Large Motion Model for Unified Multi-Modal Motion Generation

Mingyuan Zhang[⋆,1], Daisheng Jin[⋆,1], Chenyang Gu[⋆,1], Fangzhou Hong[1], Zhongang Cai[1,2], Jingfang Huang[1], Chongzhi Zhang[1], Xinying Guo[1], Lei Yang[2], Ying He[1], Ziwei Liu[†,1]

[1] S-Lab, Nanyang Technological University, Singapore
[2] SenseTime Research, China

Project Page: https://mingyuan-zhang.github.io/projects/LMM.html

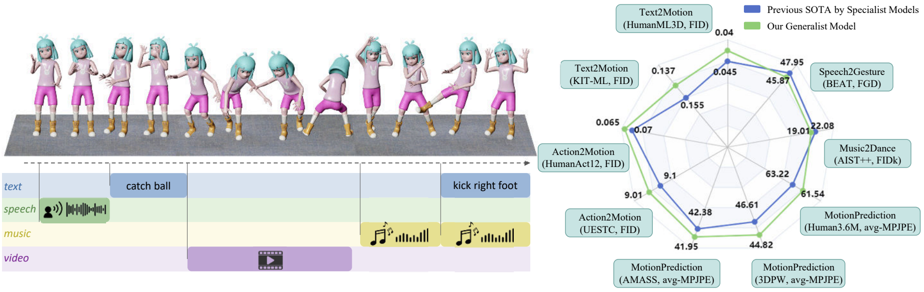


Fig. 1: We present Large Motion Model (LMM), the first generalist multi-modal motion generation model, that can perform multiple motion generation tasks simultaneously and achieve competitive performance across nine widely used benchmarks.

Abstract. Human motion generation, a cornerstone technique in animation and video production, has widespread applications in various tasks like text-to-motion and music-to-dance. Previous works focus on developing specialist models tailored for each task without scalability. In this work, we present **Large Motion Model (LMM)**, a motion-centric, multi-modal framework that unifies mainstream motion generation tasks into a generalist model. A unified motion model is appealing since it can leverage a wide range of motion data to achieve broad generalization beyond a single task. However, it is also challenging due to the heterogeneous nature of substantially different motion data and tasks. LMM tackles these challenges from three principled aspects: **1)** *Data:* We consolidate datasets with different modalities, formats and tasks into a comprehensive yet unified motion generation dataset, **MotionVerse**, comprising 10 tasks, 16 datasets, a total of 320k sequences, and 100 million frames. **2)** *Architecture:* We design an articulated attention mechanism **ArtAttention** that incorporates body part-aware modeling into Diffusion Transformer backbone. **3)** *Pre-Training:* We propose a novel pre-training strategy for LMM, which employs variable frame rates and masking forms, to better exploit knowledge from diverse training data. Extensive experiments demonstrate that our generalist LMM

⋆ co-first authors; † corresponding author

achieves competitive performance across various standard motion generation tasks over state-of-the-art specialist models. Notably, LMM exhibits strong generalization capabilities and emerging properties across many unseen tasks. Additionally, our ablation studies reveal valuable insights about training and scaling up large motion models for future research.

Keywords: Motion Generation · Unified Model · Multi-Modality

1 Introduction

Humans perform a variety of motions in response to environmental changes, personal thoughts, and emotions to achieve their goals. These intricate motions often serve as the most information-rich element within videos and animations that feature characters, making motion generation a critical component of Generative AI, significantly influencing visual experiences and content quality. Automated human motion generation, aimed at producing continuous, natural, and logical human movements based on specific commands and control conditions, has attracted considerable attention in the computer vision field.

These specific sub-tasks are characterized by their defined inputs and outputs, with clear objectives. For instance, action-to-motion generates movements based on the category of the motion [35, 104]; text-to-motion takes textual descriptions and produces corresponding movements [33, 159]; music-to-dance creates dance moves in tune with the style and beat of the music input [30, 123]. Dividing tasks allows for a more focused approach to each sub-task: constructing dedicated datasets and devising methods tailored to the task at hand. However, these approaches, when designed for a singular sub-task or modality, face challenges due to limited data quantity and a narrow data domain, which in turn can lead to models with restricted capabilities and poor generalization performance. In contrast, the objective of this work is to build a unified yet versatile foundation model for human motion generation, leveraging resources from a wide range of applications and achieving strong performance across the board.

Leveraging multi-modal and multi-task motion generation datasets presents significant challenges. First, disparate datasets feature varying motion formats and evaluation metrics, such as keypoint-based or rotation-based formats, and metrics assessing realism or diversity. Consequently, employing a single model to tackle multiple tasks and perform evaluations across different datasets is exceedingly difficult. Furthermore, transferring motion knowledge across tasks within these datasets is challenging, complicating the model's ability to integrate useful knowledge from various data sources to enhance its capabilities. For instance, differences in frame rates and the number of keypoints (sometimes even missing parts of the body) make it hard to unify the learned knowledge. Although some studies attempt to address multiple tasks simultaneously, they often utilize only two or three datasets with the same motion format, which limits their ability to achieve enhanced controllability and generalizability. In summary, integrated motion generation models for multi-modal and multi-task applications encounter the following problems: 1) Non-uniformity of motion data formats; 2) Different evaluation metrics due to varying task objectives; 3) Difficulty in transferring action knowledge across multiple tasks.

To deal with these challenges, we first amass multiple cross-modal motion datasets, encompassing 16 datasets with a total of 320k sequences and 100 million frames. These datasets span seven standard tasks: text-to-motion, action-to-motion, motion prediction, speech-to-gesture, music-to-dance, motion imitation, and motion in-betweening. Additionally, based on the standard tasks and multi-modal control signals, we introduce three new tasks: conditional motion prediction, conditional motion in-betweening, and multi-condition motion generation. Together, these datasets and tasks form our cross-modal motion benchmark, **MotionVerse**. To align the diverse formats of motion data, we employ a two-step approach: 1) We use the TOMATO representation [94] as a unified intermediary format, and then divide the entire representation into 10 parts. All kinds of motion representations are aligned to this format, with annotations indicating which body parts are present in each sequence; 2) We train a series of representation translators to convert the unified motion representation into the specific representations of each dataset during the testing phase. With MotionVerse, we can smoothly use training data from various tasks and modalities, and conduct tests across different datasets.

Building on MotionVerse, we introduce the multi-modal Large Motion Model (**LMM**), which is built on a transformer-based diffusion model. Addressing the motion format inconsistency, we developed a body part-aware motion generation model. This model divides the human body into 10 segments and employs a specialized attention mechanism **ArtAttention**, featuring multi-conditioning, spatial-temporal independence, and mask injection, allowing for distinct control over different body parts. Furthermore, body part-aware modeling decomposes motion data from various datasets into relatively independent segments, thereby enabling the model to more effectively leverage knowledge learned across different datasets. Lastly, we adopt learning strategies from large language models (LLM), proposing a training method for LMM that combines unsupervised and supervised learning. In unsupervised learning, we enhance model robustness to frame rates through random frame rate augmentation and improve control over the continuity of body part movements by applying random masks to sequences and body parts in various ways. This training approach significantly leverages large amounts of multi-modal data to bolster LMM's capabilities. In supervised learning, we refine the capabilities of models to enhance their performance on specific tasks. Experimental results show that LMM achieves state-of-the-art results across various tasks, demonstrating its exceptional generalization performance, as shown in Fig. 1. Furthermore, LMM has the ability to process multi-modal inputs simultaneously, enabling it to accomplish unseen tasks.

In summary, our core contributions are as follows:

1. We present MotionVerse, a mega-scale, multi-modal, multi-task motion generation dataset that features a unified motion representation across a wide range of tasks and motion formats.

2. We introduce a Large Motion Model (LMM) that incorporates an advanced attention mechanism ArtAttention, allowing for precise and robust control, achieving finer results.

3. We devise a pre-training strategy for the LMM, including random frame rates and various masking techniques to fully leverage extensive motion datasets and enhance the model's capabilities. Additionally, through ablation studies on our training approach, we explored certain characteristics inherent in LMM's training process, laying a foundation for future research on the LMM.

2 Related Work

2.1 Motion Synthesis

Subtasks of motion generation are differentiated based on the types of control signals they utilize. Some tasks require the algorithm to synthesize motion sequences based on the given uncompleted motion sequences. For example, in the motion prediction task [1,2,8,9,14,20,31,36,50,96,125,126,135,144], the control condition is the first several poses, and the generated motion must logically follow the preceding sequence to ensure the extended sequence appears natural and reasonable. There are also works focusing on recovering the whole body motion with the given sparse upper-body tracking signals [12,21].

Action-to-motion [13, 35, 53, 61, 88, 104, 138, 165] is an inverse task derived from the motion recognition task. While the latter identifies the action category from a motion sequence, the former involves generating a corresponding action sequence from an input action category. In the music-to-dance task [22, 45, 63, 64, 67, 69, 101, 111, 123, 124, 131, 146, 149, 171], the control signal is music, and the motion sequence (dance) should be in accordance with the style and rhythm of the music. Bailando [123] addresses crucial spatial constraints and temporal coherence in dance generation by utilizing a codebook to store standardized dancing units, confining spatial constraints. It achieves temporal fluidity through a GPT designed to detect music beats. Another task based on audio is speech-to-gesture [4,23,59,87,143,152,153,167], where the input is the speaker's audio and the output is the corresponding gestures of the speaker, taking into account the speech's pauses, emotional fluctuations, etc. GestureDiffuCLIP [4] utilizes CLIP to extract style information from the input and then employs a diffusion model to generate gestures. Text-to-motion is one of the most attracting topic in conditional motion generation [3, 5–7, 15, 18, 24, 25, 29, 30, 32, 34, 37, 39, 42, 44, 49, 51, 52, 54, 57, 58, 70, 77, 79, 80, 82, 83, 85, 93, 94, 105, 106, 108, 112–116, 118, 120, 129, 130, 132, 134, 136, 139, 141, 142, 147, 148, 150, 155–161, 163, 164, 168–170], where motions are generated based on textual descriptions. This requires the model to comprehend the meaning of the text and produce a corresponding sequence of motions. Previous works attempted to apply advanced generative model [130, 159, 163] to improve performance. While some other works focused on enhancing controllability [6, 112, 161]. Physical reality [115, 155] and out-of-domain performance [43, 85, 129] are also vital topics in this field.

In addition, some works focus on human-scene interaction generation [46,76,89,140], human-object interaction generation [19,38,62,65,66,81,103,107,121,128,137,162] and human-human interaction generation [11,16,26,74,90,122,127,166]. These tasks greatly expanded the scope of motion generation applications.

However, action generation targeting a single task often struggles with limited data volume and a singular data domain, leading to models with restricted capabilities and poor generalization. Our paper integrates various motion generation tasks, designing the Large Motion Model (LMM) to utilize multi-modality data from different tasks for model training. This enables the model to learn from various domains, enhancing its generalization capabilities.

2.2 Large Diffusion Model

As diffusion models have made remarkable strides in image generation tasks, researchers have extended their application to a broader array of fields, including video generation [56,133,145], image editing [10,17,40,55,100], and motion generation [18,130,159,161,169], etc. Moreover, given the limited functionality and control over generation provided by single-modal conditions, multi-modal inputs have been introduced into diffusion models to enhance their versatility and control capabilities. In the realm of image generation, UNIMO-G [73] takes both images and text as inputs, utilizing the subjects in pictures and textual prompts to generate realistic images that match complex semantics. For video generation, MM-Diffusion [117] incorporates audio and video in a multi-modal manner, enabling the harmonious adaptation of audio and visuals to produce realistic videos with sound. In image editing tasks, the integration of multi-modal inputs with diffusion models has made image editing more flexible and convenient. Controlcolor [75] combines text, strokes, exemplars, and other conditions to achieve interactive, multi-modal controlled image coloring. InstructAny2pix [71] uses similar conditions for multi-modal control over inpainting. Likewise, in the field of motion generation, works like MotionDiffuse [159] and MDM [130] have tackled text-to-motion and action-to-motion. MCM [82], UDE [170] has addressed text-to-motion and music-to-dance. In summary, across different domains, multi-modal diffusion models enrich the content and enable more precise control of generation tasks. Previous dual-modal motion generation methods often utilize similar annotation formats, such as SMPL [91], facilitating easy data alignment. However, the multi-modal datasets we collected cover a broader domain span, and the formats for motion annotation are extremely diverse. Thus we propose a comprehensive benchmark, MotionVerse, to unify the motion representation, ultimately leading to the development of LMM.

