M-Adaptor: Text-driven Whole-body Human Motion Generation

Alicia Li [1] Xiaodong Chen [2*] Bohao Liang [3] Qian Bao [4] Wu Liu [2]

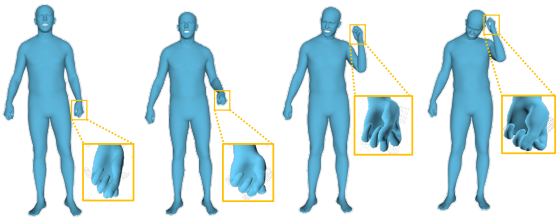
[1]Horace Mann School [2]University of Science and Technology of China
[3]University of New South Wales [4]JD Explore Academy
Aliciaysli@gmail.com, cxd1230@mail.ustc.edu.cn

Abstract

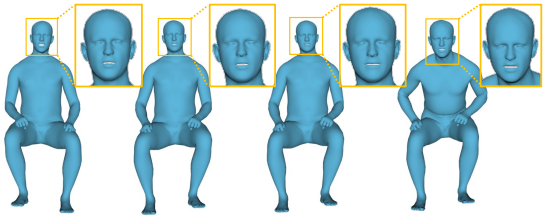
*Text-driven whole-body human motion generation, which involves the creation of motion sequences based on textual descriptions, has attracted much attention in the communities of computer vision and artificial intelligence. It aims to extend text-driven motion generation tasks to accommodate complex whole-body human motions, encompassing facial expressions and hand gestures. Researchers have recently developed large-scale 3D expressive whole-body motion datasets enriched with semantic labels and pose descriptions. Nonetheless, there remains a considerable demand within the community for a straightforward and effective framework for generating and evaluating whole-body human motion based on textual descriptions. To address the above issues, we introduce **M-Adaptor**, a two-stage Low-Rank Adaptation (LoRA)-based generator for whole-body motion generation tasks, to improve the quality and diversity of body motions, facial expressions, and hand gestures. In particular, it first generates initial coarse-grained body motion tokens from textual prompts to enhance the stability of generated motions, then iterates fine-grained facial expressions with the LoRA-based adaptor to enhance motion expressiveness. Furthermore, we extend the existing state-of-the-art CLaM model to **CLaM-H** and **CLaM-X** for evaluation of SMPL-H and SMPL-X based motion generation. Extensive qualitative and quantitative evaluations demonstrate our framework's superior performance, with a significant R-Precision improvement for text-driven whole-body motion generation.*

1. Introduction

Text-driven human motion generation [8, 15, 30, 37] focuses on creating human motion sequences based on provided textual descriptions. This technology's innovation and utility stem from its integration of natural human language understanding into body motion generation. Recently, text-driven whole-body human motion generation [22, 43] has extended to incorporating control over facial expressions and hand gestures, thus enhancing the versatility of this task. By analyzing and comprehending textual descriptions of whole-body motion and emotion, generation models convert human semantics directly into body motions, hand gestures, and facial expressions. This task has broad applications [1, 2, 23, 36], including animation making, science fiction movie production, etc.



(a) "A person uses their hand to scratch their ear, unsurely"



(b) "A person is sitting down still, sadly"

Figure 1. Illustration of text-driven whole-body human motion generation including hand gestures and facial expressions.

Although the above tasks are challenging, various meaningful attempts [8, 15, 17] have been made in related research areas. For instance, the T2M [15] and TM2T [16] generators designed by Guo *et al.* have addressed the challenge of interpreting long sentences and generating motions of varying lengths by utilizing RNN-based [12] generators. Motion Diffuse [40] and MLD [7] have enhanced motion sequence quality by introducing diffusion models into the

*Corresponding author. This work is supported by National Key Research and Development Program of China (NO. 2024YFE0203200).

generation process. Additionally, T2M-GPT [39] has significantly improved the semantic understanding of generators through the integration of Vector Quantized Variational Autoencoders (VQ-VAE) [38] and Generative Pre-trained Transformers (GPT) [33]. Furthermore, momask [17] provides a novel masked modeling framework that encodes the motion sequences as multi-layer discrete tokens, improving the realism and smoothness of text-driven motion generation with residual VQ-VAE and residual transformers.

Although the quality of generated motion sequences continues to improve, this field still faces several unavoidable challenges. Firstly, existing generators [7, 17, 41] mostly focus on the generation of body motions, neglecting the details of facial expressions and hand gestures. This limitation restricts their range of applications, particularly in complex tasks that require fine-grained coordination of expressions and gestures, such as natural interactions of virtual characters. As shown in Fig. 1, the synergy among facial expressions, hand gestures, and body motions is crucial to generating natural human motions; the lack of attention to these details often results in unnatural results. Secondly, there is currently no widely accepted evaluation standard to measure the accuracy of generated results that include facial expressions and hand gestures. Existing evaluation methods [8, 15, 32] focus mainly on the quality of body motions, failing to comprehensively evaluate the authenticity and coordination of generated facial expressions and gestures. The absence of such an evaluation standard makes it difficult to compare the performance of different generators and hinders further progress in this field. Thus, developing a universal evaluation framework that can comprehensively consider the quality of body motions, facial expressions, and hand gestures is an urgent issue to address.

To overcome these challenges, we propose **M-Adaptor**, a two-stage whole-body motion generation framework based on Low-Rank Adaptation (LoRA) [19], designed to enhance the quality and diversity of human motion sequences of body motions, facial expressions, and hand gestures. The first stage of M-Adaptor focuses on generating initial coarse-grained body motion tokens directly from textual descriptions, thereby ensuring the stability of the motion sequences. In the second stage, a LoRA-based adaptor is employed to iteratively refine these motions, adding detailed tokens for facial expressions to enrich the motion's expressiveness. Furthermore, we extend the existing CLaM [8] model to two new variants, **CLaM-H** and **CLaM-X**, precisely tailored to evaluate motion generation based on the SMPL-H [34] and SMPL-X [29] models, respectively. Specifically, CLaM-H is designed to evaluate the effectiveness of motion generation with body motions and hand gestures, while CLaM-X further evaluates whole-body motion generation, encompassing facial expressions as well. This extension allows for a more comprehensive assessment of our motion generation capabilities across different human body representations. Our framework has undergone extensive qualitative and quantitative evaluations which demonstrate its superior performance in text-driven whole-body motion generation.

