis. Section 6.2 will address the various challenges encountered in evaluating tasks of this nature.

6.1 Individual Descriptions

A pivotal issue, which indirectly contributes to several other challenges, is the suboptimal quality of individual motion descriptions. Our proposed M2T method demonstrates limited effectiveness and exhibits significant overfitting to the motion descriptions in the HumanML3D dataset. Even MotionGPT, which is the SOTA for this task, fails to consistently yield accurate results. While it generally produces diverse and plausible descriptions, its performance falters with motions that deviate significantly from the norm, often resulting in vague or inaccurate descriptions. The lack of precise and detailed individual motion descriptions is a major hindrance in fully realizing the potential of our proposal.

The initial hypothesis behind incorporating individual motion descriptions in modelling interactions was to gain control over the constituents of the interaction. For instance, the interaction "two people are dancing together" encompasses a wide range of possible individual motions. We envisioned the ability to steer these individual components. However, our experiments revealed that this level of control was not achieved.

This shortfall could be attributed to the inadequate quality of the textual descriptions for individual motions. Moreover, the diversity of individual motions represented in the interaction dataset is limited. The dataset lacks a broad spectrum of individual motions necessary for training a model capable of generating both controllable individual motions and realistic, coherent interactions.

6.2 Evaluation

The aspect of evaluation in our work is an area that requires significant improvement. We have previously noted the limitations of using LLMs for generating individual descriptions in a fair

evaluation context. These generated descriptions often fail to provide the model with meaningful information for effective motion generation. Developing a method that ensures fairness while still offering valuable data to condition the model remains a critical challenge.

Additionally, there is a pressing need to refine the evaluation metrics for generative models. It's acknowledged that evaluating these models is inherently complex, especially compared to more straightforward tasks like classification. Despite this complexity, the current metrics seem insufficient for accurately reflecting the true quality of generative models. This issue is exemplified by the fact that each new method, including ours, tends to surpass its predecessors in standard evaluation metrics. Yet, these metric improvements often do not translate into noticeably better real-world performance. This trend indicates a propensity for researchers, ourselves included, to overfit our methods to these metrics. Addressing this issue and moving towards more meaningful and representative evaluation standards is imperative for the field.

7 Conclusion

Reflecting upon the objectives set out in the initial chapter of this thesis, we can confidently affirm that we have developed a model capable of generating accurate and realistic human-human interactions driven by textual descriptions. The journey of this thesis has led to a profound understanding of this field, unravelling both its potential and its limitations.

A series of diverse experiments were conducted, substantiating our claims. Moreover, we introduced an innovative motion interaction architecture that significantly improves the generation process by integrating textual information about individual motions within an interaction. Through rigorous experimentation, it has been established that our method surpasses existing techniques in generating interactions on the InterHuman dataset.

This thesis has also been a platform for proposing additional methodologies that have impacted our results, both directly and indirectly. We devised a novel approach for creating textual descriptions from encoded motions. LLMs were employed for various tasks, such as enhancing existing dataset annotations and establishing a fair evaluation protocol to circumvent information leakage. Furthermore, we introduced a multi-weight sampling method for our diffusion model, allowing for differential weighting of various conditions.

In conclusion, this master's thesis represents a significant journey through uncharted territories in a novel field. The challenges we faced were manifold, but the outcomes were gratifying. Not only did we manage to produce notable results, but we also succeeded in advancing beyond the current SOTA in this area.

7.1 Future Lines of Work

Despite the advancements made in this thesis, there are several areas where significant improvements can be made. The primary area, as discussed in Section 6.1, is enhancing the textual descriptions of individual motions within interactions. Given that motions can be converted and rendered into other formats such as video, one approach could be to utilize more advanced video-captioning models, which are currently more developed due to greater data availability. Although potentially more time-consuming, this process could be conducted offline before training begins.

Another significant area of potential development is in the controllability of individual motions.

As we discussed in Section 6.1, our efforts to control individual motions using textual descriptions did not yield the desired results. A contributing factor might be the limited range of individual motions in the interaction dataset, which hinders the model's ability to simultaneously learn to represent both individual motions and interactions. One solution could be to pre-train our model on a large individual T2M dataset, like HumanML3D, and then fine-tune it for interaction generation. The architectural design of our proposed model offers a notable advantage in terms of flexibility. Specifically, it can be seamlessly adapted to function as a single-motion model. This adaptability is primarily enabled by the implementation of shared weights and cooperative denoisers within the architecture.

