MotionGPT: Finetuned LLMs are General-Purpose Motion Generators

Yaqi Zhang[1,2], **Di Huang**[4], **Bin Liu**[1,2*], **Shixiang Tang**[4], **Yan Lu**[4],
Lu Chen[5], **Lei Bai**[3], **Qi Chu**[1,2], **Nenghai Yu**[1,2], **Wanli Ouyang**[3]
[1]University of Science and Technology of China
[2]CAS Key Laboratory of Electromagnetic Space Information
[3]Shanghai AI Laboratory [4]The University of Sydney [5]Zhejiang University

Abstract

Generating realistic human motion from given action descriptions has experienced significant advancements because of the emerging requirement of digital humans. While recent works have achieved impressive results in generating motion directly from textual action descriptions, they often support only a single modality of the control signal, which limits their application in the real digital human industry. This paper presents a **Motion G**eneral-**P**urpose genera**T**or (MotionGPT) that can use multimodal control signals, *e.g.,* text and single-frame poses, for generating consecutive human motions by treating multimodal signals as special input tokens in large language models (LLMs). Specifically, we first quantize multimodal control signals into discrete codes and then formulate them in a unified prompt instruction to ask the LLMs to generate the motion answer. Our MotionGPT demonstrates a unified human motion generation model with multimodal control signals by tuning a mere 0.4% of LLM parameters. To the best of our knowledge, MotionGPT is the first method to generate human motion by multimodal control signals, which we hope can shed light on this new direction. Codes shall be released upon acceptance. Visit our webpage at `https://qiqiapink.github.io/MotionGPT/`.

1 Introduction

Human motion is pivotal in various applications such as video gaming, filmmaking, and virtual reality. Recent advancements in AI [41; 48; 38; 40; 39; 30; 25] have paved the way for novel approaches to motion creation, enabling various control conditions including textual descriptions, music pieces, and human poses. However, one significant shortcoming of existing works [32; 50; 43; 31; 52] is that they only target a single type of control condition, greatly limiting their applications in the real world, *e.g.,* unable to generate motion sequences conditioned on text descriptions and several keyframe human poses. To facilitate such applications, it is important to develop a unified human motion generation framework that can efficiently utilize multiple control signals simultaneously.

This paper proposes a novel and more unified framework for text-motion generation. The framework facilitates the generation of human motions using multiple control conditions, formulated as $output\_motion = f(text, task, input\_motion)$. Newly added inputs $task$ and $input\_motion$ represent the task and given motion prompts, respectively. Here, $task$ indicates the specific task the model should adapt to, while $input\_motion$ provides the keyframe poses corresponding to the given task. This framework is a departure from traditional text-motion generation models as the introduction of $input\_motion$ enables more precise control. For example, given an $input\_motion$ and set the $task$ as "generate motion given init poses", the model should compensate for the subsequent frames

*Corresponding author

Preprint. Under review.

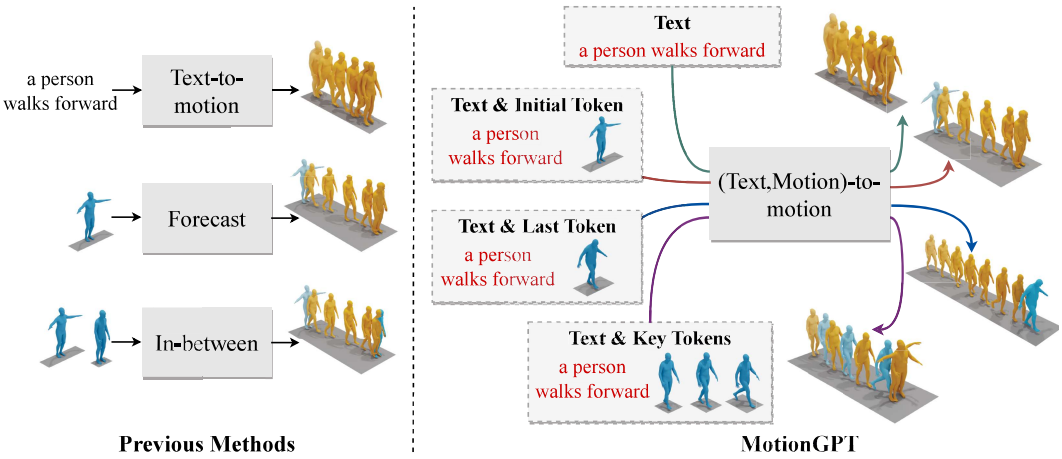

Figure 1: This work proposes a novel human motion generation method via fine-tuned LLMs, named **MotionGPT**. Compared with previous methods, MotionGPT has the unique ability to accept multiple control conditions and solve various motion generation tasks using a unified model.

of the given frames. Such a framework offers a more practical and comprehensive solution for human motion generation, where task instructions and multimodal conditions can flexibly control motion generation.

The challenge of building a model to complete such (text, motion)-motion generation task lies in understanding multimodal control conditions and generating human motions with varying motion lengths and richer patterns. We argue that these challenges can be naturally resolved by adapting from LLMs for the following reasons. First, recent studies have demonstrated that LLMs can understand multimodal inputs, *e.g.,* images [51; 6; 18; 22; 47] and videos [19], through a lightweight adapter [13]. Therefore, we expect the LLMs can also understand motion sequences with an appropriate adapter. Second, LLMs can provide diverse human motion contexts for motion generation because they have encoded diverse motion patterns from extensive large-scale text data (evidence shown in Supplementary Materials). This enables our motion generator fine-tuned from LLMs can produce motions with rich patterns. Third, since LLMs output tokens aggressively, producing human motion with flexible sequences is no longer an obstacle.

To this end, we propose a **Motion G**eneral-**P**urpose genera**T**or (MotionGPT) by fine-tuning an LLM following designed instructions. Specifically, MotionGPT first maps human poses into discrete motion codes via the pre-trained motion VQ-VAE and then generates instructions by combining codes from language prompts and motion prompts. The LLMs are fine-tuned by answering the correct human pose sequences to the instructions in an efficient way of well-known LoRA adaptation. The designed motion instruct tuning framework can incorporate pose sequence information into the fine-tuned large language model while taking advantage of strong motion priors in the original large language model.

We conduct extensive experiments on the HumanML3D [7] and KIT-ML [33] datasets, demonstrating MotionGPT has a strong ability for motion generation with multiple control conditions. Remarkably, MotionGPT achieves this with a significantly small set of training parameters (33 M), and in less training time (about 4 hours, or just 10% of the time taken by other methods). We observe that joint training under multiple control instructions outperforms training with a single type of control signal, showing the effectiveness of our unified motion generation training paradigm. Our contributions can be summarized as follows:

- We introduce a novel model, MotionGPT, for generating human motions, which allows for multiple types of control during the generation process. To the best of our knowledge, MotionGPT is the first method for using both text and poses as conditions. It supports generating subsequent, preceding, or 'in-betweening' motions using a single and unified model.
- We demonstrate that a pre-trained LLM can be readily tuned to function as a human motion generator , suggesting the potential for directly utilizing LLMs for human motion generation.

- We present a comprehensive set of experiments, showcasing the effectiveness of our proposed MotionGPT with multiple types of control signals. Experimental results also indicate that using a more powerful LLM results in superior motion generation quality, indicating that further advancements in LLM technology could substantially enhance the performance of MotionGPT in the future.

