FG-MDM: Towards Zero-Shot Human Motion Generation via ChatGPT-Refined Descriptions

Xu Shi[1], Wei Yao[1], Chuanchen Luo[2], Junran Peng[3], Hongwen Zhang[4], and Yunlian Sun[1(✉)]

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
{shixu,wei.yao,yunlian.sun}@njust.edu.cn
[2] School of Artificial Intelligence, Shandong University, Jinan, China
chuanchen.luo@sdu.edu.cn
[3] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[4] School of Artificial Intelligence, Beijing Normal University, Beijing, China
zhanghongwen@bnu.edu.cn

Abstract. Recently, significant progress has been made in text-based motion generation, enabling the generation of diverse and high-quality human motions that conform to textual descriptions. However, generating motions beyond the distribution of original datasets remains challenging, i.e., zero-shot generation. By adopting a divide-and-conquer strategy, we propose a new framework named Fine-Grained Human Motion Diffusion Model (FG-MDM) for zero-shot human motion generation. Specifically, we first parse previous vague textual annotations into fine-grained descriptions of different body parts by leveraging a large language model. We then use these fine-grained descriptions to guide a transformer-based diffusion model, which further adopts a design of part tokens. FG-MDM can generate human motions beyond the scope of original datasets owing to descriptions that are closer to motion essence. Our experimental results demonstrate the superiority of FG-MDM over previous methods in zero-shot settings. We will release our fine-grained textual annotations for HumanML3D and KIT on the project page https://sx0207.github.io/fg-mdm/

Keywords: Human Motion Generation · Diffusion Model · Zero-Shot Generation

1 Introduction

Human motion generation is an important research topic in communities of both computer vision and computer graphics. It aims to simulate and generate realistic human movements using computers. With the advancement of technologies

"A person walks happily."

"A person walks depressingly."

Arms

Neck

Legs

His arms sway freely by his sides.

His neck is held high and comfortably.

His legs move with energy, taking long strides.

His arms hang heavily by his sides.

His neck is lowered.

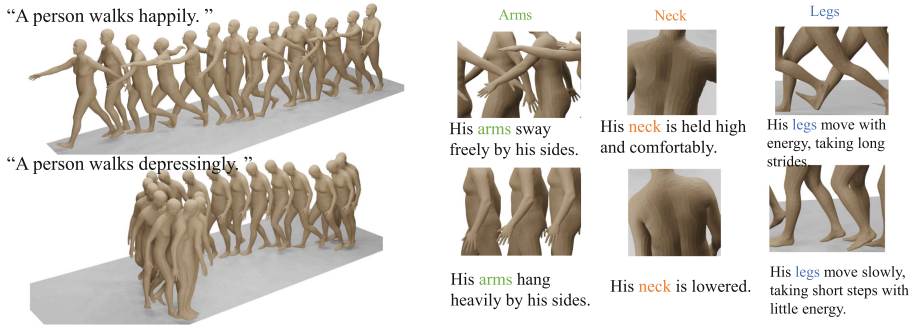His legs move slowly, taking short steps with little energy.

Fig. 1. FG-MDM can generate high-quality human motions in zero-shot settings by using fine-grained descriptions of different body parts. The two images on the left illustrate two contrasting emotional motions. Close-Up images of the arms, neck, and legs highlight these differences.

such as virtual reality, augmented reality, and movie special effects, there is a growing demand for high-quality human motion generation. In recent years, several innovative methods and techniques have emerged to tackle this challenging task [48]. Deep generative models, including GANs [1,21], VAEs [8,26,27], and diffusion models [5,15,37,45], have been widely applied to human motion generation.

However, there is relatively less research on generating motions in zero-shot settings. In order to improve the zero-shot generation capability, existing work either got help from CLIP [31] to utilize the rich semantic knowledge from CLIP (e.g., [12,36]), or attempted to explore large-scale motion datasets without textual descriptions [17] and large-scale pseudo text-pose datasets [3,20]. Compared to traditional human motion generation, generating motions beyond the distribution of the dataset is more challenging due to the limited scale and diversity of existing motion capture datasets.

Then, with only limited motion capture datasets available, can we still generate motions beyond the distribution of the dataset? For a textual description defining a motion beyond the distribution of the dataset, is there any way to associate it with motions within the dataset? For a never-before-seen motion, the entire body's motion is indeed unseen. However, motions of specific body parts might be inside the dataset. Therefore, we can adopt a divide-and-conquer strategy. By re-annotating the motion for different body parts with fine-grained descriptions, we can associate these body parts with specific body parts within the dataset. For example, a vague description "A person walks depressingly." can be reformulated as "His arms hang heavily by his sides. His legs move slowly, taking short steps with little energy...". Leg movement in this vague unseen motion may appear in "A person walks aimlessly and slowly.", of which the motion is included in the dataset. And the arm movement may appear in "His arms hang heavily by his sides.", of which the motion is included in the dataset. We give two examples in Fig. 1. On one hand, adopting fine-grained descriptions allows

the model to understand how the unseen motions are performed in detail. On the other hand, re-annotating motions for different body parts enables the model to learn the essence of motions better. Using fine-grained textual descriptions, we aim to improve the model's zero-shot understanding capability.