In summary, the contributions of this paper are threefold:

- We propose **M-Adaptor**, a two-stage whole-body motion generation framework designed to enhance the quality and diversity of human motion sequences, encompassing body motion, facial expressions, and hand gestures.
- We extend two new variants, **CLaM-H** and **CLaM-X**, from the existing CLaM model, allowing for a comprehensive evaluation of various human motion generators.
- We conduct extensive qualitative and quantitative evaluations, demonstrating the superior performance of our framework in text-driven whole-body motion generation.

2. Related Work

Human Motion Generation [4–6, 9–11, 13, 20, 25, 31, 35], particularly the subfield focusing on text-driven human motion generation, involves converting textual descriptions into 3D human motion sequences. This subfield has garnered attention due to the intuitive and accessible nature of language inputs. Early models, such as Text2Action [4] and Language2Pose [5], utilized RNN-based [12] architecture along with curriculum learning to transform text into motion sequences. However, these initial attempts often suffered from motion quality and global translation issues. To tackle these challenges, ACTOR [30] introduced the adaptation of VAEs [21] alongside additional text encoders to produce more varied motion sequences. Recently, Guo *et al.* [15] introduced HumanML3D, a comprehensive dataset, and created t2m, a method to generate human motions of reasonable lengths by estimating motion durations, along with an RNN-based evaluator to evaluate the performance of generation models. Subsequently, diffusion models like MLD [7] and Motion Diffuse [40] have been utilized to generate motions based on a variety of conditional inputs. T2M-GPT [39] and momask [17] advanced this field by developing a generative framework using VQ-VAE and generative pre-trained transformers for motion generation. Further research, including works like MDM [37] and MotionGPT [42], has concentrated on motion completion tasks, which involve generating motion sequences conditioned on partial motions, such as motion prediction or in-between motions, ensuring the continuity of segments. Despite the progress these methods have made, they mainly focus on the quality of body motions and fail to comprehensively evaluate the authenticity and coordination of generated facial expressions and hand gestures.

Whole-body Human Motion Generation [22, 26, 43] is a subfield dedicated to transforming textual descriptions

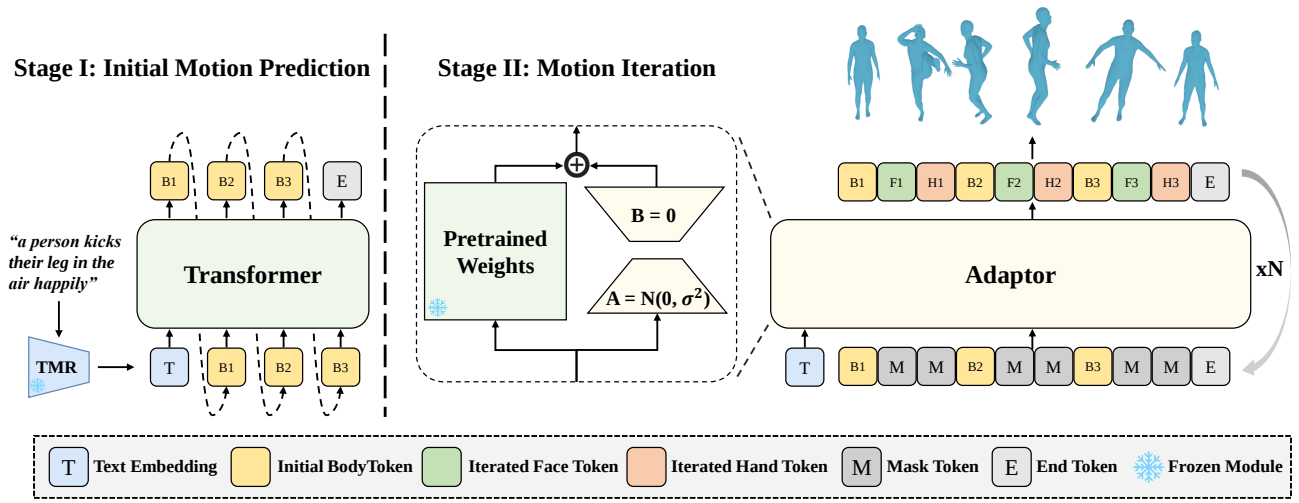


Figure 2. **The architecture of our generator M-Adaptor.** It first generates sequences of coarse-grained body motion tokens from pre-trained text embedding in stage I, then iterates the mask tokens to yield more accurate sequences of fine-grained motion tokens, including body motion, facial expressions, and hand gestures in stage II.

into whole-body 3D human motion, including body actions, hand gestures, and facial expressions. Motion-X [22, 43] collected by Lin *et al.* introduced a large-scale 3D expressive whole-body motion dataset. It provides high-precision, cost-effective, and scalable annotations derived from single- or multi-view videos, offering frame-level pose descriptions and semantic labels across various motion sequences. With 15.6 million 3D whole-body pose annotations, it significantly enhanced the expressiveness, diversity, and realism of motion generation and emerged as an essential resource in whole-body human motion generation tasks. Additionally, HUMANTOMATO [26] refined motion representation by using uniform skeletons to generate diverse joint movements. It expanded the H3D format [15] into the HUMANTOMATO format to include face expressions and hand gestures. This format highlights the importance of velocity in motion reconstruction and demonstrates that enhancements in motion representation lead to improved motion generation and reconstruction quality. However, the absence of an open-source evaluation standard and evaluator models makes it difficult to compare the performance of generators and hinders further progress in this field.

3. Method

In this section, we declare the detailed framework of our **M-Adaptor** generator, designed for the task of text-driven whole-body human motion generation. Furthermore, we introduce our open-source evaluator models, **CLaM-H** and **CLaM-X**, specifically developed for the evaluation of alignment between generated whole-body motion sequences and textual descriptions. Before describing our methods in detail, we first introduce the necessary notation and definitions.

