A Motion is Worth a Hybrid Sentence: Taming Language Model for Unified Motion Generation by Fine-grained Planning*

Ronghui Li†
Tsinghua University
Shenzhen, China
Peng Cheng Laboratory
Shenzhen, China
lrh22@mails.tsinghua.edu.cn

Lingxiao Han†
Tsinghua University
Shenzhen, China
hanlx24@mails.tsinghua.edu.cn

Shi Shu†
Tsinghua University
Shenzhen, China
shushi@whu.edu.cn

Yueyao Liu
Tsinghua University
Shenzhen, China
liuyueya24@mails.tsinghua.edu.cn

Yukang Lin
Tsinghua University
Shenzhen, China
linyk23@mails.tsinghua.edu.cn

Yue Ma
The Hong Kong University of Science and Technology
Hong Kong, China
ymacn@connect.ust.hk

Jie Guo‡
Peng Cheng Laboratory
Shenzhen, China
guoj01@pcl.ac.cn

Ziwei Liu
Nanyang Technological University
Singapore
ziwei.liu@ntu.edu.sg

Xiu Li‡
Tsinghua University
Shenzhen, China
li.xiu@sz.tsinghua.edu.cn

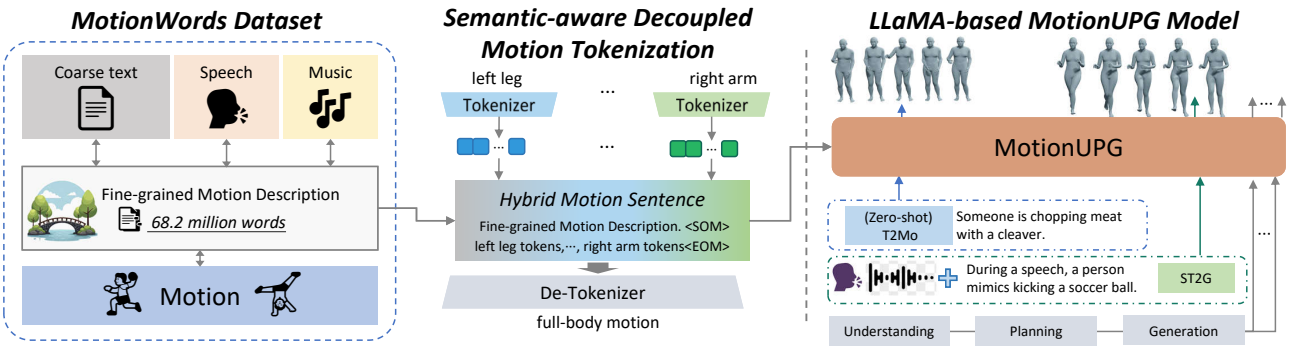

Figure 1: We propose the hybrid motion sentence to bridge the gap between motion and text, which is consistant of fine-grained motion decription and atomic body-part motion tokens. We further introduce the MotionWords dataset (68.2M words), a Semantic-Aware Decoupled Motion Tokenization method to enhance text-motion alignment, and the MotionUPG based on LLaMA, supporting unified motion generation tasks.

Abstract

Existing LLM-based motion models fail to fully leverage large models' planning capabilities for motion-related tasks, exhibiting poor generalization, limited text-motion alignment, and an inability to perform multimodal condition joint driven motion generation. We argue that these issues arise from the modality gap and the highly coupled nature of motion tokens. To address this, we proposed the *hybrid motion sentence*, which is consistant of fine-grained motion decription and atomic body-part motion token that can bridge the gap between motion and text. To obtain a large corpus of hybrid motion sentences, we introduced a novel motion-to-text generation method that combines *atomic motion operators* with GPT-4o, resulting in 68.2 million fine-grained textual descriptions across diverse modalities. To reconstruct high-quality motion from hybrid sentences and make better motion-text alignment, we introduce Semantic-Aware Decoupled Motion Tokenization. Furthermore, we propose MotionUPG based on LLaMA, leveraging MotionWords dataset for both pretraining and instruction tuning. Our method achieves strong fine-grained text-motion alignment, impressive zero-shot motion generation, and is the first to support multimodal condition joint driven motion generation tasks.

*Project page: https://motionupg.github.io/
†These authors contributed equally to this research.
‡Corresponding author.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755164

CCS Concepts

• **Computing methodologies** → **Motion processing**; **Computer graphics**.

Keywords

Motion Generation, Multi-modal Learning, Large Language Model

ACM Reference Format:

Ronghui Li, Lingxiao Han, Shi Shu, Yueyao Liu, Yukang Lin, Yue Ma, Jie Guo, Ziwei Liu, and Xiu Li. 2025. A Motion is Worth a Hybrid Sentence: Taming Language Model for Unified Motion Generation by Fine-grained Planning. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755164

1 Introduction

Motion generation is a crucial research task with widespread applications in animation, gaming, AR/VR, and embodied intelligence. In recent years, conditional motion generation has made significant progress, with diffusion-based and autoregressive models achieving remarkable results in tasks such as text-to-motion, music-to-dance, and speech-to-gesture generation. However, existing methods suffer from poor generalization in motion generation and lack fine-grained alignment with textual descriptions. Moreover, most approaches focus on single-modality tasks, making them incapable of generating motion driven by multiple modalities.

Inspired by the success of Large Language Models (LMMs) [19, 40, 49] and Multi-modal Large Language Models (MLLMs) [1, 7, 16, 25, 32], some researchers have started tokenizing motion into discrete tokens, treating them as a "foreign motion language" [18, 30, 55, 63, 64]. They expands the vocabulary of LLMs with motion-specific tokens, enabling the training of a Multi-centric Large Language Model. This all-in-tokens approach, allows a unified model to support both motion generation and understanding tasks.

However, upon reviewing works like MotionGPT [18], we find that they primarily achieve good performance on supervised fine-tuning tasks but struggle to fully leverage the motion understanding and planning capabilities of large multimodal models to enhance motion-related tasks. As a result, even with the introduction of LLMs, these methods still fail to address the challenges of poor generalization and limited text-motion alignment. Furthermore, although multimodal-driven motion generation models such as M^3GPT [39] attempt to unify different modalities through tokenization, they still fail to enable LLMs to deeply understand multimodal information. Consequently, their performance also remains confined to supervised fine-tuning tasks, and they are unable to realize motion generation driven by rich multimodal fused condition.

