

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

Jinpeng Liu^{1*}

Wenxun Dai^{1*}

Chunyu Wang^{2*}

Yiji Cheng¹

Yansong Tang^{1†}

Xin Tong²

¹Shenzhen International Graduate School, Tsinghua University

²Microsoft Research Asia

{liujp22@mails., daiwx23@mails., cyj22@mails., tang.yansong@sz.tsinghua.edu.cn}

{chnuwa, xtong}@microsoft.com

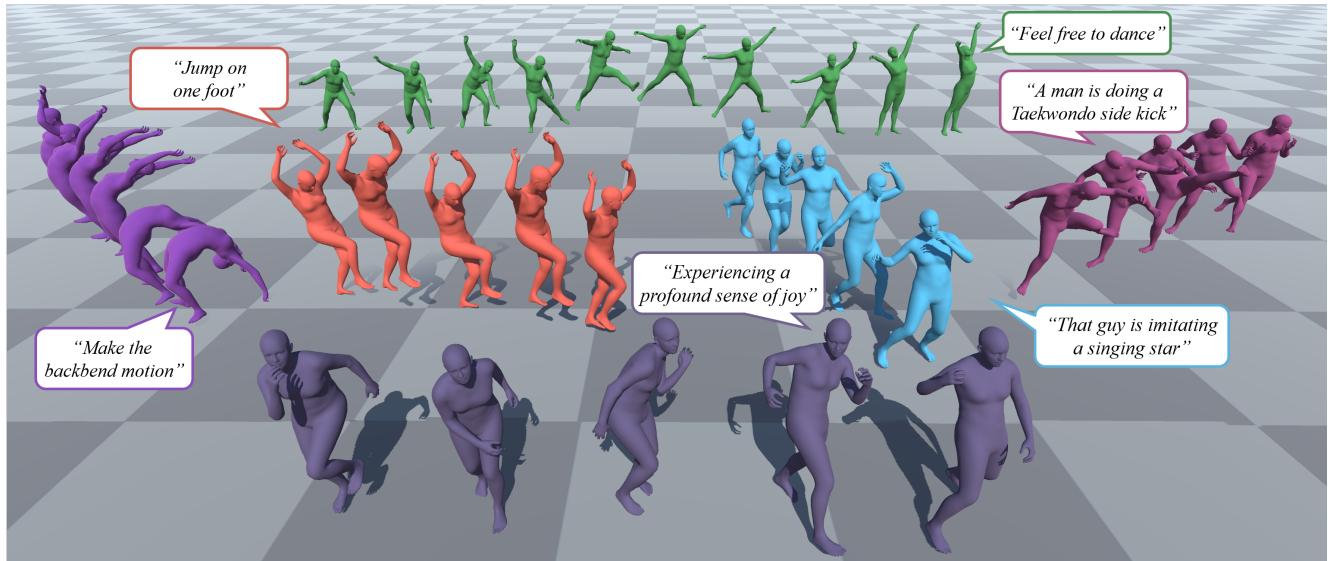


Figure 1. Exemplary motions generated by our proposed PRO-Motion system. Different from conventional models trained on paired text-motion data, our PRO-Motion can generate 3D human motion with global body translation and rotation from *open-world* text prompts, such as “Jump on one foot” and “Experiencing a profound sense of joy”.

Abstract

Conventional text-to-motion generation methods are usually trained on limited text-motion pairs, making them hard to generalize to open-world 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-diffuser and go-diffuser. The mo-

tion planner instructs Large 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-diffuser, which transforms a script to a posture, paving the way for open-world generation. Finally, go-diffuser, implemented as another diffusion model, estimates whole-body translations and rotations for all postures, resulting in realistic 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-world prompts such as “Experiencing a profound sense of joy”. The project page is available at <https://moonsliu.github.io/Pro-Motion/>.

*Equal Contribution

†Corresponding Author

¹PRO-Motion: Plan, postuRe and gO for text-to-Motion generation

1. Introduction

Text-to-motion generation has attracted rapidly increasing attention [1, 63, 85] 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 [5, 25, 62, 63] and Diffusion Models [12, 15, 76, 85, 94, 96, 97] from paired text-motion data and have achieved reasonable generation results when the text prompts are similar as those in the training set. Fig. 2 (a) illustrates this paradigm. However, they struggle to handle open-world text prompts beyond the existing datasets, which is a core challenge that has to be addressed. Otherwise, they can only generate limited “toy-like” motions.

Some recent work [34, 44, 84] 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 [68] 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 on natural languages, 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 **Plan**, **postuRe**, and **Go** for open-world 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 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 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-diffuser on a public dataset [17], which perceives the connection between structured pose descriptions and body parts leading to diverse and realistic postures. 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 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 [87] go-diffuser module is designed to capture the inner connection between key poses.

To verify the effectiveness of our proposed 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-world 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*”.

2. Related Work

Text-to-Motion Generation. Based on existing labeled motion capture datasets [25, 26, 39, 45, 52, 65, 66, 77, 83], existing works have explored various generative models for text-driven motion generation, such as GANs [1, 46], VAEs [5, 25, 62, 63, 84] and Diffusion Models [12, 15, 76, 85, 94, 96, 97]. However, these methods are constrained by the heavy reliance on limited text-motion paired datasets.

To tackle this problem, some works [34, 44] try to leverage the current powerful large-scale pre-trained models, i.e., CLIP [68], to overcome the data limitation and achieve open-vocabulary motion generation. AvatarCLIP [34] 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 [44] 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-world motion generation.

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

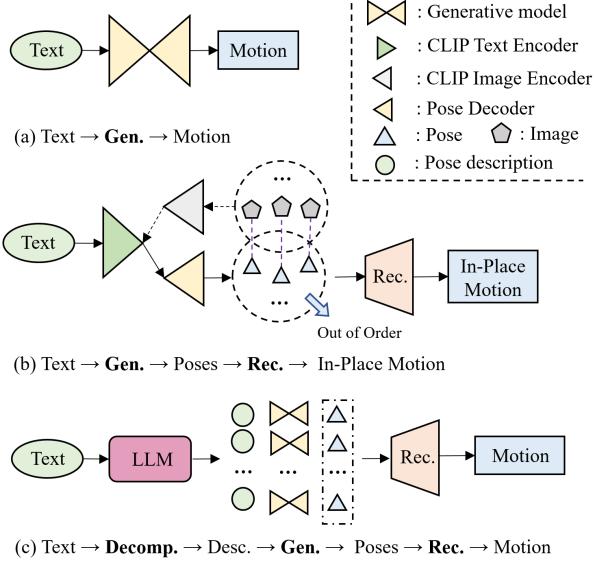


Figure 2. Comparison of different paradigms for text-to-motion generation. (a) Most existing models leverage the generative models [22, 33, 41] to construct the relationship between text and motion based on text-motion pairs. (b) Some methods render 3D poses to images and employ the image space of CLIP 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.