3.1. Preliminary

The task of text-driven human motion generation is to generate, given a textual description, corresponding human motion sequences. It usually adopts the paradigm in the following training and evaluation phases.

1) **Training Phase.** For a given text $X = (X_1, X_2, ..., X_N)$ containing N words, our aim is to generate a 3D motion sequence $M' = (m'_1, m'_2, ..., m'_{T'})$ with length T', as similar as possible to the real 3D motion sequence $M = (m_1, m_2, ..., m_T)$ with length T. Some methods set the length T' as a precondition, such as Motion Diffuse [40] and MDM [37].

2) **Evaluation Phase.** The generated motion sequence M' and the given textual description X are processed through the evaluator to extract the motion features, $f_{M'}$, and text features, f_t, respectively, and to compute metrics as follows: 1) Frechet Inception Distance (FID); 2) R-Precision; 3) Diversity; 4) Multi Modality (MModality); 5) Multi-Modal Distance (MM-Dist). Following the criteria proposed by Chen *et al.* [8], the evaluator's motion and text extractors are trained under contrastive loss and InfoNCE loss with the real motion sequence M and corresponding textual description X to produce geometrically close feature vectors for matched text-motion pairs.

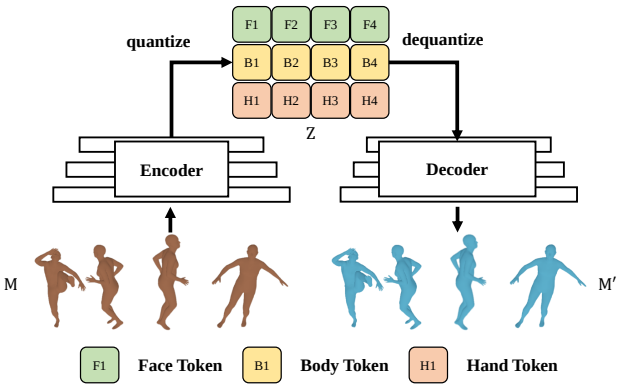


Figure 3. **The pre-processing stage of M-Adaptor.** The motion tokenizer contains a VAE-based encoder E and a decoder D for quantization and dequantization. Each token Z_i of time i can be split as body motion token B_i, facial expressions token F_i, and hand gesture token H_i

3.2. LoRA-based Generator M-Adaptor

This subsection elaborates on our LoRA-based two-stage motion generator, named M-Adaptor, for text-driven whole-body motion generation tasks. Similarly to the prediction process of the auto-regressive model, we generate sequences of coarse-grained body motion tokens from pre-trained text embeddings, as shown in Fig. 2 stage I. However, unlike the classical auto-regressive model, our M-Adaptor iterates the mask tokens to yield more accurate sequences of fine-grained motion tokens, including facial expressions and hand gestures, as shown in in Fig. 2 stage II. Before describing our M-Adaptor in detail, we introduce the pre-processing of motion tokens.

Pre-processing Stage. A motion tokenizer transforms and quantizes raw whole-body motion sequence M into a series of discrete tokens $Z = (z_1, z_2, ..., z_T)$ within latent space. As shown in Fig. 3, each token Z_i can be split into body motion token B_i, facial expression token F_i, and hand gesture token H_i. It is pre-trained using the Vector Quantized Variational Autoencoder (VQ-VAE) [38] and guided by three large motion codebooks. Specifically, the motion tokenizer contains an encoder E and a decoder D. In the pre-processing phase, the raw motion sequences M are fed into the encoder E to compute discrete motion tokens with $Z = E(M)$. For the subsequent reconstruction phase, the prediction motion sequences M' can be computed from latent space with $M' = D(Z)$.

Stage I: Initial Motion Prediction. With the pre-trained discrete motion tokens Z, including B_i, F_i and H_i from the pre-processing stage, our auto-regressive generator predicts coarse-grained body motion tokens B_i with text embedding c extracted by TMR [32] until encountering the end token [END], as shown in Fig. 2. This model is designed with causal self-attention to prevent future information leakage when making predictions. During the training stage, the optimization of the auto-regressive model can be considered as a process of maximizing the log-likelihood:

$$L_z = \mathbb{E}_{Z\sim p(Z)}[-\log p(Z|c)]\,, \tag{1}$$

where $\mathbb{E}_{Z\sim p(Z)}$ denotes the expectation over samples Z drawn from the distribution $p(Z)$, and $-\log p(Z|c)$ is the negative log-likelihood of Z with given condition c.

Stage II: Motion Iteration. Constrained by the mechanism of auto-regressive models, z_{i-1} is unable to access the information from follow-up tokens $z_{>i}$, and the generation of the motion sequence Z is irreversible. It is thus challenging to generate hand gestures and facial expressions, as these fine-grained motions often need to be synchronized with the entire sequence of body motions to avoid unnaturalness. To address these issues, we propose our motion adaptor for fine-grained motion iteration in Fig. 2 Stage II. At first, our adaptor copies and freezes the pre-trained weights from the transformer in stage I. Then, we replace the casual self-attention layers with self-attention layers to capture the relationships between each motion token. Note that the above steps do not involve any parameter changes. After that, we introduce the low-rank adaptation layer $A = N(0, \sigma^2)$ and $B = 0$ for fine-tuning without significantly increasing the number of parameters. During the training process, the ground truth tokens Z, including B_i, F_i, and H_i from the pre-processing stage, are replaced with learnable special mask tokens [MASK] with the random rate r. r is a random variable that belongs to a uniform distribution over the interval $[0, 1]$. Our adaptor is trained to predict the ground truth tokens from the mask tokens [MASK] and other known conditions with the optimization goal of maximizing log-likelihood as Eq. 1.