Rethinking these motion-centric LLM approaches, we identify two main issues: First, there remains a significant gap between different modalities. Naively adopting an "all-in-tokens" strategy only achieves superficial unification without truly bridging this gap. Second, the granularity between motion tokens and text tokens is misaligned. In previous motion tokenization approaches, a single motion token typically corresponds to multiple frames of body movement. In contrast, in LLMs, word tokens correspond to the smallest textual units—words or symbols—and different parts of speech (e.g., subject, verb, object) are composed into sentences. This highly coupled nature of motion tokens, along with the coarse-grained motion descriptions usually used in existing datasets prevent LLMs from learning fine-grained atomic motion-text alignment. Moreover, previous motion tokenizers segment motions along the temporal dimension and convert them into discrete tokens, which can only be recombined at teporal dimension. This limits the expressive motion space, making it difficult to generate novel motions, resulting in poor generalization.

To address these issues, we propose the hypothesis: A Motion is Worth a Hybrid Sentence, which blends fine-grained motion description texts with atomic body-part motion tokens. The natural language descriptions need to be detailed enough to capture rich atomic-level motion information. With this foundation, each motion token can precisely capture the atomic action described in the text, which not only matches the information density expected by large language models but also helps them better understand motion by breaking it down into atomic-level actions. *A full Hybrid Sentence is structured according to the motion syntax defined as follows:* (1) To better align with the next-token prediction paradigm of LLMs and to enable motion descriptions to serve as a bridge between multimodal conditions and generated motion, we place the fine-grained descriptive text tokens at the beginning of each hybrid motion sentence. (2) To represent motion tokens at an atomic semantic level—while also considering the joint hierarchy and the skeleton chain of human body—we divided the body into eight distinct parts: torso, left leg, right leg, head, left arm, right arm, left hand, and right hand. The corresponding body-part motion tokens are then arranged sequentially, accompanied by special start and end tokens, to form the final hybrid motion sentence.

To validate this hypothesis, we need to extract fine-grained and accurate textual descriptions from motion data. However, directly training a motion2text model is insufficient because existing datasets contain only coarse motion descriptions, making it difficult for the model to generate atomic-level text-motion dependencies. To overcome this, we design operators based on body joint movements to identify actions for each part of the body and produce short atomic descriptions. Then, we combine these atomic descriptions with motion, along with optional coarse-grained motion descriptions (available in text-motion paired datasets), and feed them into GPT-4o [40] to obtain fine-grained and accurate motion descriptions. This method enables us to obtain fine-grained annotations not only for text-to-motion datasets but also for music-to-dance and speech-to-gesture datasets. As a result, we construct MotionWords, a large-scale multi-modal motion dataset with fine-grained text-motion pairs. In total, the dataset contains 245.7 hours of motion, 60 hours of speech, and 12.9 hours of music. It includes 2.1 million words from the original coarse-grained motion descriptions and 68.2 million words from our newly collected fine-grained motion descriptions, which is approximately equivalent to the total text volume of 891 times of *Harry Potter and the Sorcerer's Stone*.

Building on MotionWords, we introduce Semantic-Aware Decoupled Motion Tokenization (SDMT). Unlike previous motion tokenizer, where only motion data is provided and the model must simultaneously learn high-level semantic features and low-level kinematic features, our SDMT first leverages textual descriptions during tokenization. Given the textual description, the encoding network can focus on capturing low-level motion details, significantly

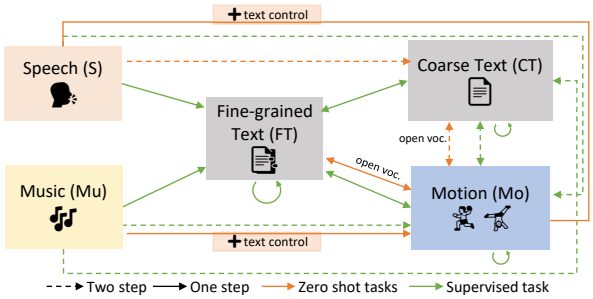

Figure 2: MotionUPG supports a wide range of tasks. It also demonstrates how fine-grained motion descriptions can bridge multimodal data. "open voc." means open vocabulary.

improving motion compression. During decoding, textual guidance enables precise and fine-grained motion control. To further enhance reconstruction motion quality, we apply residual vector quantization for each body part.

To fully leverage the rich multimodal information in MotionWords, we propose MotionUPG based on LLaMA. We extract various types of tokens, text, motion, speech, and music, from the MotionWords dataset, and design both pretraining and instruction-tuning tasks for training. Our method not only demonstrates strong fine-grained text-motion alignment but also exhibits impressive zero-shot motion generation capabilities. It can generate motions from high-level situational descriptions or open-vocabulary textual prompts. Furthermore, MotionUPGsupports motion generation jointly driven by multimodal conditions (text and audio together). Our contributions can be summarized as follows:

- We design a motion syntax to combine fine-grained text descriptions and atomic body part tokens into *hybrid motion sentences*, which serve as a bridge to reduce the modality gap between motion, text, and audio.
- We propose a method for generating fine-grained and accurate descriptions from various motions. Based on this, we collect a large-scale dataset, *MotionWords*, which includes motion, speech, music, and text data, along with 68.2 million fine-grained motion descriptions.
- We introduce *Semantic-Aware Decoupled Motion Tokenization*. We introduce the text into motion Tokenization for the first time, which effectively improves the fine-grained alignment between generated motions and textual descriptions. Besides, the decoupled tokenizer and residual vector quantization enhance the quality and diversity of reconstructed motion.
- We propose the *MotionUPG* based on LLaMA, and design a series of pretraining and instruction-tuning tasks to effectively leverage LMM priors for improving motion-related tasks. Our method supports a wide range of tasks, demonstrating strong **zero-shot generation** capabilities, fine-grained text-motion **alignment**, and **multimodal joint driven** motion generation.

2 Related Work

2.1 Multimodal Driven Motion Generation

Humans naturally respond to stimuli like text, music, and speech through movement. Multimodal motion generation seeks to synthesize realistic motion conditioned on such inputs. Most prior works focus on single-modality tasks—e.g., text-to-motion, music-to-dance, or speech-to-gesture.

Diffusion models with classifier-free guidance have advanced controllability. MDM [48] and MotionDiffuse [60] pioneered diffusion-based text-to-motion synthesis, while GMD [21] and OmniControl [56] enhanced precision. In music-to-dance, EDGE [50] provides interactive choreography; LODGE [27] uses two-level diffusion for coherence and detail. GestureDiffuCLIP [2] aligns speech-driven gestures with semantics; SynTalker [5] integrates speech and text via contrastive learning.