3 MotionVerse: Unified Motion Generation Datasets

3.1 Motivation

Large models have been extensively studied in fields such as language, image, and video. These models, by absorbing common knowledge from vast amounts of data and leveraging a unified task format, demonstrate outstanding performance across multiple tasks. However, on the path towards large motion models, a significant challenge lies ahead: the inconsistent motion formats across datasets. Specifically, there are three types of inconsistencies:

1. **Inconsistent pose representations**: For instance, the UESTC [48] benchmark adopts a 6D rotation representation based on SMPL [92], while the Human3.6M [47] motion prediction benchmark uses keypoint coordinates.
2. **Inconsistent number of keypoints**: For example, TED-Gesture++ [153] only includes upper body keypoints, while NTU-RGBD 120 [86, 119] lacks fine-grained keypoints for the hands.
3. **Inconsistent frame rates**: KIT-ML [109] operates at 12.5 fps, whereas Motion-X [78] runs at 30 fps.

Such differences not only demand models capable of handling diverse data formats but also pose significant challenges in acquiring common knowledge.

To address this challenge, we introduce the first unified and comprehensive motion-centric benchmark **MotionVerse**, the workflow of which is shown in Fig. 2. MotionVerse possesses three advantages:

1. **Unified Problem Formulation**: we describe mainstream tasks within a unified framework, reducing the need to consider task-specific properties during model design.
2. **Unified Motion Representation**: we convert the motion formats of various datasets into a unified intermediate representation, enabling the model to acquire common knowledge from the originally diverse data formats and to be evaluated on different datasets smoothly.
3. **Systematicness and Comprehensiveness**: we encompass 10 tasks across 16 datasets, comprising 320K sequences and nearly 100M frames of motion data, as shown in Tab. 2, which enables us to explore large motion models.

Task	**Mask**	**Condition**
T2M	None	Text
A2M	None	Action
M2D	None	Music
S2G	None	Speech
MIm	None	Video
MP	$m = \begin{cases} 0 & \text{if } x > k \\ 1 & \text{Otherwise} \end{cases}$	None
MIn	$m = \begin{cases} 0 & \text{if } k_1 < x \leq k_2 \\ 1 & \text{Otherwise} \end{cases}$	None
CMP	$m = \begin{cases} 0 & \text{if } x > k \\ 1 & \text{Otherwise} \end{cases}$	Any single modal
CMI	$m = \begin{cases} 0 & \text{if } k_1 < x \leq k_2 \\ 1 & \text{Otherwise} \end{cases}$	Any single modal
MMG	None	Multi-modal

Table 1: Task definitions. x and k_* are the frame indices, and k_* represents the boundary of the mask.

Dataset	**#Seq**	**#Frames**	**Repr**	**Condition**
HumanML3D [33]	14614	2M	H3D	Text
KIT-ML [109]	2485	245K	H3D	Text
Motion-X [78]	50863	9M	SMPLX	Text
BABEL [110]	5123	7M	SMPLX	Text
UESTC [48]	25600	10M	SMPL	Action
HumanAct12 [35]	1191	90K	Kpt3D	Action
NTU-RGB-D 120 [86, 119]	139656	10M	Kpt3D	Action
AMASS [95]	14244	20M	SMPLX	-
3DPW [98]	81	140K	SMPL	Video
Human3.6M [47]	210	530K	Kpt3D	Video
TED-Gesture++ [153]	34491	10M	Kpt3D	Speech
TED-Expressive [87]	27221	8M	Kpt3D	Speech
Speech2Gesture-3D [60]	1047	1M	Kpt3D	Speech
BEAT [84]	1639	18M	Kpt3D	Speech
AIST++ [69]	1408	1M	SMPL	Music
MPI-INF-3DHP [99]	16	1M	Kpt3D	Video
Total	320K	100M	-	-

Table 2: Dataset information. We collect 16 widely used dataset, process all motion data into our intermediate format.

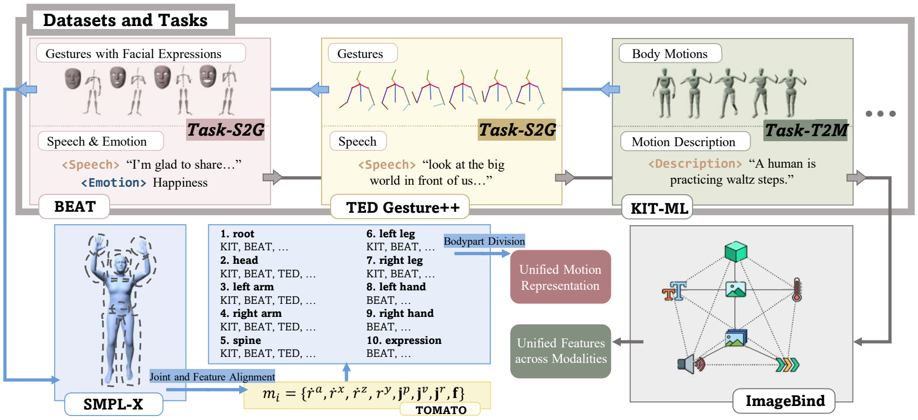


Fig. 2: MotionVerse. We preprocess distinct motion-centric datasets into a unified format. As for motion sequences, we initially convert them to the TOMATO [94] representation and then further divide them into 10 independent body parts, serving as our unified motion representation. To tackle multi-modal condition signals, we employ ImageBind [28] to transform them into unified features across modalities.

3.2 Unified Problem Formulation

Motion generation encompasses a variety of sub-tasks with differing objectives. To standardize the data for these tasks, we first formalize the input format for motion generation tasks as:

$$\Theta = M(\mathbf{x}, \mathrm{m}, \mathrm{c}), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^{F \times D}$ represents the motion data, $\mathrm{m} \in \{0,1\}^{F \times D}$ defines the model's visibility scope, which is used in motion completion tasks, such as motion prediction and motion in-betweening. Here F is the number of frames of the target motion sequence. D is the dimensionality of each pose state. c is a set of condition control signals, including text, speech, music, and video.

We have included seven standard tasks: action-to-motion (A2M), text-to-motion (T2M), music-to-dance (M2D), speech-to-gesture (S2G), motion prediction (MP), motion in-betweening (MIn) and motion imitation (MIm), along with three multi-modal tasks: conditional motion prediction (CMP), conditional motion in-betweening (CMI), and multi-condition motion generation (MMG). For each specific task, we can align their inputs using Eq.(1). Tab. 1 lists the details of these ten tasks.

3.3 Unified Motion Representation

For data formats with varying inputs and outputs across different tasks, we preprocess to ensure format consistency. The basic unit of general motion datasets can be defined as <input-output> pairs. For inputs, we consider multi-modalities including text, speech, music, and video. To align these modalities, we employ

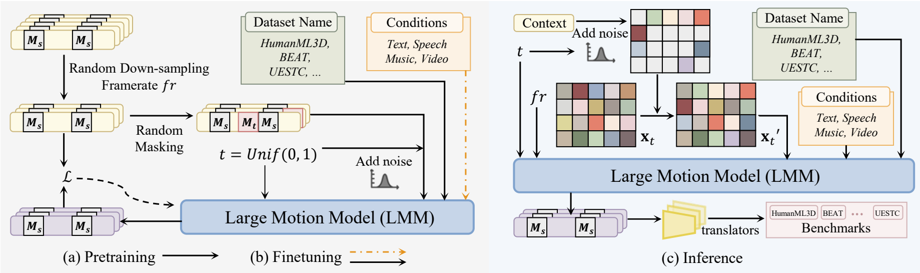


Fig. 3: Overall pipeline of LMM. Left: Our two-stage training procedure, including unsupervised pretraining and supervised fine-tuning. Random down-sampling and random mask strategies are applied to enhance knowledge absorption. **Right**: The generic inference process of LMM. The noised motion sequence and the given context are initially merged before being input into the network. LMM will then synthesize motion sequences, consistent with the provided multi-modal condition signals.

Imagebind [28] to encode different inputs into unified features of the same dimension, which ensures semantic consistency across modalities.

As for the output, it encompasses motion sequences in various formats, such as keypoints and SMPL [92]. To standardize motion representation, we define a unified format similar to TOMATO [94]. Our representation is described as:

$$m_i = \{\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y, \mathbf{j}^p, \mathbf{j}^v, \mathbf{j}^r, \mathbf{f}\}\,, \tag{2}$$

where r denotes information related to the root. Specifically, $\dot{r}^a \in \mathbb{R}$ is the angular velocity along the Y-axis, $(\dot{r}^x, \dot{r}^z \in \mathbb{R})$ represent linear velocities on the XZ-plane, and r^y indicates the root's height. $\mathbf{j}^p \in \mathbb{R}^{3(J-1)}$, $\mathbf{j}^v \in \mathbb{R}^{3J}$, and $\mathbf{j}^r \in \mathbb{R}^{6(J-1)}$ correspond to the position, velocity, and rotation of local keypoints relative to the root, with J denoting the number of joints ($J-1$ means all joints without the root joint). Here we follow SMPL-X [102] and consider 22 main body joints and 30 hand joints. Lastly, $\mathbf{f}$ denotes facial expression [72].

We further divide this representation into ten independent parts: global orientation and trajectory, face expression, head, spine, left arm, right arm, left leg, right leg, left hand, and right hand. When processing raw data, we allow for missing body parts and annotate them in the metadata. For missing keypoints, we utilize prior knowledge of the human body for completion, while extra keypoints are discarded in this process. We then train a series of motion translators to map our unified motion representation to each dataset's specific one. Thus, in the testing process of different tasks, once the model outputs in our unified format, we can map the output to the corresponding motion format through the translator, facilitating smooth metric evaluation.

4 Large Motion Model

Our model architecture closely follows the literature [130, 159], built upon a transformer-based diffusion model. We primarily reference the FineMoGen [161]

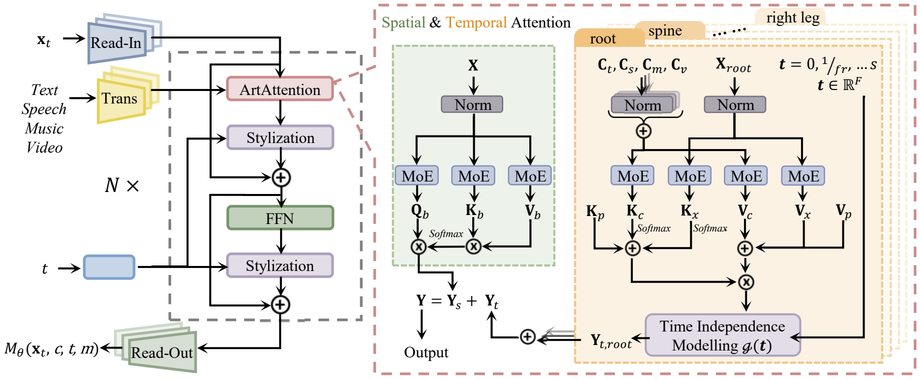


Fig. 4: Architecture of LMM. LMM is a transformer-based diffusion model. Dataset-dependent Read-In layers and Read-Out layers facilitate the conversion of the motion sequence between our intermediate representation and the latent feature space. In the stem of LMM, ArtAttention refines the feature representations through the spatial and temporal attention branches.

as our baseline and extend it to support various condition signals, multi-tasking, multiple frame rates, and various mask forms. The overall workflow is illustrated in Fig. 3 while the detailed architecture is shown in Fig. 4.