Inference. With the well-trained M-Adaptor, we first extract the text embedding from the pre-trained TMR [32] and predict the initial motion tokens with the text embedding in stage I. Then, we generate iterated motion tokens with the LoRA-based adaptor in stage II. The iteration

Component	CLaM-H	CLaM-X
Root Y-axis Velocity ($\dot{r}^a$)	✓	✓
Root XZ-plane Velocities ($\dot{r}^x, \dot{r}^z$)	✓	✓
Root Height (r^y)	✓	✓
Local Joints Positions ($\mathbf{j}^p$)	✓	✓
Local Joints Rotations ($\mathbf{j}^r$)	✓	✓
Joints Velocities ($\mathbf{j}^v$)	✓	✓
Foot Contact (g^c)	✓	✓
Facial Expressions ($\mathbf{f}$)	-	✓

Table 1. **Motion components for CLaM-H and CLaM-X.** Joints for each variant includes body joints and hand joints.

Methods	Base Model	Dataset	R-Precision (%) ↑			FID →	MM-Dist →	Diversity →
			Top-1	Top-2	Top-3			
T2M-H	Guo et al. [15]	Motion-X	$50.75^{\pm0.1}$	$72.27^{\pm0.2}$	$83.14^{\pm0.2}$	$0.001^{\pm0.000}$	$2.448^{\pm0.004}$	$10.912^{\pm0.110}$
CLaM-H (Ours)	CLaM [8]		$\mathbf{79.59}^{\pm0.1}$	$\mathbf{90.53}^{\pm0.1}$	$\mathbf{94.04}^{\pm0.1}$	$0.003^{\pm0.000}$	$4.181^{\pm0.004}$	$8.730^{\pm0.024}$
T2M-X	Guo et al. [15]	Motion-XW	$51.94^{\pm0.2}$	$73.06^{\pm0.2}$	$83.64^{\pm0.1}$	$0.001^{\pm0.000}$	$2.473^{\pm0.004}$	$11.290^{\pm0.118}$
CLaM-X (Ours)	CLaM [8]		$\mathbf{79.10}^{\pm0.1}$	$\mathbf{89.39}^{\pm0.1}$	$\mathbf{92.63}^{\pm0.1}$	$0.002^{\pm0.000}$	$4.294^{\pm0.004}$	$8.951^{\pm0.036}$

Table 2. **Comparison with different evaluators on Motion-X and Motion-X-Whole (Motion-XW) test set using ground-truth motion sequences.** The evaluation is repeated 20 times, and the mean value is reported, supplemented by a 95% confidence interval. Note that metrics on ground-truth motion sequences are not comparable, except for R-Precision.

stage is conducted N times; we use a high mask rate r_0 in the first iteration, and gradually decrease it linearly with $r_n = r_0 - (r_0 - r_b)\ \frac{n}{N}$ for the iteration n, where r_b is the lowest mask ratio. Motion tokens Z with low confidence are preferentially masked.

3.3. Evaluator CLaM-H and ClaM-X

We extend the existing CLaM [8] model to two new variants, **CLaM-H** and **CLaM-X**, specifically tailored to evaluate motion generation based on the SMPL-H [34] and SMPL-X [29] models, respectively. Although the default evaluator proposed by Guo *et al.* is the most widely used evaluation model, CLaM is recently being widely adopted in academia communities due to its robust evaluation preformance, exceeding the default evaluator by 22% Top-1 R-Precision. Specifically, CLaM-H is designed to evaluate the effectiveness of motion generation with body motions and hand gestures. Instead of using joint rotations to directly model motion as in SMPL-H, we model human motion sequences using root Y-axis angular velocity ($\dot{r}^a \in \mathbb{R}$), root XZ-plane linear velocities ($\dot{r}^x, \dot{r}^z \in \mathbb{R}$), root height $r^y \in \mathbb{R}$, local joint positions ($\mathbf{j}^p \in \mathbb{R}^{3N-1}$), local joint rotations ($\mathbf{j}^r \in \mathbb{R}^{3N-1}$), joint velocities ($\mathbf{j}^v \in \mathbb{R}^{3N}$), and foot contact with the ground ($g^c \in \mathbb{R}^4$). $N = N_B + N_H$ denotes the number of body joints N_B and hand joints N_H. Thus, we represent the body-hand motion at frame i as $\mathbf{m}_i = \{\dot{r}^a, \dot{r}^x, \dot{r}^z, \dot{r}^y, \mathbf{j}^p, \mathbf{j}^r, \mathbf{j}^v, g^c\}$. During the training of CLaM-H, the ground truth motion sequence $M = \{m_i\}$ and the given textual description X are processed using an evaluator trained with both contrastive loss [18] and InfoNCE loss [28] following [8]. Based on CLaM-H, CLaM-X further evaluates whole-body motion generation, encompassing body motions, hand gestures, and facial expressions. Due to the differences of facial expressions compared to body skeletons, we directly use the expression parameters ($\mathbf{f} \in \mathbb{R}^{50}$) from SMPL-X to control the face. The whole body motion at frame i is defined as $\mathbf{m}_i = \{\dot{r}^a, \dot{r}^x, \dot{r}^z, \dot{r}^y, \mathbf{j}^p, \mathbf{j}^r, \mathbf{j}^v, g^c, \mathbf{f}\}$.

Methods	Pretraining	R-Precision (%) ↑		
		Top-1	Top-2	Top-3
T2M-X		$48.47^{\pm0.4}$	$67.20^{\pm0.3}$	$76.77^{\pm0.2}$
T2M-X	✓	$51.94^{\pm0.2}$	$73.06^{\pm0.2}$	$83.64^{\pm0.1}$
CLaM-X		$70.53^{\pm0.3}$	$85.42^{\pm0.2}$	$90.86^{\pm0.1}$
CLaM-X	✓	$\mathbf{79.10}^{\pm0.1}$	$\mathbf{89.39}^{\pm0.1}$	$\mathbf{92.63}^{\pm0.1}$

Table 3. **Ablation studies to analyze the influence of pretraining on T2M-X and CLaM-X**. The evaluation is repeated 20 times, and the mean value is reported, supplemented by a 95% confidence interval.