Yet, these methods typically operate on a single modality and lack unified multimodal generalization.

Tokenizer + Auto-Regressive frameworks offer a unified alternative by translating discrete tokens across modalities. PoseGPT[38] is the first to discretize motion representations and perform autoregressive generation using GPT. TM2T [13] and T2M-GPT [59] handle text-to-motion, while HumanTOMATO [37] introduces Hierarchical GPT for full-body synthesis. Bailando [46] enhances rhythm alignment with Actor-Critic learning.

M^3GPT [39] encodes various modalities into discrete tokens for GPT-based generation [43], but lacks fused multimodal control—e.g., combining text with music or speech. In contrast, our method enables unified multimodal generation with fine-grained control via prompt injection.

2.2 Motion Tokenizer

Motion tokenizers represent 3D motion as continuous or discrete latent variables. Some use VAEs [23] for encoding into continuous spaces, aiding diffusion model training.

Others adopt VQ-VAE [52] to discretize motion, enabling structured multimodal learning and compatibility with Transformers. TM2T [13] uses CNN-based VQ-VAE, and T2M-GPT [59] employs EMA [47] and code reset to boost diversity.

MoMask [11] introduces RVQ-VAE for hierarchical tokenization—base layers for coarse features, residual layers for details achieving compact yet expressive encoding. TLControl [53] encodes body parts separately for part-level control. Based on part-wise motion encoding, ParCo[65] further introduces a Part Coordination Layer to enhance the quality of full-body motion generation.

However, these models lack detailed conditioning, limiting fine-grained semantics. We address this by injecting fine-grained textual prompts into the tokenizer and introducing residual token layers for richer motion representation.

2.3 Motion-Centric MLLMs

Multimodal large language models (MLLMs) like Flamingo [1], PaLM-E [9], and GPT-4o [40] have enabled powerful cross-modal reasoning across text, images, audio, and video.

In motion generation, MotionGPT [18] formulates motion as a discrete language via vector quantization, enabling unified text-motion modeling. Through large-scale pretraining and prompt tuning, it supports generation, captioning, and prediction.

MotionLLaMA [30] extends this idea by introducing a holistic tokenizer and employing LLaMA [49] for multimodal motion understanding and generation.

Yet, these methods often tokenize motion temporally, making it difficult for LLMs to align text with atomic motion units. This hinders generalization and often leads to inconsistencies between generated motions and input descriptions. Moreover, most methods only support single-modality conditioning, lacking the ability to synthesize motion from fused multimodal inputs.

2.4 LLM-Augmented Data for Motion Generation

Several recent works utilize large language models (LLMs) to enhance motion generation via textual data augmentation or instruction refinement. Action-GPT [20] enriches coarse action labels into diverse descriptions, FineMoGen [62] enables LLM-based motion editing, and SINC [3] synthesizes composite actions through text-guided motion composition. Fg-T2M++ [54] incorporates part-aware semantics for fine-grained motion control. FG-MDM[44] leverages ChatGPT to create part-wise motion descriptions, and then uses a diffusion model to generate the actual motion.

However, these methods often rely on external LLM augmentation rather than integrating LLMs into the core modeling process. Their generated descriptions are typically coarse or loosely aligned with motion structure, limiting controllability and generalization. In contrast, our method introduces a unified hybrid representation to better bridge text and motion, enabling stronger planning and fine-grained alignment.

3 Model Design

3.1 Overview

Instead of segmenting motions purely along the temporal axis as previous approaches do, we represent motions as hybrid sentences that combine atomic tokens of different body parts with fine-grained textual descriptions, guided by a specifically designed motion syntax to improve compatibility with large language models (LLMs).

To encode motions into high-quality atomic body-part tokens and enhance the fine-grained alignment between generated motions and textual descriptions, we propose Semantic-aware Decoupled Motion Tokenization (SDMT), detailed in Sec. 3.3. To obtain fine-grained descriptions for various types of motions, we introduce a new extraction method and construct the MotionWords dataset, described in Sec. 3.4. Finally, to fully leverage hybrid motion sentences and improve LLM performance on motion-related tasks, we design the MotionUPG Model, presented in Sec. 3.5.

3.2 Preliminary

Our model is designed based on our proposed MotionWords dataset, which not only contains motion data paired with text, speech, and music, but also provides fine-grained, atomic-level textual descriptions of the motions. Please refer to Section 3.4 for more details.

For the motion representation, we adopt the HumanTomato format, which extends the widely used HumanML3D representation by incorporating detailed finger movements and face expression. In this format, the motion $m \in \mathbb{R}^{T \times C}$, where T is the frame number and C is the dimension of motion feature that align with the HumanTomato representation. For the music and speech data, we use the original WAV format with a 16 kHz sampling rate. Both the speech and music are tokenized by the WavTokenizer[17].

3.3 Semantic-aware Decoupled Motion Tokenization

We first decouple the full-body motion into parts: torso, left leg, right leg, head, left arm, right arm, left/right hand (if available). Given that not all motion sequences involve detailed hand movements, we separately model the left and right hands using two vanilla VQ-VAEs [59], providing flexibility for handling diverse motion types.

For the six main body parts (excluding hands), we design six independent motion encoders, each responsible for extracting features specific to its corresponding body part. To model fine-grained details, we apply one three layers of Residual Vector Quantization (RVQ) [11] on the encoded features, resulting in six codebooks, each maintaining three layers of discrete tokens.

During decoding, the quantized feature of each body part is fused with its corresponding textual feature using a DiT block [41] to enhance part-level motion-text alignment. The features of all body parts are then concatenated and further aggregated with the global motion description feature to ensure the overall naturalness and coherence of the generated motion. A feed-forward network (FFN) layer outputs the body motion without the hands. Finally, by concatenating the hand motions reconstructed by the vanilla VQ-VAEs with the body motion, we obtain the final full-body motion.

3.4 Generate Fine-grained Motion Description

Although existing text-motion paired datasets provide motion descriptions, we argue that these descriptions are too coarse-grained, which limits the ability of multimodal large language models to learn fine-grained alignment between text and body motion.

Previous methods [18, 55] train LLMs for motion-to-text generation based on existing coarse-grained text-motion pairs. However, the generated motion descriptions remain coarse-grained. FG-MDM [45] uses a VLM by feeding the raw text-motion pairs into the VLM and prompting it to produce fine-grained motion descriptions [14]. Nevertheless, this approach can suffer from inaccurate descriptions due to hallucinations in large models [14].

