GenM^3: Generative Pretrained Multi-path Motion Model for Text Conditional Human Motion Generation

Junyu Shi[1], Lijiang Liu[1], Yong Sun[1], Zhiyuan Zhang[1], Jinni Zhou[1], Qiang Nie[1]*
[1]The Hong Kong University of Science and Technology (Guangzhou)
{jshi890, lliu135, ysun691, zzhang400}@connect.hkust-gz.edu.cn {eejinni, qiangnie}@hkust-gz.edu.cn

Abstract

Scaling up motion datasets is crucial to enhance motion generation capabilities. However, training on large-scale multi-source datasets introduces data heterogeneity challenges due to variations in motion content. To address this, we propose Generative Pretrained Multi-path Motion Model (GenM^3), a comprehensive framework designed to learn unified motion representations. GenM^3 comprises two components: 1) a Multi-Expert VQ-VAE (MEVQ-VAE) that adapts to different dataset distributions to learn a unified discrete motion representation, and 2) a Multi-path Motion Transformer (MMT) that improves intra-modal representations by using separate modality-specific pathways, each with densely activated experts to accommodate variations within that modality, and improves inter-modal alignment by the text-motion shared pathway. To enable large-scale training, we integrate and unify 11 high-quality motion datasets (approximately 220 hours of motion data) and augment it with textual annotations (nearly 10,000 motion sequences labeled by a large language model and 300+ by human experts). After training on our integrated dataset, GenM^3 achieves a state-of-the-art FID of 0.035 on the HumanML3D benchmark, surpassing state-of-the-art methods by a large margin. It also demonstrates strong zero-shot generalization on IDEA400 dataset, highlighting its effectiveness and adaptability across diverse motion scenarios.

1. Introduction

Generating diverse and accurate human motion from textual descriptions has emerged as a vital area in generative computer vision [4, 10, 13, 25, 31, 32, 39], which has broad applications across animation, virtual reality, and robotics.

Inspired by breakthroughs of pre-trained Large Language Models (LLMs) on large-scale datasets, recent research in human motion generation leverages larger models [17] and expanded datasets [10, 19] to produce more realistic and nuanced motion sequences. On the data side, datasets like HumanML3D [10] and Motion-X [19] have broadened available training samples. In terms of modeling, approaches like MotionGPT [14] fine-tunes pre-trained LLMs on the HumanML3D dataset to jointly model motion tokens and text tokens. Building upon this, MotionChain [15] extends the use of pretrained LLMs for multi-turn motion dialogue, enabling dynamic interactions over sequential motions. Another pioneering model, OMG [17], incorporates multiple motion datasets and utilizes a large-scale diffusion model to achieve strong zero-shot performance. These approaches underscore the promising role of pretrained LLMs and large-scale data integration in enhancing the realism of human motion generation.

Despite above advancements, several challenges persist in motion generation: 1) Heterogeneity in motion data distribution. The heterogeneity arising from differences in motion types, recording devices, and similar factors poses a challenge for joint training, as optimizing performance

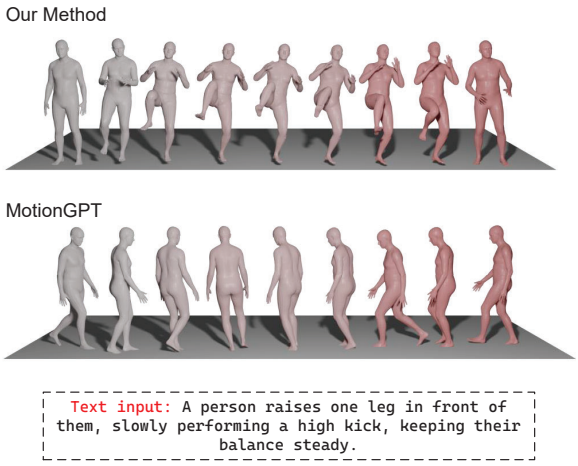

Figure 1. Due to the limited motion and text in the dataset, MotionGPT [14] is unable to respond to this relatively simple text input. We expanded the dataset in both motion and text modalities and pre-trained our model on this large-scale dataset. This enabled our model, GenM^3, to generate high-precision human motions based on diversity descriptions.

*Corresponding author

on one dataset may adversely affect the results on another dataset. 2) Lack of a dedicated pretrained motion backbone for comprehensive motion representations. While vision/language pretrained backbones have demonstrated robustness in image and text processing, efforts to adapt large language models to motion tasks [14, 15, 34, 35] are hindered by inherent structural and contextual differences between motion and text. Thus, a dedicated motion model pretrained on large-scale motion data are crucial for capturing motion data's unique representations. 3) Lack of high-quality, large-scale unified motion datasets. The high cost of capturing high-fidelity human motion results in smaller datasets compared to natural images. Moreover, Existing datasets are often limited to specific motion types (e.g., HumanML3D emphasizes self-motion, while HIMO focuses on human–object interactions), restricting generalizability across varied human activities. Figure 1 illustrates an example of motion generation. Despite the simple description, MotionGPT [14] failed to grasp the textual meaning and produced almost irrelevant motion.

To overcome these challenges, in this paper, we propose Generative Pretrained Multi-path Motion Model (GenM^3), a framework designed to learn unified motion representations to address the challenge of data heterogeneity in large-scale, multi-source motion datasets. GenM^3 comprises two novel designs: Multi-Expert VQ-VAE (MEVQ-VAE) and Multi-path Motion Transformer (MMT). As shown in Fig. 2, MEVQ-VAE discretizes continuous motion sequences through a multi-expert architecture that simultaneously activates all experts and adaptively adjusts each expert's contribution based on learned weights, capturing shared characteristics across datasets while preserving unique data attributes. Drawing inspiration from multimodal models such as VLMO [2], MMT utilizes separate modality-specific pathways, each with densely activated experts to adapt variations of data distribution within that modality. To facilitate cross-modal alignment, we also integrate a shared text-motion pathway that learns joint representations, bridging the gap between textual and motion semantics. That is, MMT employs three distinct pathways: one for the motion mode, one for the text mode, and a cross-modal pathway that handles both motion and text data. In addition, GenM^3 integrates text tokens with motion tokens through a Motion Descriptor, producing context tokens representing overarching motion summaries, providing text-guided motion insights that enhance cross-modal alignment. The Multi-path Motion Transformer concurrently processes text tokens (including context tokens) and motion tokens, learning alignments between modalities.