Although annotating fine-grained textual descriptions manually for body parts provides more accurate data, it requires significant manual work, resulting in huge costs. Fortunately, with the rapid development of large language models, OpenAI's GPT series models [25], known for their outstanding natural language processing capabilities, have gained widespread attention worldwide. In [7], Gilardi et al. demonstrated that ChatGPT performs as well as human annotators in some tasks. In [14], Action-GPT explores the excellent capability of ChatGPT in expanding human action descriptions. However, the generated content tends to be excessively redundant. For our task, we carefully design a prompt that allows ChatGPT-3.5 to provide detailed but non-redundant transcriptions of text descriptions about human motion. We then use this prompt and ChatGPT-3.5 to transcribe 44,970 short text descriptions from HumanML3D [8] and 6,353 text descriptions from KIT [29] for model training.

With these fine-grained descriptions, we propose a new framework named Fine-Grained Human Motion Diffusion Model (FG-MDM) for human motion generation. Specifically, we replace the original simple and vague text with ChatGPT-Generated fine-grained descriptions of individual body parts to guide a transformer-based diffusion model. Following MDM [37], we encode the entire fine-grained description with CLIP [31] as a global token of the transformer. Apart from this global token, we further encode descriptions of different body parts individually with CLIP as part tokens. By adopting these tokens, the model can pay attention to both the global and detailed information of human motions, thereby improving the accuracy and completeness of the denoising results.

Our contributions are summarized as follows:

– We present a novel framework that utilizes fine-grained descriptions of different body parts to guide the denoising process of the transformer-based diffusion model. This framework is capable of generating a broader range of motions that extend beyond the distribution of training datasets.
– We carefully design a prompt that enables ChatGPT to convert short and vague texts into detailed but non-redundant descriptions of different body parts. We then use this prompt to transcribe 44,970 texts from HumanML3D and 6,353 texts from KIT into fine-grained descriptions. We will make these fine-grained transcriptions publicly available.
– We conduct a series of experiments to evaluate our model's ability to not only fit the training data but also generate motions beyond the distribution of the dataset, i.e., the generalization capability.

2   Related Work

2.1   Human Motion Generation

There has been a great interest in human motion generation in recent years. Previous work has explored unconditional generative models [30,47] as well as generative models using various input conditions, such as text [5,8,37], prior motion [24], action class [9,26], and music [16,39]. In this paper, we focus on text-to-motion generation. Early work usually addressed the text-to-motion task with a sequence-to-sequence model [18]. Later on, the focus shifted beyond simple action labels. For example, Guo et al. utilized variational autoencoders to generate motions from text [8], significantly enhancing the quality and diversity of generated motions. With the success of diffusion models in AIGC, MDM [37] and other related work [5,15,45] have introduced diffusion models into the text-to-motion domain, resulting in impressive achievements.

There is a relative scarcity of work that directly focuses on the zero-shot capabilities of motion generation models. In [36], Tevet et al. proposed MotionCLIP to align human motions with the CLIP space, implicitly injecting the rich semantic knowledge from CLIP into the motion domain to enhance zero-shot generation capability. AvatarCLIP [12] also utilized CLIP to implement a zero-shot text-driven framework for 3D avatar generation and animation. Liang et al. [17] pre-trained a large-scale unconditional diffusion model to learn rich out-of-domain motion traits. In order to improve the generalization capability of motion generation models, there have been also attempts to leverage human mesh recovery approaches [38,41–43] to collect large-scale pseudo text-pose datasets [3]. As shown in Azadi et al. [3] and Lin et al. [20], the pre-training on such text-pose datasets can improve the generalization to in-the-wild descriptions. However, the static nature of text-pose data makes it difficult to well represent dynamic motions.

