






Local Action-Guided Motion Diffusion Model for Text-to-Motion Generation

Peng Jin^{1,2,3}, Hao Li^{1,2,3}, Zesen Cheng^{1,3}, Kehan Li^{1,3}, Runyi Yu^{1,3}, Chang Liu^{4*}, Xiangyang Ji⁴, Li Yuan^{1,2,3*}, and Jie Chen^{1,2,3}

¹ School of Electronic and Computer Engineering, Peking University, Shenzhen, China

² Peng Cheng Laboratory, Shenzhen, China

³ AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, Shenzhen, China

⁴ Department of Automation and BNRist, Tsinghua University, Beijing, China
jp21@stu.pku.edu.cn, liuchang2022@tsinghua.edu.cn, yuanli-ece@pku.edu.cn

Abstract. Text-to-motion generation requires not only grounding local actions in language but also seamlessly blending these individual actions to synthesize diverse and realistic global motions. However, existing motion generation methods primarily focus on the direct synthesis of global motions while neglecting the importance of generating and controlling local actions. In this paper, we propose the local action-guided motion diffusion model, which facilitates global motion generation by utilizing local actions as fine-grained control signals. Specifically, we provide an automated method for reference local action sampling and leverage graph attention networks to assess the guiding weight of each local action in the overall motion synthesis. During the diffusion process for synthesizing global motion, we calculate the local-action gradient to provide conditional guidance. This local-to-global paradigm reduces the complexity associated with direct global motion generation and promotes motion diversity via sampling diverse actions as conditions. Extensive experiments on two human motion datasets, *i.e.*, HumanML3D and KIT, demonstrate the effectiveness of our method. Furthermore, our method provides flexibility in seamlessly combining various local actions and continuous guiding weight adjustment, accommodating diverse user preferences, which may hold potential significance for the community. The project page is available at <https://jpthu17.github.io/GuidedMotion-project/>.

Keywords: Text-to-motion generation · Diffusion models

1 Introduction

Human motion generation [2, 6, 44] is a critical task in computer animation [4, 53], with the primary objective of creating realistic and dynamic motions for virtual human characters. This technology finds widespread applications in multiple

* Corresponding author: Li Yuan, Chang Liu.

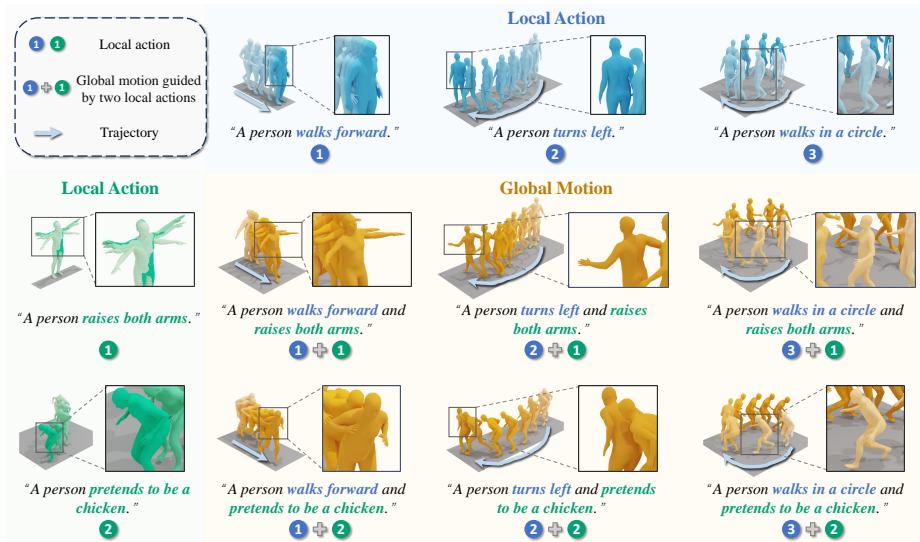


Fig. 1: Generating motion with diverse local actions. Different local actions correspond to distinct user preferences. Our method empowers users to combine preferred local actions freely, generating motions that align with their mental imagery. Furthermore, the combination of varied local actions enhances the motion diversity.

industries, such as entertainment, gaming, film production, virtual reality, and robotics. Recent developments in this field have introduced text-driven human motion generation techniques, enabling the synthesis of diverse human motion sequences based on natural language descriptions. However, text-driven human motion generation poses a series of challenges, requiring the alignment of local actions with language and the seamless integration of these individual actions to synthesize diverse and realistic global motions.

Existing text-to-motion generation methods [1, 6, 42, 69] primarily focus on directly synthesizing global motions based on language instructions. Although these methods have shown promising advancements and achieved high accuracy, they come with limitations regarding the type of control they support over the motion results. For example, precisely expressing intricate trajectories, postures, and long motion sequences involving multiple actions is challenging using text prompts alone. Typically, generating a motion that faithfully corresponds to our mental imagery requires numerous iterations of editing a prompt, reviewing the resulting motion, and then adjusting the prompt accordingly.