4. EXPERIMENTS

We first introduce benchmark text-to-motion datasets, evaluation metrics, and implementations in section 4.1 and section 4.2. Afterwards, we analyze the experiments of our new evaluator variants, ensuring that improvements in generators are accurately reflected and substantiated by the reliable evaluator. We compare the quantitative results of our CLaM-H and CLaM-X in section 4.3 and the results of our M-Adaptor in section 4.4. At last, we provide qualitative comparison results in section 4.5.

4.1. Experimental Dataset

Our experiments are conducted on primary text-driven human motion generation dataset Motion-X [22] and its extension Motion-X-Whole [15, 22]. We adopt the Motion-X dataset to train the body-hand motion generator and its corresponding CLaM-H evaluator, while we use Motion-X-Whole to train the whole-body motion generator and its corresponding CLaM-X evaluator, incorporating facial expressions.

Motion-X [22] comprises 55,705 human motion sequences and 107,522 video-level textual annotations, setting a new standard for scale and detail in human motion generation tasks. Each frame of the dataset includes expressive body-hand pose annotations using the SMPL-H model, capturing a wide range of scenarios and actions from various video sources, including games and outdoor scenes.

Generators	Base Model	R-Precision (%) ↑			FID ↓	MM-Dist ↓	Diversity →	MModality ↑
		Top-1	Top-2	Top-3				
Real motion	-	$79.59^{\pm0.1}$	$90.53^{\pm0.1}$	$94.04^{\pm0.1}$	$0.003^{\pm0.000}$	$4.181^{\pm0.004}$	$8.730^{\pm0.024}$	-
MotionGPT-H	MotionGPT [41]	$47.44^{\pm0.3}$	$64.68^{\pm0.3}$	$73.61^{\pm0.3}$	$0.899^{\pm.005}$	$5.416^{\pm.015}$	$\underline{8.868}^{\pm.042}$	$3.778^{\pm.224}$
T2M-H	T2M [15]	$57.26^{\pm0.2}$	$72.09^{\pm0.2}$	$78.71^{\pm0.2}$	$2.462^{\pm.014}$	$5.401^{\pm.005}$	$8.227^{\pm.036}$	$3.189^{\pm.138}$
T2M-GPT-H	T2M-GPT [39]	$65.51^{\pm0.5}$	$79.37^{\pm0.3}$	$85.25^{\pm0.2}$	$0.876^{\pm.009}$	$5.003^{\pm.006}$	$8.450^{\pm.045}$	$3.650^{\pm.039}$
MotionDiffuse-H§	MotionDiffuse [40]	$\underline{68.65}^{\pm0.3}$	$\underline{81.05}^{\pm0.3}$	$86.23^{\pm0.2}$	$0.702^{\pm.008}$	$\underline{4.758}^{\pm.008}$	$9.206^{\pm.043}$	$\mathbf{3.974}^{\pm.163}$
M-Adaptor (Stage I)	-	$67.07^{\pm0.2}$	$80.90^{\pm0.2}$	$\underline{86.72}^{\pm0.2}$	$\underline{0.647}^{\pm.004}$	$5.128^{\pm.006}$	$8.557^{\pm.019}$	$3.771^{\pm.025}$
M-Adaptor (Stage II)	-	$\mathbf{71.60}^{\pm0.1}$	$\mathbf{84.71}^{\pm0.2}$	$\mathbf{89.64}^{\pm0.2}$	$\mathbf{0.563}^{\pm.003}$	$\mathbf{4.718}^{\pm.004}$	$\mathbf{8.640}^{\pm.034}$	$\underline{3.802}^{\pm.028}$

Table 4. **Variants of existing methods for text-driven body-hand motion generation results on Motion-X dataset using CLaM-H model as evaluator**. § denotes results using the ground-truth motion length as a precondition. The evaluation is repeated 20 times, and the mean value is reported, supplemented by a 95% confidence interval. Bold and underlined indicate the best and the second-best results.

The Motion-X dataset is pivotal in advancing expressive whole-body motion generation, as it overcomes previous limitations related to facial expressions and hand gestures. Note that these statistics are for the Motion-X v1 version [3] updated in December 2024: it is still being updated. Following the evaluation protocol [15, 16], the dataset is divided into training, validation, and test sets at ratios of 80%, 5%, and 15% respectively. The model that performs best on the validation set is selected, and its performance on the test set is reported.

Motion-X-Whole [15, 22] is an extension of the above Motion-X dataset which contains 55,705 whole-body human motion sequences and 107,522 text descriptions. The motion sequences, originally from AMASS [27], HumanAct12 [14], etc., undergo specific pre-processing. All motions in Motion-X-Whole have parameters of body motions, hand gestures, and facial expressions. Each motion is paired with at least one accurate textual description, with an average description length of approximately 15. Although Lin *et al.* claim to have annotated this dataset with whole-body pose parameters using the SMPL-X model, only part data of the released dataset has a complete annotation of facial parameters in the updated v1 version [3]. Note that Lin *et al.* have not named Motion-X and Motion-X-Whole in the original paper [22]. In this paper, we use these two special names to facilitate differentiation between Motion-X with different parameters. In accordance with [15], the dataset is split into training, validation, and test sets at ratios of 80%, 5%, and 15%, respectively. We select the model that achieves the best performance on the validation set and report its performance on the test set.

4.2. Implementation Setup

Evaluation Metric. We use the following five distinct metrics as evaluation metrics.

1) Frechet Inception Distance (FID). We extract features from both the generated and ground truth motion sequences using the pre-trained motion encoder of the evaluator. The FID between these two distributions is calculated to measure their similarity.