To address these challenges, we propose a new approach to obtain fine-grained and accurate motion descriptions directly from motion data. Inspired by PoseScript [8], we first extract the rotation angles and relative distances between major body joints (torso, left leg, right leg, head, left arm, and right arm) from raw motion data, and then convert them into short atomic descriptions. For example, for the left arm, we calculate the angles between the left shoulder, left elbow, and left wrist, and generate an atomic description such as *"the left arm is bent greater than 90 degrees"*. We also compute the positions of key body joints, and when their L2-distance is less than 10 cm, we generate a positional description such as *"the left wrist and the right wrist are close"*. Specifically, if the joints are horizontally close but have a small vertical offset, we produce a description like *"the left wrist is nearly above the right wrist"*. Next, we feed the original motion data, the generated major body part descriptions, uniformly sampled motion frames, and

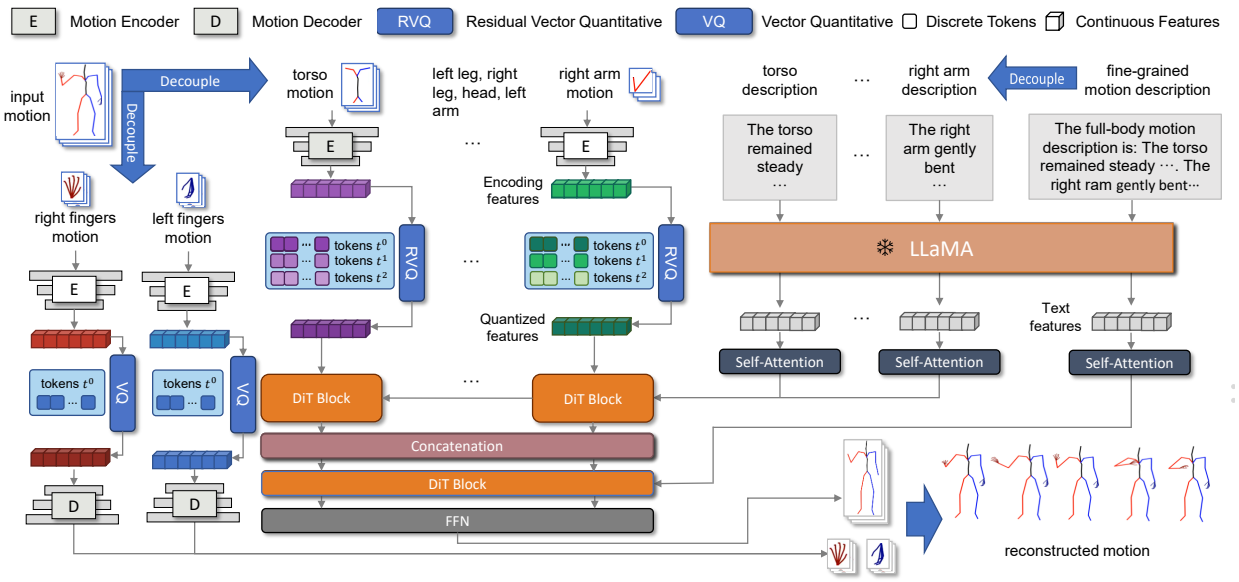

Figure 3: Overview of Semantic-aware Decoupled Motion Tokenization.

coarse-grained text captions into GPT-4o, along with a few demonstration examples, to guide GPT-4o in generating fine-grained and accurate motion descriptions. For text-paired datasets such as HumanML3D [12] and Motion-X [29], the coarse-grained captions come directly from the raw text provided in the datasets. For dance and gesture movements from AIST++ [26], FineDance [28] and BEAT2 [33], the coarse-grained text caption is simply set as *"This person is dancing/speaking"*. For each motion sequence, we generate three fine-grained motion descriptions.

Through this process, we construct the MotionWords dataset, which contains multimodal data including text, music, speech, and motion, accompanied by 68.2 million words of fine-grained motion descriptions that serve as a bridge to reduce the gap between different modalities.

3.5 MotionUPG Model

Based on the MotionWords dataset, we obtain paired multimodal condition and motion data, along with fine-grained motion descriptions. We first encode audio and text into their respective tokens using WavTokenizer[17] and LLaMA, and encode the motion into body-part-specific tokens using the proposed motion tokenization method. These motion part tokens are then combined with the fine-grained motion descriptions to form a hybrid sentence. Finally, we perform pretraining and fine-tuning based on the LLaMA model to enable motion generation conditioned on multimodal fusion.

The hybrid motion sentence. To reduce the gap between the large language model and motion tokens, and to enhance the fine-grained text-body part alignment, we propose to represent motion as a hybrid sentence. As shown in Fig. 4, we represent each motion instance using a combination of fine-grained motion descriptions and motion tokens. We introduce two special tokens, <SOM> and <EOM>, to explicitly mark the start and end of the motion tokens within the sequence. The motion tokens are ordered as torso, left leg, right leg, head, left arm, and right arm, followed by left hand and right hand. If hand motion is not available in the dataset, only the first six parts are included. Each of these six body part tokens is generated by 3 layers Residual Vector Quantization. For example, the N frame torso motion can be represented as $\{t[\text{torso}]_i^l | l = 0, 1, 2; i = 1, 2, , n\}, n = N/d$, where l is the layer index, i is the temporal index, d is the temporal down-sampling rate, $l = 3, d = 4$ in our experiments. By combining these motion tokens with the fine-grained motion descriptions, we construct hybrid motion sentences following the template in Fig. 4.

The pretraining stage. We develop our MotionUPG model based on LLaMA 3.2. We adopt the next token prediction paradigm for pretraining. The Hybrid Motion Sentences is structured as: <Fine-grained Text Tokens><SOM><Torso Tokens><Left Leg Tokens> ··· <Right Arm Tokens><Right Hand Tokens><EOM>. To further enhance the MotionUPG's understanding of the fine-grained mapping between text and motion and the motion struction, we perform pretraining on the following tasks:

- **FT2M:** Given <Fine-grained Text Tokens><SOM> as input, the network predicts the corresponding motion tokens.
- **M2FT:** Given <SOM><Motion Tokens><EOM><BOT> as input, the network predicts the corresponding fine-grained text.
- **M2M:** Given <SOM> followed by partial motion tokens, the network predicts the subsequent body motion tokens.