2 Related Work

Large language models Recently, large language models [5; 35; 36; 4; 29; 44] have been developed dramatically, *e.g.,* BERT [5], GPT [35], and Google T5 [37]. These models, such as GPT-4 [29], demonstrate exceptional performance on various linguistic tasks, thanks to the extensive training data (45 gigabytes in the case of GPT-4) and a large number of parameters they leverage. Previously, language models were task-specific, focusing on areas such as translation and sentiment analysis. However, recent developments, like ChatGPT [2], have expanded the capability of these models. Based on GPT-4, ChatGPT can interact with humans, showcasing its strong natural language understanding abilities. This effectiveness has opened up possibilities for a myriad of downstream tasks achieved through fine-tuning these LLMs. However, fine-tuning such models, considering their extensive parameters, is a challenging task. To address this issue, efficient fine-tuning strategies have been proposed, including prompt tuning [17; 23; 14], adapters [12; 11; 16], and LoRA [13]. Our work draws inspiration from the recent progress in LLMs, but it also addresses a distinct problem by introducing a new modality into the LLMs.

Human motion generation Motion generation [42; 10; 31; 21; 50; 9; 43; 32; 20] is a long-history task that can be conditioned on various conditions, such as motion description, actions, and music. For instance, HP-GAN [3] and [28] utilize a sequence-to-sequence model to anticipate future poses based on prior poses. ACTOR [31] employs a transformer VAE for both unconditional and action-based generation. TRAJEVAE [15], when supplied with an initial pose and a trajectory, can generate a motion sequence that follows the given path. In recent years, text-conditional motion generation has garnered significant attention. This approach focuses on generating human motion sequences that are conditioned on textual descriptions. TEMOS [32] proposes a VAE model that learns a shared latent space for both motion and text. MotionDiffuse [50] integrates a diffusion model into the text-to-motion generation framework and accomplishes impressive results. MDM [43], aiming to enhance motion-text consistency, uses CLIP [34] as the text encoder to incorporate more robust text priors into the model. In comparison to previous methods, our work, MotionGPT, stands out as the first unified motion generation model that supports multimodal controls.

3 MotionGPT: a Motion General-Purpose Generator

MotionGPT proposes a **Motion G**eneral-**P**urpose genera**T**or controlled by multimodal conditions, *i.e.,* texts and human poses in keyframes. Our motivation is to formulate human motion as a problem of asking the Large Language Model to generate desirable human motions according to task prompts and control conditions. Specifically, we quantize motion controls into discrete codes using the widely-used VQ-VAE [45] (Sec. 3.1). Motion discrete codes, text control conditions, and designed task instructions are then organized into a unified question template for the LoRA-finetuned LLM (Sec. 3.2) to generate a human motion sequence answer (Sec. 3.3). Following the typical framework of instruction tuning, we leverage cross-entropy loss to supervise the LoRA adapter. More importantly, our MotionGPT can address not only existing human motion generation tasks, *e.g.,* text-to-motion generation, but also new motion generation tasks by simply adjusting task instructions, showing the potential of MotionGPT as a generic baseline framework for motion generation (Sec. 3.4).

3.1 Motion Code Generation

VQ-VAE proposed in [45] enables the model to learn discrete representations for generative models. Given a human pose $\mathbf{m}$, the motion VQ-VAE can be trained by the reconstruction loss, the embedding loss and the commitment loss, *i.e.,*

$$\mathcal{L}_{\text{VQVAE}} = ||\mathcal{D}(\mathcal{E}(\mathbf{m})) - \mathbf{m}||^2 + \|\text{sg}[\mathcal{E}(\mathbf{m})] - \mathbf{e}\|_2^2 + \beta\|\mathcal{E}(\mathbf{m}) - \text{sg}[\mathbf{e}]\|_2^2, \tag{1}$$

where $\mathcal{E}$, $\mathcal{D}$ are the motion encoder and the motion decoder, respectively. Here, the estimated embedding $\mathbf{e}$ after qunatization can be found by searching the nearest embedding in a learnable

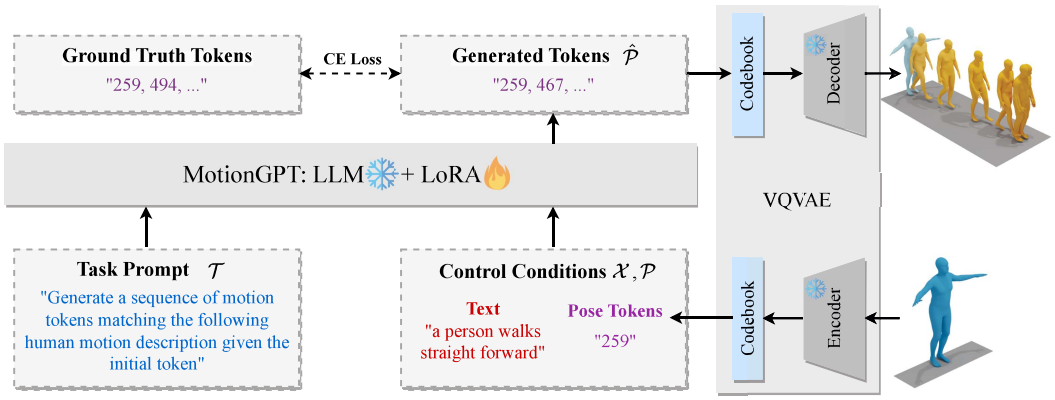

Figure 2: **The pipeline of MotionGPT**, a **Motion G**eneral-**P**urpose genera**T**or. Given text and poses as an input example, we organize task descriptions (Instruction) and multiple control conditions (Input) within a question template. MotionGPT fine-tunes an LLM to generate the corresponding motion answer, which can then be decoded into human motions using a VQ-VAE decoder.

codebook $\mathcal{B} = \{b_1, b_2, ..., b_N\}$, where N is the size of the codebook, which can be mathematically formulated as

$$\mathbf{e} = \arg\min_{b_k \in \mathcal{B}} \|\mathcal{E}(\mathbf{m}) - b_k\|_2. \tag{2}$$

Based on the estimation latent representation $\mathbf{e}$ of the motion $\mathbf{m}$, the reconstructed human pose $\hat{\mathbf{m}}$ can be produced by the decoder of VQVAE and the motion code p of human pose $\mathbf{m}$ can be calculated as the index of its nearest embedding in the codebook, *i.e.*,

$$\hat{\mathbf{m}} = \mathcal{D}(\mathbf{e}), \quad p = \arg\min_{k} \|\mathcal{E}(\mathbf{m}) - b_k\|_2. \tag{3}$$

3.2 Instruction Generation

In MotionGPT, we design instructions that combine task prompts and control conditions to enable (text, motion)-motion generation task. Specifically, given the task prompts $\mathcal{T} = \{t_1, t_2, ..., t_{n_t}\}$, text control conditions $\mathcal{X} = \{x_1, x_2, ..., x_{n_x}\}$ and pose control conditions $\mathcal{P} = \{p_1, p_2, ..., p_{n_p}\}$ where n_t, n_x and n_p is the number of discrete codes in $\mathcal{T}$, $\mathcal{X}$ and $\mathcal{P}$, the instruction $\mathcal{I}$ is formulated as

> % General control condition format
> Control Conditions: Text control condition $\mathcal{X}$ <$x_1, x_2, ..., x_{n_x}$> Pose control conditions $\mathcal{P}$ <$p_1, p_2, ..., p_{n_p}$>
> % General instruction format
> Instruction $\mathcal{I}$: {Task Prompts $\mathcal{T}$ <$t_1, t_2, ..., t_{n_t}$>} {Control Conditions}

Here, the pose control conditions $\mathcal{P} = \{p_1, p_2, ..., p_{n_p}\}$ presents pose codes, generated by using the same motion VQ-VAE as discussed in Sec. 3.1. Consequently, the entire instruction $\mathcal{I}$ can be regarded as a sequence of specialized text inputs. By generating different motion instructions, our MotionGPT can address existing human motion generation tasks and new human motion generations, which is detailed in Sec. 3.4.