4.1 Transformer-based Diffusion Model

The diffusion model is a powerful generative model, capable of producing high-quality, diverse outcomes, garnering widespread attention, and demonstrating formidable generative capabilities in many fields. Its essence lies in two intertwined processes: the forward diffusion process and the reverse diffusion process. More details are provided in the supplementary materials.

4.2 Read-In Layer and Read-Out Layer

The read-in layer is responsible for transforming noised motion data into feature representations of the form $F \times H \times D$, while the read-out layer generates clean motion data from the feature space. Here, the first dimension represents the number of frames in the motion sequence, and the second dimension represents the number of body parts. Although we standardize all motion data into a unified motion format, the differences in distribution among various datasets cannot be completely ignored. Therefore, in the backbone network's read-in and read-out stages, we employ dataset-dependent motion encoders and decoders. Additionally, to obtain more comprehensive knowledge for practical applications, during training, there is a 10% probability of replacing the original dataset name with "all". Consequently, the corresponding read-in and read-out layers can be better applied in real-world application scenarios.

4.3 ArtAttention

To achieve outstanding zero-shot continuous generation capability, our model builds upon the SAMI module from FineMoGen [161], incorporating upgrades to address three new requirements: multi-modal condition, various frame rates, and allowance for missing body parts, as shown in Fig. 4.

For multi-modal signals, we preprocess all signals into token sequences using the ImageBind [28] model. To better integrate these features into our network, we further refine them with two learnable transformer encoder layers. This process transforms text, speech, music, and video into feature sequences $\mathbf{C}_t \in \mathbb{R}^{L_t \times (H \cdot D)}, \mathbf{C}_s \in \mathbb{R}^{L_s \times (H \cdot D)}, \mathbf{C}_m \in \mathbb{R}^{L_m \times (H \cdot D)}, \mathbf{C}_v \in \mathbb{R}^{L_v \times (H \cdot D)}$, where L_t, L_s, L_m, L_v represent the lengths of the corresponding condition sequences, and $H \cdot D$ denotes the feature length of each element.

Our attention mechanism can be divided into two main components: body-part attention(spatial attention) and temporal attention. Assuming the feature representation of our motion sequence is $\mathbf{X} \in \mathbb{R}^{F \times H \times D}$. In the body-part attention section, due to the presence of inherent missing body parts in our data and the masked body parts introduced artificially during pre-training, we cannot utilize a fixed set of coefficients to determine the mutual contributions among body parts. Therefore, unlike the design in FineMoGen, for each frame, we opt to use an attention structure to obtain a set of refined features $\mathbf{Y_s} \in \mathbb{R}^{F \times H \times D}$.

In the temporal attention section, we aim to leverage the self-correlation inherent in the motion features $\mathbf{X}$ and guidance obtained from the condition signals $\mathbf{C}_t, \mathbf{C}_s, \mathbf{C}_m, \mathbf{C}_v$ to generate higher-quality features for each body part in every frame. Here, we employ a Multi-head Attention mechanism, with each head focusing on a specific body part, emphasizing the utilization of temporal correlations to optimize features. We begin by utilizing Mixture-of-Expert to obtain a set of features $\mathbf{K} \in \mathbb{R}^{L \times D}$ from these condition features, where L represents the sequence length of the corresponding feature source.

Empirically, we found that directly concatenating motion sequences and condition sequences like FineMoGen and then applying Softmax processing hinders the modeling of self-correlation in multi-condition scenarios, resulting in lower motion quality, especially when the condition feature sequence is much longer than the motion sequence. Therefore, we independently normalize the motion feature $\mathbf{K}_x \in \mathbb{R}^{F \times D}$ obtained and then normalize condition features. To support unconditional generation, we introduce 64 learnable tokens as placeholders. These tokens are concatenated with all condition signals for normalization, resulting in $\mathbf{K}_c \in \mathbb{R}^{(64+L_t+L_s+L_m+L_v) \times D}$. This approach also facilitates better blending of the model across different conditions. The remaining processing is similar to FineMoGen. After obtaining a series of time-varying signals, for each frame's pose, we use time as the sole query feature to calculate the correlation between each frame and each time-varying signal, as well as the values at each signal point. Unlike FineMoGen, which uses frame indices to represent time, we use real time to support different frame rates. More details are introduced in the supplementary material. Suppose the output of the temporal attention section is $\mathbf{Y}_t \in \mathbb{R}^{F \times H \times D}$, then the output of the entire ArtAttention is $\mathbf{Y} = \mathbf{Y}_s + \mathbf{Y}_t$.

4.4 Pre-Training and Fine-Tuning

Leveraging MotionVerse, we have collected a vast array of <input-motion> data pairs. Disregarding the inputs, the abundant motion data inherently contains valuable information, which can enable the model to comprehend numerous characteristics of human motion, such as coherence, balance, and joint-based rotations, among others. To enable the model to better acquire common knowledge across datasets, we divide the entire training process into two main parts: **Unsupervised Pre-Training** and **Supervised Fine-Tuning**. In the first stage, we require our model to learn motion priors independent of conditions. While in the second stage, the model is supposed to learn the correlation between condition signals and motion sequences.

As illustrated in Fig. 3, in the unsupervised pretraining phase, to enrich the model's prior knowledge from these motion sequences, we employ random downsampling and random masking strategies for data augmentation. Since our representation includes terms related to velocity, when downsampling, we need to recalculate the velocity values to match the downsampling rate, while the terms related to states remain unchanged. This approach enables the model to better learn from data with different original frame rates. As mentioned earlier in the data preprocessing part, some body parts in the sequences are masked out, such as the detailed keypoints of the left and right hands in the KIT dataset. We denote this original mask as $\mathbf{M}_s \in \{0,1\}^{F\times H}$. Based on this, we additionally apply masking with a certain probability to obtain a new mask $\mathbf{M}_t \in \{0,1\}^{F\times H}$. After the model performs the read-in operation, we replace the body parts marked as 1 in $\mathbf{M}_t$ with learnable empty tokens. When calculating the loss, we only ignore the parts marked by $\mathbf{M}_s$. This means that the model not only needs to restore the noised sequence to its clean parts but also utilize the visible part to infer the rest. With this modeling, the knowledge embedded in the data with missing body parts can be better absorbed by the model.

In the supervised fine-tuning phase, our goal is to enable the model to learn the relationship between condition signals and motion sequences. Here, we pass the preprocessed condition token sequences as additional inputs to the model. To support classifier-free guidance, during training, we randomly mask out the condition signals with a probability of 10%.

5 Experiments

5.1 Implementation Details

We designed four variants: LMM-Tiny, LMM-Small, LMM-Base, and LMM-Large, which have 90M, 160M, 410M, 760M parameters respectively. We use all data in MotionVerse for both pretraing and finetuning, except for the sequences used in evaluation. We maintained a fixed total batch size of 512. For the Tiny model, we conducted training directly on 8 NVIDIA V100 GPUs with 32GB memory each, with a batch size of 64 per GPU. For larger models, we FP16 and gradient accumulation to achieve training effects equivalent to a batch size of 512

Table 3: Quantitative results of text-to-motion generation on the HumanML3D test set. '↑'('↓') indicates that the values are better if the metric is larger (smaller). We run all the evaluations 20 times and report the average metric and 95% confidence interval. "MM" is MultiModality. The best scores are bold, and the second-best results are underlined.

Methods	R Precision↑			FID↓	MM Dist↓	Diversity↑	MM↑
	Top 1	Top 2	Top 3				
Real motions	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
T2M-GPT [157]	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$0.116^{\pm.004}$	$3.118^{\pm.011}$	$9.761^{\pm.081}$	$1.856^{\pm.011}$
MDM [130]	-	-	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$\mathbf{2.799^{\pm.072}}$
FineMoGen [161]	$0.504^{\pm.002}$	$0.690^{\pm.002}$	$0.784^{\pm.002}$	$0.151^{\pm.008}$	$2.998^{\pm.008}$	$9.263^{\pm.094}$	$2.696^{\pm.079}$
MoMask [32]	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$0.045^{\pm.002}$	$2.958^{\pm.008}$	-	$1.241^{\pm.040}$
LMM-Tiny	$0.496^{\pm.002}$	$0.685^{\pm.002}$	$0.785^{\pm.002}$	$0.415^{\pm.002}$	$3.087^{\pm.012}$	$9.176^{\pm.074}$	$1.465^{\pm.048}$
LMM-Small	$0.505^{\pm.002}$	$0.693^{\pm.002}$	$0.789^{\pm.002}$	$0.227^{\pm.002}$	$3.051^{\pm.012}$	$9.295^{\pm.076}$	$1.761^{\pm.049}$
LMM-Base	$0.511^{\pm.002}$	$0.710^{\pm.002}$	$0.802^{\pm.002}$	$0.138^{\pm.002}$	$2.971^{\pm.012}$	$9.573^{\pm.076}$	$2.426^{\pm.054}$
LMM-Large	$\mathbf{0.525^{\pm.002}}$	$\mathbf{0.719^{\pm.002}}$	$\mathbf{0.811^{\pm.002}}$	$\mathbf{0.040^{\pm.002}}$	$\mathbf{2.943^{\pm.012}}$	$\mathbf{9.814^{\pm.076}}$	$2.683^{\pm.054}$

Table 4: Quantitative results of motion prediction on the AMASS and 3DPW test set for different time steps (ms). We report the MPJPE error in *mm*.

Method	AMASS-BMLrub								3DPW							
	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000
LTD-10-10 [97]	10.3	**19.3**	36.6	44.6	61.5	75.9	86.2	91.2	12.0	**22.0**	38.9	46.2	59.1	69.1	76.5	81.1
SIMLPE [36]	10.8	19.6	34.3	40.5	50.5	57.3	62.4	65.7	12.1	22.1	38.1	44.5	54.9	62.4	68.2	72.2
GCNext [135]	**10.2**	**19.3**	**34.1**	40.3	50.6	57.3	62.0	65.3	**11.8**	**22.0**	37.9	44.2	55.1	62.1	67.8	72.0
LMM-Tiny	15.9	24.1	38.2	45.9	61.2	73.4	80.3	87.2	17.3	26.2	40.1	47.3	62.8	74.8	82.5	87.0
LMM-Small	14.7	23.2	37.5	43.8	58.3	69.2	75.2	81.9	16.2	25.7	39.4	45.9	60.7	71.5	79.5	82.8
LMM-Base	13.1	21.5	35.9	41.1	53.6	60.8	66.9	70.3	14.1	23.9	38.2	44.1	55.3	64.8	70.3	73.6
LMM-Large	12.8	20.9	34.3	**39.6**	**49.1**	**55.3**	**60.5**	**63.1**	13.1	22.6	**37.1**	**42.4**	**52.4**	**59.2**	**63.8**	**68.0**

without exceeding 32 V100 GPUs. During the pre-training phase, we employed the Adam optimizer with a fixed learning rate of 2×10^{-4} for 80K iterations. In the fine-tuning phase, we used the same optimizer, initially iterating for 20K steps with a learning rate of 2×10^{-4}, followed by 20K steps with a learning rate of 2×10^{-5}. For more details, please refer to the supplementary material.

5.2 Quantitative Results

We evaluate our LMM variants on three tasks: text-to-motion, music-to-dance, motion prediction, and four datasets: HumanML3D [33], 3DPW [98], AMASS [95], and AIST++ [69], as shown in Tab. 3, Tab. 4 and Tab. 5. More experimental results are reported in the supplementary material.

Text-to-Motion. Tab. 3 demonstrates that our LMM-Large surpasses other existing works in terms of accuracy and fidelity while maintaining comparable diversity. On the other hand, LMM-Tiny, which shares a similar structure with FineMoGen, performs worse than it. This discrepancy can be attributed to the significant challenges posed by the large amounts of diverse data and the trade-offs across different tasks during model training, especially for smaller models like LMM-Tiny. Additionally, it's worth noting that the representation dimension used in HumanML3D is 263, whereas ours is 669, significantly increasing the learning difficulties for LMM-Tiny.