The multimodal tuning stage. Based on the MotionWords dataset and the MLLM architecture, we can train MotionUPG capable of performing various motion understanding and generation tasks. We focus primarily on enabling the MotionUPG to generate motion under multimodal and text-conditioned settings, including Coarse/Fine-grained Text-to-Motion, Speech&Text-to-Motion, Music&Text-to-Motion, Speech-to-Motion, and Music-to-Motion. To achieve these tasks, we design a series of instruction-tuning tasks, as shown in

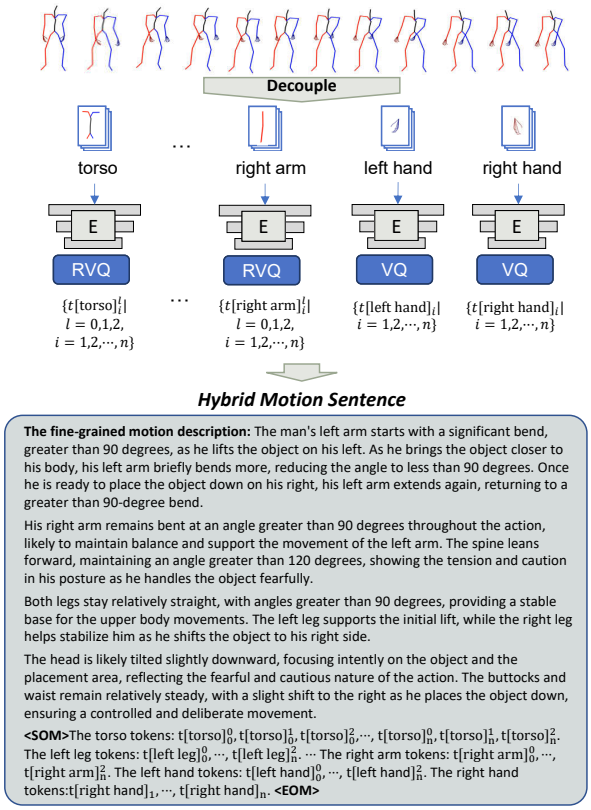

Figure 4: Illustration of the hybrid Motion Sentence. Each body part is tokenized separately. The atomic body-part motion tokens are combined with the paired fine-grained motion descriptions to construct the hybrid motion sentence.

Table 1, and fine-tune the MotionUPG based on the MotionWords dataset using LoRA [15].

The inference stage. During the generation stage, for T2Mo, S2G, and Mu2D tasks, we first combine the corresponding prompts from Table 1, convert them into a sequence of tokens, and input them into the LLM to generate fine-grained text descriptions that align with the given Coarse Text, Speech, or Music. These generated fine-grained text descriptions are then combined with the original conditions into the prompt to perform motion generation. By doing so, the model not only incorporates the original conditions but also engages in an additional reasoning step over the motion description, enabling the generation of more logically consistent motions.

For ST2G, and MuT2D tasks, We first generate fine-grained motion descriptions from the input speech or music signals. To incorporate user-specified control, we further edit the generated descriptions to reflect the desired textual instructions. The final motions are then generated based on the modified fine-grained descriptions together with the original audio input.

4 Experiment

4.1 Experimental Setup

Dataset. We train our model using the MotionWords dataset, which combines the text-to-motion datasets HumanML3D and Motion-X, the speech-to-gesture dataset BEAT2, and the music-to-dance datasets AIST++ and FineDance. All motion data are augmented with fine-grained motion descriptions using our proposed Motion2Text pipeline, enabling detailed motion-text alignment.

Implementation Details. For body parts, each RVQ-VAE adopts a three-layer residual quantization structure, with a codebook size of 1024×3 and a code dimension of 1024. For the two hands, we use a single-layer quantization, with a codebook size of 512 and a code dimension of 1024. Our LLM model is built upon the LLaMA 3.2 instruct 3b. We fine-tuning it on MotionWords dataset using LoRA with a rank of 64, alpha of 128 and dropout rate of 0.05. We train using AdamW optimizer with a learning rate of 2e-5, a batch size of 4, and a cosine learning rate scheduler. The pretraining stage runs for 800 epochs, and the instruction tuning stage for 1000 epochs cross tasks. Based on our design hybird motion sentence, the fine-grained motion descriptions can serve as a bridge representation across different modalities. This enables our method to better leverage heterogeneous multimodal data, leading to improved performance and allowing a unified model to achieve state-of-the-art results across multiple tasks.

4.2 Comparison with State-of-the-Art and Ablation Study.

Fine grained motion description production. We do quantitatively assess the quality of fine-grained motion description on the testset of HumanML3D. we uniformly sample frames from motion videos and input the corresponding textual descriptions into GPT-4o for evaluation. GPT-4o scores each description based on the following five criteria, with scores ranging from 1 (lowest) to 10 (highest):

- **Detail Accuracy:** Does the text accurately capture key action details observed in the frames? Are there any false or missing elements?
- **Body Part Coverage:** Does the description explicitly mention the relevant body parts involved in the motion? Is the coverage sufficiently specific?
- **Temporal Logic:** Does the description correctly reflect the sequence and continuity of the motion?
- **Naturalness of Expression:** Is the language fluent and naturally written?
- **Clarity and Professionalism:** Is the description precise and clear, and could it be suitable for professional contexts such as sports coaching, rehabilitation, or animation?

As shown in Table 2, compared to fine-grained motion descriptions generated directly by GPT-4o, our method achieves significant improvements across all evaluation aspects. This improvement is mainly attributed to the incorporation of body part bending descriptors and the original brief motion descriptions, which provide richer and more accurate motion cues than relying solely on GPT-4o generation.

Table 1: Some prompt templates designed for instruction tuning and inference.

Task	Input	Output
T2Mo	Can you generate a motion align with the following brief motion description and fine-grained motion description? The brief motion description is <coarse text>. The fine-grained motion description is <fine-grained text>.	<motion>
ST2G	Can you generate a motion align with the following speech and fine-grained motion description? The speech is <speech>. The fine-grained motion description is <fine-grained text>.	<motion>
MuT2D	Can you generate a motion align with the following music and fine-grained motion description? The music is <music>. The fine-grained motion description is <fine-grained text>.	<motion>
Mo2FT	Please use natural and detailed language to describe the motion illustrated in <motion>. Your description should include detailed movements of specific body parts — the torso, left leg, right leg, head, left arm, right arm.	<fine-grained text>
Mu2FT	Please provide a detailed motion description that aligns with the music illustrated in <music>. Your description should include detailed movements of specific body parts — the torso, left leg, right leg, head, left arm, right arm.	<fine-grained text>
S2FT	Please provide a detailed motion description that aligns with the speech illustrated in <speech>. Your description should include detailed movements of specific body parts — the torso, left leg, right leg, head, left arm, right arm.	<fine-grained text>

Table 2: Comparison of fine-grained motion description quality across different methods.