We integrate and unify 11 high-quality motion datasets to support GenM^3's training and provide a large-scale motion dataset across multiple scenarios. We preprocess and align these datasets in terms of data structure and pose representation to match the benchmark dataset, HumanML3D [10]. After processing, the effective motion duration is approximately 220 hours, covering scenarios including single-person motion, two-person interactions, and human-object interactions. Notably, the HIMO [21] dataset includes textual descriptions of interactions between upper limbs and objects, while the HuMMan [3] dataset provides detailed textual annotations of joint changes during movement. Additionally, we enriched the IMHD [46] dataset by manually annotating full-body interactions with objects and generated segment-level descriptions for the BABEL [28] dataset using ChatGLM [7]. These diverse textual descriptions extend HumanML3D's representation of human motion, enabling the model to respond effectively to a broader array of motion descriptions.

We pretrain GenM^3 on all motion datasets using a masked completion strategy, which harnesses the model's potential as a powerful motion representation tool by exposing it to diverse motion patterns. Subsequently, we conduct text-conditional generative training on datasets containing motion-text pairs, adapting GenM^3's motion representation capability to text-driven generation. In summary, our contributions include the following:

- We propose GenM^3 to address data heterogeneity and facilitate cross-modal alignment during training with large-scale datasets. To this end, we integrate and unify 11 high-quality motion datasets (nearly 220 hours) across diverse scenarios, and enriching them with textual labels.
- In GenM^3, we propose Multi-Expert VQ-VAE to learn a unified discrete motion representation, and a Multi-path Motion Transformer to adapt to text and motion variations and enhance text-motion alignment.
- GenM^3 achieves 0.035 FID on HumanML3D dataset, significantly surpassing SoTA mathods, and shows superior zero-shot performance on IDEA400 dataset.

2. Related Work

2.1. Text-Conditional Motion Generation

Motion generation is a fundamental research area in computer vision and graphics, aiming to produce realistic and contextually appropriate human motions. Text-conditional motion synthesis, in particular, focuses on generating motion sequences based on textual descriptions. Early methods employ deterministic models, which could result in averaged or less dynamic motions due to the complex and stochastic nature of human movement. To capture a broader range of motion variations, researchers have turned to probabilistic models like Generative Adversarial Networks (GANs) [6, 13] and Variational Autoencoders (VAEs) [25], which help in modeling the inherent uncertainty and diversity in motion data. Diffusion models [4, 5, 32, 33, 36, 39, 41–43, 48] have also gained traction for their ability to gen-

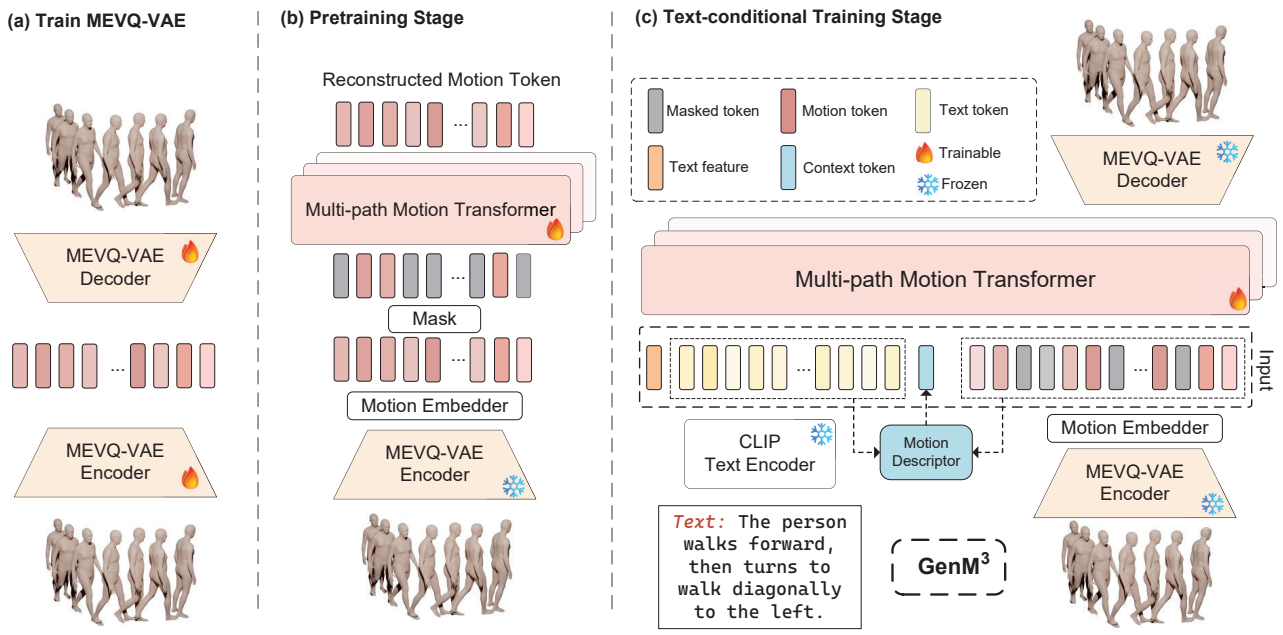

Figure 2. Overview of the GenM^3 framework. The training of GenM^3 consists of three steps: 1) Train MEVQ-VAE to extract discretized motion representations. MEVQ-VAE is frozen in the later steps. 2) Pre-train GenM^3 on a large amount of motion modality data using self-reconstruction with masked tokens. 3) Perform text-conditional training on text-motion pairs. The processed text tokens are concatenated with the embedded discretized action tokens as input to the Multi-path Motion Transformer (MMT). The output of MMT is then fed into the Decoder of MEVQ-VAE, resulting in the generated motion sequence. Note that the computations inside GenM^3 are length-independent.