3.3 Fine-tuning LLM by Motion Instructions

Instruction tuning [46] enables the LLM to handle various generation tasks by asking LLMs questions in different instructions. Therefore, we design various instructions that combine both task descriptions and control conditions to fine-tune large language model by the widely-used and efficient Low-Rank Adaptation (LoRA) [13]. Specifically, given a large language model $\mathcal{F}$, the general template of our instructions $\mathcal{I}$ and the answer of LLM $\hat{\mathcal{P}} = \mathcal{F}(\mathcal{I})$ are formulated as

> Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

% Task Prompts: Code sequences of Task Prompts
% Control Conditions: Code sequences of Control Conditions
Instruction $\mathcal{I}$: {Task Prompts $\mathcal{T}$} {Control Conditions}
Answer $\hat{\mathcal{P}}$: {Sequences of Human Motions }

The answer of LLM $\hat{\mathcal{P}} = \{\hat{p}_1, \hat{p}_2, ..., \hat{p}_{n_{\hat{p}}}\}$ is a series of generated motion codes, which can be decoded to human motion using Eq. 3.

Similar to most language models, we employ cross-entropy loss which constrains the similarity between estimated and ground-truth tokens, to finetune LLMs by LoRA, which can be presented as

$$\mathcal{L}_{lora} = \mathrm{CE}(\hat{\mathcal{P}}, \hat{\mathcal{P}^{gt}}), \tag{4}$$

where $\hat{\mathcal{P}^{gt}}$ is motion codes of ground-truth motions calculated by Eq. 3 and $\hat{\mathcal{P}}$ is the motion codes of predicted motion codes by the LLM $\mathcal{F}$.

3.4 Generalization to Existing and New Tasks

Leveraging the general template given in Sec. 3.2 and Sec. 3.3, our MotionGPT is capable of being a general-purpose motion generator, supporting various generation tasks. Specifically, for existing text-to-motion generation setting, MotionGPT address it by constructing following instruction $\mathcal{I}$:

Instruction ($\mathcal{I}$) : {Task Prompts: "Generate a sequence of motion tokens matching the following human motion description."} {Control Conditions: Text control condition $\mathcal{X}$ }

By adjusting instructions, MotionGPT can be easily adapted to multiple control conditions, *e.g.* text and an arbitrary number of human poses:

Instruction ($\mathcal{I}$) : {Task Prompts: "Generate a sequence of motion tokens matching the following human motion description given the init/last/key pose tokens."} {Control Conditions: Text control condition $\mathcal{X}$ <Motion Token> Pose control conditions $\mathcal{P}$ </Motion Token> }

4 Experiment

4.1 Datasets and Evaluation Metrics

HumanML3D HumanML3D [7] is currently the largest 3D human motion-language dataset, paired with well-annotated sequence-level textual descriptions. It contains 14,616 motion clips and 44,970 descriptions, composed from a vocabulary of 5,371 unique words. The motion sequences, sourced from the AMASS [26] and HumanAct12 [9] datasets, encompass a wide spectrum of human actions, including daily activities, sports, acrobatics, and artistic performances. Each motion clip is accompanied by 3-4 descriptive texts and has been downsampled to 20 fps, with a duration ranging from 2 to 10 seconds. The dataset is partitioned into training, validation, and test sets in an 80%, 5%, and 15% ratio, ensuring no overlap among the subsets.

KIT-ML The KIT-ML [33] dataset is comprised of 3,911 motion sequences along with 6,278 textual descriptions, averaging 9.5 words per description. This dataset is an amalgamation of selected subsets from the KIT WholeBody Human Motion Database [27] and the CMU Graphics Lab Motion Capture Database [1]. The motion sequences within KIT-ML have been downsampled to a rate of 12.5 fps, ensuring a uniform and manageable rate for analysis and experimentation.

Evaluation metrics Our evaluation comprises two categories of metrics. Firstly, to assess the quality of the generated motion, we adopt evaluation metrics consistent with previous methods. These include the *Frechet Inception Distance (FID)*, *Multi-modal Distance (MM Dist)*, *R-Precision* (calculating the Top-1/2/3 motion-to-text retrieval accuracy), and the *Diversity* metric. These metrics collectively provide a robust indication of both the realism and diversity of the generated motion.

Secondly, we introduce new metrics tailored to our proposed motion generation setting. Specifically, these metrics aim to measure the consistency between the provided pose conditions and the generated motion. For scenarios where the initial or final poses are given, the positioning of the corresponding generated poses in the motion sequence is critical. Hence, we propose the use of *Reconstruction*

Table 1: Evaluation of the effectiveness of **diverse controls** compared with state-of-the-art method MDM on **HumanML3D** and **KIT-ML** test set. **Bold** indicates the best result.

Method	Dataset	Initial token		Last token		Key tokens
		Recon ↓	Vel ↓	Recon ↓	Vel ↓	Dist ↓
MDM [43]	HumanML3D	31.04	1.370	23.45	1.577	2197
MotionGPT-13B (ours)		**13.78**	**0.549**	**6.831**	**0.397**	**55.59**
MDM [43]	KIT-ML	28.37	0.639	28.67	0.861	905.4
MotionGPT-13B (ours)		**25.17**	**0.422**	**35.43**	**0.673**	**77.58**

Table 2: Comparisons of **text-to-motion generation** with the state-of-the-art methods on **HumanML3D** and **KIT-ML** test set. MotionGPT-13B achieves comparable performance on all metrics. **Bold** and Underline indicate the best and the second best result.

Methods	HumanML3D			KIT-ML		
	FID ↓	MM Dist ↓	Diversity ↑	FID ↓	MM Dist ↓	Diversity ↑
Real motion	0.002	2.974	9.503	0.031	2.788	11.08
TEMOS [32]	3.734	3.703	8.973	3.717	3.417	10.84
TM2T [8]	1.501	3.467	8.589	1.501	3.467	8.589
T2M [7]	1.087	3.347	9.175	3.022	3.488	10.72
MotionDiffuse [50]	0.630	**3.113**	9.410	1.954	**2.958**	**11.10**
MDM [43]	**0.544**	5.566	**9.559**	**0.497**	9.191	10.85
MotionGPT-13B (Ours)	0.567	3.775	9.006	0.597	3.394	10.54

Loss (Recon) and *Velocity Loss (Vel)*, both measured by L2 loss, to evaluate the quality of pose reconstruction and its temporal continuity with neighboring poses. For scenarios where keyframe poses are provided, the positions of the corresponding generated poses within the motion sequence are unknown. Consequently, we calculate the Nearest Euclidean Distance for each key token relative to the corresponding ground truth poses, and report the *Average Distance (Dist)*. This approach allows us to quantitatively measure the accuracy of our model in reproducing the provided keyframe poses within the generated motion sequence.