Lastly, it would be beneficial to develop novel techniques for evaluating these kinds of generative models, focusing on metrics with more real-world relevance. While such results might not initially seem as compelling to some researchers, they would undoubtedly be valuable for future work and for the end-users of these models. Additionally, devising a method to generate individual textual descriptions that provide valuable information without causing data leakage would be a significant advancement. This could potentially enhance the results reported even further, making the generative models more effective and reliable for practical applications.

In summary, these future lines of work offer promising opportunities to build upon the foundations laid in this thesis. Addressing these challenges will not only advance the field of text-to-motion generation but also contribute to the broader domain of generative modelling, making it more applicable and valuable in real-world contexts.

Bibliography

[1] Chenshuang Zhang et al. *Text-to-image Diffusion Models in Generative AI: A Survey*. 2023. arXiv: 2303.07909 [cs.CV].

[2] Fengxiang Bie et al. *RenAIssance: A Survey into AI Text-to-Image Generation in the Era of Large Model*. 2023. arXiv: 2309.00810 [cs.CV].

[3] Florinel-Alin Croitoru et al. "Diffusion Models in Vision: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.9 (2023), pp. 10850–10869. DOI: 10.1109/TPAMI.2023.3261988.

[4] Mohamed Elasri et al. "Image Generation: A Review". In: *Neural Processing Letters* 54.5 (Oct. 2022), pp. 4609–4646. ISSN: 1573-773X. DOI: 10.1007/s11063-022-10777-x. URL: https://doi.org/10.1007/s11063-022-10777-x.

[5] Wayne Xin Zhao et al. *A Survey of Large Language Models*. 2023. arXiv: 2303.18223 [cs.CL].

[6] Junyi Li et al. *Pretrained Language Models for Text Generation: A Survey*. 2022. arXiv: 2201.05273 [cs.CL].

[7] Bonan Min et al. "Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey". In: *ACM Comput. Surv.* 56.2 (Sept. 2023). ISSN: 0360-0300. DOI: 10.1145/3605943. URL: https://doi.org/10.1145/3605943.

[8] Hanqing Zhang et al. "A Survey of Controllable Text Generation Using Transformer-Based Pre-Trained Language Models". In: *ACM Comput. Surv.* 56.3 (Oct. 2023). ISSN: 0360-0300. DOI: 10.1145/3617680. URL: https://doi.org/10.1145/3617680.

[9] Aditya Ramesh et al. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022. arXiv: 2204.06125 [cs.CV].

[10] Long Ouyang et al. *Training language models to follow instructions with human feedback*. 2022. arXiv: 2203.02155 [cs.CL].

[11] Lei Wang et al. *A Review of Intelligent Music Generation Systems*. 2023. arXiv: 2211.09124 [cs.SD].

[12] Miguel Civit et al. "A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends". In: *Expert Systems with Applications* 209 (2022), p. 118190. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2022.118190. URL: https://www.sciencedirect.com/science/article/pii/S0957417422013537.

[13] Chenshuang Zhang et al. *A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative AI*. 2023. arXiv: 2303.13336 [cs.SD].

[14] Jade Copet et al. *Simple and Controllable Music Generation*. 2023. arXiv: 2306.05284 [cs.SD].

[15] Felix Kreuk et al. *AudioGen: Textually Guided Audio Generation*. 2023. arXiv: 2209.15352 [cs.SD].

[16] Zhen Xing et al. *A Survey on Video Diffusion Models*. 2023. arXiv: 2310.10647 [cs.CV].

[17] Aditi Singh. "A Survey of AI Text-to-Image and AI Text-to-Video Generators". In: *2023 4th International Conference on Artificial Intelligence, Robotics and Control (AIRC)*. IEEE, May 2023. DOI: 10.1109/airc57904.2023.10303174. URL: http://dx.doi.org/10.1109/AIRC57904.2023.10303174.

[18] Uriel Singer et al. *Make-A-Video: Text-to-Video Generation without Text-Video Data*. 2022. arXiv: 2209.14792 [cs.CV].

[19] Agrim Gupta et al. *Photorealistic Video Generation with Diffusion Models*. 2023. arXiv: 2312.06662 [cs.CV].

[20] Danilo Jimenez Rezende and Shakir Mohamed. *Variational Inference with Normalizing Flows*. 2016. arXiv: 1505.05770 [stat.ML].

[21] Geoffrey E. Hinton. "Reducing the Dimensionality of Data with Neural". In: 2008. URL: https://api.semanticscholar.org/CorpusID:14820265.

[22] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].

[23] Jascha Sohl-Dickstein et al. "Deep Unsupervised Learning using Nonequilibrium Thermodynamics". In: *CoRR* abs/1503.03585 (2015). arXiv: 1503.03585. URL: http://arxiv.org/abs/1503.03585.