erate high-quality motions by iteratively refining random noise into coherent motion sequences. More recently, some approaches [5, 8, 11, 12, 26, 27, 40, 47] employ VQ-VAE to discretize motion sequences, followed by using a generative Transformer to produce motion tokens. Specifically, T2M-GPT [40] generate motion tokens in a next-token prediction paradigm, while MMM [27] and MoMask [12] utilize a masked completion approach: during training, they reconstruct randomly masked motion tokens, and in inference, they iteratively generate motions starting from a fully masked sequence.

2.2. Large Model for Motion Generation

The use of Large Model in motion generation has gained increasing attention due to their success in capturing complex, contextually rich dependencies within motion data. MotionGPT [14] applied pretrained LLMs [29] as the backbone for motion modeling. By discretizing motion sequences, motion tokens are treated as a new language, allowing joint modeling with text. Building on MotionGPT, MotionChain [15] introduced a multi-turn motion generation mechanism that enables more natural interactions. Although these approaches leverage the powerful representational capabilities of pretrained LLMs, there exists a significant domain gap between motion and text data, which can impact the reliability of the generated motions. On the other hand, the scale of the dataset is a critical factor for successfully training large motion models. HumanML3D [10], which labels textual annotations to AMASS [22] motion data, provides a benchmark dataset for motion generation. Motion-X [19] expands on this by using an automated labeling pipeline to extract a large number of motion sequences from video and generate corresponding text descriptions. However, the motion data in Motion-X contains significant noise, which limits its utility for high-precision motion generation. The OMG [17] model integrates a wide range of motion datasets and trains various sizes of diffusion-based models for motion generation, pushing forward the capacity of large-scale motion modeling. Despite the success of these methods, substantial room remains for improvement in multi-dataset joint training and large-scale motion representation models within the field of motion generation. In this paper, we aim to center the analysis on motion data, transferring the scale law of datasets to the action domain, in order to mitigate the impact of the domain gap of large language models (LLMs) on motion data, and provide community with a comprehensive motion backbone.

3. Method

3.1. Model Architecture

Our proposed GenM^3 framework comprises two main components: the Multi-Expert VQ-VAE (MEVQ-VAE) and the Multi-path Motion Transformer (MMT). The MEVQ-VAE is designed to discretize continuous motion data while adapting to distributional differences across datasets. Building on these discrete tokens, the MMT models both motion and textual tokens in a unified manner, leveraging a multi-

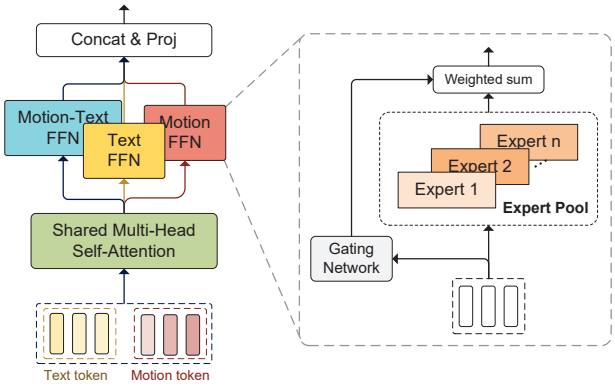

Figure 3. Architecture of multiway transformers in GenM^3. The text, motion, and motion-text (cross-modal) branches handle their respective tokens, and then their outputs are aggregated at the output end. Each branch is equipped with an expert pool.

branch strategy to enhance intra-modal and inter-modal representational capacity.

3.1.1. Multi-Expert VQ-VAE

The Multi-Expert VQ-VAE (MEVQ-VAE) is an extension of the original VQ-VAE framework, designed to discretize continuous motion data while handling variations across multiple datasets. As shown in Fig. 2 (a), given a motion sequence $\mathbf{X} = [x_1, x_2, ..., x_T]$ with $x_t \in \mathbb{R}^d$, where T is the number of frames and d is the dimensionality of each frame, the MEVQ-VAE encodes the sequence through the encoder $\mathcal{E}$ and recover the complete motion sequence by the decoder $\mathcal{D}$. A shared learnable codebook is employed, which contains K discrete codes $\mathbf{C} = c_k{}_{k=1}^{K}$ with $c_k \in \mathbb{R}^{d_c}$, where d_c is the dimensionality of each code. Each latent feature $\mathbf{Z} = \mathcal{E}(\mathbf{X}) = \left[z_1, z_2, ..., z_{T/l}\right]$, with $z_i \in \mathbb{R}^{d_c}$ and l being the downsampling rate, is quantized by selecting the closest code from $\mathbf{C}$.

The encoder $\mathcal{E}$ and decoder $\mathcal{D}$ is composed of multiple blocks, each containing three standard 1D convolutional layers followed by a specialized 1D convolutional layer with e_q experts. In this expert layer, all experts are activated simultaneously to capture diverse motion features, which can be represented as a weighted combination of each expert's output, formulated as:

$$y = \sum_{i=1}^{e_q} w_i \cdot \mathrm{Conv}_i(x) \tag{1}$$

where $\mathrm{Conv}_i(x)$ represents the output from the i-th expert's convolutional filter, and w_i is the learnable weight associated with each expert.

3.1.2. Multi-path Motion Transformer

The Multi-path Motion Transformer (MMT) framework consists of two modules: the Motion Descriptor and a Multi-path Transformer network. The Motion Descriptor is responsible for summarizing motion patterns, converting them into context tokens that provide high-level representations of the motion sequence. The Transformer network then processes three types of tokens: text tokens, context tokens (considered as a type of text token), and motion tokens.