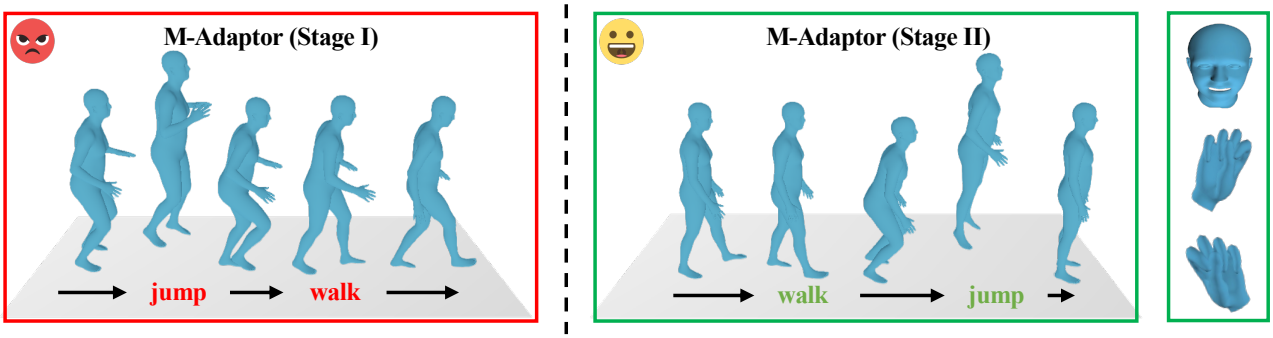
2) R-Precision. For each pair of motion sequences and descriptions, 31 other sentences are randomly selected from the test set. The well-trained contrastive evaluator extracts the motion and text embedding, ranks the Euclidean distances between them, and computes the average top-k motion-to-text retrieval.

3) Diversity. The motion sequences from the test set are randomly divided into pairs. Then, motion features are extracted and the average Euclidean distances in each pair are calculated, forming the diversity metric to measure the diversity of generated motion sequences.

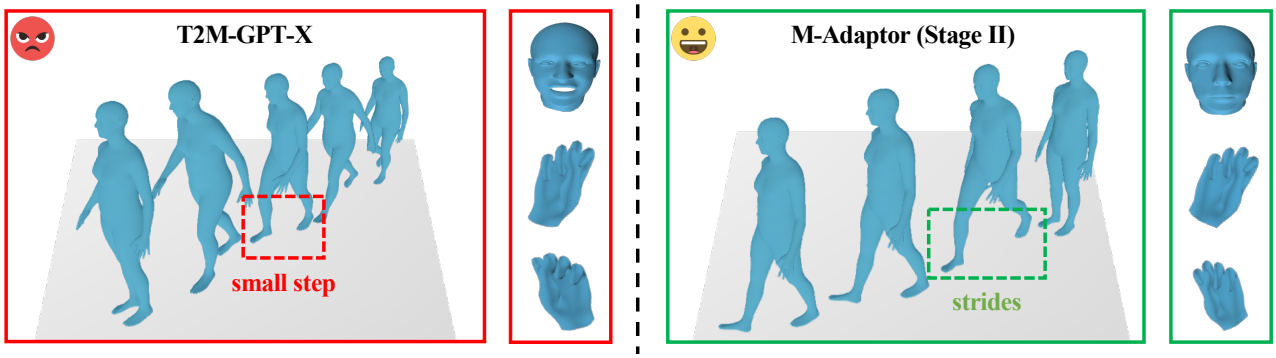
4) Multi Modality (MModality). For a single text description, we randomly generate 20 corresponding motion sequences and form 10 pairs. After that, motion features are extracted, and MModality measures the average Euclidean distances of the pairs.

5) Multi-Modal Distance (MM-Dist). With the help of the well-trained contrastive evaluator, we can calculate the Euclidean distance between the text feature from the given description and the motion feature from the motion sequence, referred to as multi-modal distance.

Implementation details. We introduce the detailed implementation as follows. Our M-Adaptor consists of 9 transformer layers for Stage I, and the LoRA module with rank $r = 64$ for Stage II. Each stage trains the first 100K iterations with a learning rate of 1e-4, and the second 100K iterations with a learning rate of 1e-5. As for CLaM-H and CLaM-X, the max context length of the text tokenizer is set to 77, and context is truncated if it is over this length. The dimension of text features is 512 to match the dimension of motion features. We train these evaluators using the AdamW [24] optimizer, with $[\beta_1, \beta_2] = [0.9, 0.98]$, a batch size of 64, and weight decay $wd = 0.01$. Training lasts 120K iterations with a learning rate of 3e-5. The backbone of CLaM-X is pre-trained using the well-trained parameters of CLaM-H.



(a) The ablation studies of stage I and II, the prompt is *"a person walks forward and then jumps up, **happily**"*



(b) The qualitative comparison with previous method, the prompt is *"a person **strides** forward, **sadly**."*

Figure 4. **Qualitative Comparison.** We provide the visualization with body motions, facial expressions, and hand gestures using different methods. To enhance the clarity of the visualization, the facial expressions and hand gestures are visualized separately.

4.3. Experiments on CLaM-H and CLaM-X

To thoroughly validate the effectiveness of our proposed CLaM-H and CLaM-X evaluators in text-driven whole-body motion generation tasks, this subsection presents a comprehensive comparison between our evaluators and several variants of the default evaluator as described in [15]. The evaluation focuses on the ability of these models to evaluate and rank generated motion sequences in comparison to ground-truth data.

As detailed in Table 2, we provide a quantitative analysis based on ground-truth motion sequences. The results highlight our CLaM-H evaluator's superior performance, achieving a Top-1 R-Precision of 79.59% and marking a substantial improvement of 28.84% over the original evaluator. Similarly, the CLaM-X evaluator demonstrates impressive performance with a Top-1 R-Precision of 79.10%, an increase of 27.16% compared to the baseline. It is important to note that while most metrics derived from ground-truth motion sequences can be influenced by scaling the features of motions and texts, the R-Precision metric remains robust and reliable for comparison purposes.

In addition, we conduct ablation studies to analyze the influence of pre-training on CLaM-X performance, as declared in subsection 4.2. The ablation studies are shown in Tab. 3. We observe a significant improvement in Top-1 R-Precision of 8.57% (70.53%→79.10%) when adding pre-training to our CLaM-X model, while only a slight improvement in Top-1 R-Precision of 3.47% (48.47%→51.94%) on the T2M-X model.