The work most closely related to ours is Action-GPT [14], which introduced, for the first time, large language models into the field of text-conditioned motion generation. Action-GPT can be integrated into any text-to-motion model. However, it enriched only the description of action classes without providing detailed descriptions of different body parts and guiding the model training. For another action generation model, SINC [2] incorporated ChatGPT to identify the body parts involved in the textual description. It achieved impressive results by generating multiple motions and concatenating them using different body parts. Specifically, SINC divides the human body into ['left arm', 'right arm', 'left leg', 'right leg', 'torso', 'neck', 'buttocks', 'waist'], which we borrow from in our work. It should be noted that both Action-GPT and SINC were designed for motion generation based on action labels, not using natural language. Therefore, directly comparing our work with them is not feasible. There is also a costly method that utilizes LLMs. By fine-tuning LLMs, MotionGPT [13,46] designed a pre-trained motion language model that supports various motion-related tasks through prompts. In contrast, our method is more efficient and can rapidly enhance the model's zero-shot generation capabilities.
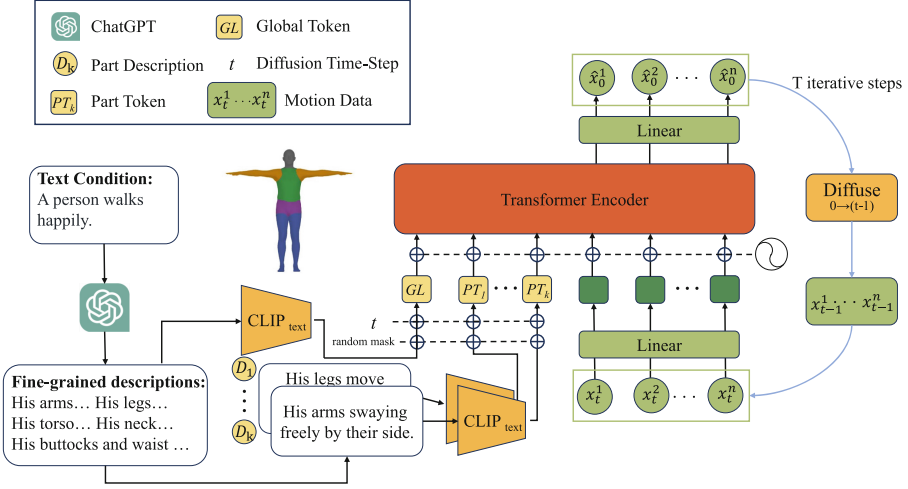
Fig. 2. The overall pipeline of FG-MDM. The model learns the denoising process of the diffusion model from the motion $x_t^{1:n}$ at time step t to the clean motion $\hat{x}_0^{1:n}$, given the text condition. The input text is first paraphrased by ChatGPT into fine-grained descriptions $D_{1:k}$ for different parts of the body, where k denotes the number of body parts. These descriptions are then fed into a pre-trained CLIP text encoder and projected, along with the time step t, onto input tokens $PT_{1:k}$ of the transformer. The overall fine-grained text is further encoded into a global input token GL, providing holistic information. In the sampling process of the diffusion model, an initial random noise $x_T^{1:n}$ is sampled, and then T iterations are performed to generate the clean motion $\hat{x}_0^{1:n}$. At each sampling step t, guided by $PT_{1:k}$ and GL, the transformer encoder predicts the clean motion $\hat{x}_0^{1:n}$ which is then noised back to $x_{t-1}^{1:n}$.

2.2   Diffusion Generative Models

The diffusion model is a neural generative model based on the stochastic diffusion process in thermodynamics [10,35]. It starts with samples from the data distribution and gradually adds noise through a forward diffusion process. Then, a neural network learns the reverse process to progressively remove the noise and restore the samples to their original states. Diffusion generative models have achieved significant success in the image generation field [32,33]. For conditional generation, [6] introduced classifier-guided diffusion, while [11] proposed a classifier-free method. Given their excellent generation quality, [15,37,45] incorporated diffusion models into the motion generation domain, leading to impressive results.

3   Method

Given a textual description, our goal is to generate a human motion $x^{1:n} = \{x^i\}_{i=1}^n$ that matches the given description. The motion consists of n frames of human poses. For each pose $x^i \in \mathbb{R}^{J \times D}$, we represent it by joint rotations

or positions, where J represents the number of joints and D represents the dimensionality of the joint representation. In Fig. 2, we give an overview of our fine-grained human motion diffusion model. First, we adopt ChatGPT to perform fine-grained paraphrasing of the vague textual description. This expands concise textual descriptions into descriptions of different body parts. FG-MDM then uses these fine-grained descriptions to guide a diffusion model for human motion generation.

3.1 Prompt Strategy

We first introduce the prompt strategy adopted for generating fine-grained descriptions. We utilize ChatGPT-3.5 to create more fine-grained descriptions based on different body parts for a given textual description of a motion. Chat-GPT is a conversational model based on a large language model that can engage in natural conversations and generate corresponding responses. The answers from ChatGPT are often directly influenced by the information and expression provided in the prompt. If the prompt offers clear and detailed questions or instructions, ChatGPT can typically provide relevant and accurate answers. However, if the prompt is too simple, ambiguous, or unclear, ChatGPT may generate unexpected responses or express unclear content. For our task, we carefully design an effective prompt by using experimental verification.

Our designed prompt is: "Translate the motion described by the given sentences to the motion of each body part only using one paragraph. The available body parts include ['arms', 'legs', 'torso', 'neck', 'buttocks', 'waist']. Here are some examples: [Q...A...]. Question: [sentence]". [sentence] is the vague textual description that needs to be refined. [Q...A...] are four examples of Q&A pairs designed manually.

3.2 Diffusion Model for Motion Generation

The basic idea of diffusion models [10,35] is to learn the reverse process of a well-defined stochastic process. Following MDM [37], we design a text-driven human motion generation model based on the diffusion model.

The diffusion model consists of the forward process and the reverse process, both of which follow the Markov chain. The forward process involves adding noise. The input is the original motion $x_0^{1:n}$ from the data distribution, and the output is the motion $x_t^{1:n}$ with adding Gaussian noise t times. When enough noise is added, the motion $x_T^{1:n}$ can approach the Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The reverse process aims to reduce the noise in the Gaussian noise $x_T^{1:n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. In the denoising process, at diffusion step t, a portion of the noise is eliminated, resulting in a less noisy motion $x_{t-1}^{1:n}$. This step is repeated iteratively until the noise is completely removed, generating a clean motion $\hat{x}_0^{1:n}$.