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 [6, 13, 20, 23, 24, 38, 53, 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, 61] and GCNs [16] are proposed to enhance the modeling of temporal and movement relationships. The emergence of Transformers [4, 10] 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 [55, 74, 90] that involve solving optimization problems, as well as statistical models [11, 54, 89]. More recently, some neural network-based methods such as RNNs [28, 82, 99], CNNs [30, 40, 102], and Transformers [18, 54, 67] 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) [8, 9, 57, 86, 95, 98] 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. [7, 27] combine large language models with diffusion-based generative models [33, 72] aimed at generating prompts for higher-quality image generation. [35, 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 [42, 48, 78, 91, 93], 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, 36, 37, 51, 79, 81, 92, 100] 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.

3. Method

3.1. Preliminaries

Denoising Diffusion Probabilistic Models (DDPMs). DDPMs, as detailed in [33, 56, 80], involve two Markov chains: a forward chain that diffuses data to noise, and a reverse chain that transforms noise back to data. Formally, the diffusion forward process iteratively adds Gaussian noise to the original signal x_0 to generate a sequence of noised samples, i.e., $\{x_t\}_{t=1}^T$, which can be formulated as follows:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I), \quad (1)$$

where $q(x_t|x_{t-1})$ represents the conditional probability distribution of x_t conditioned by x_{t-1} and $\beta_t \in (0, 1)$ denotes the level of noise at time step t . When T is sufficiently large, we can assume that $x_T \sim \mathcal{N}(0, I)$.

For the conditional synthesis setting, the denoiser is designed to model the distribution $p(x_0|c)$ by sampling a random noise from the prior distribution $\mathcal{N}(0, I)$, followed by the reversed diffusion process to convert it back to x_0 . In our setting, we follow [69] and directly predict the original sample, i.e., $\hat{x}_0 = f_\theta(x_t, t, c)$ with the simple objective [33]:

$$\min_{\theta} \mathcal{L} = \mathbb{E}_{x_0, t, c} [\|x_0 - f_\theta(x_t, t, c)\|_2^2]. \quad (2)$$

According to [33], at each time step t , we predict the original sample $\hat{x}_0 = f_\theta(x_t, t, c)$ and diffuses it back to x_{t-1} . The sampling from $p(x_0|c)$ is carried out iteratively, starting from $t = T$, until x_0 is obtained. To train our diffusion model f_θ , we employ classifier-free guidance [32] by

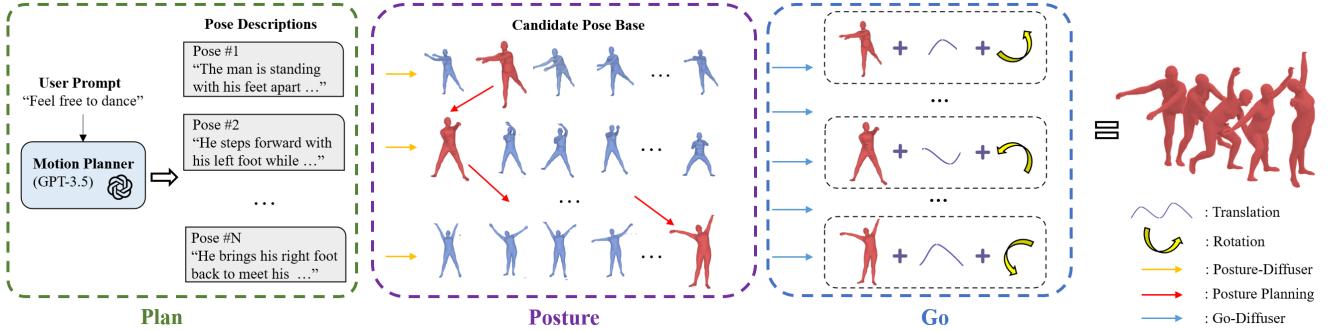


Figure 3. **Illustration of our framework for open-world text-to-motion generation from language.** Specifically, we employ the large language models as the motion planner to plan pose scripts. Then, our *Posture-Diffuser* module receives discrete pose descriptions and generates corresponding poses to construct a candidate posture base. *Posture Planning* is utilized to select the reasonable pose sequence from the candidate posture base. Finally, the *Go-Diffuser* module increases the motion frames and infers the translation and rotation.

randomly setting $c = \emptyset$ for 10% of the samples. This allows f_θ to learn both the conditioned and the unconditioned distributions, making $f_\theta(x_t, t, \emptyset)$ approximate $p(x_0)$. We use the coefficient w to balance the trade-off between sample diversity and fidelity:

$$f_\theta^w(x_t, t, c) = f_\theta(x_t, t, \emptyset) + w \cdot (f_\theta(x_t, t, c) - f_\theta(x_t, t, \emptyset)). \quad (3)$$

3.2. Motion Planner

As depicted in Fig. 3, when presented with a user prompt, such as “*Feel free to dance*”, we exploit GPT-3.5 [57] 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. For more details about the prompt design, please refer to the supplementary materials.

3.3. Posture-Diffuser

In this section, we present our *Posture-Diffuser* module, which aims to generate key poses that align with the localized body part descriptions provided by the *Motion Planner* in Sec. 3.2. 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 [87]. The intermediate residual pose feature serves as the query vector, while the text embeddings extracted from the frozen Distill-BERT [75] 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.

Due to the sampling diversity of DDPMs [33, 56, 69, 80], the *Posture-Diffuser* module can generate multiple plausible poses corresponding to each pose description, we introduce our *Posture Planning* module, which aims 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 [14], and a pose encoder Θ that uses the VPoser encoder [58]. Both the encoders produce the L2-normed embeddings for computing similarity. We employ the Viterbi algorithm [19] to search for the most reasonable pose path.

Specifically, suppose we have a set of F pose descriptions denoted as $\{t_i\}_{i=1}^F$, and for each pose description t_i , we have a collection of L generated candidate poses represented as $\{p_j^i\}_{j=1}^K$, 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:

$$A_{jk}^i = \frac{\exp\left(\Theta(p_j^{i-1})^T \Theta(p_k^i)\right)}{\sum_{l=1}^L \exp\left(\Theta(p_j^{i-1})^T \Theta(p_l^i)\right)}. \quad (4)$$

Similarly, 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 as below:

$$E_j^i = \frac{\exp\left(\Phi(t_i)^T \Theta(p_j^i)\right)}{\sum_{l=1}^L \exp\left(\Phi(t_i)^T \Theta(p_l^i)\right)}. \quad (5)$$

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

$$\underset{G}{\operatorname{argmax}} P(G) = \prod_{i=1}^F P(g_i | g_{i-1}) = E_{g_1}^1 \prod_{i=2}^F E_{g_i}^i A_{g_{i-1} g_i}^i. \quad (6)$$

3.4. Go-Diffuser

To interpolate and predict global information such as translation and rotation for the key poses obtained in Sec. 3.3, we introduce our *Go-Diffuser* module in this section. Our module is based on a diffusion model, as illustrated in Fig. 4 (b). As transformer [87] structure has been proved efficient in the field of motion generation [62–64, 85], 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.