4.2 Implement Details

Motion data pre-processing We follow the same data pre-processing method with [7]. Specifically, raw 3D motion coordinate is first transformed to make people face the Z+ direction, and subsequently pre-processed into motion features. These features include foot contact, global rotations and translations, local joint positions, velocities, and 6D rotations, having total dimensions of 263 for HumanML3D and 251 for KIT-ML.

Training details In our experiments, we utilize a frozen 13B LLaMA [44] model as the foundational LLM, which is subsequently fine-tuned using the LoRA technique. The model training process spans 37,500 epochs, starting with an initial learning rate of 3e-3. We set the batch size to 256, partitioned into micro-batches of 4 to accommodate memory constraints. We employ the AdamW optimizer [24] with a weight decay parameter of 0.01 to guide the optimization process. The training duration is approximately 4 hours for the HumanML3D dataset [7] and 3 hours for the KIT-ML dataset [33] when conducted on a single A100 GPU. These timelines highlight the efficiency of our training process compared to traditional methods. As for the pre-training of motion VQ-VAE [45], we follow the network structure and training strategy of [49], which is applied consistently across both datasets.

4.3 Comparisons for Motion Generation with Multiple Control Conditions

In this section, we conduct four different generation experiments with 1) text as the condition, 2) text and initial pose as the condition, 3) text and last pose as the condition, and 4) text and random keyframe pose as the condition. For both 2) and 3), we use 4 frame poses as the input pose condition; While for 4), we random sample 12 to 20 frame poses as the pose condition.

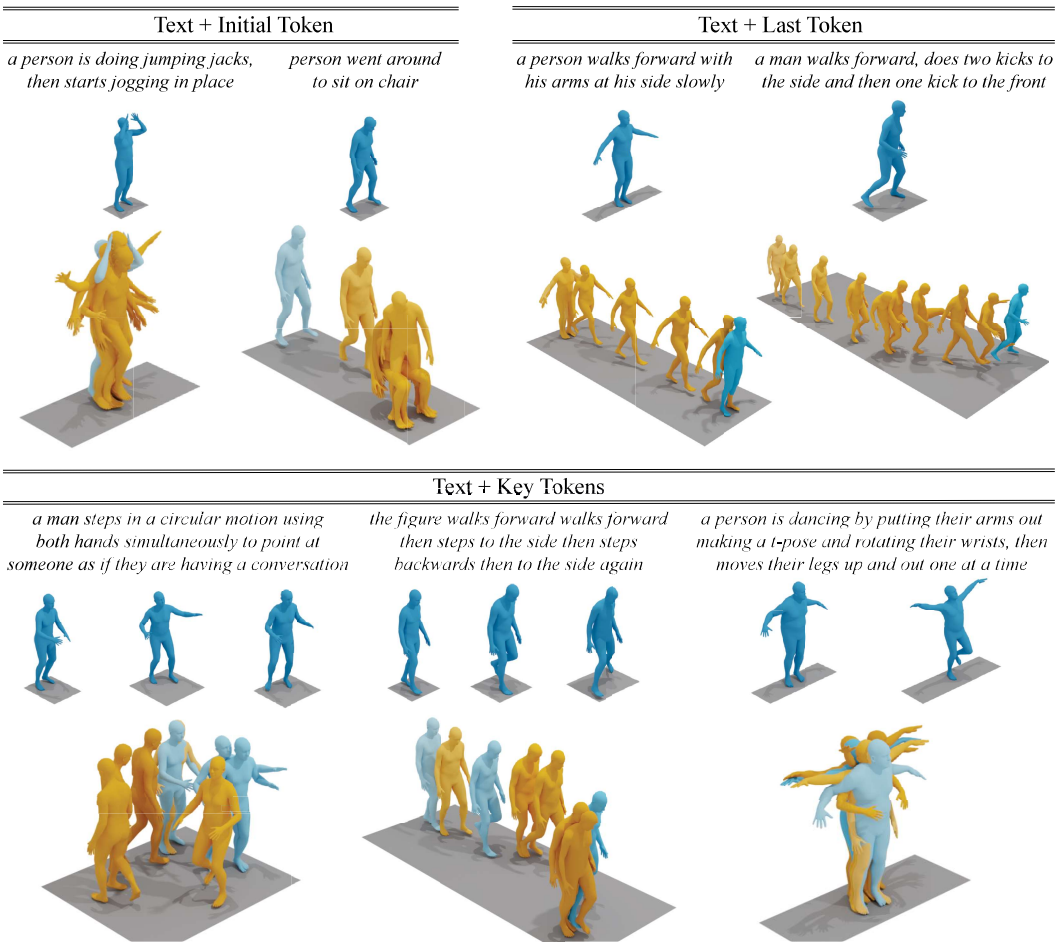

Figure 3: Generated motion by MotionGPT with multiple control conditions on HumanML3D.

Table 3: Motion generation quality on **HumanML3D** and **KIT-ML** test set for diverse control conditions.

Methods	HumanML3D			KIT-ML		
	FID ↓	MM Dist ↓	Diversity ↑	FID ↓	MM Dist ↓	Diversity ↑
Text-only	0.567	3.775	9.006	0.597	3.394	10.54
Text + Initial poses	0.520	3.844	9.588	0.664	3.445	10.39
Text + Last poses	0.591	3.718	9.251	0.856	3.336	10.58
Text + Random poses	0.367	3.598	9.176	0.671	3.411	10.76

Consistency with pose control conditions We evaluate the consistency between given pose controls and generated motion. The results are shown in Tab. 1. Our model, MotionGPT-13B, outperforms the state-of-the-art method, MDM [43], ensuring a higher degree of congruence between provided controls and generated motions across both HumanML3D [7] and KIT-ML [33] datasets. This superior performance is evidenced by the significantly reduced reconstruction and velocity errors for both initial and last tokens. Furthermore, MotionGPT-13B delivers a much lower distance error for key tokens, further illustrating its exceptional capability in accurately incorporating pose controls throughout the motion sequence. These observations highlight the efficacy of MotionGPT-13B in translating control conditions into corresponding motion sequences accurately, demonstrating its vast potential for applications in various scenarios.

Quality of generated motion The quantitative results of motion quality are depicted in Tab. 2 and Tab. 3. As illustrated in Tab. 2, our proposed model, MotionGPT, exhibits a performance that is competitive with state-of-the-art methods for text-to-motion generation. Specifically, MotionGPT

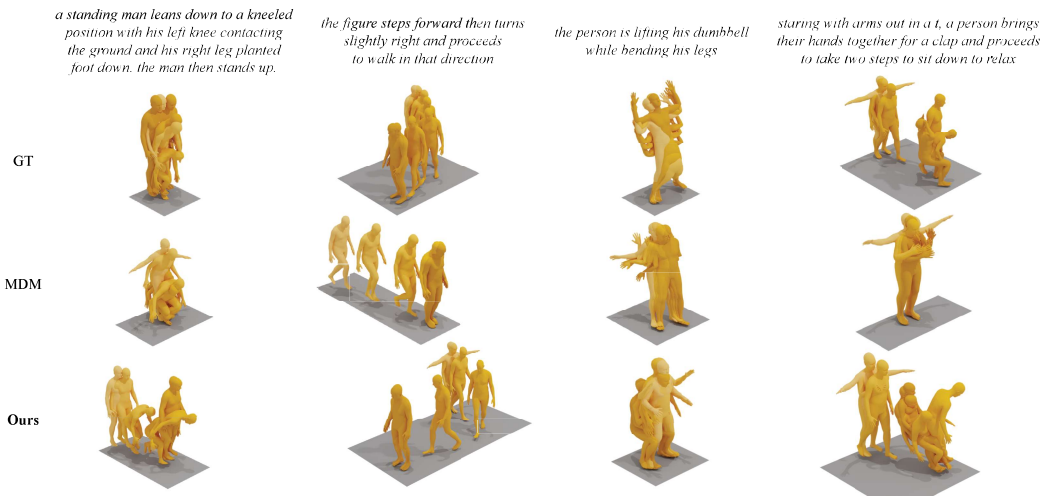

Figure 4: Qualitative comparison of the state-of-the-art motion generation method MDM with text-only conditions on HumanML3D.

