MotionGPT: Finetuned LLMs Are General-Purpose Motion Generators

Yaqi Zhang[1,2], **Di Huang**[3], **Bin Liu**[1,2*], **Shixiang Tang**[3], **Yan Lu**[3], **Lu Chen**[4], **Lei Bai**[4], **Qi Chu**[1,2], **Nenghai Yu**[1,2], **Wanli Ouyang**[4]

[1]School of Cyber Science and Technology, University of Science and Technology of China
[2]CAS Key Laboratory of Electromagnetic Space Information
[3]The University of Sydney
[4]Shanghai AI Laboratory
zhangyq99@mail.ustc.edu.cn, flowice@ustc.edu.cn

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. Visit our webpage at https://qiqiapink.github.io/MotionGPT/.

Introduction

Human motion is pivotal in various applications such as video gaming, filmmaking, and virtual reality. Recent advancements in AI (Saharia et al. 2022; Yu et al. 2022; Ramesh et al. 2022; Rombach et al. 2022; Ramesh et al. 2021; Ouyang et al. 2022; Lu et al. 2023) 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 (Petrovich, Black, and Varol 2022; Zhang et al. 2022; Tevet et al. 2023; Petrovich, Black, and Varol 2021; Zhuang et al. 2022) 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 initial poses", the model should compensate for the subsequent frames 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 (Zhu et al. 2023; Du et al. 2023; Li et al. 2023a; Liu et al. 2023; Ye et al. 2023) and videos (Li et al. 2023b), through a lightweight adapter (Hu et al. 2021a). 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. 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

*Corresponding author.

LoRA adaptation. The designed motion instruction 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 (Guo et al. 2022a) and KIT-ML (Plappert, Mandery, and Asfour 2016) 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.

Related Work

Large language models Recently, large language models (Devlin et al. 2018; Radford et al. 2018, 2019; Brown et al. 2020; OpenAI 2023; Touvron et al. 2023) have been developed dramatically, *e.g.,* BERT (Devlin et al. 2018), GPT (Radford et al. 2018), and Google T5 (Raffel et al. 2020). These models, such as GPT-4 (OpenAI 2023), demonstrate exceptional performance on various linguistic tasks, thanks to the extensive training data (45 gigabytes in the case of GPT-4) and the 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, 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 (Lester, Al-Rfou, and Constant 2021; Liu et al. 2021; Hu et al. 2021b), adapters (Houlsby et al. 2019; He et al. 2021; Le et al. 2021), and LoRA (Hu et al. 2021a). 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 (Tevet et al. 2022; Habibie et al. 2017; Petrovich, Black, and Varol 2021; Li et al. 2017; Zhang et al. 2022; Guo et al. 2020; Tevet et al. 2023; Petrovich, Black, and Varol 2022; Li et al. 2021) is a long-history task that can be conditioned on various conditions, such as motion description, actions, and music. For instance, HP-GAN (Barsoum, Kender, and Liu 2018) and (Martinez, Black, and Romero 2017) utilize a sequence-to-sequence model to anticipate future poses based on prior poses. ACTOR (Petrovich, Black, and Varol 2021) employs a transformer VAE for both unconditional and action-based generation. TRAJEVAE (Kania, Kowalski, and Trzciński 2021), 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 conditioned on textual descriptions. TEMOS (Petrovich, Black, and Varol 2022) proposes a VAE model that learns a shared latent space for both motion and text. MotionDiffuse (Zhang et al. 2022) integrates a diffusion model into the text-to-motion generation framework and accomplishes impressive results. MDM (Tevet et al. 2023), aiming to enhance motion-text consistency, uses CLIP (Radford et al. 2021) 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.

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 (Van Den Oord, Vinyals et al. 2017). Motion discrete codes, text control conditions, and designed task instructions are then organized into a unified question template for the LoRA-finetuned LLM to generate a human motion sequence answer. 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.