Network. We adopt a simple transformer [40] encoder architecture to implement our network G. Unlike the conventional diffusion model mentioned above, we follow [32] and predict the clean motion $\hat{x}_0^{1:n}$ instead of predicting the noise added in each time-step. The input of G is the noised motion $x_t^{1:n}$ obtained by adding noise t times to the original motion $x_0^{1:n}$. The noised motion $x_t^{1:n}$, together with the text condition tokens GL, $PT_{1:k}$ and the time-step t, are inputted to the transformer encoder, resulting in the clean motion $\hat{x}_0^{1:n}$. One of the reasons for directly predicting the clean motion in each time-step of the diffusion model is to incorporate human geometric losses during the training of the network, making the generated human motions more natural. For each sampling step t, from T to 1, our model predicts the clean motion $\hat{x}_0^{1:n}$, and then adds noise back to $x_{t-1}^{1:n}$. After T iterations, the final clean motion $\hat{x}_0^{1:n}$ is obtained. This form of diffusion model has become commonly adopted, as do we.

Global Token and Part Tokens. For the text condition, we encode the entire fine-grained description with CLIP [31] as a global token GL of the transformer. Apart from this global token, we further encode descriptions of different body parts individually with CLIP as part tokens $PT_{1:k}$, where k denotes the number of body parts. The global token serves as an overall condition to guide the diffusion process. Part tokens provide explicit information for fine-grained control of the movements of each body part. Part tokens effectively make up for the ambiguity of the original description text. It greatly enhances our FG-MDM's ability to understand in-the-wild text, making it outstanding on zero-shot generation tasks.

Loss Functions. For training the diffusion model, we follow [32] to predict the signal itself instead of predicting the noise, i.e., $\hat{x}_0^{1:n} = G(x_t^{1:n}, t, c)$, with the simple loss function.

$$\mathcal{L}_{\mathrm{G}} = E_{x_0^{1:n} \sim q(x_0^{1:n}|c), t \sim [1,T]}[||x_0^{1:n} - G(x_t^{1:n}, t, c)||^2] \tag{1}$$

In order to generate more natural and kinematically plausible motions, we employ the same geometric losses as MDM [37] from [26,34], i.e., positions, foot contact, and velocities.

$$\mathcal{L}_{\mathrm{pos}} = \frac{1}{n} \sum_{i=1}^{n} ||FK(x_0^i) - FK(\hat{x}_0^i)||_2^2, \tag{2}$$

$$\mathcal{L}_{\mathrm{foot}} = \frac{1}{n-1} \sum_{i=1}^{n-1} ||(FK(\hat{x}_0^{i+1}) - FK(\hat{x}_0^i)) \cdot f_i||_2^2, \tag{3}$$

$$\mathcal{L}_{\mathrm{vel}} = \frac{1}{n-1} \sum_{i=1}^{n-1} ||(x_0^{i+1} - x_0^i) - (\hat{x}_0^{i+1} - \hat{x}_0^i)||_2^2 \tag{4}$$

where $FK(\cdot)$ represents the forward kinematic function that converts joint rotations into joint positions. For each frame i, $f_i \in \{0, 1\}^J$ is the binary foot contact mask.

Overall, our training loss is

$$\mathcal{L} = \mathcal{L}_{\mathrm{G}} + \lambda_{\mathrm{pos}}\mathcal{L}_{\mathrm{pos}} + \lambda_{\mathrm{vel}}\mathcal{L}_{\mathrm{vel}} + \lambda_{\mathrm{foot}}\mathcal{L}_{\mathrm{foot}}. \tag{5}$$

where λ_{pos}, λ_{vel}, λ_{foot} are balancing coefficients for the three geometric losses.

4   Experiments

In this section, we first elaborate the datasets, evaluation metrics, and implementation details in Sect. 4.1. We then conduct quantitative experiments to compare FG-MDM with current state-of-the-art approaches in Sect. 4.2. To show the generalization capability of our model, we further perform quantitative experiments, qualitative experiments, and a user study to examine FG-MDM's ability to generate motions beyond the distribution of training datasets. Finally, to evaluate our method comprehensively, we design two additional ablation experiments in Sect. 4.3.

4.1   Experimental Details

Datasets. We utilize the HumanML3D [8] dataset and the KIT [29] dataset to train and evaluate our model. The HuMMan [4] dataset and Kungfu dataset from the Motion-X dataset [19, 22] are employed to assess the models' zero-shot performance. HumanML3D is a recently proposed large-scale dataset of motion-text pairs. It consists of 14,616 motion sequences from the AMASS [23] and HumanAct12 [9] datasets, with multiple ways of describing each motion, resulting in a total of 44,970 text annotations. The KIT dataset, on the other hand, is relatively smaller and contains 3,911 motion sequences along with their corresponding 6,353 text descriptions. For both datasets, we follow the default settings, using 80% of the data for training and the remaining for testing. Motion-X is a large-scale dataset of whole-body motions and whole-body pose annotations, integrating several existing datasets and additional online videos. For zero-shot testing, we utilize 100% of the HuMMan and Kungfu subsets from it. HuMMan is a multi-modal human dataset, containing 744 motion sequences and their corresponding 744 texts descriptions. Kungfu encompasses many human motions related to martial arts, with a total of 1040 motion sequences and their corresponding 1040 texts descriptions.