consistently achieves comparable results across all metrics on both HumanML3D [7] and KIT-ML [33] datasets. In addition to text conditions, MotionGPT can also incorporate human poses as a secondary control modality and the motion quality results are demonstrated in Tab. 3. The adoption of additional control conditions, such as initial, last, or key tokens, does not compromise the quality of the generated motions. In some instances, such as when provided with initial or key tokens, MotionGPT even outperforms its text-only counterpart from 0.567 to 0.520 or 0.367 under FID metric on HumanML3D, demonstrating its robustness and flexibility in handling diverse control modalities. Nevertheless, a slight decrease in performance is observed when the model is given the final pose as input, which is in line with our expectations, as generating motions with a predetermined end pose presents an inherently greater challenge. Despite this, MotionGPT's performance remains commendable, further affirming its capability to generate high-quality, diverse motions under various control conditions.

We present visualization results in Fig. 3 and Fig. 4. As the Fig. 3 shown, the motions generated by our model exhibit a notable alignment with the provided poses, while also displaying a consistent adherence to the textual descriptions. For the text-to-motion generation task, we compare our model, MotionGPT, with the MDM, as depicted in Fig. 4. Our model demonstrates superior text-consistency and text-completeness compared to MDM [43]. The motions generated by the MDM model often tend to align with only the initial segment of the description, ignoring the latter half. In contrast, our approach exhibits a more comprehensive understanding of the motion descriptions by leveraging the powerful capabilities of LLMs, thus generating more complete and nuanced motion sequences.

4.4 Ablation Study

Additionally, extensive ablation studies are conducted on HumanML3D [7] validation set. Throughout the experiments, the MotionGPT-7B model is utilized unless otherwise specified.

Capability of pre-trained LLM Pre-trained LLMs can provide robust priors about human motion from texts. In this context, we experiment with base models pre-trained to varying degrees, including LLaMA-7B, LLaMA-13B, and LLaMA without pre-training. For the un-pretrained LLaMA, we adopt the same network structure as LLaMA-7B without loading the pre-trained weights. The randomly initialized LLaMA is tuned by LoRA as well, fixing weights during training. As demonstrated in Tab. 4, our results show a strong correlation between the level of pre-training in LLMs and the performance of our model in the text-to-motion generation task. This highlights the significant influence of motion prior extracted from LLM. Note that the training parameters of LoRA are same.

Hyper-parameters of LoRA During training, all the trainable parameters are sourced from LoRA [13], which has two hyper-parameters: r and α. The rank of LoRA parameters is represented by r, with smaller values indicating a fewer number of parameters. α controls the scale of

Table 4: Evaluation of text-to-motion generation using **different pre-trained LLaMA** on HumanML3D validation set. **Bold** indicates the best result.

Pre-trained Model	FID ↓	MM Dist ↓	R-Precision ↑			Diversity ↑
			Top-1	Top-2	Top-3	
LLaMA w/o pre-trained	26.01	8.445	0.032	0.067	0.106	**9.745**
LLaMA-7B	0.590	3.796	0.376	0.553	0.657	9.048
LLaMA-13B	**0.542**	**3.584**	**0.411**	**0.594**	**0.696**	9.311

Table 5: Evaluation of text-to-motion generation for **different LoRA parameters** on HumanML3D validation set using MotionGPT-7B. **Bold** and Underline indicate the best and the second best result.

r	α	FID ↓	MM Dist ↓	R-Precision ↑			Diversity ↑
				Top-1	Top-2	Top-3	
8	16	0.837	4.142	0.315	0.491	0.600	8.847
16	16	0.977	4.139	0.324	0.492	0.615	**9.745**
32	16	**0.576**	3.982	0.330	0.507	0.618	8.801
8	2	1.148	4.103	0.323	0.505	0.610	9.056
16	4	0.815	3.969	0.340	0.515	0.622	8.995
32	8	0.819	3.850	0.372	**0.555**	0.652	9.420
64	8	1.869	4.614	0.267	0.419	0.529	8.438
64	32	0.773	4.181	0.321	0.482	0.602	8.824
64	16	0.590	**3.796**	**0.376**	0.553	**0.657**	9.048

Table 6: **Comparisons between separate training for each task and joint training for multiple tasks** on HumanML3D validation set using MotionGPT-7B. Red and Blue indicate the improvement and decrement in the metric, respectively. Joint training can achieve better performance for all tasks.

Task	Training Strategy	FID ↓	MM Dist ↓	R-Precision ↑			Diversity ↑
				Top-1	Top-2	Top-3	
Text	Separate	0.670	4.267	0.299	0.469	0.577	9.745
+ Initial token		0.756	3.802	0.374	0.556	0.658	9.148
+ Last token		1.409	4.516	0.290	0.446	0.564	8.771
+ Key tokens		0.702	3.690	0.370	0.546	0.668	8.974
Text	Joint	$0.590^{-.180}$	$3.796^{-.471}$	$0.376^{+.077}$	$0.553^{+.084}$	$0.657^{+.080}$	$9.048^{-.697}$
+ Initial token		$0.493^{-.263}$	$3.750^{-.052}$	$0.384^{+.010}$	$0.564^{+.008}$	$0.666^{+.008}$	$9.378^{+.230}$
+ Last token		$0.646^{-.763}$	$3.675^{-.841}$	$0.393^{+.103}$	$0.577^{+.131}$	$0.681^{+.117}$	$9.030^{+.259}$
+ Key tokens		$0.390^{-.663}$	$3.492^{-.198}$	$0.416^{+.046}$	$0.597^{+.051}$	$0.713^{+.045}$	$9.621^{+.647}$

the outputs derived from the dense layer of LoRA. As illustrated in Tab. 5, we observe that the performance of our model improves across almost all metrics when we increase the value of r, keeping α constant. By maintaining the scale factor $\frac{\alpha}{r}$, which is comparable to the learning rate, we demonstrate that an increase in r leads to superior performance. Additionally, when α is modified while r is kept stable, we find that the optimal performance is achieved when α is set to 16.

Comparison with separate training To further evaluate the effectiveness of our unified motion generation approach, we conduct separate training for each task on the HumanML3D dataset [7]. The aim is to investigate if multi-task learning could improve the performance of individual control conditions. The comparison results are depicted in Table 6. We find that joint training across all tasks yields significant improvements in all metrics. This effect is especially pronounced when text and last poses are used as conditions. These findings underscore the utility of our unified motion generation approach. It appears that the model's ability to generate motions under a specific control type is boosted by the knowledge derived from other related control conditions.