Motion Code Generation

VQ-VAE proposed in (Van Den Oord, Vinyals et al. 2017) enables the model to learn discrete representations for generative models. Given a human pose $\mathbf{m}$, the motion VQ-VAE

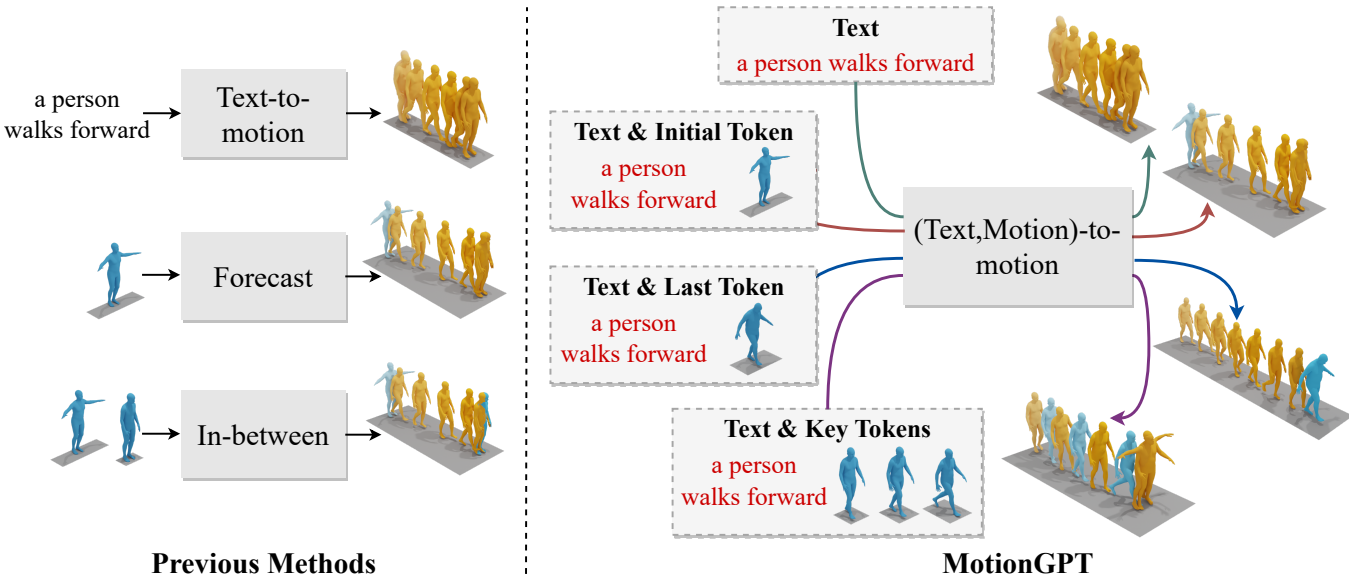

Figure 1: This work proposes a novel human motion generation method via fine-tuned LLMs, named MotionGPT. Compared with previous methods, MotionGPT has the unique ability to accept multiple control conditions and solve various motion generation tasks using a unified model.

can be trained by the reconstruction loss, the embedding loss and the commitment loss, *i.e.,*

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

where $\mathcal{E}, \mathcal{D}$ are the motion encoder and the motion decoder, respectively. sg indicates the stop gradient operation. Here, the estimated embedding $\mathbf{e}$ after qunatization can be found by searching the nearest embedding in a learnable codebook $\mathcal{B} = \{b_1, b_2, ..., b_N\}$, where N is the size of the codebook, which can be mathematically formulated as

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

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 VQ-VAE 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.,*

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

Instruction Generation

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

> % General control conditions format
> Control Conditions: {Text control conditions $\mathcal{X}$ $<x_1, x_2, ..., x_{n_x}>$} {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 mentioned earlier. 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.

Fine-tuning LLM by Motion Instructions

Instruction tuning (Wei et al. 2021) enables LLMs to handle various generation tasks by asking the LLM 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) (Hu et al. 2021a). Specifically, given a large language model $\mathcal{F}$, the general template of our instructions $\mathcal{I}$ and the answer of the 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