We preprocess the 44,970 text descriptions from HumanML3D and 6,353 text descriptions from KIT using ChatGPT-3.5. This preprocessing extends these descriptions into fine-grained ones for our model training.

Evaluation Metrics. We employ three evaluation metrics for quantitative experiments to evaluate our model's ability to fit the training data: FID, Multimodal Dist, and Diversity. Multimodal Dist assesses the correlation between
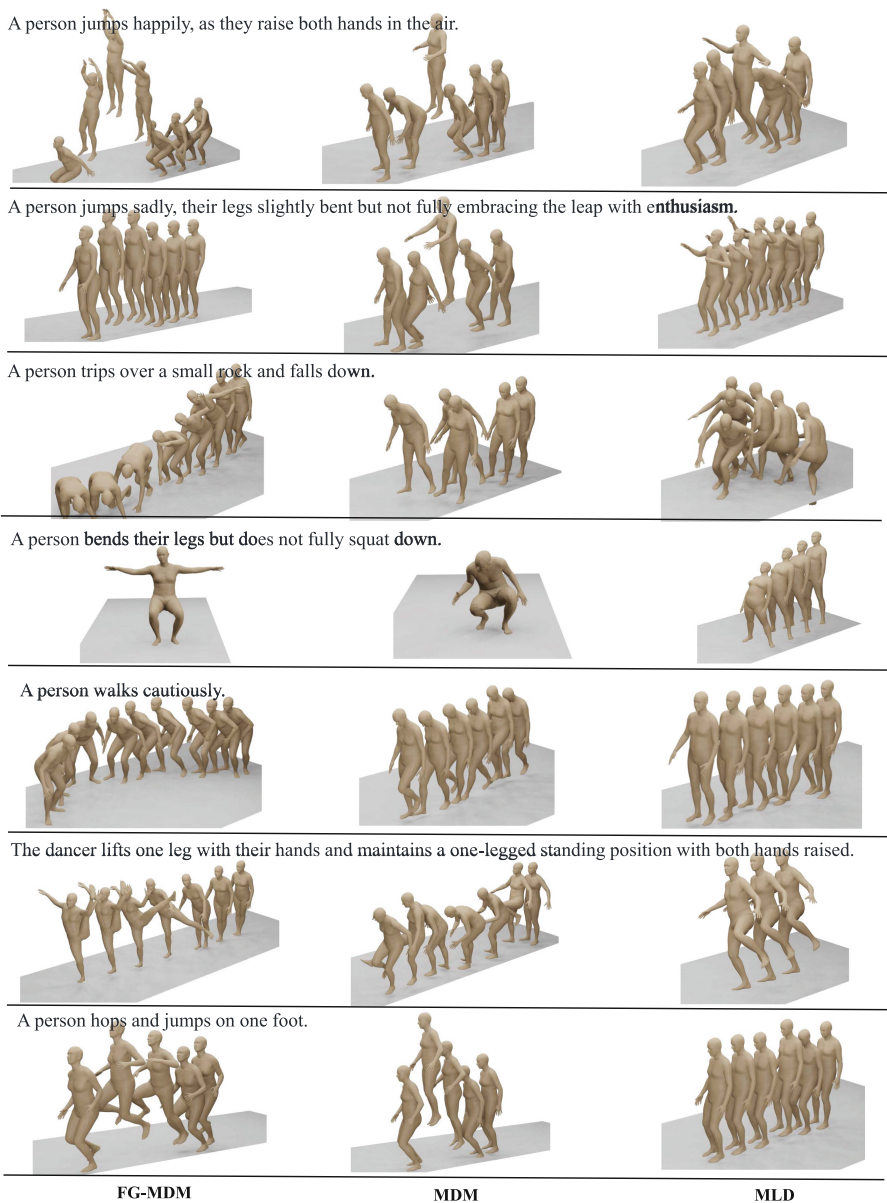
A person jumps happily, as they raise both hands in the air.

A person jumps sadly, their legs slightly bent but not fully embracing the leap with enthusiasm.

A person trips over a small rock and falls down.

A person bends their legs but does not fully squat down.

A person walks cautiously.

The dancer lifts one leg with their hands and maintains a one-legged standing position with both hands raised.

A person hops and jumps on one foot.

FG-MDM	MDM	MLD

Fig. 3. Qualitative results with unseen motions. We compare our FG-MDM with MDM [37] and MLD [5]. All three models are trained on HumanML3D. For better visualization, some pose frames are shifted to prevent overlap. Please refer to supplementary materials for more video demos.

generated motions and input text. Diversity is utilized to evaluate the diversity of generated motions. FID measures the difference in feature distribution between generated motions and ground truth in latent space, which is used to evaluate the quality of generated motions.