In this work, we propose to employ reference local actions as control signals in the global motion generation process. As illustrated in Fig. 1, an overall motion comprises a sequence of local actions, such as “*walks forward*” and “*raises both arms*”. These reference local actions can serve as control signals during the global motion generation process, facilitating the generation of global motions with similar characteristics, including movement trajectories and human body

postures, to those of the local actions. More importantly, local action serves as a more intuitive control signal than text. Users can seamlessly combine their preferred local actions, exerting precise control over the resulting global motion to align with the characteristics of those chosen local actions.

To this end, we introduce GuidedMotion, a local action-guided motion diffusion model designed for controllable text-to-motion generation. Moreover, we provide an automatic local action sampling method, which deconstructs the original motion description into multiple local action descriptions and uses a text-to-motion model to generate the reference local actions. In practical applications, the same reference local action can be sampled multiple times to suit diverse user preferences, allowing users to conveniently select their preferred action from these choices. Subsequently, we leverage graph attention networks to estimate the guiding weight of each local motion in the overall motion synthesis. To enhance generation stability, we divide the motion diffusion process for synthesizing global motion into three stages: (i) In the initial diffusion stage, we de-noise the Gaussian noise based on the original motion description to provide a good initial value for the subsequent stage. (ii) In the second diffusion stage, we apply local-action gradients based on the energy function [72] to offer conditional guidance for aligning the generated motion with the characteristics of the reference local actions. (iii) In the final diffusion stage, we fine-tune the generated results further to conform to the original motion description, rather than solely adhering to a reference local action.

The proposed GuidedMotion has three distinct advantages: **First**, compared to the direct generation of global motion, our local-to-global paradigm, leveraging local actions as a prior, simplifies the complexity associated with global motion generation, especially when generating complex motions with multiple local actions. **Second**, through the automatic sampling of diverse local actions, our method has the capability to generate a variety of motions to suit different user preferences. **Third**, our method provides flexibility in adjusting the guiding weight of each local action, enabling fine-grained and continuous control over global motion, *e.g.*, the control of movement trajectories and human body postures. Extensive experiments on two datasets for text-to-motion generation, including HumanML3D [17] and KIT [43], demonstrate the advantages of GuidedMotion. The main contributions are summarized as follows:

- We propose local action-guided motion synthesis for fine-grained controllable text-to-motion generation. It allows users to seamlessly combine their preferred local actions, enabling them to exert control over the resulting global motion to align with the characteristics of their chosen local actions.
- The proposed local-to-global paradigm, utilizing local actions as a prior, reduces the complexity associated with direct global motion generation. Experimental results demonstrate that our method has an advantage in generating complex motions comprising multiple local actions.
- More encouragingly, our method allows for continuous guiding weight adjustment, facilitating the refinement of the results to match the preferences of users, which may hold potential significance for the community.

2 Related Work

Diffusion Models. Diffusion models [14, 23, 25, 26, 49, 50, 70], rooted in thermodynamics, utilize a stochastic diffusion process to complete the generation task. In recent years, diffusion models have exhibited potential across diverse tasks, including image generation [14, 23, 24, 50, 59], natural language generation [3, 16], and visual tasks [11]. Some other works [30] have applied diffusion models to cross-modal retrieval [29]. Inspired by the success of diffusion generative models, some works [33, 53, 61, 65] have explored the application of diffusion models in human motion tasks [5, 8, 37]. Although existing text-to-motion generation methods have shown promising advancements and achieved high accuracy, they come with limitations regarding the type of control they support over the motion results. In this paper, we propose to employ reference local actions as fine-grained control signals in the global motion generation process.

Text-driven Human Motion Generation. The goal of text-driven human motion generation [13, 20, 73] is to generate human motion based on text descriptions. Due to the user-friendly and convenient nature of natural language [28], text-driven human motion generation has garnered significant attention. Recently, motion generation methods can be classified into three categories: joint-latent models [2, 42], such as TEMOS [42], which typically learn a motion variational autoencoder and a text variational autoencoder; the second category [10, 31, 47, 68], such as MDM [53], introduces a conditional diffusion model for human motion generation; and the last category [27, 71], such as T2M-GPT [67], utilizes generative pre-trained transformer for motion generation. In this work, leveraging the iterative refinement of diffusion models, we employ the diffusion model method to enhance control over the motion generation process.

Controllable Human Motion Generation. Controllable human motion generation [33, 60, 63, 66] aims to generate motions following designated control signals, offering enhanced interactivity and interpretability to humans. Existing methods predominantly focus on controlling trajectory and key points within the diffusion process through techniques such as imputation and inpainting. However, these low-level control signals lack the capability for high-level control over motions, such as adjusting the amplitude of arms. What is worse, existing methods lack support for continuous motion adjustment, limiting the ability to refine motions until they align with the expectations of users. In contrast, our method employs local actions with high-level semantics as control signals, enabling not only trajectory control but also manipulation of human body postures.

3 Methodology

In this work, we tackle the challenges associated with controllable text-driven human motion generation. Concretely, given a motion description and other fine-grained control signals, such as reference local actions, our goal is to synthesize a human motion sequence $\mathbf{x}^{1:L} = \{x^i\}_{i=1}^L$ of length L .