4. Experiments

4.1. Datasets

In our experiments, we utilize the pose data, motion data, and textual descriptions sourced from AMASS [52], PoseScript [17], Motion-X [45], and HumanML3D [26]. AMASS unifies various optical marker-based mocap

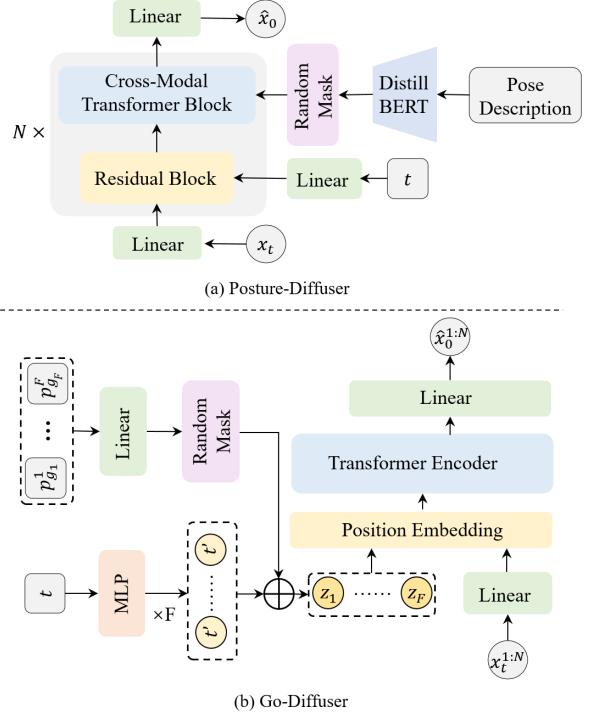


Figure 4. **Illustration of our Dual-Diffusion model.** (a) *Posture-Diffuser* module is designed to predict the original pose conditioned by the pose description. The model consists of N identical layers, with each layer featuring a residual block for incorporating time step information and a cross-modal transformer block for integrating the condition text. (b) *Go-Diffuser* 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. 3.3 are regarded as independent tokens. We perform attention operations [87] between these tokens and noised motion independently, which can significantly improve the perception ability between every condition pose and motion sequence.

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 language descriptions. We select two subsets from Motion-X for conducting open-world text-to-motion generation experiments. Specifically, we employ sentence transformers [70, 71] to compute the similarity between the text descriptions in IDEA-400 [45], a high-quality motion language subset within Motion-X, and the text descriptions in HumanML3D. We filter out pairs with similarity greater than a specified threshold α , e.g., 0.45, yielding a motion language dataset comprising 368 text-motion pairs as

Table 1. Comparison of our method with previous methods on the subsets of the IDEA-400 [45] dataset, i.e., *ood368* and *kungfu*. We achieve superior performance on R precision, and the MultiModal Dist. MDM [85] is for supervised learning. MotionCLIP [84], Codebook+Interpolation [34], Avatarclip [34] and OOHMG [44] are designed for open vocabulary text-to-motion generation.

	Text-motion				FID ↓	MultiModal Dist ↓	Smooth →
	R@10 ↑	R@20 ↑	R@30 ↑	MedR ↓			
“test on ood368 subset”							
MDM [85]	17.81	34.06	48.75	31.20	3.500541	2.613644	0.000114
MotionCLIP [84]	16.25	35.62	52.81	28.90	2.227522	2.288905	0.000073
Codebook+Interpolation [34]	15.62	31.25	46.56	32.80	4.084785	2.516041	0.000146
AvatarCLIP [34]	15.31	31.56	47.19	32.60	4.181952	2.449695	0.000146
OOHMG [44]	15.62	34.06	48.75	29.80	3.982753	2.149275	0.000758
Ours	20.25	36.56	53.14	26.10	1.488678	1.534521	0.001312
“test on kungfu subset”							
MDM [85]	12.50	29.69	42.19	37.50	12.060187	3.725436	0.000735
MotionCLIP [84]	15.62	29.69	46.88	32.50	17.414746	4.297871	0.000123
Codebook+Interpolation [34]	10.94	20.31	29.69	37.50	2.521690	2.764137	0.000138
AvatarCLIP [34]	15.62	31.25	46.88	32.50	1.966764	2.497678	0.000171
OOHMG [44]	14.06	32.81	48.44	32.50	4.904853	2.471666	0.000847
Ours	20.31	34.38	50.00	31.00	4.124218	2.374380	0.001559

our first test dataset. Moreover, we choose the *kungfu* subset of Motion-X as our second test dataset.

4.2. Open-World Motion Generation

In this section, we first introduce the supervised learning baseline [84, 85], open-world baselines [34, 44, 84], and the evaluation metrics [26, 62]. Then we discuss the comparative experimental results with these baselines.

MDM Baseline: MDM [85] employs a supervised learning approach utilizing the diffusion model. However, its performance often degenerates when applied to a new setting, i.e., open-vocabulary motion generation. We trained it on the SMPL-H [49, 59, 73] version data of AMASS with HumanML3D annotation.

MotionCLIP Baseline: MotionCLIP [84] is a supervised open-vocabulary method which trained on AMAS data with BABEL [66] 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 [59] 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 [34] is an optimizer-based method. It also includes a text-to-pose stage via matching and uses the matched poses to search the most related motion in the latent space of a motion VAE [41] trained on the AMASS dataset.

Evaluation Metrics is adopted from [26, 62], 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 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 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.

As shown in Fig. 5, for the motion description “bends over”, the supervised-based method MDM [85] generated the reasonable motion sequence, which shows the process from upright to prone. While OOHMG /citeooohmg involves more leg movement, which is not accurate for the description. Other methods utilizing CLIP [68], such as MotionCLIP [84], Codebook+Interpolation [34], and AvatarCLIP [34] mistake the key pose sequence, their spine first curves to a certain degree, and then the degree of bending decreases. Due to the lack of temporal priors in CLIP, the generated pose sequence in the first stage does not follow the right order. Moreover, for the motion description “bury one’s head and cry, and finally crouched down”, methods such as MDM [85] based on the supervised learning paradigm often fail in similar cases and cannot generate un-seen motion. Due to the gap between motion description and image description, matching text and motion via the language space of CLIP is not effective. They struggle to deal with detailed and