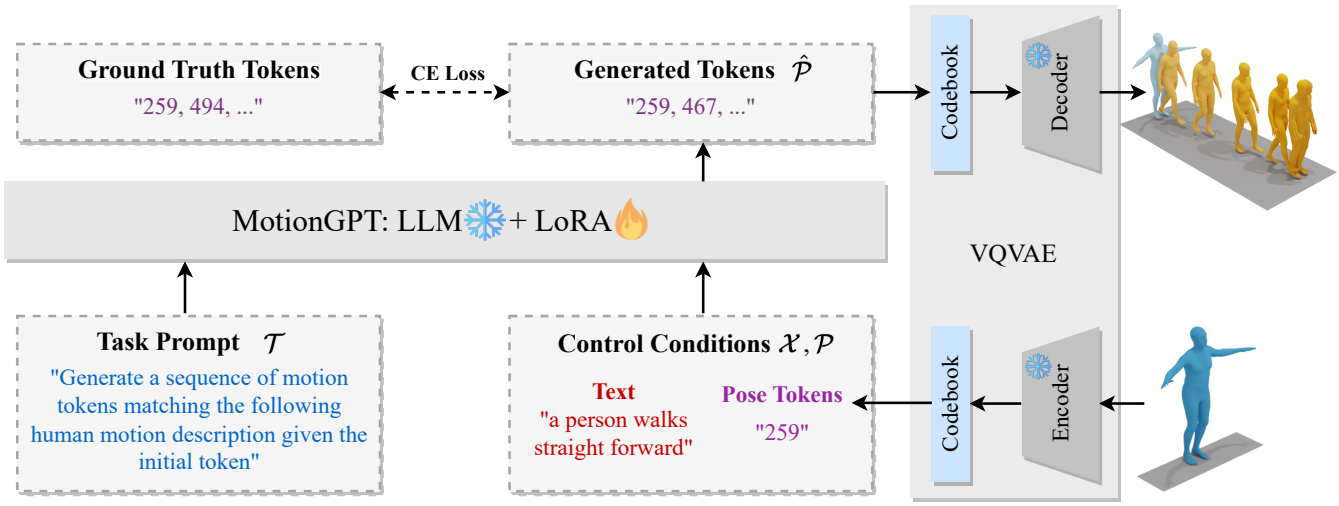

Figure 2: The pipeline of MotionGPT, a Motion General-Purpose generaTor. Given text and poses as an input example, we organize task descriptions (Instruction) and multiple control conditions (Input) within a question template. MotionGPT fine-tunes an LLM to generate the corresponding motion answer, which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

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 fine-tune LLMs by LoRA, which can be presented as

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

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

Generalization to Existing and New Tasks

Leveraging the general template given before, 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>}

Experiment

Datasets and Evaluation Metrics

Datasets We apply two widely-used datasets, HumanML3D (Guo et al. 2022a) and KIT-ML (Plappert, Mandery, and Asfour 2016) for evaluation.

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, including *Reconstruction Loss (Recon)* and *Velocity Loss (Vel)*. Specifically, these metrics aim to measure the consistency between the provided pose conditions and the generated motion.

More information about datasets, proposed new metrics, and implementation details are included in the supplementary material (Zhang et al. 2023b).

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.

The quantitative results of motion quality are depicted in Tab. 1 and Tab. 2. As illustrated in Tab. 1, our proposed model, MotionGPT, exhibits a performance that is competitive with state-of-the-art methods for text-to-motion generation. Specifically, MotionGPT consistently achieves comparable results across all metrics on both HumanML3D (Guo et al. 2022a) and KIT-ML (Plappert, Mandery, and Asfour 2016) datasets.

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 (Petrovich, Black, and Varol 2022)	3.734	3.703	8.973	3.717	3.417	10.84
TM2T (Guo et al. 2022b)	1.501	3.467	8.589	1.501	3.467	8.589
T2M (Guo et al. 2022a)	1.087	3.347	9.175	3.022	3.488	10.72
MotionDiffuse (Zhang et al. 2022)	0.630	**3.113**	9.410	1.954	**2.958**	**11.10**
MDM (Tevet et al. 2023)	0.544	5.566	9.559	0.497	9.191	10.85
MLD (Xin et al. 2023)	0.473	3.196	9.724	**0.404**	3.204	10.80
T2M-GPT (Zhang et al. 2023a)	**0.116**	3.118	**9.761**	0.514	3.007	10.92
MotionGPT-13B (Ours)	0.567	3.775	9.006	0.597	3.394	10.54

Table 1: 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.

