Plan, Posture and Go: Towards Open-vocabulary Text-to-Motion Generation

Jinpeng Liu*[1], Wenxun Dai*[1], Chunyu Wang*[2], Yiji Cheng[1], Yansong Tang†[1], and Xin Tong[2]

*equal contribution, †corresponding author
[1]Shenzhen Key Laboratory of Ubiquitous Data Enabling, Shenzhen International Graduate School, Tsinghua University
[2] Microsoft Research Asia
{liujp22@mails, tang.yansong@sz}.tsinghua.edu.cn
Project page: https://moonsliu.github.io/Pro-Motion/

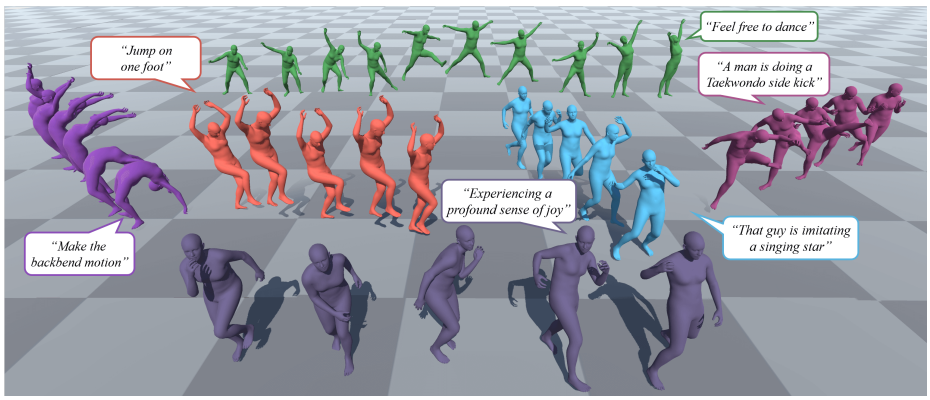

Fig. 1: Exemplary motions generated by our PRO-Motion system. Different from conventional models trained on paired text-motion data, our PRO-Motion can generate 3D human motion with global body translation and rotation from *open-vocabulary* text prompts, such as *"Jump on one foot"* and *"Experiencing a profound sense of joy"*. Furthermore, our approach is the first to address the common issue with previous methods in similar formulations [33,45,83], which is their inability to generate global body translation and rotation. This limitation often leads to unrealistic in-place motions.

Abstract. Conventional text-to-motion generation methods are usually trained on limited text-motion pairs, making them hard to generalize to open-vocabulary scenarios. Some works use the CLIP model to align the motion space and the text space, aiming to enable motion generation from natural language motion descriptions. However, they are still constrained to generate limited and unrealistic in-place motions. To address these issues, we present a divide-and-conquer framework named **PRO-Motion**[1], which consists of three modules as motion planner, posture-denoiser and go-denoiser. The motion planner instructs Large

[1] **PRO-Motion**: **P**lan, postu**R**e and g**O** for text-to-**Motion** generation

Language Models (LLMs) to generate a sequence of scripts describing the key postures in the target motion. Differing from natural languages, the scripts can describe all possible postures following very simple text templates. This significantly reduces the complexity of posture-denoiser, which transforms a script to a posture, paving the way for open-vocabulary text-to-motion generation. Finally, the go-denoiser, implemented as another diffusion model, not only increases the motion frames but also estimates the whole-body translations and rotations for all postures, resulting in more dynamic motions. Experimental results have shown the superiority of our method with other counterparts, and demonstrated its capability of generating diverse and realistic motions from complex open-vocabulary prompts such as "Feel free to dance".

Keywords: Open-vocabulary · Text-to-Motion · Motion Generation

1 Introduction

The conditioned 3D generation has attracted rapidly increasing attention [1, 14, 42, 62, 84, 87, 95, 96, 102] due to its important roles in many applications such as virtual reality, video games, and the film industry. The prior models usually train GANs [1, 46], VAEs [4, 27, 61, 62] and Diffusion Models [11, 15, 16, 75, 84, 93, 97, 98] from paired text-motion data and have achieved reasonable generation results when the text prompts are similar to those in the training set. Fig. 2 (a) illustrates this paradigm. However, they struggle to handle open-vocabulary text prompts beyond the existing datasets and can only generate limited "toy-like" motions. This situation includes not only body descriptions like "Raise arms" but also human emotion descriptions like "Experiencing a profound sense of joy".

Some recent work [33, 45, 83] propose to enhance their model's ability to handle natural language motion descriptions beyond the training data. To that end, they leverage the pre-trained vision-language model CLIP [67] to align the poses in the training motions with the motion descriptions, hoping to generate poses from natural languages. This is depicted in Fig. 2 (b). However, the text space of CLIP, which is learned in natural languages about image content, is largely different from motion descriptions, making it ineffective to connect natural languages and motions. As a result, these methods are still constrained to generate motions from limited text prompts. Besides, due to the lack of temporal priors in CLIP, these methods have difficulty in generating motions with correct chronological order. As a result, they can only generate unrealistic in-place motions.

In this paper, we present a divide-and-conquer framework named **PRO-Motion**, which consists of three steps as **P**lan, postu**R**e, and **G**o for open-vocabulary text-to-**Motion** generation, as shown in Fig. 2 (c). In the first "plan" stage, we introduce a motion planner that translates complex natural language motion descriptions into a sequence of posture scripts that describe body part relationships following a simple template, such as *"The man is standing upright, his torso is vertical. His left foot is slightly above the ground. His arms are relaxed at his sides"*. This is realized by leveraging the motion commonsense in

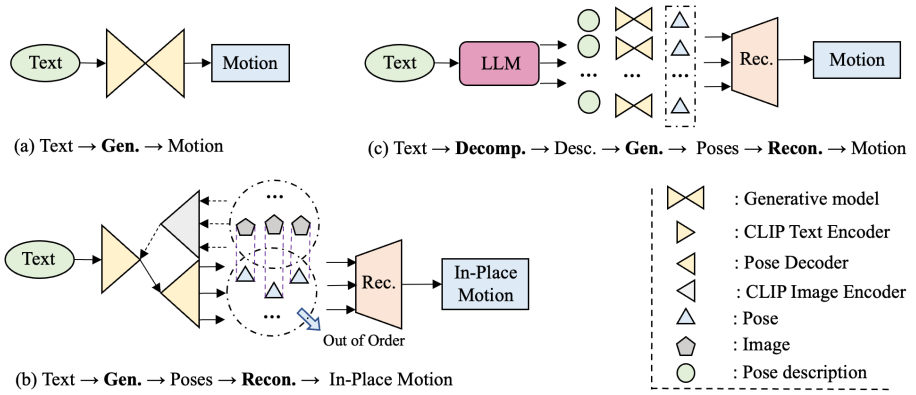

Fig. 2: Comparison of different paradigms for text-to-motion generation. (a) Most existing models leverage the generative models [23, 31, 40] to construct the relationship between text and motion based on text-motion pairs. (b) Some methods render 3D poses to images and employ the image space of CLIP to align text with poses. Then they reconstruct the motion in the local dimension based on the poses. (c) Conversely, we decompose motion descriptions into structured pose descriptions. Then we generate poses based on corresponding pose descriptions. Finally, we reconstruct the motion in local and global dimensions. "Gen.", "Decomp.", "Desc.", "Rec." stand for "Generative model", "Decompose", "Pose Description" and "Reconstruction" respectively.