Motion Prediction. Tab. 4 reports the performance on AMASS and 3DPW test splits. The superior performance of LMM-Large can be attributed to its robust motion prior, which is obtained from the mega-scale data. It is worth noting that due to the errors introduced by the motion translation step, the accuracy of

Table 5: Quantitative results for Music-conditioned Dance Generation. Quantitative results on AIST++ test set.

Methods	Motion Quality		Motion Diversity		Freezing		Best Align Score↑
	$\text{FID}_k \downarrow$	$\text{FID}_g^{\dagger} \downarrow$	$\text{Div}_k \uparrow$	$\text{FID}_g^{\dagger} \uparrow$	PFF↓	$\text{AUC}_f \downarrow$	
Ground-truth	17.10	10.60	8.19	7.45	0.00	0.00	0.2374
DanceNet [171]	69.18	25.49	2.86	2.85	**0.00**	<u>0.98</u>	0.1430
DanceRevolution [45]	73.42	25.52	3.52	4.87	<u>11.01</u>	12.22	0.1950
Bailando [123]	28.16	**9.62**	7.83	6.34	14.91	13.25	**0.2332**
TM2D [30]	**19.01**	<u>20.09</u>	<u>9.45</u>	6.36	**0.00**	**0.00**	0.2049
LMM-Tiny	37.62	28.95	6.92	5.94	**0.00**	**0.00**	0.1736
LMM-Small	34.18	27.53	7.46	6.17	**0.00**	**0.00**	0.1791
LMM-Base	25.43	24.18	9.05	<u>6.55</u>	**0.00**	**0.00**	0.2084
LMM-Large	<u>22.08</u>	21.97	**9.85**	**6.72**	**0.00**	**0.00**	<u>0.2249</u>

LMM-Large is still lower than other methods in short-distance prediction. However, it exhibits a significant advantage in long-distance prediction. Furthermore, we observed that the advantage of LMM-Large is more pronounced on 3DPW. This is because the 3DPW benchmark demands higher generalization ability from the model. After extensive learning of motion priors, our LMM-Large exhibits a more prominent performance on out-of-distribution tests.

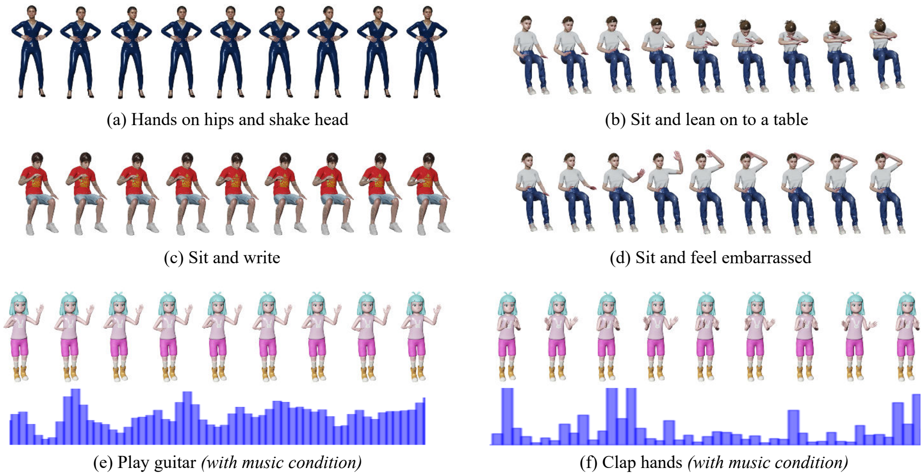
Music-to-Dance. Our Large model achieves comparable performance to the current state-of-the-art, as shown in Tab. 5. In terms of diversity-related metrics, our approach demonstrates a significant advantage. Our performance in the metrics FID_k and FID_g did not surpass existing methods. One possible reason could be the relatively small proportion of the music2dance dataset in the current dataset composition, leading to the model not fully grasping the correlation between the music condition and motion.

5.3 Ablation Study

Table 6: Ablation of the pretraining strategy. All experiments utilized LMM-Base as the base model.

	Downsample	Random Mask	Attention	HumanML3D			AMASS-BMLrub		
				Top 1	FID	MModality	80	400	1000
1)	-	-	ArtAttention	$0.031^{\pm.001}$	$32.814^{\pm.176}$	$5.293^{\pm.129}$	17.3	48.4	89.3
2)	✓	-	ArtAttention	$0.028^{\pm.001}$	$31.365^{\pm.171}$	$5.714^{\pm.147}$	16.2	46.5	78.4
3)	-	✓	ArtAttention	$0.515^{\pm.002}$	$0.151^{\pm.002}$	$2.214^{\pm.051}$	14.5	45.5	76.1
4)	✓	✓	SAMI [161]	$0.400^{\pm.009}$	$1.866^{\pm.009}$	$2.983^{\pm.071}$	15.9	47.2	80.9
5)	✓	✓	ArtAttention	$0.511^{\pm.002}$	$0.138^{\pm.002}$	$2.426^{\pm.054}$	14.1	44.1	73.6

Tab. 6 shows the ablation results. We observed that random masking is a necessary component. When the model's expressive power is strong enough, it can directly recover the clean motion sequence from the noised motion sequence. Consequently, during the fine-tuning stage, our condition signal may not play its expected role. Introducing random masking during training will make it more difficult for the model to solely restore the original sequence from the motion sequence, leading it to rely more on the additional information provided by the condition signal. Additionally, we found that both downsampling and random masking strategies are beneficial for improving the multimodality metrics in the text-to-motion task. This implies that the model can better absorb knowledge from different datasets with the help of these two strategies. These strategies also



(a) Hands on hips and shake head

(b) Sit and lean on to a table

(c) Sit and write

(d) Sit and feel embarrassed

(e) Play guitar *(with music condition)*

(f) Clap hands *(with music condition)*

Fig. 5: Visualization results of LMM-Large. Figure a)-d) show examples of text-driven motion generation. Figure e) and f) show synthesized motion sequences under both textual and musical constraints.

significantly impact the effectiveness of motion prediction. Finally, we compared our proposed ArtAttention with the original SAMI and found that our proposed method is more suitable for the scenario of large motion models.

5.4 Qualitative Results

As shown in Fig. 5 (a)-(d), LMM-Large can response to diverse textual descriptions with fine-grained control, which benefits from the large-scale training data and the well-designed architecture. In addition, Fig. 5 (e)-(f) provide examples for motion generation under both text description and music rhythms. Our generated motions successfully execute the given commands and follow the music beats simultaneously. For more visualization results, please kindly refer to the demo video in our homepage.

5.5 More Applications

In Figure 6, we show two videos that are generated based on our synthesized motion sequences. As a vital application direction, users can leverage our large motion model to customize their desired motion data by providing personalized condition signals, such as text commands or accompanying music. With the assistance of off-the-shelf motion-guided video generation technology, users can freely create videos for their favorite characters.

6 Conclusion and Discussion

In this paper, we establish a comprehensive motion-centric benchmark, MotionVerse, comprising then conditional motion generation and motion completion

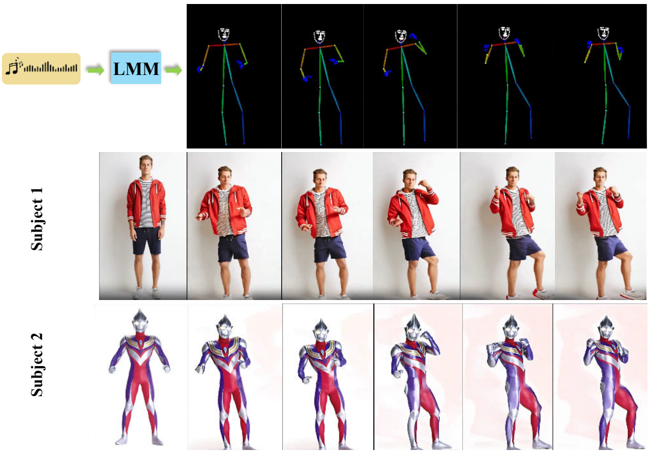


Fig. 6: Video Generation with our synthesized motion sequence. After generating a sequence of motions conditioned on music by our LMM-Large, we map the 3D keypoints to a 2D plane, serving as guidance for video generation.

tasks. We align all motion data to a unified intermediate format and convert all condition signals into token sequences that are closer in feature space. Building upon this foundation, we introduce the first large motion model, LMM, capable of generating high-quality actions under multi-condition guidance. We identify and address three challenges encountered in constructing large motion models through careful model structure design, especially our used novel attention module, ArtAttention. Our proposed LMM model achieves comparable performance and even surpasses existing state-of-the-art methods.

Limitation. The intermediate representation we propose can only address scenarios where entire body parts are missing, but it struggles to effectively handle cases where individual keypoints within a body part are missing. Our method of using motion translators introduces additional noise in downstream tasks, leading to a decrease in motion quality. A more flexible approach to motion representation and modeling needs to be explored and researched. Additionally, due to practical limitations in memory, our model needs to employ zero-shot methods for long-sequence motion generation, which may pose challenges for users in practical applications.

Boarder Impact. The ability to generate natural human motion under flexible condition signals can highly enhance productivity. However, it may also be misused for malicious activities such as creating deceptive deepfake videos or generating realistic-looking but false evidence in legal cases.

Acknowledegment. This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOET2EP20221- 0012), NTU NAP, and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

1. Ahn, H., Mascaro, Valls Esteve an Lee, D.: Can we use diffusion probabilistic models for 3d motion prediction? In: 2023 IEEE International Conference on Robotics and Automation (ICRA) (May 2023) 4
2. Ahn, H., Mascaro, E.V., Lee, D.: Can we use diffusion probabilistic models for 3d motion prediction? arXiv preprint arXiv:2302.14503 (2023) 4
3. Ahuja, C., Morency, L.P.: Language2pose: Natural language grounded pose forecasting. In: 2019 International Conference on 3D Vision (3DV). pp. 719–728. IEEE (2019) 4
4. Ao, T., Zhang, Z., Liu, L.: Gesturediffuclip: Gesture diffusion model with clip latents. ACM Trans. Graph. 4, 32
5. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: Teach: Temporal action composition for 3d humans. In: 2022 International Conference on 3D Vision (3DV). pp. 414–423. IEEE (2022) 4
6. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: SINC: spatial composition of 3d human motions for simultaneous action generation. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 9950–9961 (2023) 4
7. Azadi, S., Shah, A., Hayes, T., Parikh, D., Gupta, S.: Make-an-animation: Large-scale text-conditional 3d human motion generation. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 14993–15002 (2023) 4
8. Barquero, G., Escalera, S., Palmero, C.: Belfusion: Latent diffusion for behavior-driven human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2317–2327 (2023) 4
9. Barsoum, E., Kender, J., Liu, Z.: Hp-gan: Probabilistic 3d human motion prediction via gan. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 1418–1427 (2018) 4
10. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023) 5
11. Cai, Z., Jiang, J., Qing, Z., Guo, X., Zhang, M., Lin, Z., Mei, H., Wei, C., Wang, R., Yin, W., et al.: Digital life project: Autonomous 3d characters with social intelligence. arXiv preprint arXiv:2312.04547 (2023) 5
12. Castillo, A., Escobar, M., Jeanneret, G., Pumarola, A., Arbeláez, P., Thabet, A., Sanakoyeu, A.: Bodiffusion: Diffusing sparse observations for full-body human motion synthesis (2023) 4
13. Cervantes, P., Sekikawa, Y., Sato, I., Shinoda, K.: Implicit neural representations for variable length human motion generation. In: European Conference on Computer Vision. pp. 356–372. Springer (2022) 4, 32
14. Chen, L.H., Zhang, J., Li, Y., Pang, Y., Xia, X., Liu, T.: Humanmac: Masked motion completion for human motion prediction (2023) 4
15. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18000–18010 (2023) 4
16. Chopin, B., Tang, H., Daoudi, M.: Bipartite graph diffusion model for human interaction generation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5333–5342 (2024) 5

17. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. The Eleventh International Conference on Learning Representations (2022) 5
18. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9760–9770 (2023) 4, 5
19. Diller, C., Dai, A.: Cg-hoi: Contact-guided 3d human-object interaction generation. arXiv preprint arXiv:2311.16097 (2023) 5
20. Diller, C., Funkhouser, T., Dai, A.: Forecasting characteristic 3d poses of human actions (2022) 4
21. Du, Y., Kips, R., Pumarola, A., Starke, S., Thabet, A., Sanakoyeu, A.: Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 481–490 (2023) 4
22. Gao, X., Hu, L., Zhang, P., Zhang, B., Bo, L.: Dancemeld: Unraveling dance phrases with hierarchical latent codes for music-to-dance synthesis. arXiv preprint arXiv:2401.10242 (2023) 4
23. Ghorbani, S., Ferstl, Y., Holden, D., Troje, N.F., Carbonneau, M.A.: Zeroeggs: Zero-shot example-based gesture generation from speech. In: Computer Graphics Forum. vol. 42, pp. 206–216. Wiley Online Library (2023) 4
24. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1396–1406 (2021) 4
25. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Text-based motion synthesis with a hierarchical two-stream rnn. In: ACM SIGGRAPH 2021 Posters, pp. 1–2 (2021) 4
26. Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., Slusallek, P.: Remos: Reactive 3d motion synthesis for two-person interactions. arXiv preprint arXiv:2311.17057 (2023) 5
27. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3497–3506 (2019) 32
28. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15180–15190 (June 2023) 7, 8, 10
29. Goel, P., Wang, K.C., Liu, C.K., Fatahalian, K.: Iterative motion editing with natural language. arXiv preprint arXiv:2312.11538 (2023) 4
30. Gong, K., Lian, D., Chang, H., Guo, C., Jiang, Z., Zuo, X., Mi, M.B., Wang, X.: Tm2d: Bimodality driven 3d dance generation via music-text integration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9942–9952 (2023) 2, 4, 13
31. Gopalakrishnan, A., Mali, A., Kifer, D., Giles, L., Ororbia, A.G.: A neural temporal model for human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12116–12125 (2019) 4
32. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. arXiv preprint arXiv:2312.00063 (2023) 4, 12, 32

33. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (2022) 2, 6, 12, 32
34. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision. pp. 580–597. Springer (2022) 4
35. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020) 2, 4, 6, 32
36. Guo, W., Du, Y., Shen, X., Lepetit, V., Alameda-Pineda, X., Moreno-Noguer, F.: Back to mlp: A simple baseline for human motion prediction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4809–4819 (2023) 4, 12, 33
37. Han, B., Peng, H., Dong, M., Xu, C., Ren, Y., Shen, Y., Li, Y.: Amd autoregressive motion diffusion. arXiv preprint arXiv:2305.09381 (2023) 4
38. Hao, Y., Zhang, J., Zhuo, T., Wen, F., Fan, H.: Hand-centric motion refinement for 3d hand-object interaction via hierarchical spatial-temporal modeling. arXiv preprint arXiv:2401.15987 (2024) 5
39. He, X., Huang, S., Zhan, X., Wen, C., Shan, Y.: Semanticboost: Elevating motion generation with augmented textual cues. arXiv preprint arXiv:2310.20323 (2023) 4
40. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022) 5
41. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020) 29
42. Hoang, N.M., Gong, K., Guo, C., Mi, M.B.: Motionmix: Weakly-supervised diffusion for controllable motion generation. arXiv preprint arXiv:2401.11115 (2024) 4
43. Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., Liu, Z.: Avatarclip: zero-shot text-driven generation and animation of 3d avatars. ACM Transactions on Graphics (TOG) **41**(4), 1–19 (2022) 4
44. Hu, V.T., Yin, W., Ma, P., Chen, Y., Fernando, B., Asano, Y.M., Gavves, E., Mettes, P., Ommer, B., Snoek, C.G.: Motion flow matching for human motion synthesis and editing. arXiv preprint arXiv:2312.08895 (2023) 4
45. Huang, R., Hu, H., Wu, W., Sawada, K., Zhang, M., Jiang, D.: Dance revolution: Long-term dance generation with music via curriculum learning. arXiv preprint arXiv:2006.06119 (2020) 4, 13
46. Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16750–16761 (2023) 5
47. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence **36**(7), 1325–1339 (2013) 6
48. Ji, Y., Xu, F., Yang, Y., Shen, F., Shen, H.T., Zheng, W.S.: A large-scale rgb-d database for arbitrary-view human action recognition. In: Proceedings of the 26th ACM international Conference on Multimedia. pp. 1510–1518 (2018) 6

49. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems **36** (2024) 4
50. Jiang, C., Cornman, A., Park, C., Sapp, B., Zhou, Y., Anguelov, D., et al.: Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9644–9653 (2023) 4
51. Jin, P., Wu, Y., Fan, Y., Sun, Z., Wei, Y., Yuan, L.: Act as you wish: Fine-grained control of motion diffusion model with hierarchical semantic graphs. In: NeurIPS (2023) 4
52. Jing, B., Zhang, Y., Song, Z., Yu, J., Yang, W.: Amd: Anatomical motion diffusion with interpretable motion decomposition and fusion. arXiv preprint arXiv:2312.12763 (2023) 4
53. Kalakonda, S.S., Maheshwari, S., Sarvadevabhatla, R.K.: Action-gpt: Leveraging large-scale language models for improved and generalized action generation. In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 31–36. IEEE (2023) 4
54. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2151–2162 (2023) 4
55. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Conference on Computer Vision and Pattern Recognition 2023 (2023) 5
56. Kim, G., Shim, H., Kim, H., Choi, Y., Kim, J., Yang, E.: Diffusion video autoencoders: Toward temporally consistent face video editing via disentangled video encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6091–6100 (2023) 5
57. Kim, J., Kim, J., Choi, S.: Flame: Free-form language-based motion synthesis & editing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 8255–8263 (2023) 4
58. Kong, H., Gong, K., Lian, D., Mi, M.B., Wang, X.: Priority-centric human motion generation in discrete latent space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14806–14816 (2023) 4
59. Kucherenko, T., Hasegawa, D., Henter, G.E., Kaneko, N., Kjellström, H.: Analyzing input and output representations for speech-driven gesture generation. In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents. pp. 97–104 (2019) 4
60. Kucherenko, T., Hasegawa, D., Kaneko, N., Henter, G.E., Kjellström, H.: Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. International Journal of Human–Computer Interaction (2021). https://doi.org/10.1080/10447318.2021.1883883 6
61. Kulal, S., Mao, J., Aiken, A., Wu, J.: Programmatic concept learning for human motion description and synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13843–13852 (2022) 4
62. Kulkarni, N., Rempe, D., Genova, K., Kundu, A., Johnson, J., Fouhey, D., Guibas, L.: Nifty: Neural object interaction fields for guided human motion synthesis. arXiv preprint arXiv:2307.07511 (2023) 5
63. Lee, H.Y., Yang, X., Liu, M.Y., Wang, T.C., Lu, Y.D., Yang, M.H., Kautz, J.: Dancing to music. Advances in Neural Information Processing Systems **32** (2019) 4

64. Li, B., Zhao, Y., Shi, Z., Sheng, L.: Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In: AAAI (2022) 4
65. Li, J., Clegg, A., Mottaghi, R., Wu, J., Puig, X., Liu, C.K.: Controllable human-object interaction synthesis. arXiv preprint arXiv:2312.03913 (2023) 5
66. Li, J., Wu, J., Liu, C.K.: Object motion guided human motion synthesis. ACM Transactions on Graphics (TOG) **42**(6), 1–11 (2023) 5
67. Li, J., Yin, Y., Chu, H., Zhou, Y., Wang, T., Fidler, S., Li, H.: Learning to generate diverse dance motions with transformer. arXiv preprint arXiv:2008.08171 (2020) 4
68. Li, J., Kang, D., Pei, W., Zhe, X., Zhang, Y., He, Z., Bao, L.: Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11293–11302 (2021) 32
69. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13401–13412 (2021) 4, 6, 12
70. Li, S., Zhuang, S., Song, W., Zhang, X., Chen, H., Hao, A.: Sequential texts driven cohesive motions synthesis with natural transitions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9498–9508 (2023) 4
71. Li, S., Singh, H., Grover, A.: Instructany2pix: Flexible visual editing via multimodal instruction following. arXiv preprint arXiv:2312.06738 (2023) 5
72. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. **36**(6), 194–1 (2017) 8
73. Li, W., Xu, X., Liu, J., Xiao, X.: Unimo-g: Unified image generation through multimodal conditional diffusion. arXiv preprint arXiv:2401.13388 (2024) 5
74. Liang, H., Zhang, W., Li, W., Yu, J., Xu, L.: Intergen: Diffusion-based multi-human motion generation under complex interactions. arXiv preprint arXiv:2304.05684 (2023) 5
75. Liang, Z., Li, Z., Zhou, S., Li, C., Loy, C.C.: Control color: Multimodal diffusion-based interactive image colorization. arXiv preprint arXiv:2402.10855 (2024) 5
76. Lim, D., Jeong, C., Kim, Y.M.: Mammos: Mapping multiple human motion with scene understanding and natural interactions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4278–4287 (2023) 5
77. Lin, A.S., Wu, L., Corona, R., Tai, K., Huang, Q., Mooney, R.J.: Generating animated videos of human activities from natural language descriptions. In: NeurIPS Workshop (2018) 4
78. Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motion-x: A large-scale 3d expressive whole-body human motion dataset. Advances in Neural Information Processing Systems (2023) 6
79. Lin, J., Chang, J., Liu, L., Li, G., Lin, L., Tian, Q., Chen, C.w.: Ohmg: Zero-shot open-vocabulary human motion generation. arXiv preprint arXiv:2210.15929 (2022) 4
80. Lin, J., Chang, J., Liu, L., Li, G., Lin, L., Tian, Q., Chen, C.w.: Being comes from not-being: Open-vocabulary text-to-motion generation with wordless training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23222–23231 (2023) 4
81. Lin, P., Xu, S., Yang, H., Liu, Y., Chen, X., Wang, J., Yu, J., Xu, L.: Handdiffuse: Generative controllers for two-hand interactions via diffusion models. arXiv preprint arXiv:2312.04867 (2023) 5

82. Ling, Z., Han, B., Wong, Y., Kangkanhalli, M., Geng, W.: Mcm: Multi-condition motion synthesis framework for multi-scenario. arXiv preprint arXiv:2309.03031 (2023) 4, 5
83. Liu, C., Zhao, M., Ren, B., Liu, M., Sebe, N., et al.: Spatio-temporal graph diffusion for text-driven human motion generation. In: British Machine Vision Conference (2023) 4
84. Liu, H., Zhu, Z., Iwamoto, N., Peng, Y., Li, Z., Zhou, Y., Bozkurt, E., Zheng, B.: Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In: European Conference on Computer Vision. pp. 612–630. Springer (2022) 6, 32
85. Liu, J., Dai, W., Wang, C., Cheng, Y., Tang, Y., Tong, X.: Plan, posture and go: Towards open-world text-to-motion generation. arXiv preprint arXiv:2312.14828 (2023) 4
86. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2684–2701 (2019) 6
87. Liu, X., Wu, Q., Zhou, H., Xu, Y., Qian, R., Lin, X., Zhou, X., Wu, W., Dai, B., Zhou, B.: Learning hierarchical cross-modal association for co-speech gesture generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10462–10472 (2022) 4, 6
88. Liu, X., Chen, G., Tang, Y., Wang, G., Lim, S.N.: Language-free compositional action generation via decoupling refinement. arXiv preprint arXiv:2307.03538 (2023) 4
89. Liu, X., Hou, H., Yang, Y., Li, Y.L., Lu, C.: Revisit human-scene interaction via space occupancy. arXiv preprint arXiv:2312.02700 (2023) 5
90. Liu, Y., Chen, C., Yi, L.: Interactive humanoid: Online full-body motion reaction synthesis with social affordance canonicalization and forecasting. arXiv preprint arXiv:2312.08983 (2023) 5
91. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015) 5
92. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015) 6, 8
93. Lou, Y., Zhu, L., Wang, Y., Wang, X., Yang, Y.: Diversemotion: Towards diverse human motion generation via discrete diffusion. arXiv preprint arXiv:2309.01372 (2023) 4
94. Lu, S., Chen, L.H., Zeng, A., Lin, J., Zhang, R., Zhang, L., Shum, H.Y.: Humantomato: Text-aligned whole-body motion generation. arXiv preprint arXiv:2310.12978 (2023) 3, 4, 7, 8
95. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision. pp. 5442–5451 (Oct 2019) 6, 12
96. Mao, W., Liu, M., Salzmann, M.: History repeats itself: Human motion prediction via motion attention. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. pp. 474–489. Springer (2020) 4
97. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9489–9497 (2019) 12

98. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: European Conference on Computer Vision (ECCV) (sep 2018) 6, 12
99. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 3D Vision (3DV), 2017 Fifth International Conference on. IEEE (2017). https://doi.org/10.1109/3dv.2017.00064, http://gvv.mpi-inf.mpg.de/3dhp_dataset 6
100. Nguyen, T., Li, Y., Ojha, U., Lee, Y.J.: Visual instruction inversion: Image editing via visual prompting. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=l9BsCh8ikK 5
101. Okamura, M., Kondo, N., Sakamoto, T.F.M., Ochiai, Y.: Dance generation by sound symbolic words. arXiv preprint arXiv:2306.03646 (2023) 4
102. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019) 8
103. Peng, X., Xie, Y., Wu, Z., Jampani, V., Sun, D., Jiang, H.: Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. arXiv preprint arXiv:2312.06553 (2023) 5
104. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10985–10995 (2021) 2, 4, 32
105. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision. pp. 480–497. Springer (2022) 4
106. Petrovich, M., Litany, O., Iqbal, U., Black, M.J., Varol, G., Peng, X.B., Rempe, D.: Multi-track timeline control for text-driven 3d human motion generation. arXiv preprint arXiv:2401.08559 (2024) 4
107. Pi, H., Peng, S., Yang, M., Zhou, X., Bao, H.: Hierarchical generation of human-object interactions with diffusion probabilistic models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15061–15073 (2023) 5
108. Pinyoanuntapong, E., Wang, P., Lee, M., Chen, C.: Mmm: Generative masked motion model. arXiv preprint arXiv:2312.03596 (2023) 4
109. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data **4**(4), 236–252 (2016) 6
110. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: Babel: Bodies, action and behavior with english labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 722–731 (2021) 6
111. Qi, Q., Zhuo, L., Zhang, A., Liao, Y., Fang, F., Liu, S., Yan, S.: Diffdance: Cascaded human motion diffusion model for dance generation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1374–1382 (2023) 4
112. Qian, Y., Urbanek, J., Hauptmann, A.G., Won, J.: Breaking the limits of text-conditioned 3d motion synthesis with elaborative descriptions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2306–2316 (2023) 4
113. Qing, Z., Cai, Z., Yang, Z., Yang, L.: Story-to-motion: Synthesizing infinite and controllable character animation from long text. In: SIGGRAPH Asia 2023 Tech-

nical Communications, SA Technical Communications 2023, Sydney, NSW, Australia, December 12-15, 2023. pp. 28:1–28:4 (2023) 4

114. Raab, S., Leibovitch, I., Tevet, G., Arar, M., Bermano, A.H., Cohen-Or, D.: Single motion diffusion. arXiv preprint arXiv:2302.05905 (2023) 4
115. Ren, J., Zhang, M., Yu, C., Ma, X., Pan, L., Liu, Z.: Insactor: Instruction-driven physics-based characters. Advances in Neural Information Processing Systems **36** (2024) 4
116. Ribeiro-Gomes, J., Cai, T., Milacski, Z.A., Wu, C., Prakash, A., Takagi, S., Aubel, A., Kim, D., Bernardino, A., De La Torre, F.: Motiongpt: Human motion synthesis with improved diversity and realism via gpt-3 prompting. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5070–5080 (2024) 4
117. Ruan, L., Ma, Y., Yang, H., He, H., Liu, B., Fu, J., Yuan, N.J., Jin, Q., Guo, B.: Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10219–10228 (2023) 5
118. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418 (2023) 4
119. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1010–1019 (2016) 6
120. Shi, X., Luo, C., Peng, J., Zhang, H., Sun, Y.: Generating fine-grained human motions using chatgpt-refined descriptions. arXiv preprint arXiv:2312.02772 (2023) 4
121. Shimada, S., Mueller, F., Bednarik, J., Doosti, B., Bickel, B., Tang, D., Golyanik, V., Taylor, J., Theobalt, C., Beeler, T.: Macs: Mass conditioned 3d hand and object motion synthesis. arXiv preprint arXiv:2312.14929 (2023) 5
122. Siyao, L., Gu, T., Yang, Z., Lin, Z., Liu, Z., Ding, H., Yang, L., Loy, C.C.: Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment. In: The Twelfth International Conference on Learning Representations (2023) 5
123. Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C.C., Liu, Z.: Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11050–11059 (2022) 2, 4, 13
124. Sun, G., Wong, Y., Cheng, Z., Kankanhalli, M.S., Geng, W., Li, X.: Deepdance: music-to-dance motion choreography with adversarial learning. IEEE Transactions on Multimedia **23**, 497–509 (2020) 4
125. Sun, J., Lin, Z., Han, X., Hu, J.F., Xu, J., Zheng, W.S.: Action-guided 3d human motion prediction. Advances in Neural Information Processing Systems **34**, 30169–30180 (2021) 4
126. Sun, J., Chowdhary, G.: Towards globally consistent stochastic human motion prediction via motion diffusion. arXiv preprint arXiv:2305.12554 (2023) 4
127. Tanaka, M., Fujiwara, K.: Role-aware interaction generation from textual description. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15999–16009 (2023) 5
128. Tendulkar, P., Surís, D., Vondrick, C.: Flex: Full-body grasping without full-body grasps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21179–21189 (2023) 5

129. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: European Conference on Computer Vision. pp. 358–374. Springer (2022) 4
130. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2022) 4, 5, 8, 12, 32
131. Tseng, J., Castellon, R., Liu, K.: Edge: Editable dance generation from music. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 448–458 (2023) 4
132. Voas, J., Wang, Y., Huang, Q., Mooney, R.: What is the best automated metric for text to motion generation? In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–11 (2023) 4
133. Voleti, V., Jolicoeur-Martineau, A., Pal, C.: Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. Advances in Neural Information Processing Systems **35**, 23371–23385 (2022) 5
134. Wan, W., Dou, Z., Komura, T., Wang, W., Jayaraman, D., Liu, L.: Tlcontrol: Trajectory and language control for human motion synthesis. arXiv preprint arXiv:2311.17135 (2023) 4
135. Wang, X., Cui, Q., Chen, C., Liu, M.: Gcnext: Towards the unity of graph convolutions for human motion prediction. arXiv preprint arXiv:2312.11850 (2023) 4, 12, 33
136. Wang, Y., Leng, Z., Li, F.W., Wu, S.C., Liang, X.: Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22035–22044 (2023) 4
137. Wang, Y., Lin, J., Zeng, A., Luo, Z., Zhang, J., Zhang, L.: Physhoi: Physics-based imitation of dynamic human-object interaction. arXiv preprint arXiv:2312.04393 (2023) 5
138. Wang, Z., Yu, P., Zhao, Y., Zhang, R., Zhou, Y., Yuan, J., Chen, C.: Learning diverse stochastic human-action generators by learning smooth latent transitions. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12281–12288 (2020) 4
139. Wei, D., Sun, X., Sun, H., Li, B., Hu, S., Li, W., Lu, J.: Enhanced fine-grained motion diffusion for text-driven human motion synthesis (2023) 4
140. Xiao, Z., Wang, T., Wang, J., Cao, J., Zhang, W., Dai, B., Lin, D., Pang, J.: Unified human-scene interaction via prompted chain-of-contacts. arXiv preprint arXiv:2309.07918 (2023) 5
141. Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation. arXiv preprint arXiv:2310.08580 (2023) 4
142. Xie, Z., Wu, Y., Gao, X., Sun, Z., Yang, W., Liang, X.: Towards detailed text-to-motion synthesis via basic-to-advanced hierarchical diffusion model. arXiv preprint arXiv:2312.10960 (2023) 4
143. Xu, Z., Zhang, Y., Yang, S., Li, R., Li, X.: Chain of generation: Multi-modal gesture synthesis via cascaded conditional control. arXiv preprint arXiv:2312.15900 (2023) 4, 32
144. Yan, H., Hu, Z., Schmitt, S., Bulling, A.: Gazemodiff: Gaze-guided diffusion model for stochastic human motion prediction. arXiv preprint arXiv:2312.12090 (2023) 4

145. Yang, S., Zhou, Y., Liu, Z., , Loy, C.C.: Rerender a video: Zero-shot text-guided video-to-video translation. In: ACM SIGGRAPH Asia 2023 Conference Proceedings (2023) 5
146. Yang, S., Yang, Z., Wang, Z.: Longdancediff: Long-term dance generation with conditional diffusion model. arXiv preprint arXiv:2308.11945 (2023) 4
147. Yang, Z., Su, B., Wen, J.R.: Synthesizing long-term human motions with diffusion models via coherent sampling. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 3954–3964 (2023) 4
148. Yao, H., Song, Z., Zhou, Y., Ao, T., Chen, B., Liu, L.: Moconvq: Unified physics-based motion control via scalable discrete representations. arXiv preprint arXiv:2310.10198 (2023) 4
149. Yao, S., Sun, M., Li, B., Yang, F., Wang, J., Zhang, R.: Dance with you: The diversity controllable dancer generation via diffusion models. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 8504–8514 (2023) 4
150. Yazdian, P.J., Liu, E., Cheng, L., Lim, A.: Motionscript: Natural language descriptions for expressive 3d human motions. arXiv preprint arXiv:2312.12634 (2023) 4
151. Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: Generating holistic 3d human motion from speech. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 469–480 (2023) 32
152. Yin, L., Wang, Y., He, T., Liu, J., Zhao, W., Li, B., Jin, X., Lin, J.: Emog: Synthesizing emotive co-speech 3d gesture with diffusion model. arXiv preprint arXiv:2306.11496 (2023) 4
153. Yoon, Y., Cha, B., Lee, J.H., Jang, M., Lee, J., Kim, J., Lee, G.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics (TOG) **39**(6), 1–16 (2020) 4, 6, 32
154. Yoon, Y., Ko, W.R., Jang, M., Lee, J., Kim, J., Lee, G.: Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 4303–4309. IEEE (2019) 32
155. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16010–16021 (2023) 4
156. Zhai, Y., Huang, M., Luan, T., Dong, L., Nwogu, I., Lyu, S., Doermann, D., Yuan, J.: Language-guided human motion synthesis with atomic actions. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5262–5271 (2023) 4
157. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. arXiv preprint arXiv:2301.06052 (2023) 4, 12, 32
158. Zhang, J., Huang, S., Tu, Z., Chen, X., Zhan, X., Yu, G., Shan, Y.: Tapmo: Shape-aware motion generation of skeleton-free characters. arXiv preprint arXiv:2310.12678 (2023) 4
159. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024) 2, 4, 5, 8, 32
160. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 364–373 (2023) 4, 32