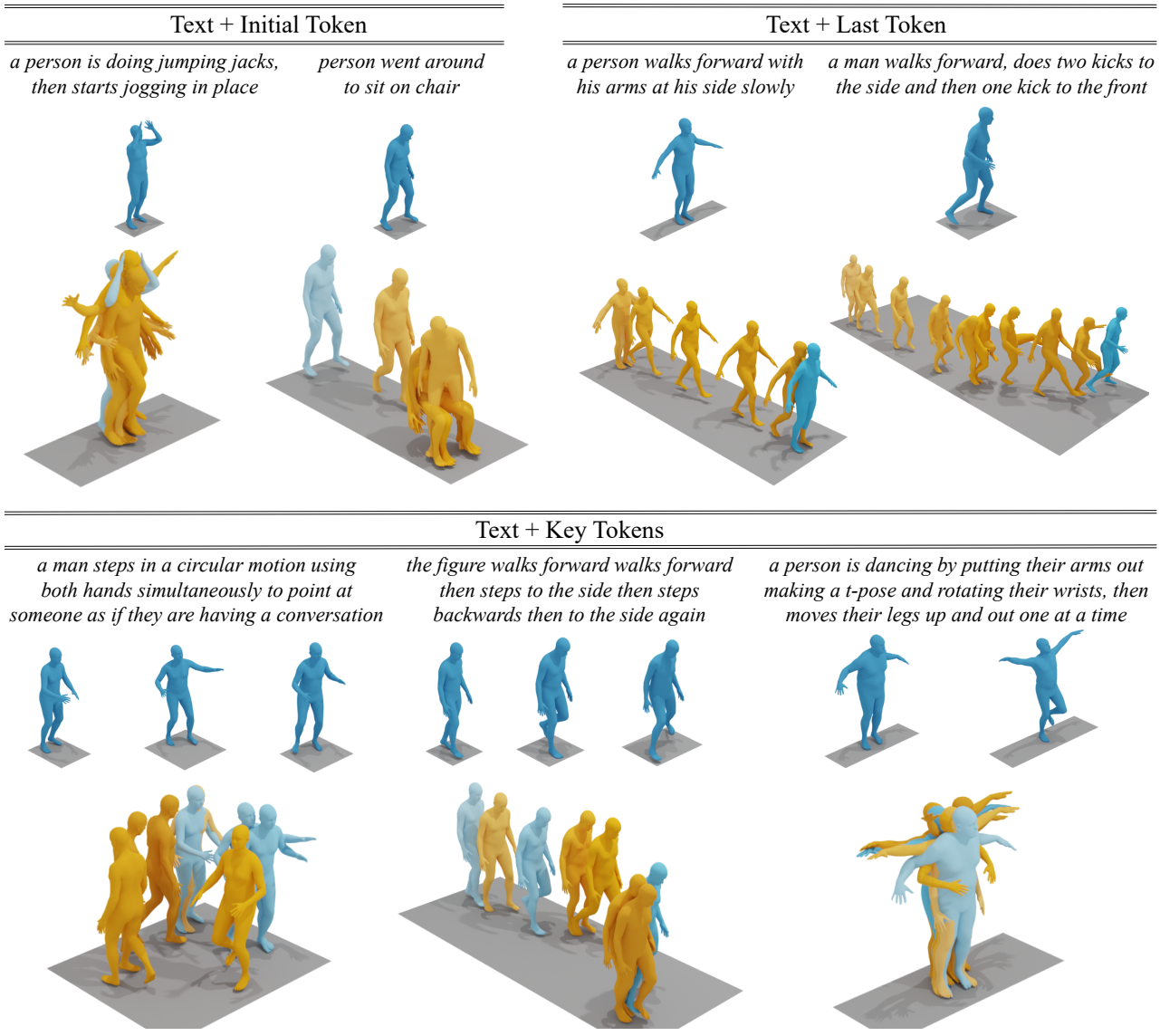

Figure 3: Generated motion by MotionGPT with multiple control conditions on HumanML3D.