Method	Detail Accuracy ↑	Body Part Coverage ↑	Temporal Logic ↑	Naturalness ↑	Clarity ↑
GPT-4o	6.87	6.83	7.17	6.79	6.65
Ours	**7.62**	**7.97**	**7.82**	**7.67**	**7.76**

Table 3: Comparisons and ablation studies of our SDMT. We report the reconstruction FID on the Humanml3D test set.

Compare vs. SOTAs

Method	FID↓
TM2T [12]	0.307
M2DM [24]	0.063
T2M-GPT [59]	0.070
MoMask [11]	0.019
SDMT	0.012

Ablation of SDMT

VQ layers	Decouple	Semantic	FID↓
0			0.091
2			0.049
2	✓		0.017
2		✓	0.032
2	✓	✓	0.012

Ablation studies. As shown in Tab. 3, we conduct ablation studies on the proposed Semantic-aware Decoupled Motion Tokenization (SDMT). The results demonstrate that both the Decouple and Semantic components contribute significantly to reducing the FID score. Besides, our SDMT achieves a better reconstruction FID compared to the state-of-the-art MoMask. The ablation studies of MotionUPG can be found in Tables 4 and 5. MotionUPG(w/o FT, w/o SMT) denotes the variant that removes both the Fine-grained text and the Semantic-aware Tokenization. MotionUPG (w/ FT, w/o SMT) refers to the variant where text features are not injected into the decoder during the Semantic-aware Motion Tokenization stage, only use the decoupled Tokenization with residual layers.

Text-to-Motion. As shown in Table 4, we perform both comparison and ablation studies following the HumanML3D benchmark. R Precision and MultiModal Dist evaluates the alignment between motion and text, FID measures the overall motion quality, and MultiModality reflects the model's ability to generate diverse results under the same conditions. Our MotionUPG exhibits a significant advantage in R Precision and MultiModal Dist. According to the ablation study, this improvement is mainly attributed to the fine-grained motion descriptions. Semantic-aware Tokenization further enhances semantic alignment and improves motion quality. Among multi-modal methods, our approach achieves significant improvements across all metrics by incorporating fine-grained motion descriptions, which help reduce modality gaps.

Zero-shot Motion Generation. Our method effectively supports zero-shot motion generation. We observe that many atomic-level human motions are shared across different contexts. Leveraging the reasoning ability of large language models, we can generate fine-grained motion descriptions based on high-level textual inputs, and subsequently use our MotionUPGto produce condition-consistent motions. The high-level input can describe a scenario, such as "a person being chased by a bear," or specify an action outside the training distribution, such as "a person imitating a bird flying." Through this approach, we significantly expand the boundary of text-to-motion generation, enabling our model to handle diverse and unseen motion concepts.

Speech-to-Gesture. We follow the BEAT2 [33] benchmark to compare our method with other state-of-the-art approaches. FGD and Diversity evaluate the quality and diversity of the generated gestures, respectively. BC assesses speech-motion synchrony. As shown in Table 5, our method achieves the best performance on both FGD and Diversity, outperforming both single-modal and multi-modal baselines. Although the BC score is slightly lower than the best-performing method ProbTalk, our model maintains a strong level of speech-motion synchrony. This minor drop may be due to the incorporation of fine-grained textual information, which introduces additional semantic conditioning that could slightly weaken the dominance of audio features in motion generation. Nevertheless, our approach strikes an effective balance between gesture quality, diversity, and cross-modal alignment.

Music-to-Dance. We follow the FineDance [28] benchmark to compare our method with other state-of-the-art approaches. Specifically, FID_k and DIV_k are used to evaluate the quality and diversity of the generated dances, respectively [46]. BAS measures the degree of beat alignment between the music and the generated dance [26]. As shown in Table 6, our method achieves highly competitive performance across all evaluation metrics. Specifically, we achieve the

Table 4: Quantitative results of text-to-motion generation. The best scores are bold, and the second-best results are underlined. M indicates that the model supports multi-modal generation, while S indicates that the model only supports single-modal generation. The gray background highlights the best results among multi-modal generation methods.

Method	R Precision ↑			FID↓	MultiModal Dist↓	MultiModality↑
	Top 1	Top 2	Top 3			
Ground Truth	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	-
S MDM [61]	-	-	$0.611^{\pm.007}$	$0.544^{\pm.44}$	$5.566^{\pm.027}$	$2.799^{\pm.072}$
S ReMoDiffuse [61]	$0.510^{\pm.005}$	$0.698^{\pm.006}$	$0.795^{\pm.004}$	$0.103^{\pm.004}$	$2.974^{\pm.016}$	$1.795^{\pm.043}$
S MMM [42]	$0.504^{\pm.003}$	$0.696^{\pm.003}$	$0.794^{\pm.002}$	$0.080^{\pm.003}$	$2.998^{\pm.007}$	$1.164^{\pm.041}$
S MoMask [11]	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$\mathbf{0.045}^{\pm.002}$	$2.958^{\pm.008}$	$1.241^{\pm.040}$
M TM2D [10]	$0.319^{\pm.000}$	-	-	$1.021^{\pm.000}$	$4.099^{\pm.000}$	$\mathbf{4.139}^{\pm.000}$
M TM2T [13]	$0.424^{\pm.003}$	$0.618^{\pm.003}$	$0.729^{\pm.003}$	$1.501^{\pm.000}$	$3.467^{\pm.000}$	$2.424^{\pm.000}$
M MotionGPT [63]	$0.364^{\pm.005}$	$0.533^{\pm.003}$	$0.629^{\pm.004}$	$0.805^{\pm.002}$	$3.914^{\pm.013}$	$2.473^{\pm.041}$
M MotionGPT [18]	$0.492^{\pm.003}$	$0.681^{\pm.003}$	$0.733^{\pm.006}$	$0.232^{\pm.008}$	$3.096^{\pm.008}$	$2.008^{\pm.084}$
M MotionAgent [55]	$0.482^{\pm.004}$	$0.672^{\pm.003}$	$0.770^{\pm.002}$	$0.491^{\pm.019}$	$3.138^{\pm.010}$	-
M MotionCraft [4]	$0.501^{\pm.003}$	$0.697^{\pm.003}$	$0.796^{\pm.002}$	$0.173^{\pm.002}$	$3.025^{\pm.008}$	-
M MotionUPG(w/o FT, w/o SMT)	$0.516^{\pm.005}$	$0.704^{\pm.004}$	$0.812^{\pm.005}$	$0.094^{\pm.007}$	$2.937^{\pm.009}$	$2.515^{\pm.036}$
M MotionUPG(w/ FT, w/o SMT)	$\underline{0.532}^{\pm.004}$	$\underline{0.723}^{\pm.005}$	$\underline{0.835}^{\pm.006}$	$0.076^{\pm.006}$	$2.956^{\pm.011}$	$2.734^{\pm.045}$
M MotionUPG	$\mathbf{0.549}^{\pm.004}$	$\mathbf{0.736}^{\pm.003}$	$\mathbf{0.836}^{\pm.007}$	$\underline{0.069}^{\pm.006}$	$\mathbf{2.802}^{\pm.010}$	$\underline{2.821}^{\pm.055}$