161. Zhang, M., Li, H., Cai, Z., Ren, J., Yang, L., Liu, Z.: Finemogen: Fine-grained spatio-temporal motion generation and editing. Advances in Neural Information Processing Systems **36** (2024) 4, 5, 8, 10, 12, 13, 32
162. Zhang, X., Bhatnagar, B.L., Starke, S., Guzov, V., Pons-Moll, G.: Couch: Towards controllable human-chair interactions. In: European Conference on Computer Vision. pp. 518–535. Springer (2022) 5
163. Zhang, Y., Huang, D., Liu, B., Tang, S., Lu, Y., Chen, L., Bai, L., Chu, Q., Yu, N., Ouyang, W.: Motiongpt: Finetuned llms are general-purpose motion generators. arXiv preprint arXiv:2306.10900 (2023) 4
164. Zhang, Y., Tsipidi, E., Schriber, S., Kapadia, M., Gross, M., Modi, A.: Generating animations from screenplays. In: Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019. pp. 292–307 (2019) 4
165. Zhao, M., Liu, M., Ren, B., Dai, S., Sebe, N.: Modiff: Action-conditioned 3d motion generation with denoising diffusion probabilistic models. arXiv preprint arXiv:2301.03949 (2023) 4
166. Zhao, W., Hu, L., Zhang, S.: Diffugesture: Generating human gesture from two-person dialogue with diffusion models. In: Companion Publication of the 25th International Conference on Multimodal Interaction. pp. 179–185 (2023) 5
167. Zhi, Y., Cun, X., Chen, X., Shen, X., Guo, W., Huang, S., Gao, S.: Livelyspeaker: Towards semantic-aware co-speech gesture generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20807–20817 (2023) 4
168. Zhong, C., Hu, L., Zhang, Z., Xia, S.: Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 509–519 (2023) 4
169. Zhou, W., Dou, Z., Cao, Z., Liao, Z., Wang, J., Wang, W., Liu, Y., Komura, T., Wang, W., Liu, L.: Emdm: Efficient motion diffusion model for fast, high-quality motion generation. arXiv preprint arXiv:2312.02256 (2023) 4, 5
170. Zhou, Z., Wang, B.: Ude: A unified driving engine for human motion generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5632–5641 (2023) 4, 5
171. Zhuang, W., Wang, C., Chai, J., Wang, Y., Shao, M., Xia, S.: Music2dance: Dancenet for music-driven dance generation. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **18**(2), 1–21 (2022) 4, 13

Appendix A MotionVerse

In this section, we offer additional details about the construction of **MotionVerse** benchmark.

A.1 Dataset Preprocess

Based on the characteristics of each dataset, we employ different preprocessing methods. To avoid overlap between the training and test sets of different datasets, we **excluded** sequences from the training set of each benchmark that intersect with any test set motion sequences. Below, we provide detailed explanations of the processing methods for each dataset. After aligning each dataset to SMPL-X joints and processing them into TOMATO format using the same set of scripts, they are decomposed into 10 body parts. Therefore, our main focus will be on how each dataset is aligned to SMPL-X 3D joints and which body parts are included in each dataset.

HumanML3D. The HumanML3D dataset is annotated from two sources: AMASS data and HumanAct12. The former provides native SMPL-X annotations, while the latter offers 22 keypoint annotations based on the SMPL format. Additionally, the MotionX dataset provides facial motion data corresponding to each action sequence in HumanML3D. Therefore, there are two overall annotation formats. For data from AMASS, it includes all 10 body parts and does not require keypoint mapping. We use SMPL-X model to convert the SMPL-X beta parameters and theta parameters into 3D coordinates. For data from HumanAct12, it includes 7 body parts (excluding face, left hand, and right hand).

KIT-ML. We located the AMASS data corresponding to KIT-ML and utilized this portion of the AMASS data to generate the motion sequences for KIT-ML. Since there is no additional face motion data available, it comprises a total of 9 body parts.

Motion-X. Motion-X provides key points based on the SMPL-X format and facial expressions based on FLAME. Here, we don't need to perform additional key point conversion, and it includes all 10 body parts.

BABEL. BABEL is also annotated based on AMASS. Since there is no face motion available, we only consider its 9 body parts.

UESTC. For the UESTC dataset, we follow the processing method used in ACTOR. We use the SMPL parameters estimated from VIBE as the raw data. We use default betas parameters to obtain the 3D coordinates of each joint. Since we do not consider global orientation and global translation during evaluation, and due to the significant noise in the estimation from VIBE, we do not consider four body parts: left hand, right hand, face expression, and global configuration.

HumanAct12. For the HumanAct12 dataset, we employ the pre-processing method used in HumanML3D. Here, HumanAct12 does not include three body parts: left hand, right hand, and facial expression.

NTU-RGB-D 120. For the NTU-RGBD 120 dataset, it comes with native 3D keypoint annotations, but due to their poor accuracy, we only consider the inherent motion captured by these keypoints. Regarding the spine, we use an interpolation method to map the keypoint data from NTU-RGBD 120 to the SMPLX format for the spine. Finally, we only consider four body parts: spine, left hand, right hand, and head.

AMASS. AMASS provides annotations based on the SMPL-X format. We use the provided beta and theta parameters to obtain the corresponding 3D keypoints. Here, we do not consider the body part of facial expression.

3DPW. 3DPW provides SMPL parameters, allowing us to obtain the 3D keypoint positions in the SMPL format. Since the motion prediction task involved in 3DPW does not consider global translation, we only consider six body parts: spine, left arm, right arm, left leg, right leg, and head.

Human3.6M. Similar to 3DPW, we obtain keypoint sequences using SMPL parameters, which are ultimately converted into six body parts: spine, left arm, right arm, left leg, right leg, and head.

TED-Gesture++. TED-Gesture++ only provides keypoints for the upper body, so we consider only the spine, left arm, right arm, and head as the five body parts. For the spine, we utilize interpolation to obtain a keypoint set that conforms to SMPL-X.

TED-Expressive. The keypoint annotations of TED-Expressive++ are almost identical to SMPL-X. We directly selected the corresponding keypoints and removed the redundant parts. It includes all body parts except for facial expressions.

Speech2Gesture-3D. Similar to TED-Expressive, we directly selected the corresponding keypoints and removed the redundant parts. It includes all body parts except for facial expressions.

BEAT. The keypoint set of BEAT completely covers the keypoints of SMPL-X and provides facial expression, so we consider all body parts and discard the keypoints that do not exist in SMPL-X.

AIST++. AIST++ provides annotations based on SMPL parameters, corresponding to 7 body parts excluding face expression, left hand, and right hand.

MPI-INF-3DHP. Similar to 3DPW, we obtain keypoint sequences using SMPL parameters, which are ultimately converted into six body parts: spine, left arm, right arm, left leg, right leg, and head.

A.2 Motion Translator

During evaluation, there are three types of estimation. The first type is based on H3D vectors, primarily used in the Text2Motion datasets. For this type, we train an MLP to convert our frame representations into the corresponding representations required for evaluation. The second type is based on keypoint sequences, without considering global translation and global orientation, such as in the UESTC evaluation. Here, we also directly train an MLP for mapping. The third type considers global translation and global orientation, and is based on keypoint sequence evaluation. In this case, we first convert our representations into the keypoint sequence format and then train an MLP for mapping.

Appendix B Large Motion Model

To facilitate a deeper understanding of the LMM approach, this chapter provides additional technical details.

B.1 Diffusion Model

This paper utilizes the Denoising Diffusion Probabilistic Model (DDPM) [41], a probability generative model based on the Markov chain. Its essence lies in two intertwined processes: the forward diffusion process and the reverse diffusion process.

The forward diffusion process systematically injects noise into the original distribution, progressively disrupting the data's initial distribution. Starting from the original distribution $x_0 \sim q(x_0)$, noise is added over T steps to generate x_1, x_2,..., x_T. This process employs an efficient, tractable noise addition method, with Gaussian perturbation being a classic approach. The specific formula is:

$$q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right), \tag{3}$$

where β controls the amount of noise added. In the context of motion generation tasks, x can be considered a series of poses. To streamline the forward process, the noise-added result at any step can be approximately calculated from x_0: $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_t$.

The reverse diffusion process is the inverse operation of adding noise, aiming to restore the original distribution from a noisy distribution. Given the difficulty and critical nature of this process, we employ deep learning models to learn the denoising process, defined as: $q(x_{t-1} \mid x_t) = M(\mathbf{x}, \mathrm{m}, \mathrm{c})$. During the model training phase, the supervisory objective is to minimize the difference between the predicted distribution $\hat{x}_0$ and the ground truth x_0.

B.2 ArtAttention

Following a similar approach to FineMoGen, we incorporate the temporal aspect to account for the influence of motion sequences and other condition signals across different time intervals. Specifically, we introduce the notion of time explicitly into this process. Formally, we present the following approximation for refining temporal features:

$$\mathbf{Y}_{k,i} \approx \mu_i(x_k) = \sum_{j=1}^{N_g} \mathbf{G}'_{i,j}(x_k) \cdot \mathbf{G}^*_{i,j}(x_k) \tag{4}$$

where x_k represents the time position of kth element in the motion sequence. $\mathbf{G}'_{i,j}(x)$ indicates the time-varied signal we derive from the feature vector $\mathbf{G}_{i,j}$

and $\mathbf{G}^*_{i,j}(x_k)$ denotes the relative significance of this template for the kth position. $\mathbf{G}_{i,j}$ is the j-th global template in the i-th attention head. We construct $\mathbf{G}^*_{i,j}(x_k)$ as:

$$\mathbf{G}^*_{i,j}(x_k) = \frac{e^{-(x_k - \mathbf{G}^t_{i,j})^2/\sigma^2}}{\sum_{l\in[1,N_g]} e^{-(x_k - \mathbf{G}^t_{i,l})^2/\sigma^2}}, \tag{5}$$

In this setup, the jth global template of the ith group is considered as a set of signals propagating outward from the temporal center $\mathbf{G}^t_{i,j}$. As for $\mathbf{G}'_{i,j}(x)$, we consider its Taylor expansion at $\mathbf{G}^t_{i,j}$:

$$\mathbf{G}'_{i,j}(x) = \sum_{n=0}^{k} \frac{\mathbf{G}^{(n)}_{i,j}}{n!}(x - \mathbf{G}^t_{i,j})^n. \tag{6}$$

We use linear projections to process the original $\mathbf{G}_{i,j}$ and acquire all coefficients $\mathbf{G}^t_{i,j}, \mathbf{G}^{(n)}_{i,j}, n \in [0, k]$. We perceive a global template as an anchor with its initial state defined as $\mathbf{G}^{(0)}_{i,j}$, velocity as $\mathbf{G}^{(1)}_{i,j}$, acceleration as $\mathbf{G}^{(2)}_{i,j}$ and so on. Therefore we name this method a kinetic modelling on the latent feature space.

Moreover, to integrate influences from all signals, we adopt the square of the time difference as a metric to assess the significance of each global template. We employ a Softmax operation to standardize their weights. An immediate benefit of this modeling strategy is its flexibility in appending a new stage subsequent to the current one. This can be achieved by adjusting a bias term in $\mathbf{G}^t_{i,j}$ accordingly, facilitating our method to execute zero-shot temporal combination.

B.3 Stylization Block

The primary function of the stylization block is to inject the information of the timestamp t into the features, thereby informing the model about the current step in the reverse process. This enhancement aids in improving the model's denoising capability. The stylization block injects information about frame rate, dataset name, and current timestep into the feature representation. Drawing inspiration from FineMoGen, we convert the timestamp t into a vector $\mathbf{e_t}$. In each stylization block, $\mathbf{e_t}$ undergoes two linear transformations to generate two features $\mathbf{e_w} \in \mathbb{R}^{H\times D}$ and $\mathbf{e_b} \in \mathbb{R}^{H\times D}$. Every pose feature θ inputted into this module is optimized as $\theta' = \theta \cdot \mathbf{e_w} + \mathbf{e_b}$, where $(\cdot)$ denotes Hadamard product.