Implementation Details. In our study, the transformer accepts tokens whose feature dimension is 512 as input. We use four attention heads and apply a dropout rate of 0.1. The transformer encoder consists of 8 stacked encoder layers to capture complex relationships and hierarchies in the data. For ChatGPT, we adopt the gpt-3.5-turbo API provided by OpenAI. For text encoding, we employ the frozen CLIP-ViT-B/32 model as the encoder. Our batch size is set to 64. Additionally, we set the diffusion step to 1000. On a single NVIDIA GeForce RTX3090 GPU, it takes about six days to train our model.

Table 1. Quantitative results on HumanML3D and HuMMan. The model marked with * indicates that both the ChatGPT-Refined text and the manually annotated text provided by HumanML3D are used during training. **Bold text** is the best result, underlined text is the second-best result. Zero-Shot means that the models are evaluated directly on HuMMan after training on HumanML3D.

Methods	HumanML3D			HuMMan(zero-shot)		
	FID↓	MM Dist↓	Diversity↑	FID↓	MM Dist↓	Diversity↑
Real	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	$0.032^{\pm.002}$	$23.019^{\pm.042}$	$4.709^{\pm.097}$
MAA [3]	$0.774^{\pm.007}$	–	$8.230^{\pm.064}$	–	–	–
T2M-GPT [44]	$\mathbf{0.116}^{\pm.004}$	$\underline{3.118}^{\pm.011}$	$\mathbf{9.761}^{\pm.081}$	$\mathbf{9.631}^{\pm.203}$	$27.582^{\pm.073}$	$5.149^{\pm.145}$
MLD [5]	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$\underline{9.724}^{\pm.082}$	$14.970^{\pm.472}$	$27.104^{\pm.024}$	$5.493^{\pm.101}$
MotionDiffuse [45]	$0.630^{\pm.001}$	$\mathbf{3.113}^{\pm.001}$	$9.410^{\pm.049}$	$30.138^{\pm.712}$	$28.747^{\pm.041}$	$5.357^{\pm.045}$
MDM [37]	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$13.375^{\pm.408}$	$27.689^{\pm.055}$	$5.585^{\pm.089}$
FG-MDM	$0.663^{\pm.012}$	$5.649^{\pm.024}$	$9.476^{\pm.068}$	$17.180^{\pm.272}$	$\underline{26.867}^{\pm.030}$	$\underline{5.589}^{\pm.124}$
FG-MDM*	$0.618^{\pm.009}$	$5.274^{\pm.048}$	$9.563^{\pm.097}$	$\underline{12.460}^{\pm.330}$	$\mathbf{26.814}^{\pm.019}$	$\mathbf{5.626}^{\pm.100}$

4.2   Comparison with Prior Work

To evaluate the performance of FG-MDM in handling zero-shot text-conditioned motion generation, we compare our work with five recent motion generation approaches: MAA [3], T2M-GPT [44], MLD [5], MotionDiffuse [45], and MDM [37]. In Table 1 and Table 2, we provide experimental results on the HumanML3D, HuMMan, and Kungfu datasets, respectively. For all experiments, We run the evaluation five times, and "±" indicates the 95% confidence interval. For the six SOTA methods, on HumanML3D, we directly cite their results reported in their original papers. To examine the generalization ability of the methods, we use HumanML3D as the training set and HuMMan and Kungfu as the test sets. To

Table 2. Quantitative results on Kungfu. Zero-Shot means that the models are evaluated directly on Kungfu after training on HumanML3D.

Methods	Kungfu(zero-shot)		
	FID↓	MM Dist↓	Diversity↑
Real	$0.133^{\pm.010}$	$22.164^{\pm.041}$	$5.351^{\pm.312}$
MAA [3]	–	–	–
T2M-GPT [44]	$\mathbf{12.652}^{\pm.429}$	$\underline{25.826}^{\pm.041}$	$5.702^{\pm.428}$
MLD [5]	$18.524^{\pm.352}$	$27.182^{\pm.020}$	$5.598^{\pm.356}$
MotionDiffuse [45]	$26.363^{\pm.337}$	$26.320^{\pm.035}$	$\mathbf{6.117}^{\pm.691}$
MDM [37]	$16.396^{\pm.466}$	$26.280^{\pm.095}$	$5.468^{\pm.590}$
FG-MDM	$19.340^{\pm.797}$	$26.845^{\pm.052}$	$5.142^{\pm.759}$
FG-MDM*	$\underline{15.892}^{\pm.567}$	$\mathbf{25.325}^{\pm.035}$	$\underline{5.814}^{\pm.479}$

do so, we train TMR [27,28] using the HuMMan and Kungfu datasets to obtain a pair of text encoder and motion encoder for calculating the MM Dist metric. For SOTA methods, we apply their released pre-trained models on HumanML3D to HuMMan and Kungfu to evaluate their zero-shot generation performance. Since MAA [3] does not release the pre-trained model, we cannot test its zero-shot generation performance.