Motion Descriptor Inspired by LLaMA-VID's approach [16] to text information compression, we design the Motion Descriptor to enhance motion representation. After processing the motion sequence $\mathbf{X}$ through MEVQ-VAE, it is mapped to a sequence of indices $\mathbf{S} = \left[s_1, s_2, ..., s_{T/l}, s_{End}\right]$ within a learned codebook. Passing $\mathbf{S}$ through a motion embedder Emb_m, we obtain the motion embedding $\mathbf{E}_m$. Simultaneously, the input text is encoded by a CLIP encoder to generate both a text embedding $\mathbf{E}_t$ and a global text feature $\mathbf{e}_t$.

To construct a contextual summary embedding $\mathbf{E}_{ctx}$ for motion, we aggregate motion features by scoring their response to the text query:

$$\mathbf{E}_{ctx} = \mathrm{mean}\left(\mathrm{softmax}\left(\mathbf{E}_m \mathbf{E}_t\right) \mathbf{E}_t\right) \tag{2}$$

This selective aggregation captures a high-level motion summary within $\mathbf{E}_{ctx}$, which integrates both motion and textual perspectives and serves as an enriched contextual input within the textual branch of MMT.

Multi-path Transformer We concatenate text features $\mathbf{e}_t$, text embeddings $\mathbf{E}_t$, context embeddings $\mathbf{E}_{ctx}$, and motion embeddings $\mathbf{E}_m$, adding positional encoding to form the input for our multi-path Transformer. The first half of the Transformer follows the standard Transformer architecture, utilizing self-attention and FeedForward layers to compute cross-modal attention:

$$\mathbf{E}_u = \mathrm{concat}\left(\mathbf{e}_t, \mathbf{E}_t, \mathbf{E}_{ctx}, \mathbf{E}_m\right) + \mathbf{P} \tag{3}$$

$$\mathbf{Y} = \mathrm{FFN}\left(\mathrm{Attention}\left(\mathbf{E}_u\right)\right) \tag{4}$$

where $\mathbf{P}$ and FFN represent the Position Encoding and FeedForward Network, respectively.

As shown in Fig. 3, In the whole second half of the model, we introduce specialized pathways within the FeedForward layer. These pathways include a motion pathway, a text pathway, and a cross-modal shared pathway, which are parallel. To further adapt to varying data distributions and diverse textual descriptions across datasets, we incorporate a multi-expert strategy within each pathway. Mathematically, each pathway p consists of multiple experts, and for each expert i, we compute:

$$\mathbb{E}_{p,i}(x) = \mathbf{W}_{p,i}^2 \sigma\left(\mathbf{W}_{p,i}^1 x + b^1\right) + b^2 \tag{5}$$

The output from each expert is weighted and combined based on the learned attention weight $\mathrm{g}_{p,i}$ for each expert,

yielding:

$$\mathbb{E}_p(x) = \sum_i \mathrm{g}_{p,i}(x)\,\mathbb{E}_{p,i}(x) \tag{6}$$

where $\mathbb{E}_{p,i}$ denotes the output of the i^{th} expert in p^{th} pathway to input x. $g_{p,i}(x)$ is a gating function that calculates the weight of each expert. The outputs from the motion, text, and cross-modal pathways are concatenated to produce the final representation:

$$\text{Output} = \mathbf{W}_{proj}\left([\mathbb{E}_{motion}; \mathbb{E}_{text}; \mathbb{E}_{cross-modal}]\right) + b_{proj} \tag{7}$$

Notably, all experts are activated simultaneously, with their contributions to each token dynamically adjusted by learned weights. This design benefits from building modality-sharing channels within each pathway, leveraging common features across datasets while supporting specialized learning.

3.2. Training Strategy

Stage 1: Train the MEVQ-VAE In the first stage, we train the MEVQ-VAE on raw motion data to learn a quantized representation of motion sequences. Following previous methods [27, 40], the MEVQ-VAE model optimizes two losses, a reconstruction loss $\mathcal{L}_{rec}$ and a commitment loss $\mathcal{L}_{commit}$. The total loss can be expressed as:

$$\mathcal{L}_q = \mathcal{L}_{rec} + \beta\mathcal{L}_{commit} \tag{8}$$

where β is the hyper-parameter to balance the losses. Additionally, we utilize moving averages for codebook updates and resets for inactive codes [38].

Stage 2: Pretrain the GenM^3 As shown in Fig. 2 (b), in the pretraining phase, the GenM^3 model is trained exclusively on motion data to establish a robust foundational representation of movement patterns. During this stage, text, motion, and cross-modal (motion-text) pathways all take motion tokens as inputs and utilize motion feature to form the prior knowledge that is useful in the next stage. We adopt a masked modeling approach on the motion token level to enhance GenM^3's ability to reconstruct missing parts of motion data, following the method described in prior works [27]. In each discrete motion token sequence, random indices are replaced by learnable special-purpose tokens [Mask], encouraging the model to infer and regenerate these occluded motion parts. The objective is to minimize the negative log-likelihood of the predicted masked tokens conditioned on visible tokens:

$$\mathcal{L} = -\sum_{i\in\mathcal{M}} \log P\left(x_i | x_{\setminus\mathcal{M}}\right) \tag{9}$$

where $\mathcal{M}$ denotes the set of masked indices, and $x_{\setminus\mathcal{M}}$ represents the visible (non-masked) tokens in the sequence.

Stage 3: Text-conditional Training As shown in Fig. 2 (c), in the text-conditional training phase, GenM^3 is trained on both motion and textual data, enabling multimodal alignment between motion sequences and corresponding textual descriptions. During this stage, all pathways of the GenM^3–the motion, text, and cross-modal shared pathways–are active, allowing the model to fully leverage interactions within and across modalities. The motion pathway remains in a masked modeling framework, but depends on both textual and visible information.