Appendix C Experiments

C.1 Implementation Details

Batch Formation Overall, we determine the sampling probability of motion sequences in each dataset based on the quality of the dataset and the diversity of the actions.

1. **Text-to-Motion (40%)**: In the task of text-to-motion, there is a wide variety of motion types, mostly consisting of high-quality motion capture data from AMASS, which is beneficial for the model. Therefore, the sampling proportion is set relatively high at 40%. Within this category, HumanML3D provides high-quality and semantically rich motions, accounting for 15%; Motion-X, although lower in quality, offers high diversity in motions, also at 15%; KIT-ML and BABEL each contribute 5%.
2. **Unconditional Motion Generation (25%)**: We primarily focus on the AMASS dataset, where the motion quality is generally high, aiding the model in learning motion priors. Hence, we set a relatively high proportion for this dataset.
3. **Action-to-Motion (10%)**: We uniformly sample sequences from the HumanAct12, UESTC, and NTU-RGBD 120 datasets.
4. **Speech-to-Gesture (10%)**: As BEAT is selected as the test set, we assign it half of the weight. The remaining portion is evenly distributed among TED-Gesture++, TED-Expressive, and Speech2Gesture-3D.
5. **Music-to-Dance (5%)**: For Music2dance, there is only one AIST++ dataset, which accounts for all the weight.
6. **Motion Imitation (10%)**: During training, we exclude the 3DPW dataset and only consider MPI-INF-3DHP and H36M, with both datasets equally sharing the weight.

Model	#Latent Dim	#Layers	#Experts	#Params
LMM-Tiny	64	4	16	90M
LMM-Small	64	8	16	160M
LMM-Base	128	12	16	410M
LMM-Large	128	20	32	760M

Table 7: Model card.

Model Card. Tab. 7 shows the hyperparameter of each variant.
Mask strategy. Considering that larger models have stronger capabilities to fit motions, to enhance the control ability of conditions, we use mask probabilities of 0.1, 0.2, 0.3, and 0.4 for LMM-Tiny, LMM-Small, LMM-Base, and LMM-Large, respectively.

C.2 More Quantitative Results

Text-to-Motion. We observed that compared to its performance on HumanML3D, LMM-Large performs slightly worse on KIT-ML, which could be related to the proportion of the two datasets in batch formation. However, overall, KIT-ML also achieves accuracy comparable to the state-of-the-art, especially achieving a new state-of-the-art in terms of FID.

Table 8: Quantitative results on the KIT-ML test set.

Methods	R Precision↑ Top 1	 Top 2	 Top 3	FID↓	MM Dist↓	Diversity↑	MM↑
Real motions	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$0.031^{\pm.004}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
Guo *et al.* [33]	$0.370^{\pm.005}$	$0.569^{\pm.007}$	$0.693^{\pm.007}$	$2.770^{\pm.109}$	$3.401^{\pm.008}$	$10.91^{\pm.119}$	$1.482^{\pm.065}$
T2M-GPT [157]	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$0.514^{\pm.029}$	$3.007^{\pm.023}$	$10.921^{\pm.108}$	$1.570^{\pm.039}$
MDM [130]	-	-	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$9.191^{\pm.022}$	$10.847^{\pm.109}$	$\mathbf{1.907^{\pm.214}}$
MotionDiffuse [159]	$0.417^{\pm.004}$	$0.621^{\pm.004}$	$0.739^{\pm.004}$	$1.954^{\pm.062}$	$2.958^{\pm.005}$	$\underline{11.10^{\pm.143}}$	$0.730^{\pm.013}$
ReMoDiffuse [160]	$0.427^{\pm.014}$	$0.641^{\pm.004}$	$0.765^{\pm.055}$	$\mathbf{0.155^{\pm.006}}$	$2.814^{\pm.012}$	$10.80^{\pm.105}$	$1.239^{\pm.028}$
FineMoGen [161]	$\underline{0.432^{\pm.006}}$	$0.649^{\pm.005}$	$0.772^{\pm.006}$	$0.178^{\pm.007}$	$2.869^{\pm.014}$	$10.85^{\pm.115}$	$1.877^{\pm.093}$
MoMask [32]	$\mathbf{0.433^{\pm.007}}$	$\mathbf{0.656^{\pm.005}}$	$\mathbf{0.781^{\pm.005}}$	$0.204^{\pm.011}$	$\mathbf{2.779^{\pm.022}}$	-	$1.131^{\pm.043}$
LMM-Tiny	$0.419^{\pm.018}$	$0.627^{\pm.014}$	$0.748^{\pm.019}$	$0.817^{\pm.015}$	$2.904^{\pm.022}$	$10.85^{\pm.087}$	$1.607^{\pm.110}$
LMM-Small	$0.421^{\pm.015}$	$0.634^{\pm.021}$	$0.755^{\pm.017}$	$0.471^{\pm.017}$	$2.851^{\pm.021}$	$10.94^{\pm.101}$	$1.625^{\pm.114}$
LMM-Base	$0.428^{\pm.015}$	$0.648^{\pm.017}$	$0.769^{\pm.017}$	$0.239^{\pm.015}$	$2.810^{\pm.018}$	$11.05^{\pm.097}$	$1.804^{\pm.130}$
LMM-Large	$0.430^{\pm.015}$	$\underline{0.653^{\pm.017}}$	$\underline{0.779^{\pm.014}}$	$\underline{0.137^{\pm.023}}$	$\underline{2.791^{\pm.018}}$	$\mathbf{11.24^{\pm.103}}$	$\underline{1.885^{\pm.127}}$

Table 9: Quantitative results for Action-conditioned Motion Generation. As for UESTC dataset, we report FID on the test split. MM: MultiModality.

Methods	HumanAct12 FID↓	 Accuracy↑	 Diversity→	 MM→	UESTC FID↓	 Accuracy↑	 Diversity→	 MM→
Real motions	$0.020^{\pm.010}$	$0.997^{\pm.001}$	$6.850^{\pm.050}$	$2.450^{\pm.040}$	$2.79^{\pm.29}$	$0.988^{\pm.001}$	$33.34^{\pm.320}$	$14.16^{\pm.06}$
Action2Motion [35]	$0.338^{\pm.015}$	$0.917^{\pm.003}$	$6.879^{\pm.066}$	$2.511^{\pm.023}$	-	-	-	-
ACTOR [104]	$0.12^{\pm.00}$	$0.955^{\pm.008}$	$6.84^{\pm.03}$	$2.53^{\pm.02}$	$23.43^{\pm 2.20}$	$0.911^{\pm.003}$	$31.96^{\pm.33}$	$14.52^{\pm.09}$
INR [13]	$0.088^{\pm.004}$	$0.973^{\pm.001}$	$6.881^{\pm.048}$	$2.569^{\pm.040}$	$15.00^{\pm.09}$	$0.941^{\pm.001}$	$31.59^{\pm.19}$	$14.68^{\pm.07}$
MotionDiffuse [159]	$\underline{0.07^{\pm.00}}$	$\mathbf{0.992^{\pm.13}}$	$\mathbf{6.85^{\pm.02}}$	$\underline{2.46^{\pm.02}}$	$\underline{9.10^{\pm.437}}$	$\underline{0.950^{\pm.000}}$	$\underline{32.42^{\pm.214}}$	$14.74^{\pm.07}$
LMM-Tiny	$0.105^{\pm.00}$	$0.992^{\pm.008}$	$6.819^{\pm.025}$	$\mathbf{2.457^{\pm.018}}$	$20.16^{\pm 1.78}$	$0.917^{\pm.002}$	$30.80^{\pm.228}$	$\mathbf{14.29^{\pm.066}}$
LMM-Small	$0.094^{\pm.00}$	$0.963^{\pm.008}$	$6.827^{\pm.028}$	$2.498^{\pm.022}$	$14.28^{\pm 1.14}$	$0.922^{\pm.002}$	$31.25^{\pm.231}$	$\underline{14.42^{\pm.067}}$
LMM-Base	$0.087^{\pm.00}$	$\underline{0.985^{\pm.007}}$	$\underline{6.848^{\pm.030}}$	$2.551^{\pm.022}$	$10.36^{\pm 0.60}$	$0.948^{\pm.000}$	$32.39^{\pm.236}$	$14.65^{\pm.065}$
LMM-Large	$\mathbf{0.065^{\pm.00}}$	$\mathbf{0.992^{\pm.008}}$	$6.871^{\pm.031}$	$2.560^{\pm.019}$	$\mathbf{9.01^{\pm 0.54}}$	$\mathbf{0.952^{\pm.000}}$	$\mathbf{32.58^{\pm.254}}$	$14.81^{\pm.064}$

Action-to-Motion. On the action-conditioned motion generation task, each LMM-Large model achieves the best performance in terms of both FID and Accuracy. Additionally, due to exposure to more data, it exhibits higher diversity and multimodality. However, because of the nature of the action-to-motion task, an increase in both aspects does not necessarily indicate better performance.

Table 10: Quantitative results on Speech-to-Gesture on the BEAT dataset.

Methods	FGD↓	SRGR↑	BeatAlign↑
Seq2Seq [154]	261.3	0.173	0.729
Speech2Gesture [27]	256.7	0.092	0.751
MultiContext [153]	176.2	0.195	0.776
Audio2Gesture [68]	223.8	0.097	0.766
CaMN [84]	123.7	0.239	0.783
TalkShow [151]	91.0	-	0.840
GestureDiffuCLIP [4]	85.17	-	-
CoG [143]	**45.87**	**0.308**	**0.931**
LMM-Tiny	92.51	0.142	0.825
LMM-Small	86.94	0.169	0.836
LMM-Base	57.18	0.228	0.879
LMM-Large	$\underline{47.95}$	$\underline{0.277}$	$\underline{0.913}$

Speech-to-Gesture. In MotionVerse, we introduce multiple speech-to-gesture datasets, and overall, LMM-Large performs impressively on the BEAT dataset as well.

Motion Imitation We evaluate our method on the test set of 3DPW, and obtain PA-MPJPE scores of 95.7, 91.2, 76.3, and 71.5 for LMM-Tiny, LMM-Small, LMM-Base, and LMM-Large, respectively. For reference, the PA-MPJPE scores for HMR and VIBE are 81.3 and 51.9, respectively. The performance for video-conditioning is relatively low; we will focus on addressing this issue in future work.

Table 11: Quantitative results of motion prediction on the Human3.6M test set for different time steps (ms). We report the MPJPE error in *mm*.

Method	Human3.6M							
	80	160	320	400	560	720	880	1000
siMLPe [36]	9.6	21.7	46.3	57.3	75.7	90.1	101.8	109.4
GCNext [135]	**9.3**	**21.5**	45.5	56.4	74.7	88.9	100.8	108.7
LMM-Tiny	14.8	28.6	48.3	59.2	79.3	93.6	105.9	112.0
LMM-Small	14.1	27.4	47.2	58.1	78.1	91.5	103.4	110.3
LMM-Base	12.9	25.9	44.9	55.0	74.8	87.6	99.5	107.1
LMM-Large	11.8	23.6	**43.7**	**53.1**	**73.6**	**85.0**	**96.9**	**104.6**

Motion Prediction Similar to the conclusion we found in 3DPW and AMASS dataset, LMM-Large performs worse than the existing work in short-term prediction and better than these work in long-term prediction.

Table 12: Quantitative results of conditional motion completion on the HumanML3D test set. We report the MPJPE error in *mm*. We use LMM-Large for all experiments.

Condition	First 25 frames	Last 25 frames	avg-MPJPE
No	Yes	No	63.8
No	Yes	Yes	59.1
Yes	Yes	No	54.7
Yes	Yes	Yes	51.9

Conditional Motion Completion To facilitate the conditional motion completion task, we selected motion sequences from the HumanML3D test set with lengths ranging from 80 to 150 frames. We experimented with various settings and observed that the difficulty of motion inbetweening is significantly lower than motion prediction. Furthermore, introducing text conditions proved advantageous in reducing prediction errors.