Table 5: Quantitative results on BEAT2 test set. We report FGD $\times 10^{-1}$, BC $\times 10^{-1}$, and diversity.

Method	FGD ↓	BC ↑	Diversity ↑
Ground-Truth	–	0.703	11.97
S HA2G [35]	12.32	6.779	8.626
S DisCo [31]	9.417	6.439	9.912
S CaMN [34]	6.644	6.769	10.86
S TalkShow [58]	6.209	6.947	13.47
S ProbTalk [36]	6.170	**8.099**	10.43
S EMAGE [33]	5.512	7.724	13.06
S MambaTalk [57]	<u>5.366</u>	7.812	13.05
M SynTalker [6]	6.413	<u>7.971</u>	12.721
M MotionUPG (w/o FT, w/o SMT)	5.978	7.695	<u>12.751</u>
M MotionUPG (w/ FT, w/o SMT)	5.443	7.228	13.114
M MotionUPG	**5.224**	7.198	**13.219**

Table 6: Quantitative results of music-to-dance generation on the FineDance test set.

Method	FID_k ↓	DIV_k ↑	BAS↑
Ground Truth	-	-	0.2120
S FACT [26]	113.38	3.36	0.1831
S MNET [22]	104.71	3.12	0.1864
S Bailando [46]	82.81	7.74	0.2029
S EDEG [51]	94.34	<u>8.13</u>	0.2116
M M^3GPT [39]	86.47	7.75	0.2158
S Lodge [27]	**50.00**	5.67	**0.2269**
M MotionUPG	<u>73.18</u>	**8.43**	<u>0.2172</u>

second-best FID_k score, significantly outperforming most baselines such as M^3GPT and Bailando. In terms of diversity, thanks to our decoupled tokenization and the fine-grained text, we significantly improve the diversity of the generated dance.

5 Conclusion

In this work, we proposed the hypothesis that "A Motion is Worth a Hybrid Sentence." To validate it, we introduced the MotionWords dataset for collecting large-scale fine-grained motion descriptions, developed a Semantic-aware Decoupled Motion Tokenization method, and constructed Hybrid Motion Sentences to enable precise motion reconstruction. Furthermore, we presented MotionUPG, which is pre-trained on Hybrid Sentences and instruction-tuned on the multimodal data provided by MotionWords. Our experiments demonstrate that MotionUPG achieves strong fine-grained text-motion alignment, and exhibits impressive zero-shot generalization capabilities. It also successfully enables multimodal joint driven motion generation tasks, such as Text&Speech-to-Gesture and Text&Music-to-Dance. Ablation studies further demonstrate the superior reconstruction quality of our proposed SDMT module, as well as the significant improvements in motion generation tasks achieved by training the LMM with fine-grained motion descriptions. This work propose a large-scale dataset with fine-establishes a new paradigm for motion generation by effectively leveraging the reasoning capabilities of large language models, enhancing performance across motion generation tasks, and expanding the boundaries of motions that can be generated under the constraints of existing datasets.

Acknowledgments

This work was partly supported by the Peng Cheng Laboratory (PCL2023A10-2), and partly supported by the Shenzhen Key Laboratory of Next Generation Interactive Media Innovative Technology (ZDSYS20210623092001004).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.

[2] Tenglong Ao, Zeyi Zhang, and Libin Liu. 2023. GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents. *ACM Trans. Graph.* (2023), 18 pages. doi:10.1145/3592097

[3] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. 2023. SINC: Spatial Composition of 3D Human Motions for Simultaneous Action Generation. *ICCV* (2023).

[4] Yuxuan Bian, Ailing Zeng, Xuan Ju, Xian Liu, Zhaoyang Zhang, Wei Liu, and Qiang Xu. 2024. MotionCraft: Crafting Whole-Body Motion with Plug-and-Play Multimodal Controls. *arXiv preprint arXiv:2407.21136* (2024).

[5] Bohong Chen, Yumeng Li, Yao-Xiang Ding, Tianjia Shao, and Kun Zhou. 2024. Enabling Synergistic Full-Body Control in Prompt-Based Co-Speech Motion Generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. ACM, New York, NY, USA, 10. doi:10.1145/3664647.3680847

[6] Bohong Chen, Yumeng Li, Yao-Xiang Ding, Tianjia Shao, and Kun Zhou. 2024. Enabling synergistic full-body control in prompt-based co-speech motion generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 6774–6783.

[7] Ting Chen, Mostafa Dehghani, Vladimir Mirkes, Aditya Ramesh, Sean Welleck, et al. 2022. PaLI: A Jointly-Scaled Multimodal Model of Language and Vision. *arXiv preprint arXiv:2209.06794* (2022). https://arxiv.org/abs/2209.06794

[8] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. 2024. Posescript: Linking 3d human poses and natural language. *IEEE transactions on pattern analysis and machine intelligence* (2024).

[9] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. 2023. Palm-e: An embodied multimodal language model. (2023).

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

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

[12] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5152–5161.

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

[14] Xin He, Shaoli Huang, Xiaohang Zhan, Chao Weng, and Ying Shan. 2023. Semanticboost: Elevating motion generation with augmented textual cues. *arXiv preprint arXiv:2310.20323* (2023).