3.3. Inference Strategy

During inference, all motion tokens are initialized as [Mask] to form an empty sequence. Then, using parallel decoding, all tokens are generated simultaneously in each step. In subsequent iterations, tokens with lower confidence are re-masked and regenerated.

4. Integration of Motion Data

Data Standardization To train a motion learning backbone capable of comprehensive representation, we collected a diverse set of highly accurate motion capture datasets. Given the disparities across datasets in terms of format (such as spatial coordinates versus SMPL parameters), body shape of performers, joint count, and frame rate, we standardized all datasets accordingly. Specifically, we normalized the joint count to 22, aligning with the HumanML3D [10] standard, and retargeted each dataset to a default human skeletal template. To prevent issues with undersampling, we set the frame rate to 30 fps.

Text Modality Our dataset includes HumanML3D [10], HuMMan [3], and HIMO [21], each with textual descriptions for every motion sequence. However, motion data inherently differs from text in structure and contextual relationships, and there are distribution differences in motion/text data across different datasets. HumanML3D emphasizes holistic human movement. HuMMan provides detailed motion patterns at the joint level. While HIMO focuses on interactions between the human upper body and objects. Additionally, the BABEL dataset provides action categories for each sub-segment. We divide the sequences based on sub-segment duration and generated sequence descriptions for each complete segment using ChatGLM [7]. In order to further increase the diversity of text data, we manually divide sequences and label motion descriptions on the human-body-object interaction dataset HDMI.

Sequence Segmentation During pretraining, for datasets without predefined or suitable sequence segmentation (e.g., sequences over 10 seconds), we sample training sequences with overlapping windows, ensuring overlap rates were kept

Table 1. Comparison of Text-to-Motion on HumanML3D [10] (using our retrained evaluator on 30FPS motion data) and Zero-shot Text-to-Motion on IDEA400 [19]. The highlight and underline represent the best and the second-best. Noting that GenM^3 uses text pairs from HumanML3D for text-conditional training, while GenM^{3*} denotes the use of all text-motion pairs from our integrated dataset.

Dataset	Methods	R-Precision↑			FID↓	MMDist↓	Diversity↑
		Top1	Top2	Top3			
HumanML3D	Real	$0.492^{\pm.004}$	$0.687^{\pm.003}$	$0.785^{\pm.003}$	$0.002^{\pm.000}$	$2.982^{\pm.010}$	$9.458^{\pm.090}$
	T2M-GPT [40]	$0.487^{\pm.002}$	$0.678^{\pm.003}$	$0.770^{\pm.002}$	$0.160^{\pm.006}$	$3.083^{\pm.009}$	$9.653^{\pm.065}$
	MMM [27]	$0.496^{\pm.003}$	$0.686^{\pm.002}$	$0.784^{\pm.002}$	$0.110^{\pm.005}$	$2.951^{\pm.009}$	$9.484^{\pm.113}$
	Ours GenM^{3*}	$\underline{0.510}^{\pm.002}$	$\underline{0.702}^{\pm.002}$	$\underline{0.802}^{\pm.002}$	$\underline{0.053}^{\pm.002}$	$\underline{2.860}^{\pm.009}$	$\underline{9.629}^{\pm.077}$
	Ours GenM^3	$\mathbf{0.511}^{\pm.003}$	$\mathbf{0.705}^{\pm.002}$	$\mathbf{0.804}^{\pm.002}$	$\mathbf{0.046}^{\pm.002}$	$\mathbf{2.852}^{\pm.009}$	$\mathbf{9.675}^{\pm.087}$
IDEA400	Real	$0.124^{\pm.003}$	$0.221^{\pm.005}$	$0.308^{\pm.006}$	$0.001^{\pm.000}$	$4.615^{\pm.004}$	$6.001^{\pm.108}$
	T2M-GPT [40]	$0.122^{\pm.003}$	$0.219^{\pm.004}$	$0.301^{\pm.004}$	$7.947^{\pm.012}$	$5.488^{\pm.034}$	$7.636^{\pm.110}$
	MMM [27]	$0.126^{\pm.003}$	$0.221^{\pm.005}$	$0.307^{\pm.004}$	$6.001^{\pm.010}$	$4.980^{\pm.028}$	$7.730^{\pm.103}$
	Ours GenM^{3*}	$\mathbf{0.136}^{\pm.004}$	$\mathbf{0.244}^{\pm.002}$	$\mathbf{0.338}^{\pm.003}$	$\mathbf{4.232}^{\pm.008}$	$\mathbf{4.520}^{\pm.012}$	$\underline{7.750}^{\pm.101}$
	Ours GenM^3	$\underline{0.134}^{\pm.003}$	$\underline{0.241}^{\pm.005}$	$\underline{0.335}^{\pm.003}$	$\underline{4.430}^{\pm.060}$	$\underline{4.732}^{\pm.007}$	$\mathbf{7.851}^{\pm.131}$

Table 2. Comparison of Text-to-Motion on HumanML3D [10] (using the evaluator trained on 20FPS motion data [9]).