When evaluated on the test set of HumanML3D, all five methods achieve state-of-the-art performance. For FG-MDM, the ChatGPT-Refined fine-grained textual description doesn't match the manually annotated textual description well. Therefore, under within-dataset settings, our model does not exceed those SOTA models on HumanML3D. However, on the HuMMan and Kungfu datasets, FG-MDM captures most of the best and second-best results. Note that the size of our training dataset is much smaller than some SOTA methods like [3], but we still demonstrate solid zero-shot capabilities.

In addition, we provide some qualitative results to let readers intuitively feel the superiority of our method. In Fig. 3, we show motions generated by MDM [37], MLD [5] and our FG-MDM. Note that for all three methods, we use models trained on HumanML3D to generate motions. In comparison, our method generates motions more consistent with the details described in the fine-grained textual descriptions. This shows that our divide-and-conquer method works. Motion generation models require clear and specific conditions to generate the motions needed.

4.3   Ablation Study

To validate our contribution, we conduct two ablation studies. As shown in Table 3, the first row shows our baseline. The first study examines the contribution of ChatGPT-Generated fine-grained texts, which is performed by replacing the original short text with the fine-grained description. The improvement can be said to be huge. We cleverly utilize the powerful reasoning capabilities of

Table 3. Ablation study results on HumanML3D and KIT. "Fine-Grained" denotes using ChatGPT-Generated fine-grained descriptions. "Part" represents adopting part tokens. Note that the models are trained on HumanML3D.

Fine-Grained	Part	HumanML3D			KIT		
		FID↓	MM Dist↓	Diversity↑	FID↓	MM Dist↓	Diversity↑
		4.363	7.298	8.432	16.372	10.502	8.758
✓		1.050	6.778	**9.509**	0.549	9.826	**10.829**
✓	✓	**0.663**	**5.649**	9.476	**0.344**	**9.352**	10.707

LLMs and let them help our generative model better understand the nature of text conditions, bringing a leap to the zero-shot performance. The second study checks the contribution of part tokens when fine-grained descriptions are used. As observed, a reasonable framework also improves the quality of generated motions. However, perhaps more conditions bring more constraints to generations, leading to a decrease in diversity. But this drop is acceptable. So, we finally adopt the design of part tokens.
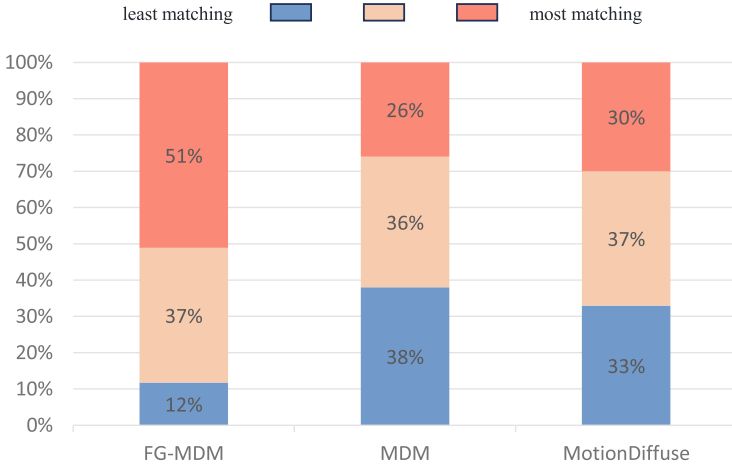

Fig. 4. User study results. For each method, a color bar ranging from blue to red represents the percentage of text-to-motion match levels, with blue indicating the least match and red indicating the most match. (Color figure online)

4.4 User Study

To further examine FG-MDM's generalization capability, we conduct a user study to evaluate the quality of motions generated by our model based on human visual perception. We customize a total of 40 textual descriptions beyond the distribution of the dataset. With these descriptions, we generate motions by using

MDM [37], MotionDiffuse [45], and our FG-MDM. We then recruit 10 users for the study. In each question, participants are asked to rate the degree of matching between the generated motion and the textual description on a scale of 0 to 2. The results are given in Fig. 4. Apparently, FG-MDM matches texts much better in generating motions beyond the distribution of the dataset than the other two methods. Nearly half of the generated motions get the highest score. In contrast, MDM and MotionDiffuse perform poorly. Most of the generated motions are not satisfactory.

5   Conclusion

In this study, we used LLMs to perform fine-grained paraphrasing on the textual annotations of HumanML3D and KIT. With these fine-grained descriptions, we explored a Fine-Grained Human Motion Diffusion Model. It utilizes fine-grained descriptions of different body parts to guide the training of a diffusion model. This enables it to learn the essence of motions and thus generate motions beyond the distribution of training datasets. In the future, we would like to improve the quality of fine-grained annotations of human motions. Having high-quality text labels will greatly promote research on human motion generation.