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[16] Puxin Huang, Noah Shinn, Zhenhai Zhang, Yujia Xiao, Aniruddha Sharma, Xuezhi Han, Xiaohua Zhan, Xinlei Li, Jianwei Zhang, Kai-Wei Chang, et al. 2023. Language Is Not All You Need: Aligning Perception with Language Models. *arXiv preprint arXiv:2302.14045* (2023). https://arxiv.org/abs/2302.14045

[17] Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, et al. 2024. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532* (2024).

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

[19] Yinhan Jiang, Maud Bibi, Naman Patel, Sharan Mathur, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[20] Sai Shashank Kalakonda, Shubh Maheshwari, and Ravi Kiran Sarvadevabhatla. 2023. Action-GPT: Leveraging Large-scale Language Models for Improved and Generalized Action Generation. arXiv:2211.15603 [cs.CV] https://arxiv.org/abs/2211.15603

[21] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. 2023. Guided Motion Diffusion for Controllable Human Motion Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2151–2162.

[22] Jinwoo Kim, Heeseok Oh, Seongjean Kim, Hoseok Tong, and Sanghoon Lee. 2022. A Brand New Dance Partner: Music-Conditioned Pluralistic Dancing Controlled by Multiple Dance Genres. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3490–3500.

[23] Diederik P Kingma, Max Welling, et al. 2013. Auto-encoding variational bayes.

[24] Hanyang Kong, Kehong Gong, Dongze Lian, Michael Bi Mi, and Xinchao Wang. 2023. Priority-centric human motion generation in discrete latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14806–14816.

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597* (2023). https://arxiv.org/abs/2301.12597

[26] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF international conference on computer vision*. 13401–13412.

[27] Ronghui Li, YuXiang Zhang, Yachao Zhang, Hongwen Zhang, Jie Guo, Yan Zhang, Yebin Liu, and Xiu Li. 2024. Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1524–1534.

[28] Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. 2023. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10234–10243.

[29] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. 2023. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems* 36 (2023), 25268–25280.

[30] Zeyu Ling, Bo Han, Shiyang Li, Hongdeng Shen, Jikang Cheng, and Changqing Zou. 2024. MotionLLaMA: A Unified Framework for Motion Synthesis and Comprehension. *arXiv preprint arXiv:2411.17335* (2024).

[31] Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis. In *Proceedings of the 30th ACM international conference on multimedia*. 3764–3773.

[32] Haotian Liu, Chunyuan Zhang, Yinan Du, and Jason Wang. 2023. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485* (2023). https://arxiv.org/abs/2304.08485

[33] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. 2024. 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*. 1144–1154.

[34] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*. Springer, 612–630.

[35] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. 2022. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10462–10472.

[36] Yifei Liu, Qiong Cao, Yandong Wen, Huaiguang Jiang, and Changxing Ding. 2024. Towards variable and coordinated holistic co-speech motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1566–1576.

[37] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. 2023. HumanTOMATO: Text-aligned Whole-body Motion Generation. *arxiv:2310.12978* (2023).

[38] Thomas Lucas*, Fabien Baradel*, Philippe Weinzaepfel, and Grégory Rogez. 2022. PoseGPT: Quantization-based 3D Human Motion Generation and Forecasting. In *European Conference on Computer Vision (ECCV)*.

[39] Mingshuang Luo, Ruibing Hou, Zhuo Li, Hong Chang, Zimo Liu, Yaowei Wang, and Shiguang Shan. 2024. M^3GPT: An Advanced Multimodal, Multitask Framework for Motion Comprehension and Generation. *arXiv preprint arXiv:2405.16273* (2024).

[40] OpenAI. 2023. GPT-4 Technical Report. https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

[41] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.

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

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

[44] Xu Shi, Chuanchen Luo, Junran Peng, Hongwen Zhang, and Yunlian Sun. 2023. Generating Fine-Grained Human Motions Using ChatGPT-Refined Descriptions. *arXiv preprint arXiv:2312.02772* (2023).

[45] Xu Shi, Wei Yao, Chuanchen Luo, Junran Peng, Hongwen Zhang, and Yunlian Sun. 2023. FG-MDM: Towards Zero-Shot Human Motion Generation via Fine-Grained Descriptions. *arXiv preprint arXiv:2312.02772* (2023).

[46] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. 2022. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11050–11059.

[47] James W Taylor. 2003. Exponential smoothing with a damped multiplicative trend. *International journal of Forecasting* 19, 4 (2003), 715–725.

[48] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=SJ1kSyO2jwu

[49] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Baptiste Rozière, Naman Goyal, Olivier Siméoni, Thomas Massart, Ludovic Denoyer, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023).

[50] Jonathan Tseng, Rodrigo Castellon, and C Karen Liu. 2022. EDGE: Editable Dance Generation From Music. *arXiv preprint arXiv:2211.10658* (2022).

[51] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. 2023. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 448–458.

[52] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).

[53] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. 2024. Tlcontrol: Trajectory and language control for human motion synthesis. In *European Conference on Computer Vision*. Springer, 37–54.

[54] Yin Wang, Mu Li, Jiapeng Liu, Zhiying Leng, Frederick W. B. Li, Ziyao Zhang, and Xiaohui Liang. 2025. Fg-T2M++: LLMs-Augmented Fine-Grained Text Driven Human Motion Generation. arXiv:2502.05534 [cs.CV] https://arxiv.org/abs/2502.05534

[55] Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. 2024. Motion-Agent: A Conversational Framework for Human Motion Generation with LLMs. *arXiv preprint arXiv:2405.17013* (2024).

[56] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. 2024. OmniControl: Control Any Joint at Any Time for Human Motion Generation. In *The Twelfth International Conference on Learning Representations*.

[57] Zunnan Xu, Yukang Lin, Haonan Han, Sicheng Yang, Ronghui Li, Yachao Zhang, and Xiu Li. 2024. Mambatalk: Efficient holistic gesture synthesis with selective state space models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[58] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. 2023. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 469–480.

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

[60] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. *arXiv preprint arXiv:2208.15001* (2022).

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

[62] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. 2023. FineMoGen: Fine-Grained Spatio-Temporal Motion Generation and Editing. *NeurIPS* (2023).

[63] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. 2024. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 7368–7376.

[64] Zixiang Zhou and Baoyuan Wang. 2023. Ude: A unified driving engine for human motion generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5632–5641.

[65] Qiran Zou, Shangyuan Yuan, Shian Du, Yu Wang, Chang Liu, Yi Xu, Jie Chen, and Xiangyang Ji. 2024. ParCo: Part-Coordinating Text-to-Motion Synthesis. *arXiv preprint arXiv:2403.18512* (2024).