LLMs which is further enhanced by in-context demonstrations. It is important to understand that although the scripts are simple and limited to a small space, they are expressive enough to cover all possible postures due to their compositional nature. The motion planner bridges the gap between natural languages and pose descriptions and effectively addresses out-of-distribution problems.

Benefiting from the merits above, during the second "posture" stage, we can train a generative model to achieve script-to-posture generation only using a relatively small labeled dataset. We conjecture and demonstrate that the model has strong generalization capability and can cover extensive postures and scripts, considering that a novel posture or script can be decomposed into multiple familiar body parts. In the implementation, we developed a diffusion-based model called posture-denoiser, which perceives the connection between structured pose descriptions and body parts leading to diverse and realistic postures. It is designed to predict the sample rather than the noise. This facilitates the utilization of established losses of the poses. Then we further utilize a posture planning module to select key poses, taking into account the consistency of adjacent poses and the semantic alignment between text and poses.

Furthermore, in the last "go" stage, we have observed that we can predict both translation and rotation by analyzing multiple consecutive body postures. For example, in a sequence where the initial pose depicts a standing pose followed by a left-foot step in the second pose and a right-foot step in the third pose, we

can estimate a forward translation. Additionally, learning interpolation between adjacent key poses is straightforward and requires only a small amount of motion data to capture such priors effectively. Accordingly, a transformer-based [86] go-denoiser module is designed to capture the inner connection between key poses.

To verify the effectiveness of our PRO-Motion, we conduct experiments on a variety of datasets. Both quantitative and qualitative results have shown the advantage of our method compared with the state-of-the-art approaches for open-vocabulary text-to-motion generation and demonstrated its capability of generating diverse and realistic motions from complex prompts such as *"Jump on one foot"* and *"Experiencing a profound sense of joy"*. Pro-Motion possesses several distinctive properties and accordingly has the following contributions:

1. We propose a novel paradigm transferring motion generation into pose and script alignment (see Fig. 2). Infinite numbers of motions can be decomposed into articulated poses. And poses are limited to a low-dimensional manifold [20]. The property makes poses very suitable for text-to-poses-to-motion generation. It avoids the challenge of collecting large-scale text-motion training data, which is prohibitively expensive.
2. We introduce the posture-denoiser module to transform scripts into poses and the posture planning module to select key poses from the above poses, paving the way for open-vocabulary text-to-motion generation.
3. We propose a diffusion-based generative model go-denoiser to address reconstruction tasks that reconstruct translation, rotation and poses.

2 Related Work

Text-to-Motion Generation. Based on labeled motion capture datasets [26, 27,38,44,51,64,65,76,82], existing works have explored various generative models for text-driven motion generation, such as GANs [1, 46], VAEs [4, 27, 61, 62, 83] and Diffusion Models [11, 15, 75, 84, 93, 97, 98]. However, these methods are constrained by the heavy reliance on limited text-motion paired datasets. To tackle this problem, some works [33, 45] try to leverage the current powerful large-scale pre-trained models, *i.e.*, CLIP [67], to overcome the data limitation and achieve open-vocabulary motion generation. AvatarCLIP [33] generates motions for given textual descriptions through online matching and optimization. Nevertheless, matching is unable to generate out-of-distribution candidate poses, which limits the ability to generate complex motions, and online optimization is time-consuming and unstable. OOHMG [45] uses CLIP image features to generate candidate poses and performs motion generation via mask learning. However, this method cannot capture the chronological order of actions due to the lack of temporal priors in CLIP, leading to inaccurate or even completely opposite motion. Our approach takes a different step to probe the powerful prior knowledge of human body pose and motion in LLMs to enhance text-motion alignment capability and enable open-vocabulary motion generation.

Keyframe-based Motion Generation. Given that motion can be viewed as a composition of a sequence of poses, keyframe-based motion generation has

attracted lots of interest. Motion prediction involves generating unrestricted motion continuation when provided with one or more keyframes of animation as context. Early efforts [5,12,22,24,25,37,52,60] employed RNNs to model human motion sequence, motivated by the powerful capability in capturing temporal dynamics. Besides RNNs, other network architectures like CNNs [47, 59] and GCNs [17] are proposed to enhance the modeling of temporal and movement relationships. The emergence of Transformers [3, 9] has further facilitated the modeling of long-range dependencies within a motion sequence. Close to our method is motion in-betweening, which is constrained by both past and future keyframes. Early methods include physically-based approaches [54, 73, 89] that involve solving optimization problems, as well as statistical models [10, 53, 88]. More recently, some neural network-based methods such as RNNs [29, 81, 100], CNNs [30, 39, 104], and Transformers [19, 53, 66] have gained dominance in this field. Unlike motion in-betweening methods that explicitly provide translation and rotation, we achieve the prediction of translation and rotation, as well as pose interpolation, by having the model learn priors between adjacent key poses.

LLM aided Visual Content Generation. In recent years, large language models (LLMs) [7, 8, 56, 85, 94, 99] have attracted substantial interest in the field of natural language processing (NLP) and artificial general intelligence (AGI) owing to their remarkable proficiency in tasks such as language generation, reasoning, world knowledge, and in-context learning. [6,28] combine large language models with diffusion-based generative models [31, 71] aimed at generating prompts for higher-quality image generation. [34, 43] leverage large language models to plan the generation of visual content and identify the pivotal actions, enabling complex dynamic video generation. Another line of works, including [41,48,77,90,92], have proposed to integrate visual APIs with language models to facilitate decision-making or planning based on visual information, which further connects vision and language models. Close to our method are works that utilize LLMs as a planner for embodied agents [2, 35, 36, 50, 78, 80, 91, 101] to generate executable plans in real-world environments. Unlike works focus on robots, we introduce LLMs to manipulate the generation of key poses of motion, enabling fine-granularity control over motion.