4.4. Experiments on M-Adaptor

In this subsection, we evaluate M-Adaptor's performance by comparing it with various existing methods, including T2M [15] and T2M-GPT [39], regarding the task of whole-body text-driven human motion generation. The evaluation is conducted on two datasets: the comprehensive Motion-X dataset and its subset, the Motion-X-Whole dataset. Following the evaluation criteria established in [15], each evaluation is repeated 20 times to ensure statistical reliability, with the mean values reported alongside a 95% confidence interval. This rigorous approach allows for a robust comparison of performance.

The results of whole-body text-driven motion generation on the Motion-X dataset are shown in Table 4, while the results on the Motion-X-Whole dataset are shown in Table 5.

Generators	Base Model	R-Precision (%) ↑			FID ↓	MM-Dist ↓	Diversity →	MModality ↑
		Top-1	Top-2	Top-3				
Real motion	-	$79.10^{\pm0.1}$	$89.39^{\pm0.1}$	$92.63^{\pm0.1}$	$0.002^{\pm0.000}$	$4.294^{\pm0.004}$	$8.951^{\pm0.036}$	-
MotionGPT-X	MotionGPT [41]	$47.01^{\pm0.2}$	$64.07^{\pm0.2}$	$72.55^{\pm0.2}$	$0.908^{\pm.007}$	$5.613^{\pm.020}$	$8.763^{\pm.038}$	$3.771^{\pm.201}$
T2M-X	T2M [15]	$57.40^{\pm0.1}$	$71.74^{\pm0.2}$	$78.20^{\pm0.2}$	$2.804^{\pm.015}$	$5.461^{\pm.005}$	$8.148^{\pm.029}$	$3.161^{\pm.111}$
T2M-GPT-X	T2M-GPT [39]	$65.55^{\pm0.5}$	$78.63^{\pm0.3}$	$84.88^{\pm0.2}$	$0.931^{\pm.008}$	$5.102^{\pm.005}$	$8.480^{\pm.042}$	$3.593^{\pm.033}$
MotionDiffuse-X§	MotionDiffuse [40]	$\underline{68.27}^{\pm0.2}$	$80.67^{\pm0.2}$	$85.89^{\pm0.2}$	$0.713^{\pm.007}$	$\underline{4.803}^{\pm.009}$	$\underline{9.127}^{\pm.036}$	$\mathbf{3.856}^{\pm.138}$
M-Adaptor (Stage I)	-	$66.77^{\pm0.2}$	$\underline{80.72}^{\pm0.2}$	$\underline{86.01}^{\pm0.2}$	$\underline{0.690}^{\pm.003}$	$5.412^{\pm.005}$	$8.683^{\pm.021}$	$3.691^{\pm.019}$
M-Adaptor (Stage II)	-	$\mathbf{71.29}^{\pm0.1}$	$\mathbf{84.35}^{\pm0.2}$	$\mathbf{89.23}^{\pm0.2}$	$\mathbf{0.569}^{\pm.002}$	$\mathbf{4.758}^{\pm.004}$	$\mathbf{8.803}^{\pm.023}$	$\underline{3.780}^{\pm.025}$

Table 5. **Variants of existing methods for text-driven whole-body motion generation results on Motion-X-Whole dataset using CLaM-X model as evaluator**. § denotes results using the ground-truth motion length as a precondition. The evaluation is repeated 20 times, and the mean value is reported, supplemented by a 95% confidence interval. Bold and underlined indicate the best and the second-best results.

r_0	r_b	R-Precision (%) ↑		
		Top-1	Top-2	Top-3
0.7	0.5	$69.16^{\pm0.2}$	$82.54^{\pm0.2}$	$87.73^{\pm0.2}$
0.7	0.3	$69.76^{\pm0.2}$	$83.14^{\pm0.2}$	$87.73^{\pm0.2}$
0.5	0.3	$70.68^{\pm0.2}$	$82.84^{\pm0.2}$	$88.03^{\pm0.2}$
0.7	0.1	$70.37^{\pm0.1}$	$83.74^{\pm0.2}$	$88.63^{\pm0.2}$
0.5	0.1	$\mathbf{71.29}^{\pm0.1}$	$\mathbf{84.35}^{\pm0.2}$	$\mathbf{89.23}^{\pm0.2}$
0.3	0.1	$69.76^{\pm0.2}$	$82.54^{\pm0.1}$	$88.03^{\pm0.1}$

Table 6. **Ablation studies of the highest and lowest mask rate r_0 and r_b for stage II of our M-Adaptor on Motion-X-Whole dataset using CLaM-X model as evaluator**.

N	R-Precision (%) ↑			Time
	Top-1	Top-2	Top-3	
1	$68.86^{\pm0.2}$	$83.14^{\pm0.3}$	$87.44^{\pm0.2}$	+3.4%
2	$69.76^{\pm0.2}$	$83.14^{\pm0.2}$	$88.33^{\pm0.2}$	+6.9%
3	$70.37^{\pm0.2}$	$83.44^{\pm0.2}$	$88.63^{\pm0.1}$	+10.3%
4	$\mathbf{71.29}^{\pm0.1}$	$\mathbf{84.35}^{\pm0.2}$	$89.23^{\pm0.2}$	+13.8%
5	$70.68^{\pm0.2}$	$83.44^{\pm0.2}$	$\mathbf{89.53}^{\pm0.2}$	+17.2%

Table 7. **Ablation studies of the iteration number N of stage II on Motion-X-Whole dataset using CLaM-X model as evaluator**. 'Time' means the increase of inference time in percent.

Compared to previous methods, we find that our M-Adaptor is effective and outperforms the performance under most metrics, especially in R-Precision. It even outperforms the huge MotionGPT [41] model containing 770M parameters.