5 Conclusion and Limitations

Conclusion This study introduces MotionGPT, a novel method capable of generating human motion using multimodal control signals, such as text and single-frame poses. The approach effectively discretizes pose conditions and creates a unified set of instructions by combining codes from both

textual and pose prompts. With MotionGPT, we envision a path toward more practical and versatile motion generation systems, offering a fresh perspective in the field.

Limitations Although current MotionGPT may support any control modalities beyond current human poses and text, this paper only validates the effectiveness on text and human poses. Validating our MotionGPT on a broader spectrum of possible modalities, such as music pieces, would be highly beneficial to more applications in the real world.

References

[1] Cmu graphics lab motion capture database. `http://mocap.cs.cmu.edu/`.

[2] Openai. (2023). chatgpt (mar 14 version) [large language model]. `https://chat.openai.com/chat/`.

[3] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1418–1427, 2018.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Yuqing Du, Ksenia Konyushkova, Misha Denil, Akhil Raju, Jessica Landon, Felix Hill, Nando de Freitas, and Serkan Cabi. Vision-language models as success detectors. *arXiv preprint arXiv:2303.07280*, 2023.

[7] 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.

[8] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 580–597. Springer, 2022.

[9] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020.

[10] Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. A recurrent variational autoencoder for human motion synthesis. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[11] Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. On the effectiveness of adapter-based tuning for pretrained language model adaptation. *arXiv preprint arXiv:2106.03164*, 2021.

[12] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[14] Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. *arXiv preprint arXiv:2108.02035*, 2021.

[15] Kacper Kania, Marek Kowalski, and Tomasz Trzciński. Trajevae: Controllable human motion generation from trajectories. *arXiv preprint arXiv:2104.00351*, 2021.

[16] Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. Lightweight adapter tuning for multilingual speech translation. *arXiv preprint arXiv:2106.01463*, 2021.

[17] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[19] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

[20] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021.

[21] Zimo Li, Yi Zhou, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. *arXiv preprint arXiv:1707.05363*, 2017.

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[23] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[25] Zeyu Lu, Di Huang, Lei Bai, Xihui Liu, Jingjing Qu, and Wanli Ouyang. Seeing is not always believing: A quantitative study on human perception of ai-generated images. *arXiv preprint arXiv:2304.13023*, 2023.

[26] 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.

[27] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The kit whole-body human motion database. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 329–336. IEEE, 2015.

[28] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017.

[29] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[30] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[31] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021.

[32] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 480–497. Springer, 2022.

[33] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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

[36] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[37] 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. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[42] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022.

[43] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[44] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

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

[46] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[47] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023.

[48] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[49] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052*, 2023.

[50] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[51] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[52] Wenlin Zhuang, Congyi Wang, Jinxiang Chai, Yangang Wang, Ming Shao, and Siyu Xia. Music2dance: Dancenet for music-driven dance generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2):1–21, 2022.

Supplementary Material

In this supplementary material, we provide additional experiments (Sec. A) and more visualization results (Sec. B).

A Additional Experiments

To further demonstrate the effectiveness of our model, we conducted several additional experiments on the HumanML3D validation set for text-to-motion generation, employing the MotionGPT-7B model architecture.

A.1 Evaluation of batch size

We conducted an evaluation of the performance of the MotionGPT-7B model trained with different batch sizes, and the results are presented in Table 7. It can be observed that the performances for batch sizes of 128 and 512 are comparable, while the batch size of 256 significantly outperforms the others across nearly all metrics.

Table 7: Evaluation of text-to-motion generation for MotionGPT-7B training with different batch sizes on HumanML3D validation set.

Batch Size	FID ↓	MM Dist ↓	R-Precision ↑			Diversity ↑
			Top-1	Top-2	Top-3	
128	0.752	4.063	0.314	0.491	0.612	**9.100**
256	**0.590**	**3.796**	**0.376**	**0.553**	**0.657**	9.048
512	0.684	4.010	0.311	0.495	0.611	8.947

A.2 Evaluation of prompt design

LLMs are known to be sensitive to prompts, emphasizing the criticality of carefully designing prompts to optimize model performance. In this section, we delve into the impact of employing two alternative prompts and assess their respective performances. Denoting the prompt used in our model as V_0, we also introduce two additional prompts, namely V_1 and V_2, as follows:

% Prompts V_1
Human motion can be represented by token indices by VQ-VAE. Below is an instruction that describes human motion generation condition types, paired with an input that provides specific conditions. Write a sequence of tokens matching with given conditions.

Instruction ($\mathcal{I}$) : {Task Prompts: "Motion description(and the init/last/key pose tokens)."} {Control Conditions: Text control condition $\mathcal{X}$(<Motion Token> Pose control conditions $\mathcal{P}$ </Motion Token>) }

% Prompts V_2
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction ($\mathcal{I}$) : {Task Prompts: "Generate the token sequence of the given human motion description(under the premise of the given init/last/key pose tokens)."} {Control Conditions: Text control condition $\mathcal{X}$(<Motion Token> Pose control conditions $\mathcal{P}$ </Motion Token>) }

For the prompts V_1, we incorporated specific human motion generation details into the overall descriptions, while simplifying the task prompts to only include condition types. On the other hand, for the prompts V_2, we modified the expression of the task prompts. The comparison results between these prompts are presented in Tab. 8, highlighting the efficiency and effectiveness of our proposed prompt designs. These findings underscore the significance of well-designed prompts in enhancing the performance of our model.

Table 8: Evaluation of text-to-motion generation for MotionGPT-7B applying different prompts on HumanML3D validation set.

Prompts	FID ↓	MM Dist ↓	R-Precision ↑			Diversity ↑
			Top-1	Top-2	Top-3	
V_1	8.506	5.490	0.200	0.331	0.447	7.566
V_2	3.018	4.858	0.249	0.402	0.508	8.237
V_0 (Ours)	**0.590**	**3.796**	**0.376**	**0.553**	**0.657**	**9.048**

B Qualitative Results

In this section, we showcase additional qualitative results generated by MotionGPT-13B for all four different control conditions. These results are presented in Figure 5, Figure 6, Figure 7, and Figure 8, respectively. The motion descriptions are sourced from the HumanML3D test set, and the pose control conditions are highlighted in blue. These visual examples offer further insights into the capabilities and performance of our model in generating motions based on different control conditions.