3 Method

Based on two key findings: (1) despite an infinite number of potential motions, the underlying postures are limited to a smaller space [20]; (2) mapping between this small space and natural language can be achieved with current dataset [18]. We divide the task into three steps and leverage LLMs to implement open-vocabulary generation. Specifically, PRO-Motion first instructs LLMs to generate a sequence of scripts describing the key postures in the target motion. The scripts follow simple patterns that focus on the relationships of body parts, allowing the generation of postures from scripts using a simple diffusion model. Finally, with key postures as conditions, we train another diffusion model to estimate whole-body translations and rotations for all postures.

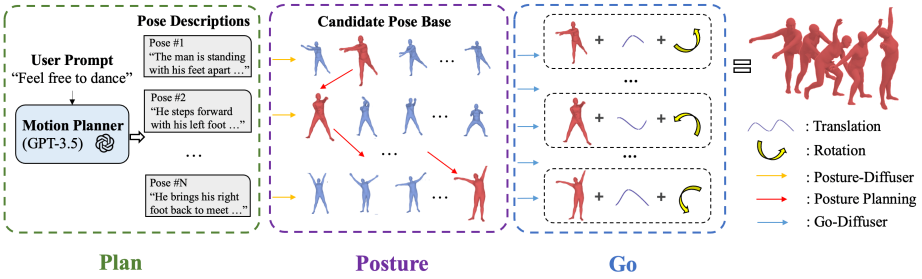

Fig. 3: Illustration of our framework for open-vocabulary text-to-motion generation. Specifically, we employ large language models (LLMs) as the *Motion Planner* to generate a sequence of scripts describing the key postures. Then, our *Posture-Denoiser* module receives discrete pose scripts and generates corresponding poses to construct a candidate posture base. Posture planning module is utilized to select reasonable pose sequences from the candidate posture base. Finally, the Go-Denoiser module increases the motion frames and infers the translation and rotation.

3.1 Motion Planner

As depicted in Fig. 3, when presented with a user prompt, such as *"Feel free to dance"*, we exploit GPT-3.5 [56] to create a plan for describing key poses based on the prior knowledge about body parts that are involved in the motion. To ensure that GPT-3.5 generates descriptions for these key poses while maintaining consistency throughout the motion, we provide GPT-3.5 with a user prompt indicating the expected motion and a task description that guarantees the temporal consistency and control over various motion attributes like frames per second (FPS) and the number of frames. Beyond governing the overall motion, we have established five fundamental rules to guide GPT-3.5 in describing key poses: (1) Characterize the degree of bending of body parts, *e.g.*, *"left elbow"* using descriptors like *"completely bent"*, *"slightly bent"*, *"straight"*. (2) Classify the relative distances between different body parts, *e.g.*, two hands, as *"close"*, *"shoulder width apart"*, *"spread"* or *"wide"* apart. (3) Describe the relative positions of different body parts, *e.g.*, *"left hip"* and *"left knee"*, using terms like *"behind"*, *"below"* or *"at the right of"*. (4) Determine whether a body part is oriented *"vertical"* or *"horizontal"*, *e.g.*, *"right knee"*. (5) Identify whether a body part is in contact with the ground, such as *"left knee"* and *"right foot"*. Furthermore, we offer GPT-3.5 some reference pose descriptions to guide its generation process. Through this rule-based approach, we can guide the LLM to generate precise key pose descriptions, achieving fine-grained control over poses. For more details about the prompt design, please refer to the supplementary materials.

3.2 Posture-Denoiser

In this section, we present our *Posture-Denoiser* module, which aims to generate key poses that align with the localized body part descriptions provided by

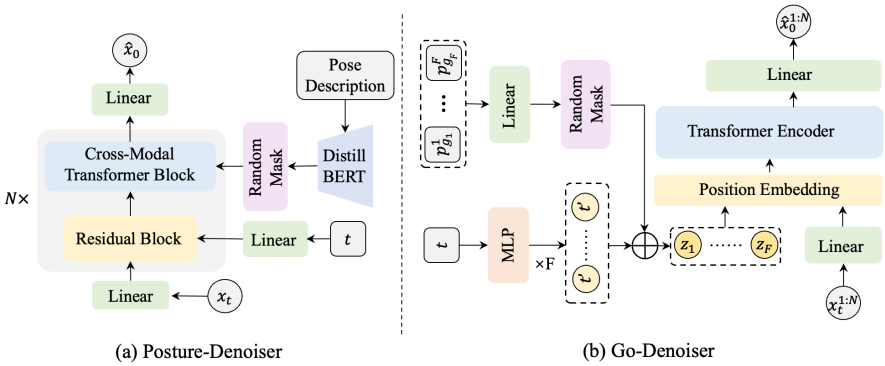

Fig. 4: Illustration of our Dual-Diffusion model. (a) *Posture-Denoiser* module is designed to predict the original pose conditioned by the pose description. The model consists of N identical layers, with each layer featuring a residual block for incorporating time step information and a cross-modal transformer block for integrating the condition text. (b) *Go-Denoiser* module serves the function of obtaining motion with translation and rotation from discrete key poses without global information. In this module, the key poses obtained from Sec. 3.2 are regarded as independent tokens. We perform attention [86] between these tokens and noised motion independently, which can significantly improve the perception ability between every condition pose and the motion sequence.

the *Motion Planner* in Sec. 3.1. As shown in Fig. 4 (a), we utilize a denoising diffusion model, which is composed of a stack of N identical layers. Each layer has two sub-blocks. The first is a residual block, which incorporates the time embedding generated by passing the sinusoidal time embedding through a two-layer feed-forward network. The second is a cross-modal transformer block, which integrates the conditioning signal, *i.e.*, text, via a standard cross-attention mechanism [86]. The intermediate residual pose feature serves as the query vector, while the text embeddings extracted from the frozen DistillBERT [74] act as the key and value vectors. Furthermore, we randomly mask the text embeddings for classifier-free learning. This module enables us to generate key poses that align with the pose descriptions precisely.

Posture Planning Due to the sampling diversity of DDPMs [31,55,68,79], the *Posture-Denoiser* module can generate multiple plausible poses corresponding to each pose description, we introduce our *Posture Planning* module, aiming to select the most reasonable key poses from the candidate poses. We propose two objectives: (1) minimizing differences between poses in adjacent frames and (2) maximizing the similarity between poses and corresponding descriptions. To achieve the objectives, we design two encoders: a text encoder Φ composed of a single layer of bi-GRU [13], and a pose encoder Θ that uses the VPoser encoder [58]. The encoders produce the L2-normed embeddings for computing similarity. We employ the Viterbi algorithm [21] to search for reasonable paths.