References

1. Ahn, H., Ha, T., Choi, Y., Yoo, H., Oh, S.: Text2action: generative adversarial synthesis from language to action. In: IEEE International Conference on Robotics and Automation, pp. 5915–5920. IEEE (2018)
2. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: SINC: spatial composition of 3d human motions for simultaneous action generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9984–9995 (2023)
3. Azadi, S., Shah, A., Hayes, T., Parikh, D., Gupta, S.: Make-an-animation: large-scale text-conditional 3D human motion generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
4. Cai, Z., et al.: HuMMan: multi-modal 4D human dataset for versatile sensing and modeling. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13667, pp. 557–577. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20071-7_33
5. Chen, X., et al.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18000–18010 (2023)
6. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. Adv. Neural. Inf. Process. Syst. **34**, 8780–8794 (2021)
7. Gilardi, F., Alizadeh, M., Kubli, M.: ChatGPT outperforms crowd workers for text-annotation tasks. Proc. Natl. Acad. Sci. **120**(30), e2305016120 (2023)
8. Guo, C., et al.: Generating diverse and natural 3D human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5152–5161 (2022)
9. Guo, C., et al.: Action2motion: conditioned generation of 3D human motions. In: Proceedings of the ACM International Conference on Multimedia, pp. 2021–2029 (2020)

10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
11. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv:2207.12598 (2022)
12. Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., Liu, Z.: AvatarClip: zero-shot text-driven generation and animation of 3D avatars. ACM Trans. Graph. **41**(4), 1–19 (2022)
13. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: MotionGPT: human motion as a foreign language. Adv. Neural Inf. Process. Syst. **36** (2024)
14. Kalakonda, S.S., Maheshwari, S., Sarvadevabhatla, R.K.: Action-GPT: leveraging large-scale language models for improved and generalized action generation. In: IEEE International Conference on Multimedia and Expo (2023)
15. Kim, J., Kim, J., Choi, S.: Flame: free-form language-based motion synthesis & editing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 8255–8263 (2023)
16. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: AI choreographer: music conditioned 3D dance generation with AIST++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13401–13412 (2021)
17. Liang, H., et al.: OMG: towards open-vocabulary motion generation via mixture of controllers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
18. Lin, A.S., Wu, L., Corona, R., Tai, K., Huang, Q., Mooney, R.J.: Generating animated videos of human activities from natural language descriptions. Learning **2018**(1) (2018)
19. Lin, J., et al.: Motion-X: a large-scale 3D expressive whole-body human motion dataset. Adv. Neural Inf. Process. Syst. (2023)
20. Lin, J., et al.: Being comes from not-being: open-vocabulary text-to-motion generation with wordless training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23222–23231 (2023)
21. Lin, X., Amer, M.R.: Human motion modeling using DVGANs. arXiv:1804.10652 (2018)
22. Lu, S., et al.: HumanTOMATO: text-aligned whole-body motion generation. arXiv:2310.12978 (2023)
23. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5442–5451 (2019)
24. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2891–2900 (2017)
25. Ouyang, L., et al.: Training language models to follow instructions with human feedback. Adv. Neural. Inf. Process. Syst. **35**, 27730–27744 (2022)
26. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3D human motion synthesis with transformer VAE. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10985–10995 (2021)
27. Petrovich, M., Black, M.J., Varol, G.: TEMOS: generating diverse human motions from textual descriptions. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13682, pp. 480–497. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20047-2_28
28. Petrovich, M., Black, M.J., Varol, G.: TMR: text-to-motion retrieval using contrastive 3D human motion synthesis. In: International Conference on Computer Vision (2023)

29. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data **4**(4), 236–252 (2016)
30. Raab, S., et al: MoDI: unconditional motion synthesis from diverse data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13873–13883 (2023)
31. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
32. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. **1**(2), 3 (2022). arXiv:2204.06125
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
34. Shi, M., et al.: MotioNet: 3D human motion reconstruction from monocular video with skeleton consistency. ACM Trans. Graph. **40**(1), 1–15 (2020)
35. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, pp. 2256–2265. PMLR (2015)
36. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: MotioncLIP: exposing human motion generation to clip space. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13682, pp. 358–374. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20047-2_21
37. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023)
38. Tian, Y., Zhang, H., Liu, Y., Wang, L.: Recovering 3D human mesh from monocular images: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **45**, 15406–15425 (2023)
39. Tseng, J., Castellon, R., Liu, K.: EDGE: editable dance generation from music. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 448–458 (2023)
40. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
41. Yao, W., Zhang, H., Sun, Y., Tang, J.: STAF: 3D human mesh recovery from video with spatio-temporal alignment fusion. IEEE Trans. Circ. Syst. Video Technol., 1 (2024). https://doi.org/10.1109/TCSVT.2024.3410400
42. Zhang, H., et al.: PyMAF-X: towards well-aligned full-body model regression from monocular images. IEEE Trans. Pattern Anal. Mach. Intell. **45**, 12287–12303 (2023)
43. Zhang, H., et al.: PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11446–11456 (2021)
44. Zhang, J., et al.: T2M-GPT: generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
45. Zhang, M., et al.: MotionDiffuse: text-driven human motion generation with diffusion model. IEEE Trans. Pattern Anal. Mach. Intell. **46**, 4115–4128 (2024)
46. Zhang, Y., et al.: MotionGPT: finetuned LLMs are general-purpose motion generators. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 7368–7376 (2024)

47. Zhao, R., Su, H., Ji, Q.: Bayesian adversarial human motion synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6225–6234 (2020)
48. Zhu, W., et al.: Human motion generation: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **46**, 2430–2449 (2023)