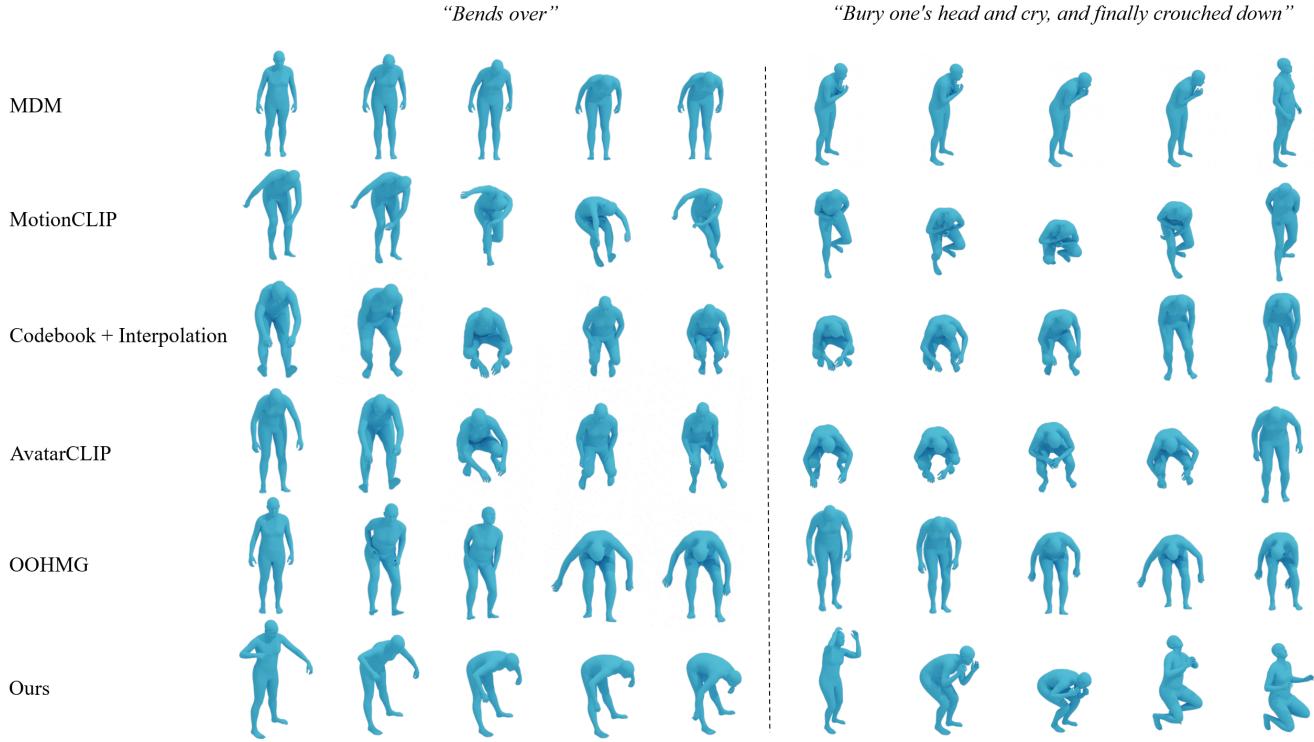


Figure 5. Comparation of our methods with previous text-to-motion generation methods.

Table 2. Comparison of our method with baseline methods on AMASS[52] dataset. We achieve state-of-the-art performance on Average Positional Error and Average Variance Error. *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
Regression	5.878673	5.53344	0.642252	5.919954	35.387340	35.386562	0.147606	35.483219
Baseline[62]	0.384152	0.373394	0.183978	0.469322	0.114308	0.113845	0.015207	0.126049
Ours	0.365327	0.354685	0.128763	0.418265	0.111131	0.110855	0.008708	0.118334

precise motion descriptions. In Tab. 1, quantitative metrics demonstrate the superiority of our model over other methods in terms of semantic consistency and motion rationality.

4.3. Ablation Study

Posture-Diffuser In comparison to the zero-shot open-vocabulary text-to-pose generation methods [34, 44], we first utilize LLMs to translate the pose description into localized body part description, which is fed to our pose generator in Sec. 3.3 to generate pose precisely. As shown in Fig. 6, 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-world text-based pose synthesis.

Go-Diffuser. 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 perception power. Based on our observation, the translation and rotation of the people could be estimated by analyzing the variations in body parts between adjacent poses. We treat

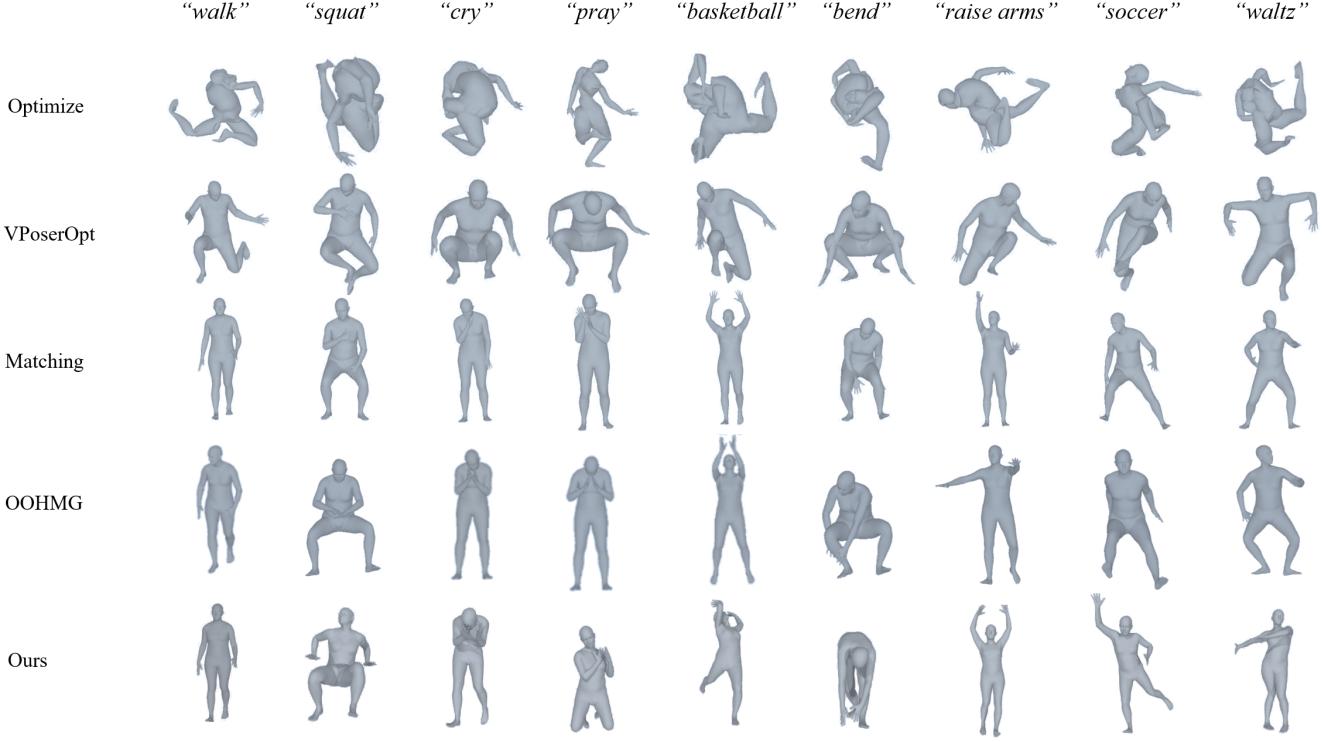


Figure 6. Comparison of our method with previous text-to-pose generation methods.