Specifically, suppose we have a set of F pose descriptions denoted as $\{d_i\}_{i=1}^{F}$, and for each pose description d_i, we have a collection of L generated candidate poses represented as $\{p_j^i\}_{j=1}^{L}$, serving as pose observations at each frame. The transition probability matrix A^i for the i-th ($i > 1$) frame is for the first objective, where the selection of poses for adjacent frames should preferably consider pairs with higher similarity as follows:

$$A_{jk}^i = \frac{\exp\left(\Theta(p_j^{i-1})^T \Theta(p_k^i)\right)}{\sum_{l=1}^{L} \exp\left(\Theta(p_j^{i-1})^T \Theta(p_l^i)\right)}. \tag{1}$$

The emission probability matrix E^i for the i-th ($i \geq 1$) frame is to satisfy the second objective, where the selection of the current frame's key pose should preferably consider poses with a higher matching degree to the description:

$$E_j^i = \frac{\exp\left(\Phi(d_i)^T \Theta(p_j^i)\right)}{\sum_{l=1}^{L} \exp\left(\Phi(d_i)^T \Theta(p_l^i)\right)}. \tag{2}$$

The overall objective of the algorithm is to generate a pose path $G = \{g_i\}_{i=1}^{F}$ that maximizes the joint probability:

$$\underset{G}{\operatorname{argmax}} P(G) = \prod_{i=1}^{F} P(g_i|g_{i-1}) = E_{g_1}^1 \prod_{i=2}^{F} E_{g_i}^i A_{g_{i-1}g_i}^i. \tag{3}$$

3.3 Go-Denoiser

To interpolate and predict global information such as translation and rotation for the key poses obtained in Sec. 3.2, we introduce our *Go-Denoiser* module Ψ in this section. Our module is based on a diffusion model, as illustrated in Fig. 4 (b). As transformer [86] structure has been proved efficient in the field of motion generation [61–63, 84], we adopt it with the transformer encoder architecture in our implementation. The module is fed a noised motion sequence $x_t^{1:N}$ in a time step t, as well as t itself and the condition, *i.e.*, key poses $\{p_{g_i}^i\}_{i=1}^{F}$. To enhance the modeling of relationships between key poses and better capture global information, we treat them as discrete tokens rather than a unified feature. In practice, the key poses are projected and then randomly masked for classifier-free learning. The noised input $x_t^{1:N}$ is projected and integrates positional information. The transformer encoder output is projected back to the original motion dimension, yielding the predicted motion sequence $\hat{x}_0^{1:N}$. This module enables us to interpolate the key poses smoothly and assign global properties to motions.

We do the sampling step from $p(x_0|G)$ in an iterative manner according to [31, 84]. The target of the reverse step is to predict the clean sample $x_0^{1:N} = \Psi(x_t^{1:N}, t, G)$ and noise it back to $x_{t-1}^{1:N}$. It is repeated from $t = T$ until x_0 is achieved. Go-Denoiser is trained using classifier-free guidance [32]:

$$\Psi(x_t^{1:N}, t, G) = \Psi(x_t^{1:N}, t, \emptyset) + s * (\Psi(x_t^{1:N}, t, G) - \Psi(x_t^{1:N}, t, \emptyset)) \tag{4}$$

4 Experiments

4.1 Datasets and Motion Representation

Dataset. In our experiments, we utilize the pose data, motion data, and texts sourced from AMASS [51], PoseScript [18], Motion-X [44], and HumanML3D [26]. AMASS unifies various optical marker-based mocap datasets, offering over 40 hours of motion data without textual descriptions. PoseScript consists of static 3D human poses extracted from AMASS, together with fine-grained semantic human-written annotated descriptions (PoseScript-H) and automatically generated captions (PoseScript-A). HumanML3D is a widely used motion language dataset that provides captions for motion data sourced from AMASS. Motion-X is a large-scale 3D expressive whole-body motion dataset with detailed descriptions. When training and test sets are of high similarity, models that overfit the training set can exhibit impressive performance. To ensure the fairness of comparisons between surprised-learning method [83, 84] and open-vocabulary methods [33, 45, 83], we select two subsets from Motion-X as the testing set. In the open-vocabulary setting, all the methods are trained using the data of AMASS and HumanML3D. Specifically, we employ sentence transformers [69,70] to compute the similarity between the text in IDEA-400 [44] (a high-quality motion language subset within Motion-X) and the text in HumanML3D. We filter out pairs with similarity greater than a specified threshold α, *e.g.*, 0.45, yielding a dataset comprising 368 text-motion pairs as our first test dataset. Moreover, we choose the *kungfu* subset of Motion-X as our second test dataset.

Motion Representation. We follow the motion representation of TEMOS [62] and construct the feature vector for SMPL data. It consists of three parts, including "Translation", "Root Orientation" and "SMPL local rotations". "Translation" consists of two parts. The first part is the velocity of the root joint in the global coordinate system. The second part is the position of the root joint for the Z axis. Root Orientation contains one rotation, and we utilize the 6D continuous representation [103] to store it. Pose Body is from the SMPL-H [49,72] version of AMASS. Because we focus on the movement of the human body, we removed all rotations on the hands, resulting in 21 rotations). The same as root orientation, we utilize the 6D continuous representation [103]. The translation of neighboring poses is subtracted and represented by the instantaneous velocity as the translation attribute of the current frame.

4.2 Open-vocabulary Motion Generation

In this section, we first introduce the supervised learning baseline [83, 84], open-vocabulary baselines [33, 45, 83], and the evaluation metrics [26, 61]. Then we discuss the comparative experimental results with these baselines.

MDM Baseline: MDM [84] employs a supervised learning approach utilizing the diffusion model. However, its performance often degenerates when applied

Table 1: Comparison of our method with previous methods on the subsets of the IDEA-400 [44] dataset, *i.e.*, *ood368* and *kungfu*. We achieve superior performance on R precision, FID, and the MultiModal Dist. MDM [84] is for supervised learning. MotionCLIP [83], Codebook+Interpolation [33], Avatarclip [33] and OOHMG [45] are designed for open vocabulary text-to-motion generation.

	Text-motion				FID ↓	MM Dist. ↓	Smooth
	R@10 ↑	R@20 ↑	R@30 ↑	MedR ↓			
"test on ood368 subset"							
MDM [84]	17.81	34.06	48.75	31.20	3.5005	2.6136	0.0114
MotionCLIP [83]	16.25	35.62	52.81	28.90	2.2275	2.2889	0.0073
Codebook+Interpolation [33]	15.62	31.25	46.56	32.80	4.0847	2.5160	0.0146
AvatarCLIP [33]	15.31	31.56	47.19	32.60	4.1819	2.4496	0.0156
OOHMG [45]	15.62	34.06	48.75	29.80	3.9827	2.1492	0.0758
Ours	**20.25**	**36.56**	**53.14**	**26.10**	**1.4886**	**1.5345**	0.1312
"test on kungfu subset"							
MDM [84]	12.50	29.69	42.19	37.50	12.0601	3.7254	0.0735
MotionCLIP [83]	15.62	29.69	46.88	32.50	17.4147	4.2978	0.0123
Codebook+Interpolation [33]	10.94	20.31	29.69	37.50	2.5216	2.7641	0.0138
AvatarCLIP [33]	15.62	31.25	46.88	32.50	**1.9667**	2.4976	0.0171
OOHMG [45]	14.06	32.81	48.44	32.50	4.9048	2.4716	0.0847
Ours	**20.31**	**34.38**	**50.00**	**31.00**	4.1242	**2.3743**	0.1559

to a new setting, *i.e.*, open-vocabulary motion generation. We trained it on the SMPL-H [49, 57, 72] version data of AMASS with HumanML3D annotation.