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *CoRR* abs/2006.11239 (2020). arXiv: 2006.11239. URL: https://arxiv.org/abs/2006.11239.

[25] Prafulla Dhariwal and Alex Nichol. "Diffusion Models Beat GANs on Image Synthesis". In: *CoRR* abs/2105.05233 (2021). arXiv: 2105.05233. URL: https://arxiv.org/abs/2105.05233.

[26] Robin Rombach et al. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *CoRR* abs/2112.10752 (2021). arXiv: 2112.10752. URL: https://arxiv.org/abs/2112.10752.

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising Diffusion Implicit Models". In: *CoRR* abs/2010.02502 (2020). arXiv: 2010.02502. URL: https://arxiv.org/abs/2010.02502.

[28] Jonathan Ho and Tim Salimans. "Classifier-Free Diffusion Guidance". In: *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*. 2021. URL: https://openreview.net/forum?id=qw8AKxfYbI.

[29] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[31] Wentao Zhu et al. "Human Motion Generation: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023), pp. 1–20. DOI: 10.1109/TPAMI.2023.3330935.

[32] Guy Tevet et al. *MotionCLIP: Exposing Human Motion Generation to CLIP Space*. en. arXiv:2203.08063 [cs]. Mar. 2022. URL: http://arxiv.org/abs/2203.08063 (visited on 04/06/2023).

[33] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *CoRR* abs/2103.00020 (2021). arXiv: 2103.00020. URL: https://arxiv.org/abs/2103.00020.

[34] Mathis Petrovich, Michael J. Black, and Gül Varol. *TEMOS: Generating diverse human motions from textual descriptions*. en. arXiv:2204.14109 [cs]. July 2022. URL: http://arxiv.org/abs/2204.14109 (visited on 03/15/2023).

[35] Chuan Guo et al. *TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts*. en. arXiv:2207.01696 [cs]. Aug. 2022. URL: http://arxiv.org/abs/2207.01696 (visited on 03/15/2023).

[36] Jianrong Zhang et al. *T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations*. en. arXiv:2301.06052 [cs]. Feb. 2023. URL: http://arxiv.org/abs/2301.06052 (visited on 03/15/2023).

[37] Biao Jiang et al. *MotionGPT: Human Motion as a Foreign Language*. en. arXiv:2306.14795 [cs]. June 2023. URL: http://arxiv.org/abs/2306.14795 (visited on 07/05/2023).

[38] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. *FLAME: Free-form Language-based Motion Synthesis & Editing*. en. arXiv:2209.00349 [cs]. Jan. 2023. URL: http://arxiv.org/abs/2209.00349 (visited on 04/17/2023).

[39] Mingyuan Zhang et al. *MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model*. en. arXiv:2208.15001 [cs]. Aug. 2022. URL: http://arxiv.org/abs/2208.15001 (visited on 03/15/2023).

[40] Guy Tevet et al. *Human Motion Diffusion Model*. en. arXiv:2209.14916 [cs]. Oct. 2022. URL: http://arxiv.org/abs/2209.14916 (visited on 03/02/2023).

[41] Ye Yuan et al. *PhysDiff: Physics-Guided Human Motion Diffusion Model*. en. arXiv:2212.02500 [cs]. Dec. 2022. URL: http://arxiv.org/abs/2212.02500 (visited on 05/18/2023).

[42] Yusuke Goutsu and Tetsunari Inamura. "Linguistic Descriptions of Human Motion with Generative Adversarial Seq2Seq Learning". en. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. Xi'an, China: IEEE, May 2021, pp. 4281–4287. ISBN: 978-1-72819-077-8. DOI: 10.1109/ICRA48506.2021.9561519. URL: https://ieeexplore.ieee.org/document/9561519/ (visited on 07/05/2023).

[43] Chuan Guo et al. *TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts*. en. arXiv:2207.01696 [cs]. Aug. 2022. URL: http://arxiv.org/abs/2207.01696 (visited on 07/05/2023).

[44] Adel Jalal Yousif and Mohammed H. Al-Jammas. "Exploring deep learning approaches for video captioning: A comprehensive review". In: *e-Prime - Advances in Electrical Engineering, Electronics and Energy* 6 (2023), p. 100372. ISSN: 2772-6711. DOI: https://doi.org/10.1016/j.prime.2023.100372. URL: https://www.sciencedirect.com/science/article/pii/S277267112300267X.

[45] Moloud Abdar et al. *A Review of Deep Learning for Video Captioning*. 2023. arXiv: 2304.11431 [cs.CV].