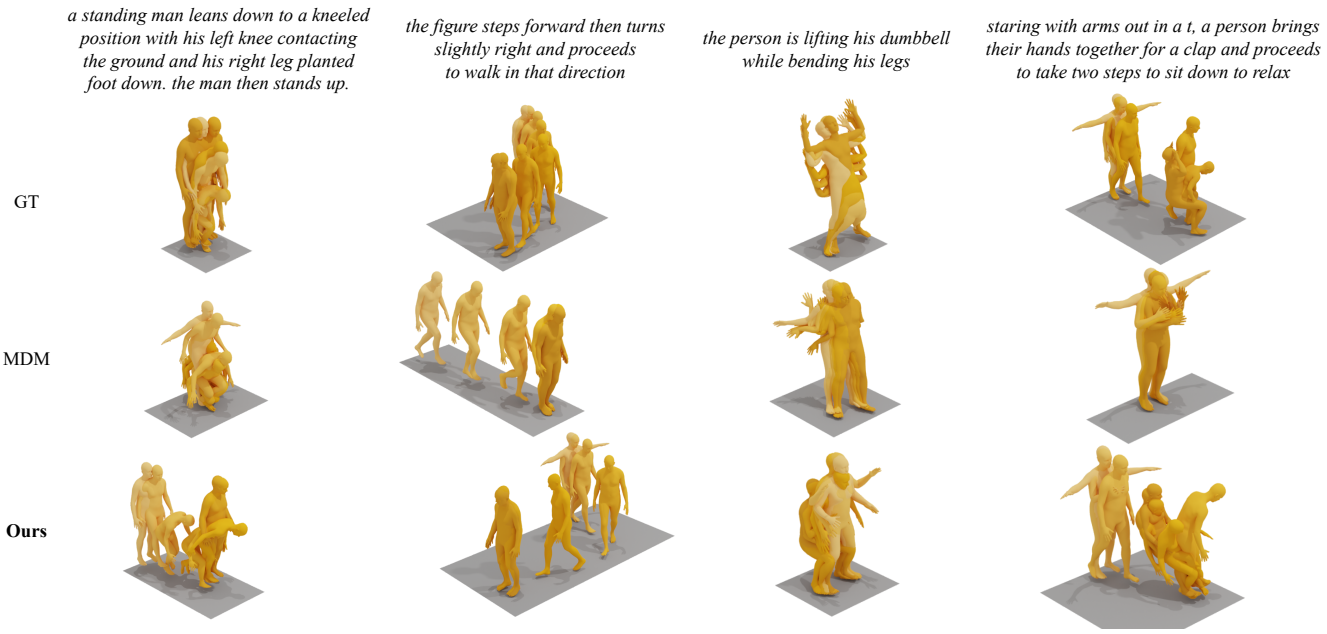

Figure 4: Qualitative comparison of the state-of-the-art motion generation method MDM with text-only conditions on HumanML3D.

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

Table 2: Motion generation quality on HumanML3D and KIT-ML test set for diverse control conditions.

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. 2. 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 (Tevet et al. 2023). 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.

Ablation Study

Additionally, extensive ablation studies are conducted on HumanML3D (Guo et al. 2022a) dataset to indicate the effectiveness of our MotionGPT. More ablation studies are included in the supplementary material (Zhang et al. 2023b).

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. 3, our results show a strong correlation between the level of pre-training in LLMs and the

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

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		$0.390^{-.663}$	$3.492^{-.198}$	$0.416^{+.046}$	$0.597^{+.051}$	$0.713^{+.045}$	$9.621^{+.647}$

Table 4: Comparisons between separate training for each task and joint training for multiple tasks on HumanML3D validation set using MotionGPT-7B. Superscripts indicate the improvement or decrement in the metric. Joint training can achieve better performance for all tasks.

Methods	Recon ↓	Vel ↓
Initial token		
Text-only	24.70	1.095
Text + Initial poses	**13.78**	**0.549**
Last token		
Text-only	19.70	1.172
Text + Last poses	**6.831**	**0.397**
Key tokens		
Text-only	8.035	3.813
Text + Random poses	**5.383**	**2.423**

Table 5: Evaluation of the effectiveness of pose control conditions on HumanML3D test set using MotionGPT-13B model.

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.

Consistency with pose control conditions We demonstrate the effectiveness of pose control conditions by assessing the consistency between pose controls and generated motion on the HumanML3D test set. For each task (initial/last/key), we generate motion with and without pose controls using (text+pose)-to-motion and text-to-motion methods, respectively. The results are shown in Tab. 5. In comparison to text-only generation, better keyframe pose consistency arises from generating under pose conditions, showcasing (text+pose)-to-motion's effectiveness with pose control.

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 (Guo et al. 2022a). The aim is to investigate if multi-task learning could improve the performance of individual control conditions. The comparison results are depicted in Table 4. 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.

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.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62121002 and Grant No. 62272430).

References

Barsoum, E.; Kender, J.; and Liu, Z. 2018. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 1418–1427.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Du, Y.; Konyushkova, K.; Denil, M.; Raju, A.; Landon, J.; Hill, F.; de Freitas, N.; and Cabi, S. 2023. Vision-language models as success detectors. arXiv:2303.07280.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022a. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5152–5161.