MotionCLIP Baseline: MotionCLIP [83] is a supervised open-vocabulary method which trained on AMASS data with BABEL [65] annotation. We use the pre-trained model provided by the authors and test the model's performance in the open-vocabulary setting.

Codebook+Interpolation Baseline: In the pose generation stage, we utilize VPoserCodebook [57] as the pose generator and select the most similar pose in the pose generation stage. For the motion generation stage, we just use the interpolation method to generate the motion.

AvatarCLIP Baseline: AvatarCLIP [33] is an optimizer-based method. It also includes the first text-to-pose stage via matching between text and poses. Then it uses the matched poses to search the most related motion in the latent space of a motion VAE [40] trained on the AMASS dataset.

Evaluation Metrics is adopted from [26,61], which includes R precision, Frechet Inception Distance(FID), and MultiModal Distance. For quantitative evaluation, a motion feature extractor and a text feature extractor are trained using contrastive loss to produce geometrically close feature vectors for matched text-motion pairs. For more details about the above metrics as well as the design of the text and motion feature extractor, please refer to the supplementary materials. Consider R precision: for each generated motion, its ground-truth text description and n-1 randomly selected mismatched descriptions from the test dataset form a description pool, followed by calculating and ranking the Euclidean distances between the motion feature and the text feature of each description in

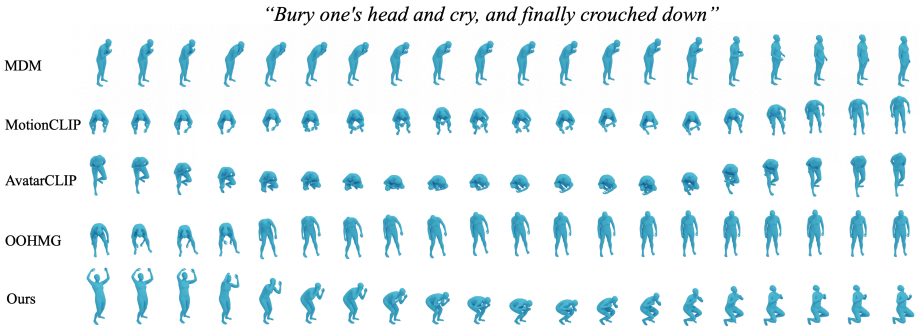

Fig. 5: Comparation of our methods with previous text-to-motion generation methods.

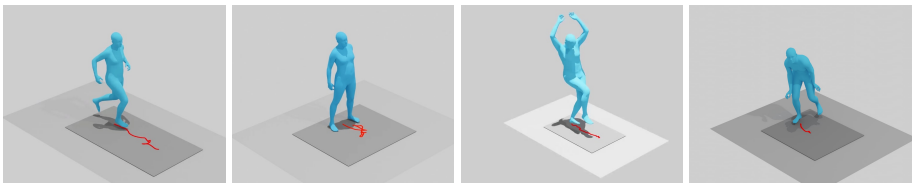

Fig. 6: (a) The guy is imitating a singing star. (b) Experiencing a profound sense of joy. (c) Jump on one foot. (d) Failure case: Lift left leg and walk forward. Best viewed in Adobe Reader where (a)-(d) should play as videos. They are also included in Supp.

the pool. Meanwhile, MultiModal distance is computed as the average Euclidean distance between the motion feature of each generated motion and the text feature of its corresponding description in the test dataset. Pose-based methods may precipitate a lack of cohesion between poses. So we have devised such a metric. $smooth = \frac{1}{n-1}\sum_{2}^{n}||p_i - p_{i-1}||^2$. A diminutive smooth does not necessarily signify optimal model performance. If the value is excessively minimal, it implies that the alteration in the movement is not substantial.

As shown in Fig. 5, for the motion description *"bury one's head and cry, and finally crouched down"*, methods such as MDM [84] based on the supervised learning paradigm often fail in similar cases and cannot generate un-seen motion. Moreover, due to the gap between motion description and image description, matching text and motion via the language space of CLIP [67] is not effective. MotionCLIP [83], AvatarCLIP [33] and OOHMG [45] struggle to deal with detailed and precise motion descriptions. We also show more visualization results in the form of gifs, as shown in Fig. 6. (a) and (b) show that our model can deal with descriptions that include emotions or other abstract situations rather than specific body movements. (c) shows our model can generate some atypical motion. While (d) is a failure case, foot sliding occurs since the representation in our method didn't deal well with foot contact information. In Tab. 1, quantitative metrics demonstrate the superiority of our model over other methods in terms of semantic consistency and motion rationality.

Table 2: Comparison of our method with PoseScript [18]. We follow the settings in PoseScript and compare the results on the datasets of PoseScript-A and PoseScript-H.

	FID↓	mRecall (R/G)↑	mRecall (G/R)↑
evaluation on automatic captions (PoseScript-A)			
PoseScript [18]	$0.48_{\pm 0.01}$	$29.37_{\pm 1.84}$	$48.97_{\pm 4.39}$
Ours	$\mathbf{0.24}_{\pm 0.02}$	$\mathbf{36.16}_{\pm 2.41}$	$\mathbf{59.13}_{\pm 2.62}$
evaluation on human captions (PoseScript-H)			
PoseScript [18]	$0.48_{\pm 0.01}$	$18.23_{\pm 1.72}$	$28.27_{\pm 1.53}$
Ours	$\mathbf{0.31}_{\pm 0.01}$	$\mathbf{24.40}_{\pm 2.73}$	$\mathbf{30.94}_{\pm 3.95}$

4.3 Ablation Study

Posture-Denoiser In comparison to the zero-shot open-vocabulary text-to-pose generation methods [33, 45], we first utilize LLMs to translate the pose description into localized body part description, which is fed to our pose generator in Sec. 3.2 to generate pose precisely. As shown in Fig. 7, the Matching method demonstrates superior pose generation results compared to Optimize and VPoserOptimize, which suggests that directly using CLIP for matching is more effective than optimization through the complex pipeline. However, "Matching" fails to generate more precise poses for diverse texts, exhibiting a limitation in preserving textual information in the generated poses. For instance, in cases like *"cry"* and *"pray"*, "Matching" generates identical poses for texts with distinct meanings. When generating poses that require more precise control over body parts, such as *"dance the waltz"* or *"kick soccer"*, both OOHMG and "Matching" fail to achieve satisfactory results. In contrast, by employing LLMs to precisely describe the expected pose, we achieve accurate control over pose generation, enabling more effective open-vocabulary text-based pose synthesis.