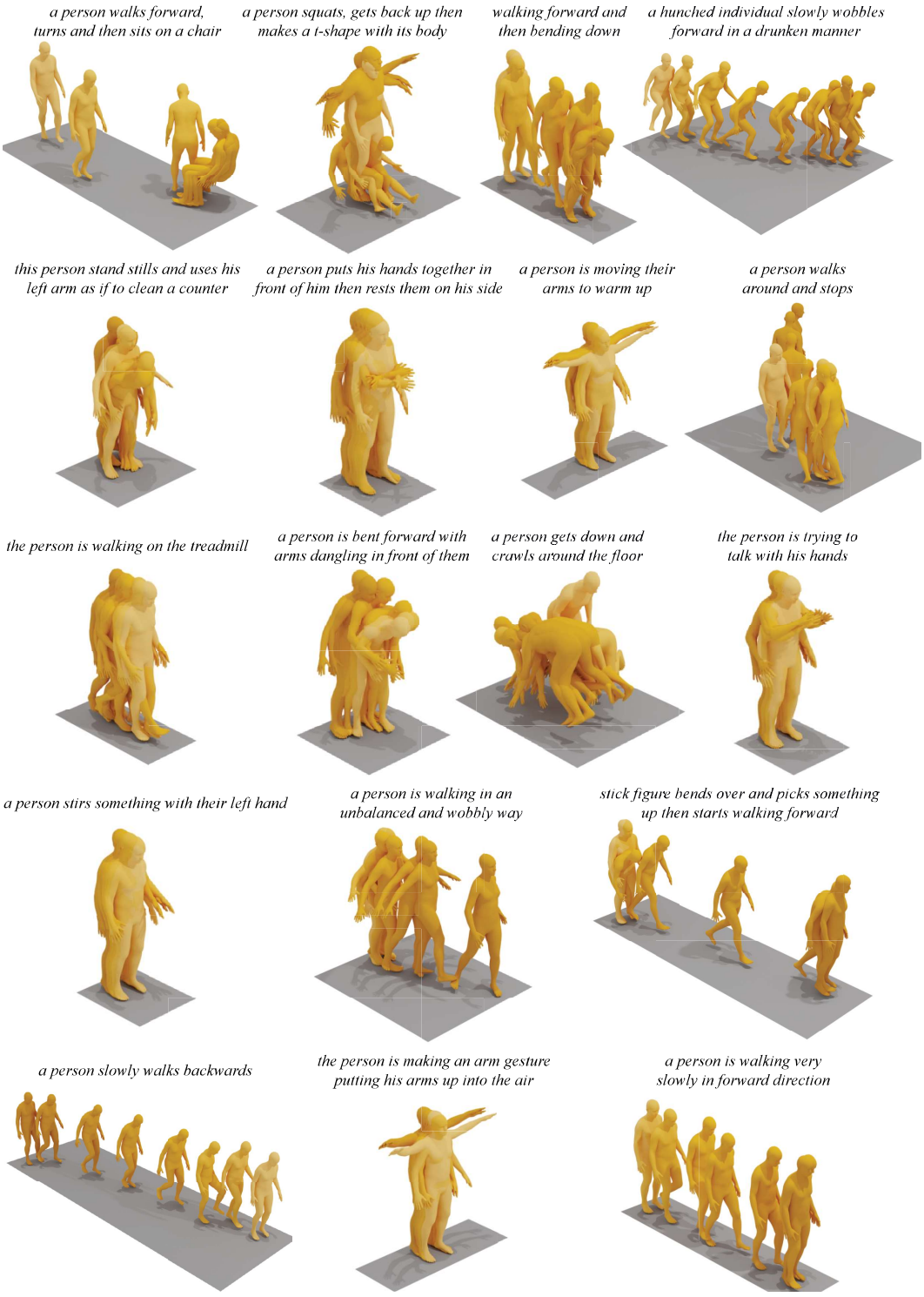

Figure 5: More text-to-motion samples generated by MotionGPT-13B using texts from the HumanML3D test set.

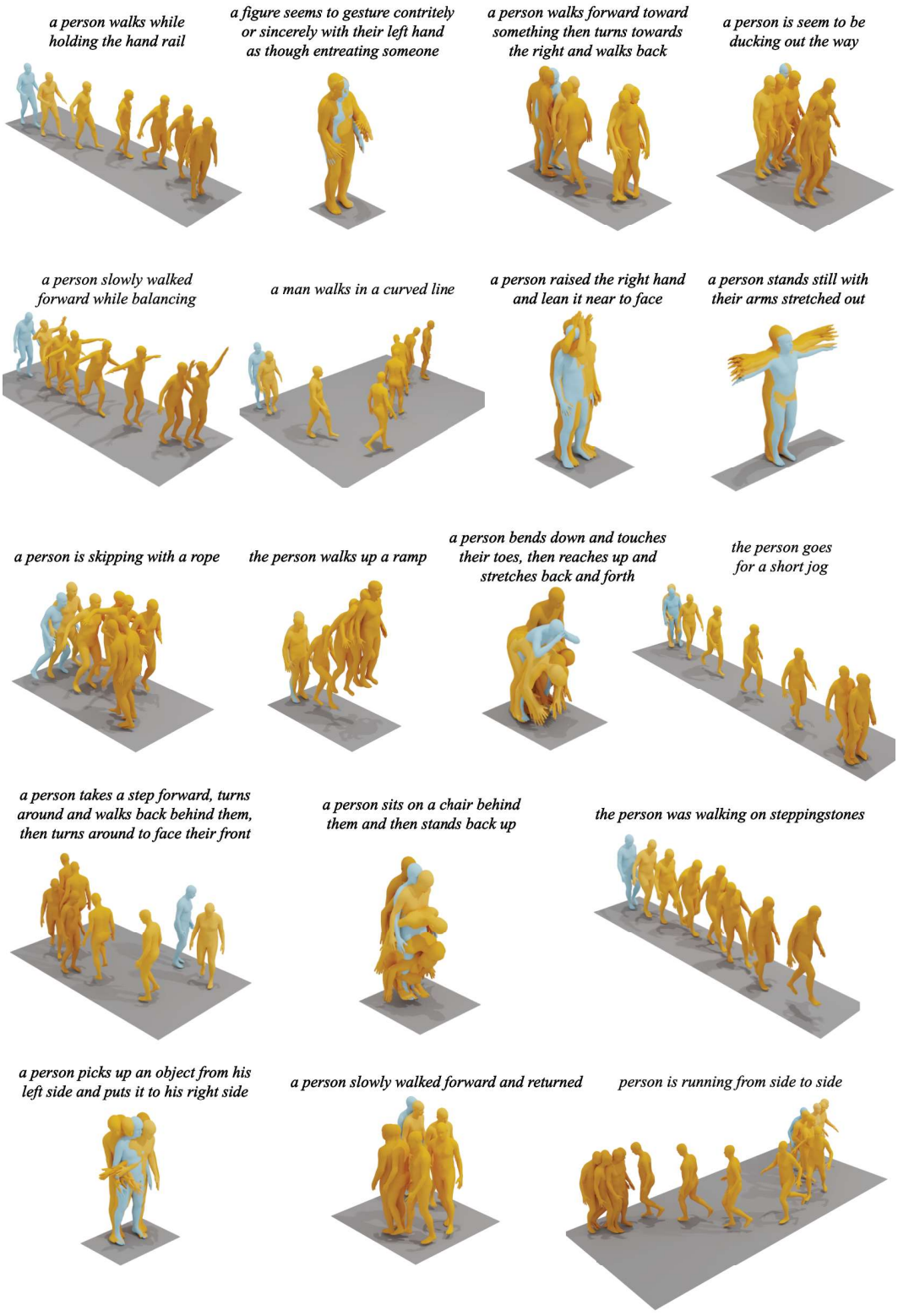

Figure 6: More (text+initial token)-to-motion samples generated by MotionGPT-13B using texts from the HumanML3D test set. The initial pose condition is highlighted in blue.