Methods	FID↓	R-Precision (Top3)↑	Diversity↑
Real	$0.002^{\pm.000}$	$0.797^{\pm.002}$	$9.503^{\pm.065}$
MotionDiff [43]	$0.630^{\pm.001}$	$0.782^{\pm.001}$	$9.410^{\pm.049}$
MDM [32]	$0.544^{\pm.044}$	$0.611^{\pm.007}$	$9.559^{\pm.086}$
MLD [4]	$0.473^{\pm.013}$	$0.772^{\pm.002}$	$9.724^{\pm.082}$
T2M-GPT [40]	$0.141^{\pm.005}$	$0.775^{\pm.002}$	$9.722^{\pm.082}$
EMDM [48]	$0.112^{\pm.019}$	$0.786^{\pm.006}$	$9.551^{\pm.078}$
FineMoGen [42]	$0.151^{\pm.008}$	$0.784^{\pm.002}$	$9.263^{\pm.094}$
MoMamba [45]	$0.281^{\pm.009}$	$0.792^{\pm.002}$	$\mathbf{9.871}^{\pm0.084}$
MoMask [12]	$\underline{0.045}^{\pm.002}$	$\mathbf{0.807}^{\pm.002}$	-
MMM [27]	$0.080^{\pm.003}$	$0.794^{\pm.002}$	$9.411^{\pm.058}$
MotionGPT [14]	$0.232^{\pm.008}$	$0.778^{\pm.002}$	$9.528^{\pm.071}$
MotionLLM [35]	$0.491^{\pm.019}$	$0.770^{\pm.002}$	$\underline{9.838}^{\pm.244}$
MotionChain [15]	$0.248^{\pm.009}$	$0.790^{\pm.003}$	$9.470^{\pm.075}$
OMG [17]	$0.381^{\pm.008}$	$0.784^{\pm.002}$	$9.657^{\pm.085}$
Our GenM^{3*}	$0.046^{\pm.002}$	$0.790^{\pm.003}$	$9.346^{\pm.076}$
Our GenM^3	$\mathbf{0.035}^{\pm.002}$	$\underline{0.795}^{\pm.002}$	$9.341^{\pm.080}$

below 25% to maintain diversity. For text-conditional generative training, we follow the original segmentation provided by each dataset or our manual partitioning.

5. Experiments

This section provides a comprehensive evaluation of our proposed GenM^3 framework. In Sec. 5.1, we introduce the datasets and experimental setup. Sec. 5.2 presents a comparative analysis of our method against *state-of-the-art* methods. In Sec. 5.3, we conduct ablation studies to investigate the contributions of individual components. We provide more information in the supplemental.

5.1. Datasets and Implementation

Datasets for Pretraining and Text-conditioned Training We collected and processed a total of 11 motion capture datasets [1, 3, 10, 18, 20, 21, 23, 28, 30, 44, 46], aligning each to the HumanML3D dataset's format for consistency in training and evaluation. We use all the motion data for pretraining. For text-conditional training, we used datasets that include text modality: HumanML3D [10], HuMMan [3], HIMO [21], BABEL [28], and IMHD [46].

Implementation Details Our model is trained on Nvidia RTX-4090 GPUs. In MEVQ-VAE, motion sequences are downsampled by a factor of 4, and both encoder and decoder use a four-expert multi-expert architecture. The codebook has 8192 entries of 32 dimensions, with the loss weight β set to 1. The MMT contains 18 layers: the first nine layers follow standard attention, while the latter nine layers implement multi-path transformer with multiple experts. For pretraining and text-conditional training, we set the batch size to 160 and use AdamW optimizer to train the model for 120,000 iterations with a learning rate of 2×10^{-4} following a warm-up and cosine annealing decay strategy.

5.2. Comparison

Quantitative Comparison We first perform a quantitative evaluation on the HumanML3D [10]. We report the results evaluated by the evaluator from prior works [10, 27] in Tab. 2. However, this evaluator was trained on 20 FPS motion samples, which may not be suitable to evaluate on our constructed dataset. Thus, we re-trained the evaluator following the prior settings [10] and trained MMM and T2M-GPT on 30FPS HumanML3D for fair comparisons. The

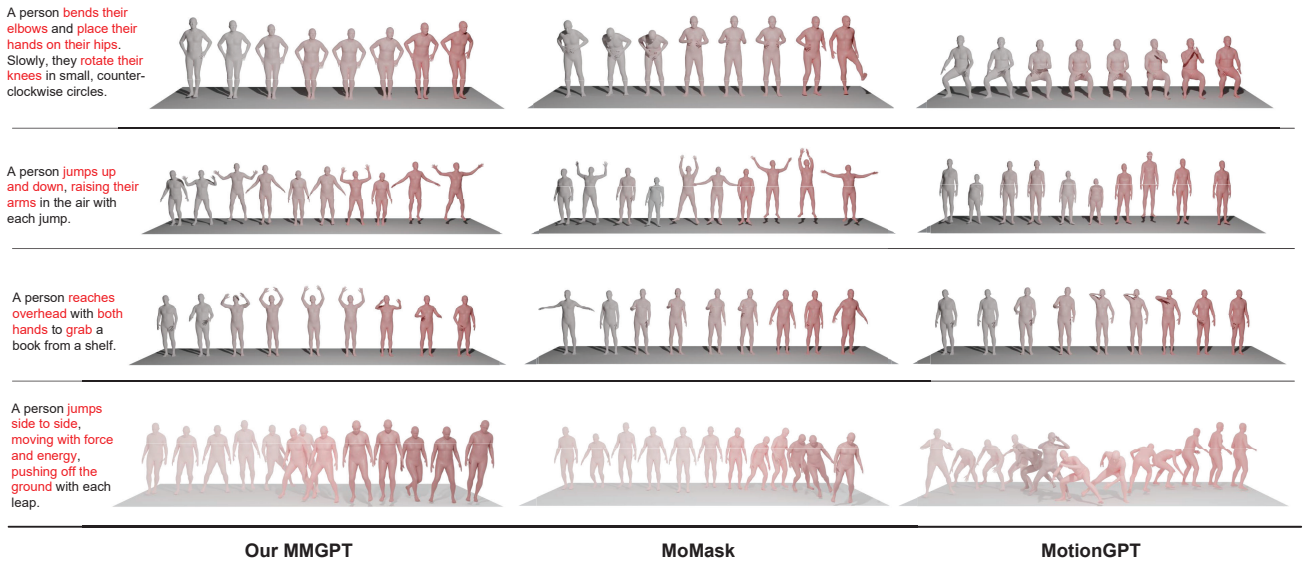

Figure 4. Motion generation results based on text inputs.