To evaluate the effects of our posture-denoiser quantitatively, we follow the setting of PoseScript [18] and test the effects in PoseScript-A and PoseScript-H. As shown in Tab. 2, our method achieved state-of-the-art results both on the FID, mRecall (R/G) and mRecall (G/R). Text-to-pose retrieval is evaluated by ranking the whole set of poses for each of the query texts. We then compute the recall@K (R@K), which is the proportion of query texts for which the corresponding pose is ranked in the top-K retrieved poses. We proceed similarly to evaluate poseto-text retrieval. We use K = 1, 5, 10 and report the mean recall (mRecall) as the average overall recall@K values from both retrieval directions.

Go-Denoiser. While this is the first effort at predicting spatial information of motion in a zero-shot local pose-driven manner to our best knowledge, we have developed appropriate baseline methods to evaluate the translation and rotation reconstruction and frame interpolation effects. Based on our observation, the translation and rotation of the body could be estimated by analyzing the variations in body parts between adjacent poses. We treat the process of estimating

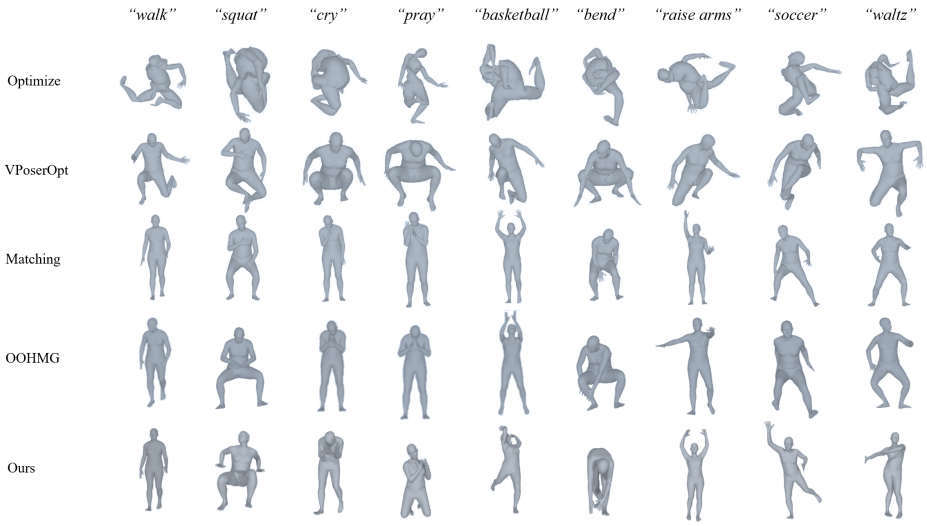

Fig. 7: Comparison of our method with previous text-to-pose generation methods.

global information from the key poses as a reconstruction task. Thus, we design several methods to evaluate the efforts. The model input is key poses and the model output is motion. "Reg." consists of two networks; one predicts rotation and the other predicts translation. Each part is composed of LeakyReLU, BatchNorm1D, and three linear layers. The transformer structure has been proven correct in the motion generation filed [61, 62], thus we design the baseline method by extracting pose sequence features utilizing TEMOS Motion Encoder [62] as the condition to inject into the diffusion model. As shown in Fig. 8, from the four images in the top left corner, it can be observed that a simple MLP network is capable of predicting motion translation information to some extent. As indicated in the bottom-left image, extracting pose sequences as features using existing motion encoders may overlook the internal relationships within the pose sequences, thereby leading to confusing details. In the top right image, our method exhibits better fidelity in capturing fine details such as knee flexion. Moreover, the four images in the bottom right corner reveal that when dealing with similar adjacent poses, our model demonstrates a finer-grained perceptual capability, thus imparting appropriate motion trends to digital avatars.

As shown in Tab. 3, our method achieved state-of-the-art results both on the APE. and the AVE. of global trajectory, rotation and local pose joints.

$$APE[j] = \frac{1}{NF}\sum_{n\in N}\sum_{f\in F}\left\|H_f\left[j\right] - \hat{H}_f\left[j\right]\right\|_2, AVE[j] = \frac{1}{N}\sum_{n\in N}\left\|\delta\left[j\right] - \hat{\delta}\left[j\right]\right\|_2 \tag{5}$$

The average positional error(APE.) and Average Variance Error (AVE.) are adopted from [62]. APE. for a specific joint j is determined by computing the mean of the L2 distances between the generated and ground truth joint positions

Table 3: Comparison of our method with baseline methods on AMASS [51] dataset. We achieve state-of-the-art performance on APE. and AVE. . *Root joint, global traj. and mean local* metrics represent the performance of translation and rotation. *Mean global* represents the performance of body joints and global translation.

Methods	Average Positional Error ↓				Average Variance Error ↓			
	root joint	global traj.	mean local	mean global	root joint	global traj.	mean local	mean global
Reg.	5.8786	5.5334	0.6422	5.9199	35.3873	35.3865	0.1476	35.4832
B/L [61]	0.3841	0.3733	0.1839	0.4693	0.1143	0.1138	0.0152	0.1260
Ours	**0.3653**	**0.3546**	**0.1287**	**0.4182**	**0.1111**	**0.1108**	**0.0087**	**0.1183**

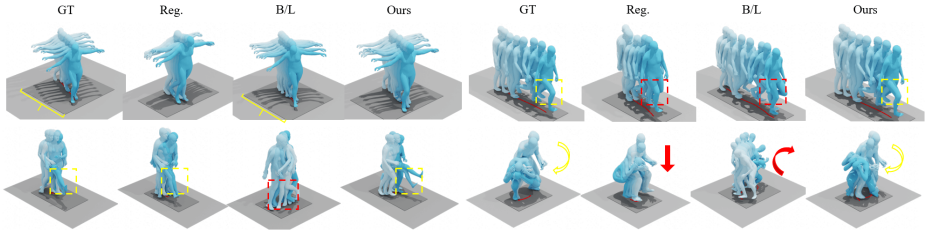

Fig. 8: Comparison of different methods. Yellow color represents details that should be paid attention to, and the red color represents inaccuracies.

across the frames (F) and samples (N). The AVE. quantifies the distinction in variations. This metric is defined as the mean of the L2 distances between the generated samples and ground truth samples' variances for the joint j.