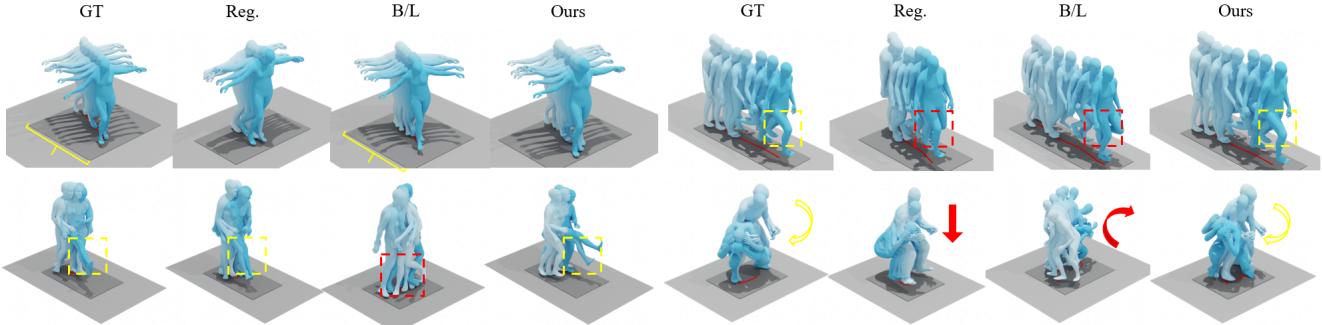


Figure 7. Comparison of different methods. Yellow color represents details that need attention, and red color represents inaccuracies.

the process of estimating global information from the key poses as a reconstruction task. Thus, we designed a simple MLP-based network to achieve this first. The transformer structure has been proven correct in the motion generation filed [62, 63], thus we design the baseline method by extracting pose sequence features as the condition to inject into the diffusion model. As shown in Fig. 7, 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. 2, our method achieved state-of-the-art results both on the average positional error and the average variance error of global trajectory, rotation and local pose joints.

5. Conclusion

In this paper, we introduce PRO-Motion, a model designed to tackle open-world text-to-motion generation tasks. It con-

sists of three modules: motion planner, posture-diffuser, and go-diffuser. The motion planner instructs the large language models to generate a sequence of scripts describing the key postures in the target motion. The posture-diffuser transforms a script into a posture, paving the way for open-world generation. Finally, the go-diffuser, 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.

References

- [1] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *ICRA*, pages 5915–5920, 2018. 2
- [2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 3
- [3] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *3DV*, 2019. 16
- [4] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *3DV*, pages 565–574, 2021. 3
- [5] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Güл Varol. Teach: Temporal action composition for 3d humans. In *3DV*, pages 414–423, 2022. 2
- [6] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *CVPRW*, pages 1418–1427, 2018. 3
- [7] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 3
- [8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NIPS*, pages 1877–1901, 2020. 3
- [9] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 3
- [10] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. Learning progressive joint propagation for human motion prediction. In *ECCV*, pages 226–242, 2020. 3
- [11] Jinxiang Chai and Jessica K Hodgins. Constraint-based motion optimization using a statistical dynamic model. In *SIGGRAPH*, pages 8–es. 2007. 3
- [12] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, pages 18000–18010, 2023. 2
- [13] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *WACV*, pages 1423–1432, 2019. 3
- [14] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 4
- [15] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *CVPR*, pages 9760–9770, 2023. 2
- [16] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *ICCV*, pages 11467–11476, 2021. 3
- [17] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3d human poses from natural language. In *ECCV*, pages 346–362, 2022. 2, 5, 13, 14
- [18] Yinglin Duan, Tianyang Shi, Zhengxia Zou, Yenan Lin, Zhehui Qian, Bohan Zhang, and Yi Yuan. Single-shot motion completion with transformer. *arXiv preprint arXiv:2103.00776*, 2021. 3
- [19] G David Forney. The viterbi algorithm. *IEEE*, 61(3):268–278, 1973. 4
- [20] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *ICCV*, pages 4346–4354, 2015. 3
- [21] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *ICCV*, 2021. 16
- [22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3
- [23] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G Ororbia. A neural temporal model for human motion prediction. In *CVPR*, pages 12116–12125, 2019. 3
- [24] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. Adversarial geometry-aware human motion prediction. In *ECCV*, pages 786–803, 2018. 3
- [25] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *ACM MM*, pages 2021–2029, 2020. 2
- [26] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, pages 5152–5161, 2022. 2, 5, 6
- [27] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *arXiv preprint arXiv:2212.09611*, 2022. 3

- [28] Félix G Harvey and Christopher Pal. Recurrent transition networks for character locomotion. In *SIGGRAPH Asia*, pages 1–4. 2018. 3
- [29] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 14
- [30] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *ICCV*, pages 7134–7143, 2019. 3
- [31] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 16
- [32] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3
- [33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NIPS*, 33:6840–6851, 2020. 3, 4, 14
- [34] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022. 2, 6, 7
- [35] Susung Hong, Junyoung Seo, Sunghwan Hong, Heeseong Shin, and Seungryong Kim. Large language models are frame-level directors for zero-shot text-to-video generation. *arXiv preprint arXiv:2305.14330*, 2023. 3
- [36] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *ICRA*, pages 10608–10615, 2023. 3
- [37] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *ICML*, pages 9118–9147, 2022. 3
- [38] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *CVPR*, pages 5308–5317, 2016. 3
- [39] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale rgb-d database for arbitrary-view human action recognition. In *ACM MM*, pages 1510–1518, 2018. 2
- [40] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Convolutional autoencoders for human motion infilling. In *3DV*, pages 918–927, 2020. 3
- [41] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3, 6
- [42] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 3
- [43] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023. 3
- [44] Junfan Lin, Jianlong Chang, Lingbo Liu, Guanbin Li, Liang Lin, Qi Tian, and Chang-wen Chen. Being comes from not-being: Open-vocabulary text-to-motion generation with wordless training. In *CVPR*, pages 23222–23231, 2023. 2, 6, 7
- [45] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *NeurIPS*, 2023. 2, 5, 6, 14
- [46] Xiao Lin and Mohamed R Amer. Human motion modeling using dvgans. *arXiv preprint arXiv:1804.10652*, 2018. 2
- [47] Xiaoli Liu, Jianqin Yin, Jin Liu, Pengxiang Ding, Jun Liu, and Huaping Liu. Trajectorycnn: a new spatio-temporal feature learning network for human motion prediction. *TCSVT*, 31(6):2133–2146, 2020. 3
- [48] Zhaoyang Liu, Yinan He, Wenhui Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Yang Yang, Qingyun Li, Jiashuo Yu, et al. Internchat: Solving vision-centric tasks by interacting with chatbots beyond language. *arXiv preprint arXiv:2305.05662*, 2023. 3
- [49] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 6, 13
- [50] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 14
- [51] Yujie Lu, Weixi Feng, Wanrong Zhu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. Neuro-symbolic procedural planning with commonsense prompting. *arXiv preprint arXiv:2206.02928*, 2022. 3
- [52] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451, 2019. 2, 5, 7
- [53] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, pages 2891–2900, 2017. 3
- [54] Jianyuan Min, Yen-Lin Chen, and Jinxiang Chai. Interactive generation of human animation with deformable motion models. *TOG*, 29(1):1–12, 2009. 3
- [55] J Thomas Ngo and Joe Marks. Spacetime constraints revisited. In *SIGGRAPH*, pages 343–350, 1993. 3
- [56] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171, 2021. 3, 4
- [57] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NIPS*, pages 27730–27744, 2022. 3, 4
- [58] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 4
- [59] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 6
- [60] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485*, 2018. 3