Additionally, we conduct ablation studies to investigate the influence of key hyperparameters on M-Adaptor's performance. Specifically, we examine the effects of the highest and lowest mask rates, denoted as r_0 and r_b, as well as the iteration number N in Stage II of the model. The results of these studies are detailed in Tab. 6 and Tab. 7, respectively. Our research indicates that setting r_0 to 0.5, r_b to 0.1, and N to 4 yields optimal performance in terms of R-Precision metrics. These insights provide valuable guidance for fine-tuning the M-Adaptor to achieve its best possible performance in text-driven motion generation tasks.

4.5. Qualitative Comparison

We show qualitative visualizations in Fig. 4, where we compare our M-Adaptor in each stage and with an existing method. Our two-stage M-Adaptor exhibits a more robust semantic understanding ability. As depicted in Fig. 4 (a), in stage II our adaptor can correct the wrong order of generated movements from stage I. Moreover, as shown in Fig. 4 (b), our generator is able to better understand semantic differences between similar words compared to the existing T2M-GPT-X method.

5. CONCLUSION

In this work, we focus on text-driven whole-body human motion generation, addressing the challenges of generating natural and expressive human motions that include body motions, facial expressions, and hand gestures. We introduce M-Adaptor, a novel two-stage framework, which significantly enhances the quality and diversity of generated motion sequences. Additionally, we extend the existing CLaM model to CLaM-H and CLaM-X, providing a comprehensive evaluation framework for text-driven whole-body human motion generation. These evaluators enable a more thorough evaluation of motion generation capabilities. Extensive qualitative and quantitative evaluations demonstrate the superior performance of our framework, with significant improvements in R-Precision and FID. Our contributions pave the way for more natural and expressive human motion generation, with broad applications in virtual characters, motion-guided video generation, and 3D digital humans. Future work will focus on further refining the expressiveness of generated motions and exploring new applications in interactive and immersive environments.

References

[1] aplaybox. *https://www.aplaybox.com/*, 2025. 1

[2] mixamo. *http://www.mixamo.com/*, 2025. 1

[3] Motion-X-V1. *https://github.com/IDEA-Research/Motion-X/tree/ce7c869273730152a469e564ee2df94e07117b34/*, 2025. 6

[4] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *ICRA*, pages 1–5. IEEE, 2018. 2

[5] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *3DV*, pages 719–728. IEEE, 2019. 2

[6] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *3DV*, pages 414–423. IEEE, 2022. 2

[7] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, pages 18000–18010. IEEE, 2023. 1, 2

[8] Xiaodong Chen, Kunlang He, Wu Liu, Xinchen Liu, Zheng-Jun Zha, and Tao Mei. Clam: An open-source library for performance evaluation of text-driven human motion generation. In *ACM Multimedia*, pages 11194–11197. ACM, 2024. 1, 2, 3, 5

[9] Xiaodong Chen, Wu Liu, Qian Bao, Xinchen Liu, Quanwei Yang, Ruoli Dai, and Tao Mei. Motion capture from inertial and vision sensors. *CoRR*, abs/2407.16341, 2024. 2

[10] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, pages 2694–2703. IEEE, 2023.

[11] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip H. S. Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *ICCV*, pages 20167–20177. IEEE, 2023. 2

[12] Jeffrey L. Elman. Finding structure in time. *Cogn. Sci.*, 14 (2):179–211, 1990. 1, 2

[13] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *ACM Multimedia*, pages 2021–2029. ACM, 2020. 2

[14] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *ACMMM*, pages 2021–2029, 2020. 6

[15] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, pages 5142–5151. IEEE, 2022. 1, 2, 3, 5, 6, 7, 8

[16] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. TM2T: stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV (35)*, pages 580–597. Springer, 2022. 1, 6

[17] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *CVPR*, pages 1900–1910. IEEE, 2024. 1, 2

[18] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, pages 1735–1742. IEEE, 2006. 5

[19] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net, 2022. 2

[20] Qiao Jin, Xiaodong Chen, Wu Liu, Tao Mei, and Yongdong Zhang. T-SVG: text-driven stereoscopic video generation. *CoRR*, abs/2412.09323, 2024. 2

[21] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2

[22] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 2023. 1, 2, 3, 5, 6

[23] Hengyuan Liu, Xiaodong Chen, Xinchen Liu, Xiaoyan Gu, and Wu Liu. Animateanywhere: Context-controllable human video generation with id-consistent one-shot learning. In *HCMA@ACM Multimedia*, pages 41–43. ACM, 2024. 1

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 6

[25] Qiujing Lu, Yipeng Zhang, Mingjian Lu, and Vwani Roychowdhury. Action-conditioned on-demand motion generation. In *ACM Multimedia*, pages 2249–2257. ACM, 2022. 2

[26] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: text-aligned whole-body motion generation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 32939–32977, 2024. 2, 3

[27] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *ICCV*. 6

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[29] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 5

[30] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer VAE. In *ICCV*, pages 10965–10975. IEEE, 2021. 1, 2

[31] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: generating diverse human motions from textual descriptions. In *ECCV (22)*, pages 480–497. Springer, 2022. 2

[32] Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *ICCV*, pages 9488–9497, 2023. 2, 4

[33] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2

[34] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2, 5

[35] Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. A survey of multimodal-guided image editing with text-to-image diffusion models. *CoRR*, abs/2406.14555, 2024. 2

[36] Xincheng Shuai, Henghui Ding, Zhenyuan Qin, Hao Luo, Xingjun Ma, and Dacheng Tao. Free-form motion control: A synthetic video generation dataset with controllable camera and object motions. *CoRR*, abs/2501.01425, 2025. 1

[37] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 1, 2, 3

[38] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NIPS*, pages 6306–6315, 2017. 2, 4

[39] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023. 2, 6, 7, 8

[40] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024. 1, 2, 3, 6, 8

[41] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. *arXiv preprint arXiv:2306.10900*, 2023. 2, 6, 8

[42] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7368–7376, 2024. 2

[43] Yuhong Zhang, Jing Lin, Ailing Zeng, Guanlin Wu, Shunlin Lu, Yurong Fu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x++: A large-scale multimodal 3d whole-body human motion dataset. *arXiv preprint arXiv:2501.05098*, 2025. 1, 2, 3