5 Conclusions and Discussions

In this paper, we introduce PRO-Motion, a model designed to tackle open-vocabulary text-to-motion generation tasks. It consists of three modules: motion planner, posture-denoizer, and go-denoiser. The motion planner instructs the large language models to generate a sequence of scripts describing the key postures in the target motion. The posture-denoiser transforms a script into a posture, paving the way for open-vocabulary generation. Finally, the go-denoiser, estimates whole-body translations and rotations for all postures, resulting in diverse and realistic motions. Experimental results have shown the superiority of our method compared to other counterparts.

Limitation 1) Dependent on LLMs' stability. Although it is cost-effective, employing GPT-3.5 as the motion planner may generate a few unsatisfactory scripts. This drawback may be addressed by employing more powerful LLMs or fine-tuning a language model end-to-end. 2) Limited motion duration. Pro-Motion generates a fixed number of frames in a motion sequence, similar to previous open-vocabulary text-to-motion methods [33, 45, 83]. Our prompt is designed for eight keyframes, but Pro-Motion can be generalized to a wider range of scenarios by including different numbers of keyframes with flexible durations.

Acknowledgements

The research of Yansong is supported by Shenzhen Ubiquitous Data Enabling Key Lab under grant ZDSYS20220527171406015 and CCF-Tencent Rhino-Bird Open Research Fund.

References

1. Ahn, H., Ha, T., Choi, Y., Yoo, H., Oh, S.: Text2action: Generative adversarial synthesis from language to action. In: ICRA. pp. 5915–5920 (2018)
2. Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., et al.: Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691 (2022)
3. Aksan, E., Kaufmann, M., Cao, P., Hilliges, O.: A spatio-temporal transformer for 3d human motion prediction. In: 3DV. pp. 565–574 (2021)
4. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: Teach: Temporal action composition for 3d humans. In: 3DV. pp. 414–423 (2022)
5. Barsoum, E., Kender, J., Liu, Z.: Hp-gan: Probabilistic 3d human motion prediction via gan. In: CVPRW. pp. 1418–1427 (2018)
6. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR. pp. 18392–18402 (2023)
7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NIPS pp. 1877–1901 (2020)
8. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al.: Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712 (2023)
9. Cai, Y., Huang, L., Wang, Y., Cham, T.J., Cai, J., Yuan, J., Liu, J., Yang, X., Zhu, Y., Shen, X., et al.: Learning progressive joint propagation for human motion prediction. In: ECCV. pp. 226–242 (2020)
10. Chai, J., Hodgins, J.K.: Constraint-based motion optimization using a statistical dynamic model. In: SIGGRAPH, pp. 8–es (2007)
11. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: CVPR. pp. 18000–18010 (2023)
12. Chiu, H.k., Adeli, E., Wang, B., Huang, D.A., Niebles, J.C.: Action-agnostic human pose forecasting. In: WACV. pp. 1423–1432 (2019)
13. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
14. Cong, P., Dou, Z.W., Ren, Y., Yin, W., Cheng, K., Sun, Y., Long, X., Zhu, X., Ma, Y.: Laserhuman: Language-guided scene-aware human motion generation in free environment. arXiv preprint arXiv:2403.13307 (2024)
15. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: CVPR. pp. 9760–9770 (2023)
16. Dai, W., Chen, L.H., Wang, J., Liu, J., Dai, B., Tang, Y.: Motionlcm: Real-time controllable motion generation via latent consistency model. arXiv preprint arXiv:2404.19759 (2024)

17. Dang, L., Nie, Y., Long, C., Zhang, Q., Li, G.: Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In: ICCV. pp. 11467–11476 (2021)
18. Delmas, G., Weinzaepfel, P., Lucas, T., Moreno-Noguer, F., Rogez, G.: Posescript: 3d human poses from natural language. In: ECCV. pp. 346–362 (2022)
19. Duan, Y., Shi, T., Zou, Z., Lin, Y., Qian, Z., Zhang, B., Yuan, Y.: Single-shot motion completion with transformer. arXiv preprint arXiv:2103.00776 (2021)
20. Elgammal, A., Lee, C.S.: Inferring 3d body pose from silhouettes using activity manifold learning. In: CVPR (2004)
21. Forney, G.D.: The viterbi algorithm. IEEE **61**(3), 268–278 (1973)
22. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: ICCV. pp. 4346–4354 (2015)
23. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
24. Gopalakrishnan, A., Mali, A., Kifer, D., Giles, L., Ororbia, A.G.: A neural temporal model for human motion prediction. In: CVPR. pp. 12116–12125 (2019)
25. Gui, L.Y., Wang, Y.X., Liang, X., Moura, J.M.: Adversarial geometry-aware human motion prediction. In: ECCV. pp. 786–803 (2018)
26. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: CVPR. pp. 5152–5161 (2022)
27. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: ACM MM. pp. 2021–2029 (2020)
28. Hao, Y., Chi, Z., Dong, L., Wei, F.: Optimizing prompts for text-to-image generation. arXiv preprint arXiv:2212.09611 (2022)
29. Harvey, F.G., Pal, C.: Recurrent transition networks for character locomotion. In: SIGGRAPH Asia, pp. 1–4 (2018)
30. Hernandez, A., Gall, J., Moreno-Noguer, F.: Human motion prediction via spatio-temporal inpainting. In: ICCV. pp. 7134–7143 (2019)
31. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NIPS **33**, 6840–6851 (2020)
32. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
33. Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., Liu, Z.: Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. arXiv preprint arXiv:2205.08535 (2022)
34. Hong, S., Seo, J., Hong, S., Shin, H., Kim, S.: Large language models are frame-level directors for zero-shot text-to-video generation. arXiv preprint arXiv:2305.14330 (2023)
35. Huang, C., Mees, O., Zeng, A., Burgard, W.: Visual language maps for robot navigation. In: ICRA. pp. 10608–10615 (2023)
36. Huang, W., Abbeel, P., Pathak, D., Mordatch, I.: Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In: ICML. pp. 9118–9147 (2022)
37. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-rnn: Deep learning on spatio-temporal graphs. In: CVPR. pp. 5308–5317 (2016)
38. Ji, Y., Xu, F., Yang, Y., Shen, F., Shen, H.T., Zheng, W.S.: A large-scale rgb-d database for arbitrary-view human action recognition. In: ACM MM. pp. 1510–1518 (2018)