- [61] Dario Pavllo, Christoph Feichtenhofer, Michael Auli, and David Grangier. Modeling human motion with quaternion-based neural networks. *IJCV*, 128:855–872, 2020. 3
- [62] Mathis Petrovich, Michael J Black, and Gü̈l Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *ICCV*, pages 10985–10995, 2021. 2, 5, 6, 7, 8
- [63] Mathis Petrovich, Michael J Black, and Gü̈l Varol. Temos: Generating diverse human motions from textual descriptions. In *ECCV*, pages 480–497, 2022. 2, 8, 13, 16
- [64] Mathis Petrovich, Michael J. Black, and Gü̈l Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *ICCV*, 2023. 5
- [65] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 2
- [66] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *CVPR*, pages 722–731, 2021. 2, 6
- [67] Jia Qin, Youyi Zheng, and Kun Zhou. Motion in-betweening via two-stage transformers. *TOG*, 41(6):1–16, 2022. 3
- [68] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2, 6
- [69] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3, 4
- [70] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. 5, 14
- [71] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *EMNLP*, 2020. 5, 14
- [72] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3
- [73] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *SIGGRAPH Asia*, 36(6), 2017. 6, 13
- [74] Charles Rose, Brian Guenter, Bobby Bodenheimer, and Michael F Cohen. Efficient generation of motion transitions using spacetime constraints. In *SIGGRAPH*, pages 147–154, 1996. 3
- [75] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 4
- [76] Yonatan Shafir, Guy Tevet, Roy Karon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 2
- [77] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, pages 1010–1019, 2016. 2
- [78] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueling Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023. 3
- [79] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *ICRA*, pages 11523–11530, 2023. 3
- [80] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265, 2015. 3, 4
- [81] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *ICCV*, pages 2998–3009, 2023. 3
- [82] Xiangjun Tang, He Wang, Bo Hu, Xu Gong, Ruifan Yi, Qilong Kou, and Xiaogang Jin. Real-time controllable motion transition for characters. *TOG*, 41(4):1–10, 2022. 3
- [83] Yansong Tang, Jinpeng Liu, Aoyang Liu, Bin Yang, Wenxun Dai, Yongming Rao, Jiwen Lu, Jie Zhou, and Xiu Li. Flag3d: A 3d fitness activity dataset with language instruction. In *CVPR*, pages 22106–22117, 2023. 2
- [84] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *ECCV*, pages 358–374, 2022. 2, 6
- [85] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2, 5, 6
- [86] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3
- [87] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 2017. 2, 4, 5
- [88] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval—an empirical odyssey. In *CVPR*, pages 6439–6448, 2019. 14
- [89] Jack M Wang, David J Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *TPAMI*, 30(2):283–298, 2007. 3
- [90] Andrew Witkin and Michael Kass. Spacetime constraints. *SIGGRAPH*, 22(4):159–168, 1988. 3
- [91] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 3
- [92] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. *arXiv preprint arXiv:2309.07918*, 2023. 3

- [93] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 3
- [94] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, 2023. 2
- [95] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022. 3
- [96] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 2
- [97] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. *arXiv preprint arXiv:2304.01116*, 2023. 2
- [98] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 3
- [99] Xinyi Zhang and Michiel van de Panne. Data-driven auto-completion for keyframe animation. In *SIGGRAPH*, pages 1–11, 2018. 3
- [100] K Zheng, K Zhou, J Gu, Y Fan, J Wang, Z Li, X He, and XE Wang. Jarvis: a neuro-symbolic commonsense reasoning framework for conversational embodied agents. *arXiv preprint arXiv:2208.13266*, 2022. 3
- [101] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 13
- [102] Yi Zhou, Jingwan Lu, Connelly Barnes, Jimei Yang, Sita Xiang, et al. Generative tweening: Long-term inbetweening of 3d human motions. *arXiv preprint arXiv:2005.08891*, 2020. 3

Appendix

In this supplementary material, we provide additional details and experiments not included in the main paper due to limitations in space.

- Appendix A: Qualitative results of precise control for motion generation in our formulation.
- Appendix B: Details of the motion representations.
- Appendix C: Details of the prompt engineering and the entire system message template.
- Appendix D: Details of our dual-diffusion models.
- Appendix E: Details of the evaluation metrics.

Note: Blue characters denote the main paper's reference.

A. Precise Pose Control

As shown in Fig. 8, our *Posture-Diffuser* module can accurately generate poses from edited pose descriptions, which becomes particularly beneficial when users are not content with the key pose descriptions produced by the *Motion Planner*. In such cases, users have the option to manually edit the description of specific body parts to gain precise control over the generated poses. For example, in the Control #1 of Pose Description #1, we replace the description of “their left arm is behind the right with their right elbow bent at 90 degrees.” with “the left hand is beside the right arm and the right hand is to the left of the left hand”. It’s obvious the position of the hands changes correctly. As well, when we delete the description of “The left foot is stretched backwards and is behind the right foot” in the Control #1 of Pose Description #2, the position of the left foot changes correctly. Furthermore, we could find that although the scripts are simple and limited to a small space, they are expressive to cover all possible postures due to their compositional nature.

B. Motion Representations

Skinned Multi-Person Linear model (SMPL)[49][73] is a skinned vertex-based model that accurately represents various body shapes in natural human poses. It’s widely utilized in the human motion generation field. So we adopted the SMPL-based human model as well. SMPL begins with an artist-created mesh with $N = 6890$ vertices and $K = 23$ joints. The mesh has the same topology for men and women, spatially varying resolution, a clean quad structure, a segmentation into parts, initial blend weights, and a skeletal rig. We follow the SMPL-based representation of TEMOS [63] and construct the feature vector for SMPL data.

- **Translation.** It 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. The dimension of the translation is 3.
- **Root Orientation.** It contains one rotation, and we utilize the 6D continuous representation [101] to store it. So the dimension of the root orientation is $1 * 6 = 6$.

- **Pose Body.** The motion data is from the SMPL-H [49, 73] 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 [101]. So the dimension of the pose body is $21 * 6 = 126$.

So the final dimension of the feature is $3 + 1 * 6 + 21 * 6 = 135$. In our experiments, we remove all the motion sequences that are less than 64 frames, and the motion sequences longer than 64 frames are processed as 64 frames. The data sample will be in $\mathbb{R}^{64 * 135}$.

```

1 # read motion
2 motion_parms = {
3     'trans': motion[:, :, 3], # controls the global
4         body position
5     'root_orient': motion[:, 3 : 3 + 6], # controls the global root orientation
6     'pose_body': motion[:, 9 : 9 + 21 * 6], # controls the body
7 }
```

Listing 1. motion data

When we perform the translation and rotation prediction, we only know the local attributes of the adjacent poses, but not their global location information. So we must normalize the pose position. Following [63], the translation of neighboring poses is subtracted and represented by the instantaneous velocity as the translation attribute of the current frame.

```

1 # extract the root gravity axis
2 # for smpl it is the last coordinate
3 root = trans[..., 2]
4 trajectory = trans[..., [0, 1]]
5
6 # Compute the difference of trajectory (for X and Y axis)
7 vel_trajectory = torch.diff(trajectory, dim=-2)
8 # 0 for the first one => keep the dimentionality
9 vel_trajectory = torch.cat((0 * vel_trajectory
    [..., [0], :], vel_trajectory), dim=-2)
```