[46] Daniela Moctezuma et al. "Video captioning: A comparative review of where we are and which could be the route". In: *Computer Vision and Image Understanding* 231 (2023), p. 103671. ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2023.103671. URL: https://www.sciencedirect.com/science/article/pii/S1077314223000516.

[47] Han Liang et al. *InterGen: Diffusion-based Multi-human Motion Generation under Complex Interactions*. 2023. arXiv: 2304.05684 [cs.CV].

[48] Yonatan Shafir et al. *Human Motion Diffusion as a Generative Prior*. en. arXiv:2303.01418 [cs]. Mar. 2023. URL: http://arxiv.org/abs/2303.01418 (visited on 03/15/2023).

[49] Mikihiro Tanaka. "Role-Aware Interaction Generation from Textual Description". en. In: ().

[50] Han Liang et al. *InterGen: Diffusion-based Multi-human Motion Generation under Complex Interactions*. en. arXiv:2304.05684 [cs]. Apr. 2023. URL: http://arxiv.org/abs/2304.05684 (visited on 11/17/2023).

[51] Matthew Loper et al. "SMPL: A Skinned Multi-Person Linear Model". In: *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34.6 (Oct. 2015), 248:1–248:16.

[52] Naureen Mahmood et al. "AMASS: Archive of Motion Capture as Surface Shapes". In: *International Conference on Computer Vision*. Oct. 2019, pp. 5442–5451.

[53] Javier Romero, Dimitrios Tzionas, and Michael J. Black. "Embodied hands: modeling and capturing hands and bodies together". In: *ACM Transactions on Graphics* 36.6 (Nov. 2017), pp. 1–17. ISSN: 1557-7368. DOI: 10.1145/3130800.3130883. URL: http://dx.doi.org/10.1145/3130800.3130883.

[54] Abhinanda R. Punnakkal et al. "BABEL: Bodies, Action and Behavior with English Labels". In: *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 722–731.

[55] Chuan Guo et al. "Generating Diverse and Natural 3D Human Motions From Text". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 5152–5161.

[56] Chuan Guo et al. "Action2motion: Conditioned generation of 3d human motions". In: *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, pp. 2021–2029.

[57] Timo von Marcard et al. "Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera". In: *European Conference on Computer Vision (ECCV)*. 2018.

[58] Matthias Plappert, Christian Mandery, and Tamim Asfour. "The KIT Motion-Language Dataset". In: *Big Data* 4.4 (Dec. 2016), pp. 236–252. DOI: 10.1089/big.2016.0028. URL: http://dx.doi.org/10.1089/big.2016.0028.

[59] Amir Shahroudy et al. "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis". In: *CoRR* abs/1604.02808 (2016). arXiv: 1604.02808. URL: http://arxiv.org/abs/1604.02808.

[60] N.P. van der Aa et al. "Utrecht Multi-Person Motion (UMPM) benchmark: a multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction". In: *Proceedings of the Workshop on Human Interaction in Computer Vision (HICV), in conjunction with ICCV 2011*. 2011.

[61] Evonne Ng et al. "You2Me: Inferring Body Pose in Egocentric Video via First and Second Person Interactions". In: *CoRR* abs/1904.09882 (2019). arXiv: 1904.09882. URL: http://arxiv.org/abs/1904.09882.

[62] Aditya Ramesh et al. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022. arXiv: 2204.06125 [cs.CV].

[63] Alex Nichol and Prafulla Dhariwal. *Improved Denoising Diffusion Probabilistic Models*. 2021. arXiv: 2102.09672 [cs.LG].

[64] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. 2019. arXiv: 1711.05101 [cs.LG].

List of Acronyms

AI Artificial Intelligence. 1, 5, 13

AR Augmented Reality. 1

CNNs Convolutional Neural Netwoks. 13, 14

GANs Generative Adversarial Networks. 5, 13, 14

LLM Large Language Model. 13, 15

LLMs Large Language Models. 3, 25, 26, 38, 39, 43, 45

M2T Motion-to-Text. 3, 14, 15, 23, 27, 33, 37, 43

MLPs Multi-Layer Perceptrons. 13

MoCap Motion Capture. 1, 13

NLP Natural Language Processing. 9, 10

RNNs Recurrent Neural Netwoks. 13, 14

SOTA state-of-the-art. 33, 37, 39, 43, 45

T2I Text-to-Interaction. 3, 15, 37

T2M Text-to-Motion. 14, 15, 33, 46

VAEs Variational Auto Encoders. 5, 13, 14

VR Virtual Reality. 1