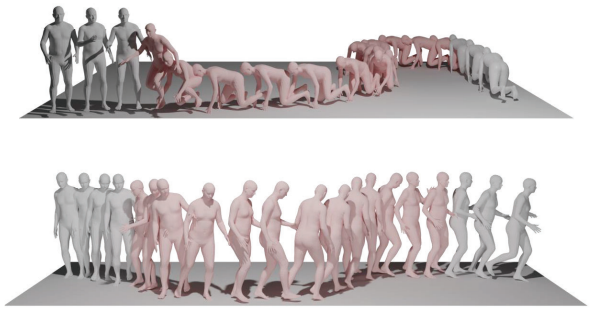

Figure 5. Visualizations of motion in-between.

results are shown in Tab. 1. Following the standard evaluation protocols, we report the results from 20 repeated trials, with each result shown as the mean with a 95% confidence interval. We trained two versions of GenM^3: GenM^3, using only HumanML3D text pairs for text-conditional training, and GenM^{3*}, using all motion-text data. GenM^3 performs better on HumanML3D, but GenM^{3*} exhibits stronger generalization ability (will be discussed later). It can be seen that Our method outperforms other approaches in FID and achieves comparable performances in other metrics.

Qualitative Comparison Fig. 4 presents qualitative results showcasing the generated motions. We used different types of text as conditions, including: 1) descriptions based on body joint movements (e.g., "place their hands on their hips"); and 2) descriptions based on motion types (e.g., "jump"). Due to the limited diversity and quantity of text in the HumanML3D dataset, methods trained on this dataset often fail to generate effective motion sequences for some simple and common action descriptions. Our approach, however, demonstrates a clear advantage to a wide range of descriptions and high-quality generations.

Table 3. Comparison of different VQ methods.

Method	VQ	MEVQVAE	RVQ [12]	MERVQ	G-RVQ [37]	FSQ [24]
FID	0.098	0.048	0.043	0.032	0.045	0.057

Table 4. Results of motion in-betweening. R. means R-precision. Div. means Diversity.

Type	Method	R.(Top 3)↑	FID↓	Div.↑
Prefix	MMM [27]	0.780	0.098	9.432
	Ours	**0.797**	**0.040**	**9.583**
Suffix	MMM [27]	0.788	0.142	9.455
	Ours	**0.800**	**0.065**	**9.537**
Infix	MMM [27]	0.784	0.133	9.398
	Ours	**0.803**	**0.056**	**9.506**

Comparison of VQ Method Tab. 3 compares different quantization methods, including RVQ [12], G-RVQ [37] and FSQ [24]. Our multi-expert design adapts to diverse dataset distributions, substantially outperforming standard VQ. RVQ achieves the best scores but requires an extra Residual Transformer to model residuals [12]. Integrating our Multi-Expert into RVQ (MERVQ) further boosts performance, demonstrating its generalizability.

Motion Completion We conducted comparative experiments with MMM [27] on the motion completion task on the HumanML3D test set, evaluating three distinct modes: prefix-based (generating the last 50% based on the first 50% of observed motion), suffix-based (generating the first 50% based on the last 50% of observed motion), infix-based

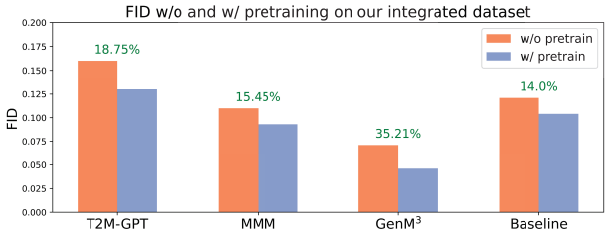

Figure 6. The impact of pretraining on different methods on our integrated dataset.

(generating the middle 50% based on the initial and final 25% of observed motion) completion. As shown in the Tab. 4, our method achieves better metrics especially the FID. Fig. 5 visualizes some of the motion completion results, where it can be seen that our method excels in both overall semantic coherence and motion transition.

Zero-shot Inference We tested the zero-shot capability of GenM^3 on IDEA400 dataset. GenM^3* refers to the model trained on all motion-text pairs in our integrated dataset. As can be seen in Tab. 1, thanks to the strong representational power learned during pre-training, the GenM^3 model trained only on HumanML3D still exhibits robust zero-shot reasoning ability. GenM^3*, trained on more motion-text pairs, shows even stronger generalization ability.

5.3. Ablation Study

Pretraining Improvement We conducted experiments by pretraining MMM [27], T2M-GPT [40] and our baseline (refers to our method without the Multi-path, Multi Expert, and Motion Descriptor components) on our integrated dataset in Fig. 6. It can be seen that pretraining on our integrated dataset yields the greatest improvement (35.21%) for our method. This can be attributed to the multi-expert design in our MEVQ-VAE, which enables the model to adapt to the heterogeneity of various datasets. Moreover, the Motion Pathway in MMT leverages multiple experts to simultaneously learn motion patterns from different datasets.

Effective of Multi-path Transformer in GenM^3 We conducted ablation experiments on the text and cross-modal branches in GenM^3, with the results shown in Tab. 6. As can be seen, when both branches are used simultaneously, the model is able to more effectively learn the features both within and across modalities.

Impact of Expert Count We analyzed the effect of expert numbers on three components. In MEVQ-VAE (Fig. 7 (a)), using eight experts produced the best reconstruction performance. However, due to computational costs, we set up four experts to get nearly the same result. In the motion and text pathways (Fig. 7 (b)), increasing the number of experts enhances the model's ability to handle a broader range

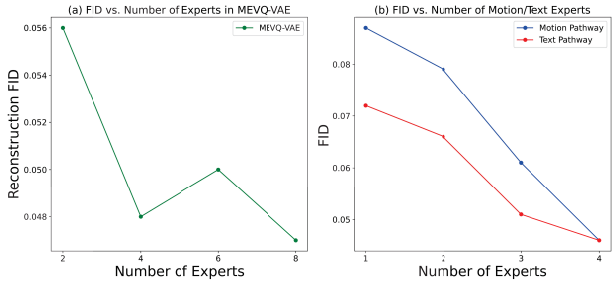

Figure 7. The impact of the number of experts in MEVQ-VAE on reconstruction performance.