39. Kaufmann, M., Aksan, E., Song, J., Pece, F., Ziegler, R., Hilliges, O.: Convolutional autoencoders for human motion infilling. In: 3DV. pp. 918–927 (2020)
40. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
41. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
42. Li, R., Zhao, J., Zhang, Y., Su, M., Ren, Z., Zhang, H., Tang, Y., Li, X.: Finedance: A fine-grained choreography dataset for 3d full body dance generation. In: CVPR (2023)
43. Lin, H., Zala, A., Cho, J., Bansal, M.: Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. arXiv preprint arXiv:2309.15091 (2023)
44. Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motion-x: A large-scale 3d expressive whole-body human motion dataset. Advances in Neural Information Processing Systems (2023)
45. Lin, J., Chang, J., Liu, L., Li, G., Lin, L., Tian, Q., Chen, C.w.: Being comes from not-being: Open-vocabulary text-to-motion generation with wordless training. In: CVPR. pp. 23222–23231 (2023)
46. Lin, X., Amer, M.R.: Human motion modeling using dvgans. arXiv preprint arXiv:1804.10652 (2018)
47. Liu, X., Yin, J., Liu, J., Ding, P., Liu, J., Liu, H.: Trajectorycnn: a new spatio-temporal feature learning network for human motion prediction. TCSVT **31**(6), 2133–2146 (2020)
48. Liu, Z., He, Y., Wang, W., Wang, W., Wang, Y., Chen, S., Zhang, Q., Yang, Y., Li, Q., Yu, J., et al.: Internchat: Solving vision-centric tasks by interacting with chatbots beyond language. arXiv preprint arXiv:2305.05662 (2023)
49. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. SIGGRAPH Asia (2015)
50. Lu, Y., Feng, W., Zhu, W., Xu, W., Wang, X.E., Eckstein, M., Wang, W.Y.: Neuro-symbolic procedural planning with commonsense prompting. arXiv preprint arXiv:2206.02928 (2022)
51. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: ICCV. pp. 5442–5451 (2019)
52. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: CVPR. pp. 2891–2900 (2017)
53. Min, J., Chen, Y.L., Chai, J.: Interactive generation of human animation with deformable motion models. TOG **29**(1), 1–12 (2009)
54. Ngo, J.T., Marks, J.: Spacetime constraints revisited. In: SIGGRAPH. pp. 343–350 (1993)
55. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML. pp. 8162–8171 (2021)
56. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. NIPS pp. 27730–27744 (2022)
57. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR (2019)
58. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR. pp. 10975–10985 (2019)

59. Pavllo, D., Feichtenhofer, C., Auli, M., Grangier, D.: Modeling human motion with quaternion-based neural networks. IJCV **128**, 855–872 (2020)
60. Pavllo, D., Grangier, D., Auli, M.: Quaternet: A quaternion-based recurrent model for human motion. arXiv preprint arXiv:1805.06485 (2018)
61. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: ICCV. pp. 10985–10995 (2021)
62. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: ECCV. pp. 480–497 (2022)
63. Petrovich, M., Black, M.J., Varol, G.: TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In: ICCV (2023)
64. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data **4**(4), 236–252 (2016)
65. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: Babel: Bodies, action and behavior with english labels. In: CVPR. pp. 722–731 (2021)
66. Qin, J., Zheng, Y., Zhou, K.: Motion in-betweening via two-stage transformers. TOG **41**(6), 1–16 (2022)
67. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
68. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
69. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (2019)
70. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: EMNLP (2020)
71. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
72. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. SIGGRAPH Asia **36**(6) (2017)
73. Rose, C., Guenter, B., Bodenheimer, B., Cohen, M.F.: Efficient generation of motion transitions using spacetime constraints. In: SIGGRAPH. pp. 147–154 (1996)
74. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
75. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418 (2023)
76. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: CVPR. pp. 1010–1019 (2016)
77. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y.: Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. arXiv preprint arXiv:2303.17580 (2023)
78. Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., Garg, A.: Progprompt: Generating situated robot task plans using large language models. In: ICRA. pp. 11523–11530 (2023)
79. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML. pp. 2256–2265 (2015)

80. Song, C.H., Wu, J., Washington, C., Sadler, B.M., Chao, W.L., Su, Y.: Llm-planner: Few-shot grounded planning for embodied agents with large language models. In: ICCV. pp. 2998–3009 (2023)
81. Tang, X., Wang, H., Hu, B., Gong, X., Yi, R., Kou, Q., Jin, X.: Real-time controllable motion transition for characters. TOG **41**(4), 1–10 (2022)
82. Tang, Y., Liu, J., Liu, A., Yang, B., Dai, W., Rao, Y., Lu, J., Zhou, J., Li, X.: Flag3d: A 3d fitness activity dataset with language instruction. In: CVPR. pp. 22106–22117 (2023)
83. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: ECCV. pp. 358–374 (2022)
84. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
85. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
86. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NIPS (2017)
87. Wan, W., Dou, Z., Komura, T., Wang, W., Jayaraman, D., Liu, L.: Tlcontrol: Trajectory and language control for human motion synthesis. arXiv preprint arXiv:2311.17135 (2023)
88. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models for human motion. TPAMI **30**(2), 283–298 (2007)
89. Witkin, A., Kass, M.: Spacetime constraints. SIGGRAPH **22**(4), 159–168 (1988)
90. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023)
91. Xiao, Z., Wang, T., Wang, J., Cao, J., Zhang, W., Dai, B., Lin, D., Pang, J.: Unified human-scene interaction via prompted chain-of-contacts. arXiv preprint arXiv:2309.07918 (2023)
92. Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., Wang, L.: Mm-react: Prompting chatgpt for multimodal reasoning and action. arXiv preprint arXiv:2303.11381 (2023)
93. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
94. Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al.: Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414 (2022)
95. Zhang, B., Cheng, Y., Wang, C., Zhang, T., Yang, J., Tang, Y., Zhao, F., Chen, D., Guo, B.: Rodinhd: High-fidelity 3d avatar generation with diffusion models (2024)
96. Zhang, B., Cheng, Y., Yang, J., Wang, C., Zhao, F., Tang, Y., Chen, D., Guo, B.: Gaussiancube: Structuring gaussian splatting using optimal transport for 3d generative modeling. arXiv preprint arXiv:2403.19655 (2024)
97. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022)
98. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. arXiv preprint arXiv:2304.01116 (2023)

99. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
100. Zhang, X., van de Panne, M.: Data-driven autocompletion for keyframe animation. In: SIGGRAPH. pp. 1–11 (2018)
101. Zheng, K., Zhou, K., Gu, J., Fan, Y., Wang, J., Li, Z., He, X., Wang, X.: Jarvis: a neuro-symbolic commonsense reasoning framework for conversational embodied agents. arXiv preprint arXiv:2208.13266 (2022)
102. Zhou, W., Dou, Z., Cao, Z., Liao, Z., Wang, J., Wang, W., Liu, Y., Komura, T., Wang, W., Liu, L.: Emdm: Efficient motion diffusion model for fast, high-quality motion generation. arXiv preprint arXiv:2312.02256 (2023)
103. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: CVPR (2019)
104. Zhou, Y., Lu, J., Barnes, C., Yang, J., Xiang, S., et al.: Generative tweening: Long-term inbetweening of 3d human motions. arXiv preprint arXiv:2005.08891 (2020)