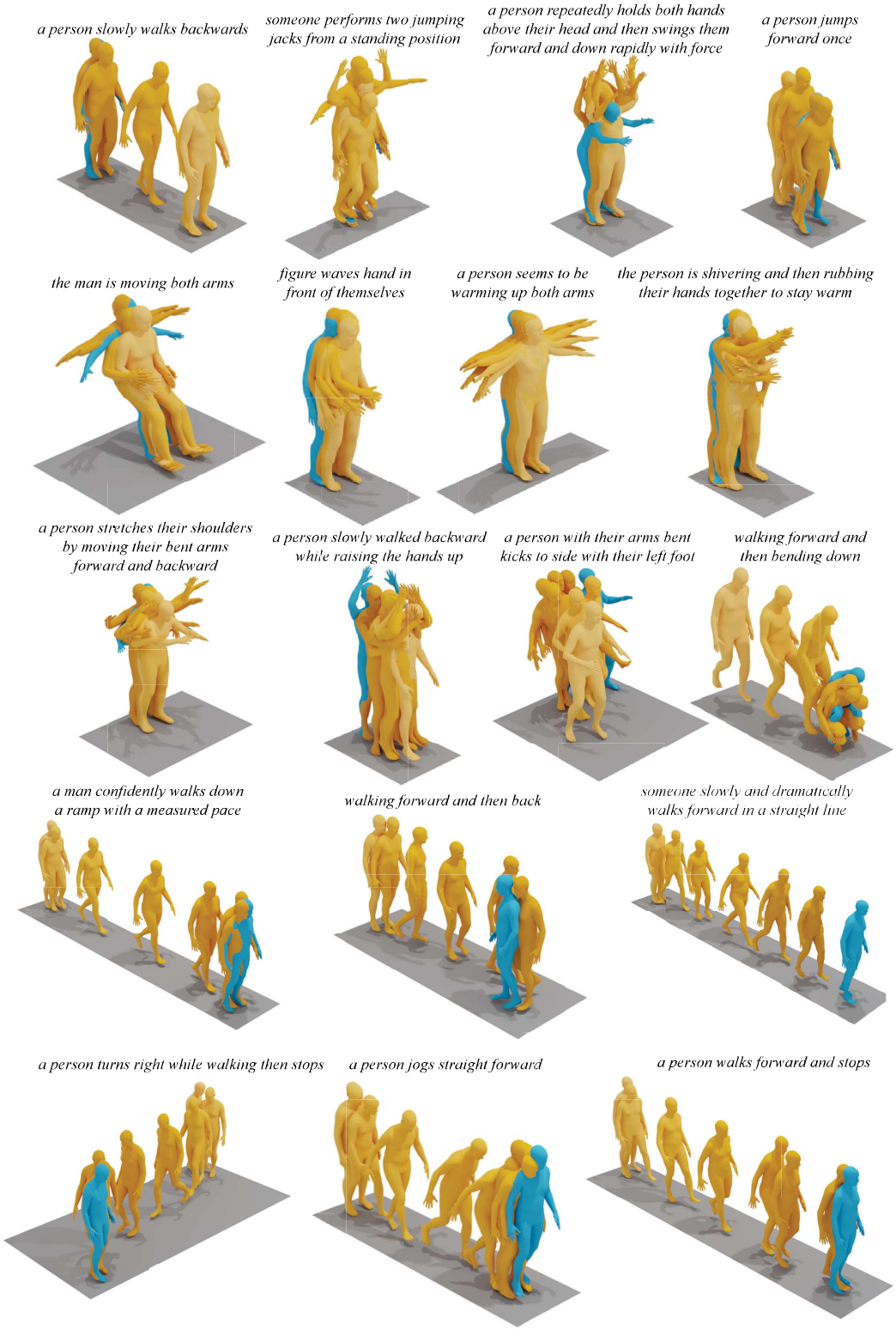

Figure 7: More (text+last token)-to-motion samples generated by MotionGPT-13B using texts from the HumanML3D test set. The last pose condition is highlighted in blue.

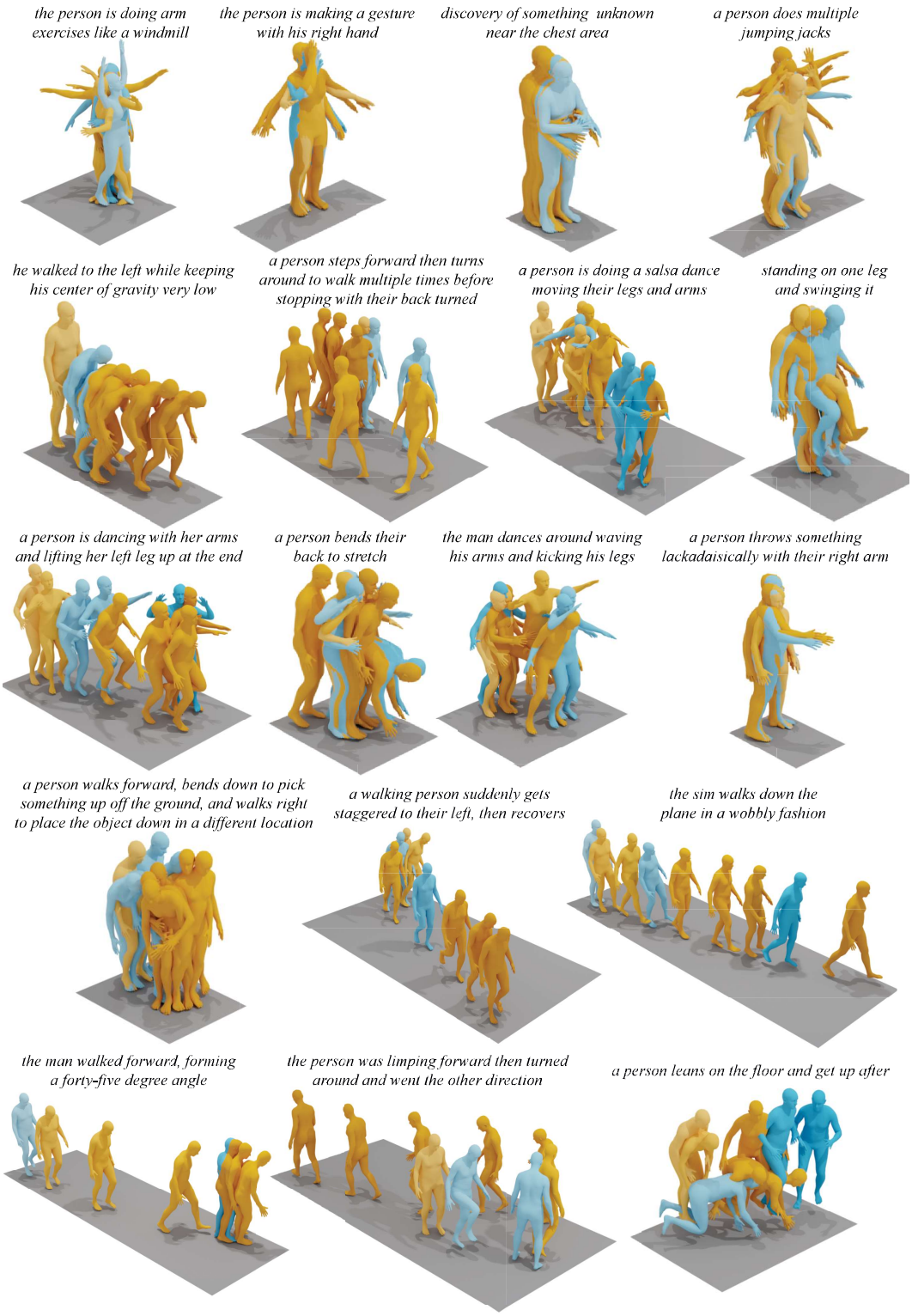

Figure 8: More (text+key tokens)-to-motion samples generated by MotionGPT-13B using texts from the HumanML3D test set. The key pose conditions are highlighted in blue.