Table 5. Comparisons of dense MoE and sparse MoE.

	FID↓	R-Precision (Top3)↑	Diversity↑
Sparse MoE	0.058	0.799	**9.676**
Dense MoE	**0.046**	**0.804**	9,675

Table 6. The impact of the multiway transformer in GenM^3 on generation performance.

motion	text	cross-modal	FID↓	Diversity↑
✓	-	-	0.058	**9.400**
✓	✓	-	0.045	9.282
✓	-	✓	<u>0.044</u>	<u>9.344</u>
✓	✓	✓	**0.035**	9.341

of data distributions (e.g. diverse textual descriptions and diverse motion patterns present across datasets).

Dense MoE *v.s.* Sparse MoE We conducted comparisons of dense MoE (fully activated) and sparse MoE (partially activated) in the motion pathway. As shown in Tab. 5, fully activated experts better capture commonalities across datasets and adapt different dataset from a higher dimension. While Sparse MoE may overlook some shared patterns and failed adapt the data in a narrower action space.

6. Conclusion

In this paper, we propose GenM^3, which consists of a multi-expert VQ-VAE to learn the unified discrete motion representation, and a Multi-path Motion Transformer that enhances intra-modal representations by employing dedicated pathways for each modality, with densely activated experts that capture the inherent variations, and bolsters inter-modal alignment through a shared text-motion pathway. We integrated and unified 11 high-quality motion datasets to support the training of GenM^3. Experimental results show that our method achieves state-of-the-art performance on the Text-to-Motion task while demonstrating remarkable generalization capabilities. In the future, we plan to further expand this dataset to train a more generalizable motion backbone for the research community.

7. Acknowledgment

This work is partially supported by the Guangzhou-HKUST(GZ) Joint Funding Program (No. 2025A03J3656).

References

[1] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21211–21221, 2023. 6

[2] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022. 2

[3] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *17th European Conference on Computer Vision*, pages 557–577. Springer, 2022. 2, 5, 6

[4] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 1, 2, 6

[5] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. 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*, pages 481–490, 2023. 2, 3

[6] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1396–1406, 2021. 2

[7] Team GLM. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024. 2, 5

[8] Kehong Gong, Dongze Lian, Heng Chang, Chuan Guo, Zihang Jiang, Xinxin Zuo, Michael Bi Mi, and Xinchao Wang. Tm2d: Bimodality driven 3d dance generation via music-text integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9942–9952, 2023. 3

[9] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 6

[10] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 1, 2, 3, 5, 6

[11] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 3

[12] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 3, 6, 7

[13] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020. 1, 2

[14] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023. 1, 2, 3, 6

[15] Biao Jiang, Xin Chen, Chi Zhang, Fukun Yin, Zhuoyuan Li, Gang YU, and Jiayuan Fan. Motionchain: Conversational motion controllers via multimodal prompts. In *European Conference on Computer Vision*, 2024. 1, 2, 3, 6

[16] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 4

[17] Han Liang, Jiacheng Bao, Ruichi Zhang, Sihan Ren, Yuecheng Xu, Sibei Yang, Xin Chen, Jingyi Yu, and Lan Xu. Omg: Towards open-vocabulary motion generation via mixture of controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–493, 2024. 1, 3, 6

[18] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 132(9):3463–3483, 2024. 6

[19] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36, 2023. 1, 3, 6

[20] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1154, 2024. 6

[21] Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, et al. Himo: A new benchmark for full-body human interacting with multiple objects. In *European Conference on Computer Vision*, pages 300–318. Springer, 2025. 2, 5, 6

[22] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 3

[23] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 6

[24] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *The Twelfth International Conference on Learning Representations*, 2024. 7

[25] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 1, 2

[26] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. Bamm: Bidirectional autoregressive motion model. 2024. 3

[27] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2024. 3, 5, 6, 7, 8

[28] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, 2021. 2, 6

[29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 3

[30] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020. 6

[31] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 1

[32] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 6

[33] Yin Wang, Zhiying Leng, Frederick WB Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22035–22044, 2023. 2

[34] Yuan Wang, Di Huang, Yaqi Zhang, Wanli Ouyang, Jile Jiao, Xuetao Feng, Yan Zhou, Pengfei Wan, Shixiang Tang, and Dan Xu. Motiongpt-2: A general-purpose motion-language model for motion generation and understanding. *arXiv preprint arXiv:2410.21747*, 2024. 2

[35] Qi Wu, Yubo Zhang, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Motion-agent: A conversational framework for human motion generation with llms. *arXiv preprint arXiv:2405.17013*, 2024. 2, 6

[36] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[37] Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*, 2023. 7

[38] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQGAN. In *International Conference on Learning Representations*, 2022. 5

[39] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16010–16021, 2023. 1, 2

[40] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 3, 5, 6, 8

[41] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 364–373, 2023. 2

[42] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. *Advances in Neural Information Processing Systems*, 36:13981–13992, 2023. 6

[43] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4115–4128, 2024. 2, 6

[44] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision*, pages 180–200. Springer, 2022. 6

[45] Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion mamba: Efficient and long sequence motion generation. In *European Conference on Computer Vision*, pages 265–282. Springer, 2025. 6

[46] Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I'm hoi: Inertia-aware monocular capture of 3d human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 729–741, 2024. 2, 6

[47] Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 509–519, 2023. 3

[48] Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*, pages 18–38. Springer, 2025. 2, 6