Listing 2. translation normalization

C. Prompt Design

Fig. 9 illustrates the complete prompt used by our *Motion Planner* in Sec. 3.2. We first define the overall objective and task requirements and then establish five fundamental rules of body parts including bending degree, relative distance, relative position, orientation, and ground contact. Additionally, we specify the body parts to which each rule applies. Through this rule-based approach, we can guide the LLM to generate precise key pose descriptions, achieving fine-grained control over poses. Next, we define the output format of the LLM and provide examples of pose descriptions for the LLM to reference. These examples are sourced from Posescript [17]. Finally, a user prompt of motion description is provided to instruct the LLM to generate the

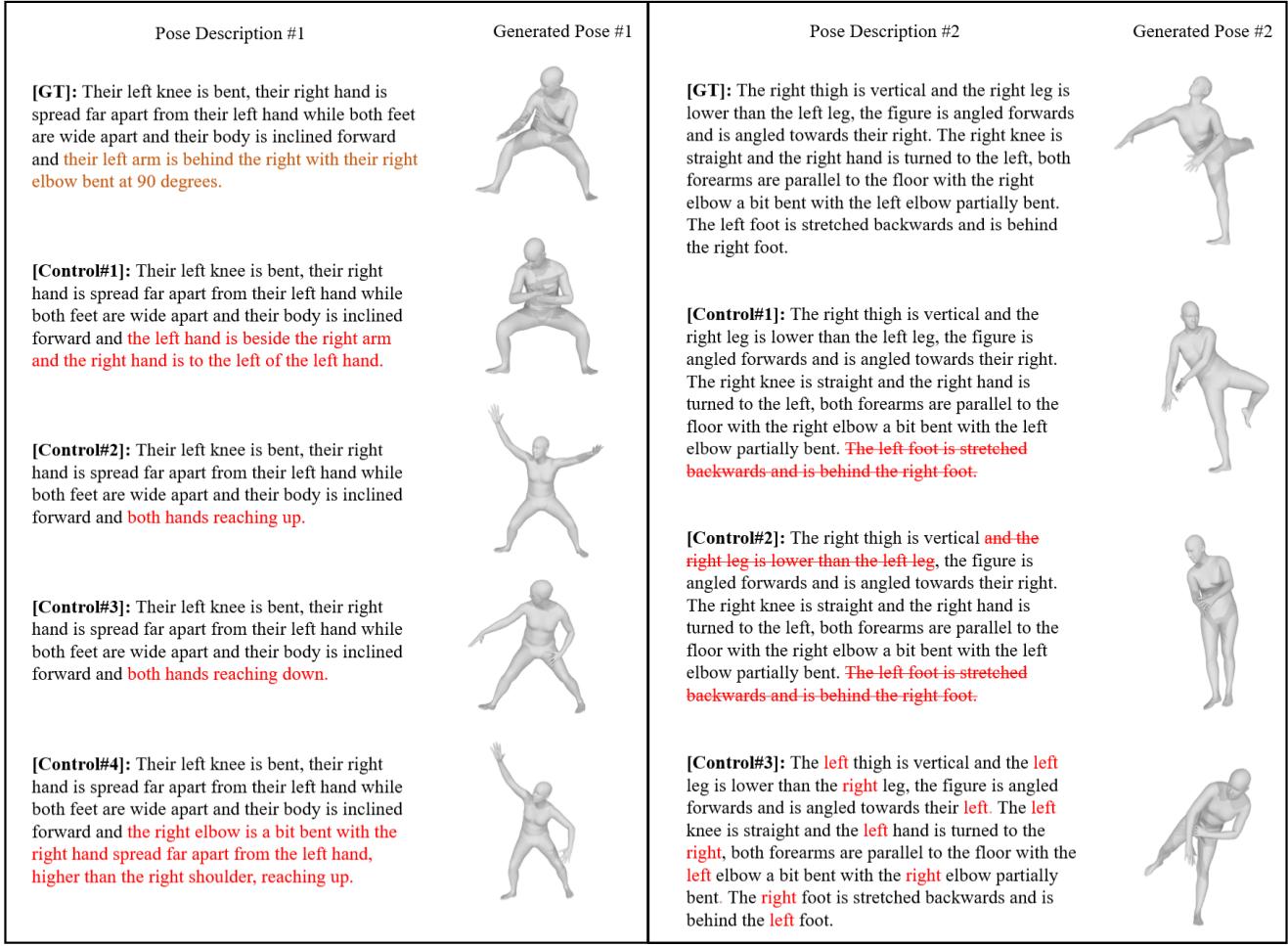


Figure 8. Examples of Precise Pose Control. In the first example #1, we control the position of hands, with the original and modified hand descriptions represented in brown and red respectively. In the second example #2, “**The**” represents deleting the description. To determine whether the description of a body part is the reason for the proper positioning of that body part, we removed the corresponding descriptions to see if the body joints would change as a result of this operation. We test the model’s understanding of the left and right sides of the body.

key pose descriptions. By harnessing the powerful LLM, the user prompts are no longer confined to specific formats but are open-world, allowing for expressions like “*Jump on one foot*” and “*Experiencing a profound sense of joy*”.

D. Detailed Implementations

Posture-diffuser. We employ an AdamW [50] optimizer for 1000 epochs with a batch size of 512 and a learning rate of 1e-4. The number of layers is $N = 3$ and the latent dim is set to 512. Following [33], we use the *linear* schedule, where $\beta_{start} = 0.00085$ and $\beta_{end} = 0.012$. Our model is trained using poses with automatically generated captions (PoseScript-A [17]) and adopts the last checkpoint as our final model. For the two encoders used in *Posture Planning* module, we use the pretrained text-to-pose retrieval model from [17], which is trained on PoseScript-A using the Batch-

Based Classification loss [88].

Go-diffuser. We employ an AdamW [50] optimizer with a batch size of 64 and a learning rate of 1e-4. The latent dim is 512, the number of transformer layers is 8, and the number of heads is 4. We utilize GELU [29] as the activation function and the dropout is 0.1. The mask probability of the model’s condition is 0.1. We use the cosine schedule and the number of diffusion steps is 100.

E. Evaluation details

We will detail our evaluation metrics on the Experiments part in this section. In open-world text-to-motion experiments, we employ sentence transformers [70, 71] to compute the similarity between the text descriptions in IDEA-400 [45], a high-quality motion language subset within Motion-X, and the text descriptions in HumanML3D. We filter out pairs

You are a helpful assistant that help me to describe the body pose in a motion. I will give you a motion description. Envision a motion scene and create eight distinct pose descriptions at a frame rate of 0.5 fps. Ensure that each description is self-contained. The difference between two adjacent descriptions must be small, considering the small interval. Your response should be detailed as possible. Before you write each description, you must follow these instructions. These are primary rules:

Rule 1. Characterize the degree of bending of body parts.

- 1.1 You should select the description word from the list: ['completely bent', 'almost completely bent', 'bent at right angle', 'partially bent', 'slightly bent', 'straight'].
- 1.2 You should select the body part form the list: [left knee, right knee, left elbow, right elbow].

Rule 2. Classify the relative distances between different body parts.