Guo, C.; Zuo, X.; Wang, S.; and Cheng, L. 2022b. 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*, 580–597. Springer.

Guo, C.; Zuo, X.; Wang, S.; Zou, S.; Sun, Q.; Deng, A.; Gong, M.; and Cheng, L. 2020. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2021–2029.

Habibie, I.; Holden, D.; Schwarz, J.; Yearsley, J.; and Komura, T. 2017. A recurrent variational autoencoder for human motion synthesis. In *Proceedings of the British Machine Vision Conference (BMVC)*.

He, R.; Liu, L.; Ye, H.; Tan, Q.; Ding, B.; Cheng, L.; Low, J.-W.; Bing, L.; and Si, L. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation. arXiv:2106.03164.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2790–2799. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021a. Lora: Low-rank adaptation of large language models. arXiv:2106.09685.

Hu, S.; Ding, N.; Wang, H.; Liu, Z.; Wang, J.; Li, J.; Wu, W.; and Sun, M. 2021b. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. arXiv:2108.02035.

Kania, K.; Kowalski, M.; and Trzciński, T. 2021. TrajeVAE: Controllable Human Motion Generation from Trajectories. arXiv:2104.00351.

Le, H.; Pino, J.; Wang, C.; Gu, J.; Schwab, D.; and Besacier, L. 2021. Lightweight adapter tuning for multilingual speech translation. arXiv:2106.01463.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The power of scale for parameter-efficient prompt tuning. arXiv:2104.08691.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv:2301.12597.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023b. VideoChat: Chat-Centric Video Understanding. arXiv:2305.06355.

Li, R.; Yang, S.; Ross, D. A.; and Kanazawa, A. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13401–13412.

Li, Z.; Zhou, Y.; Xiao, S.; He, C.; Huang, Z.; and Li, H. 2017. Auto-conditioned recurrent networks for extended complex human motion synthesis. arXiv:1707.05363.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. arXiv:2304.08485.

Liu, X.; Ji, K.; Fu, Y.; Tam, W. L.; Du, Z.; Yang, Z.; and Tang, J. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv:2110.07602.

Lu, Z.; Huang, D.; Bai, L.; Liu, X.; Qu, J.; and Ouyang, W. 2023. Seeing is not always believing: A Quantitative Study on Human Perception of AI-Generated Images. arXiv:2304.13023.

Martinez, J.; Black, M. J.; and Romero, J. 2017. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2891–2900.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Petrovich, M.; Black, M. J.; and Varol, G. 2021. Action-conditioned 3D human motion synthesis with transformer VAE. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10985–10995.

Petrovich, M.; Black, M. J.; and Varol, G. 2022. 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*, 480–497. Springer.

Plappert, M.; Mandery, C.; and Asfour, T. 2016. The KIT motion-language dataset. *Big data*, 4(4): 236–252.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.;

et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. arXiv:2204.06125.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Tevet, G.; Gordon, B.; Hertz, A.; Bermano, A. H.; and Cohen-Or, D. 2022. 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*, 358–374. Springer.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-or, D.; and Bermano, A. H. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. arXiv:2302.13971.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. arXiv:2109.01652.

Xin, C.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; Yu, J.; and Yu, G. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; Jiang, C.; Li, C.; Xu, Y.; Chen, H.; Tian, J.; Qi, Q.; Zhang, J.; and Huang, F. 2023. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. arXiv:2304.14178.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. arXiv:2206.10789.

Zhang, J.; Zhang, Y.; Cun, X.; Huang, S.; Zhang, Y.; Zhao, H.; Lu, H.; and Shen, X. 2023a. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022. Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv:2208.15001.

Zhang, Y.; Huang, D.; Liu, B.; Tang, S.; Lu, Y.; Chen, L.; Bai, L.; Chu, Q.; Yu, N.; and Ouyang, W. 2023b. MotionGPT: Finetuned LLMs are General-Purpose Motion Generators. arXiv:2306.10900.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv:2304.10592.

Zhuang, W.; Wang, C.; Chai, J.; Wang, Y.; Shao, M.; and Xia, S. 2022. Music2dance: Dancenet for music-driven dance generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2): 1–21.