- 2.1 You should select the description word from the list: ['close', 'shoulder width apart', 'spread apart', 'wide apart'].
- 2.2 You must compare the distances between these body part pairs as much as possible:
 (left elbow, right elbow), (left hand, right hand), (left knee, right knee), (left foot, right foot),
 (left hand, left shoulder), (left hand, right shoulder), (right hand, left shoulder), (right hand, right shoulder),
 (left hand, right elbow), (right hand, left elbow), (left hand, left knee), (left hand, right knee), (right hand, left knee),
 (right hand, right knee), (left hand, left ankle), (left hand, right ankle), (right hand, left ankle), (right hand, right ankle),
 (left hand, left foot), (left hand, right foot), (right hand, left foot), (right hand, right foot)

Rule 3. Describe the relative positions of different body parts.

- 3.1 For the front-back direction, you should select the description word from the list: ['behind', 'in front of'].
 For the up-down direction, you should select the description word from the list: ['below', 'above'].
 For the left-right direction, you should select the description word from the list: ['at the right of', 'at the left of'].
- 3.2 You must compare the relative positioning between these body part pairs as much as possible:
 (left shoulder, right shoulder), (left elbow, right elbow), (left hand, right hand), (left knee, right knee), (right foot, right foot),
 (neck, pelvis), (left ankle, neck), (right ankle, neck), (left hip, left knee), (right hip, right knee), (left hand, left shoulder),
 (right hand, right shoulder), (left foot, left hip), (right foot, right hip), (left wrist, neck), (right wrist, neck),
 (left hand, left hip), (right hand, right hip), (left hand, torso), (right hand, torso), (left foot, torso), (right foot, torso)

Rule 4. Determine whether a body part is oriented 'vertical' or 'horizontal'.

- 4.1 You should select the description word from the list: ['vertical', 'horizontal'].
- 4.2 You need to determine as much as possible whether the body limb formed by the following pairs of body parts is 'vertical' or 'horizontal':
 (left hip, left knee), (right hip, right knee), (left knee, left ankle), (right knee, right ankle),
 (left shoulder, left elbow), (right shoulder, right elbow), (left elbow, left wrist), (right elbow, right wrist),
 (pelvis, left shoulder), (pelvis, right shoulder), (pelvis, neck)

Rule 5. Identify whether a body part is in contact with the ground.

- 5.1 You should select the description word from the list: ['on the ground'].
- 5.2 You should select the body part form the list: [left knee, right knee, left foot, right foot].

You should write all the pose description together. The response should follow the format:

```
{"F1":"pose description","\nF2":"pose description","\nF3":"pose description","\nF4":"pose description","\nF5":"pose description","\nF6":"pose description","\nF7":"pose description","\nF8":"pose description"}
```

Some sample pose descriptions are as follows:

"The right knee is unbent with the right leg next to the left while both hands are apart wider than shoulder width. The right upper arm, the left leg, the torso and the right thigh are straightened up while the right elbow is bent a bit. "
 "The figure is in a yoga pose. Their left elbow is nearly bent and their hands are underneath both hips while their left knee is bent at 90 degrees while the figure is leaning backwards and their feet are front with their right elbow partially bent. Their left hand is further than shoulder width apart from the other. "
 "Their left calf is straightened up with their left shoulder above their right shoulder while their right knee is straight while their left elbow and their left knee are a bit bent. Their knees are approximately shoulder width apart and the figure is angled towards the right side. Their left hand is reaching up and to the right of their right hand. It is in the right direction while their right hand is lying beneath their right hip, close to their right knee. "

Prompt: {User Prompt}

Figure 9. The complete prompt for our *Motion Planner* in Sec. 3.2.

with similarity greater than a specified threshold α , e.g., 0.45, yielding a motion language dataset comprising 368 text-motion pairs as our first test dataset.

```

1 # calculate the similarity of motion descriptions
2     and filter out pairs with similarity greater
3     than a specified threshold
4
5 import torch
6 import torch.nn as nn
7 import torch.nn.functional as F
8 from transformers import AutoTokenizer, AutoModel
9     , util
10
11 def mean_pooling(model_output, attention_mask):
12     token_embeddings = model_output[0]  # First
13     element of model_output contains all token
14     embeddings
15     input_mask_expanded = attention_mask.
16     unsqueeze(-1).expand(token_embeddings.size())
17     .float()
18     return torch.sum(token_embeddings *
19     input_mask_expanded, 1) / torch.clamp(
20     input_mask_expanded.sum(1), min=1e-9)
21
22 # calculate the features of descriptions
23 class TextToSen(nn.Module):
24     def __init__(self, device):
25         super().__init__()
26         self.device = device
27         self.tokenizer = AutoTokenizer.
28         from_pretrained('sentence-transformers/all-
29         mpnet-base-v2')
30         self.model = AutoModel.from_pretrained('
31         sentence-transformers/all-mpnet-base-v2').
32         eval().to(self.device)
33
34     def forward(self, sentences):
35         encoded_input = self.tokenizer(sentences,
36             padding=True, truncation=True,
37             return_tensors='pt').to(self.device)
38
39         # Compute token embeddings
40         with torch.no_grad():
41             model_output = self.model(**
42             encoded_input)
43
44         # Perform pooling
45         sentence_embeddings = mean_pooling(
46             model_output, encoded_input['attention_mask']
47         )
48
49         # Normalize embeddings
50         sentence_embeddings = F.normalize(
51             sentence_embeddings, p=2, dim=1)
52
53         return sentence_embeddings
54
55 # calculate the similarity
56 similarity_score = util.pytorch_cos_sim(
57     embeddings1, embeddings2)
58
59 if similarity_score > threshold:
60     False

```

Listing 3. similarity filter

R-Precision is determined by ranking the Euclidean dis-

tances between the motion and text embeddings, given one motion sequence and K text descriptions (1 ground-truth and $K - 1$ randomly selected mismatched descriptions).

Frechet Inception Distance (FID) measures the distributional difference between the generated and real motion by applying FID [31] to the extracted motion features derived from that text.

Multimodal Distance (MM-Dist) is calculated as the average Euclidean distance between each text feature and the corresponding generated motion feature.

In go-diffuser experiments, we utilize Average Position Error (APE) and Average Variance Error (AVE) to evaluate our methods [63]. We report the listed four metrics in Tab. 2.

Root joint errors. Take the 3 coordinates of the root joint.

Global trajectory errors. Tak3 only the X and Y coordinates of the root joint. It is the red trajectory on the ground in the visualizations of Fig. 7.

Mean local errors. Average the joint errors in the body's local coordinate system,

Mean global errors. Average the joint errors in the global coordinate system.

As in JL2P [3], Ghosh et al. [21] and TEMOS [63], the APE for a specific joint j is determined by computing the mean of the L2 distances between the generated and ground truth joint positions across the frames (F) and samples (N):

$$APE[j] = \frac{1}{NF} \sum_{n \in N} \sum_{f \in F} \|H_f[j] - \hat{H}_f[j]\|_2 \quad (7)$$

As introduced in Ghosh et al [21] and TEMOS [63], the Average Variance Error (AVE), quantifies the distinction in variations. This metric is defined as the mean of the L2 distances between the generated and ground truth variances for the joint j .

$$AVE[j] = \frac{1}{N} \sum_{n \in N} \|\delta[j] - \hat{\delta}[j]\|_2 \quad (8)$$

where,

$$\delta[j] = \frac{1}{F-1} \sum_{f \in F} (H_f[j] - \tilde{H}_f[j])^2 \in R^3 \quad (9)$$

denotes the variance of the joint j .