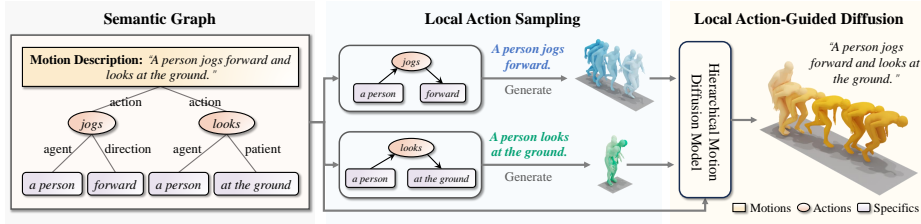


Fig. 2: The overall framework of GuidedMotion for controllable text-to-motion generation. We propose to employ reference local actions as control signals in the global motion generation process. To automatically obtain these local actions, we deconstruct the original motion description into multiple local action descriptions and utilize a text-to-motion model to generate these local actions.

3.1 Automatic Local Action Sampling

Local actions can be accessible from the repository, enabling users to choose their preferred local action as the control signal for generating the global motion. Moreover, we provide an automatic local action sampling method, which deconstructs the original motion description into multiple local action descriptions and utilizes a text-to-motion model to generate these local actions.

Semantic Graph Parsing. As shown in Fig. 2, motion descriptions inherently exhibit hierarchical structures, represented as hierarchical graphs comprising three types of abstract nodes: motions, actions, and specifics. Concretely, the complete sentence describes the global motion, encompassing multiple actions, for example, “jogs” and “looks” in Fig. 2. Each action includes various specifics, which serve as its attributes, such as the agent and patient of the action.

To obtain actions, action attributes, and the semantic role of each attribute in relation to the corresponding action, we employ a semantic parser for motion descriptions based on a semantic role parsing toolkit [9, 31, 48]. In practice, we extract three types of nodes (motions, actions, and specifics) and twelve types of edges to represent various associations among the nodes. For further details about semantic graph parsing, please refer to our supplementary material.

Local Action Sampling. Given the semantic graph, we create a local action description for each local action by considering each action node and its associated specific nodes. Subsequently, We employ a text-to-motion generation model, *i.e.*, MLD [10], to generate local actions based on these local action descriptions. To further enrich the variety of local actions, the local action descriptions can be expanded using large language models, such as GPT [7] and LLaMA [54, 55].

3.2 Local Action Diffusion Guidance

Following previous works [10, 31], we encode the motion sequence \mathbf{x} into the latent space \mathbf{z} utilizing the variational autoencoder [35]. Subsequently, we employ diffusion models to learn the noise component ϵ at every noise level t .

Energy Diffusion Guidance. In accordance with score theory [51,52,64], the core objective of conditional diffusion models [23] is to estimate the score function $\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t|\mathbf{c})$, where \mathbf{c} is the condition. The reverse process of conditional diffusion models is formulated as:

$$\mathbf{z}_{t-1} = (1 + \frac{1}{2}\beta_t)\mathbf{z}_t + \beta_t \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t|\mathbf{c}) + \sqrt{\beta_t}\boldsymbol{\epsilon}, \quad (1)$$

where $\boldsymbol{\epsilon}$ is a noise sampled from the standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. $\beta_t \in \mathbb{R}$ is a pre-defined step size which gradually increases. Based on Bayesian formula $p(\mathbf{z}_t|\mathbf{c}) = \frac{p(\mathbf{c}|\mathbf{z}_t)p(\mathbf{z}_t)}{p(\mathbf{c})}$, we rewrite the conditional score function as:

$$\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t|\mathbf{c}) = \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t) + \nabla_{\mathbf{z}_t} \log p(\mathbf{c}|\mathbf{z}_t), \quad (2)$$

where the correction gradient $\nabla_{\mathbf{z}_t} \log p(\mathbf{c}|\mathbf{z}_t)$ holds paramount significance in the conditional diffusion models. However, the correction gradient $\nabla_{\mathbf{z}_t} \log p(\mathbf{c}|\mathbf{z}_t)$ is hard to measure directly. Following previous works [64,72], we employ the energy function [36] to formulate the correction term as:

$$p(\mathbf{c}|\mathbf{z}_t) = \frac{\exp(-\mathcal{E}(\mathbf{c}, \mathbf{z}_t))}{\int_{\mathbf{c} \in \mathcal{C}} \exp(-\mathcal{E}(\mathbf{c}, \mathbf{z}_t))}, \quad (3)$$

where \mathcal{C} denotes the domain of the condition \mathbf{c} . With Eq. (3), the correction gradient can be estimated by the gradient of the energy function $\mathcal{E}(\mathbf{c}, \mathbf{z}_t)$, *i.e.*, $\nabla_{\mathbf{z}_t} \log p(\mathbf{c}|\mathbf{z}_t) \propto -\nabla_{\mathbf{z}_t} \mathcal{E}(\mathbf{c}, \mathbf{z}_t)$. Therefore, the reverse process of conditional diffusion models can be rewritten as:

$$\mathbf{z}_{t-1} = \tilde{\mathbf{z}}_{t-1} - \lambda_t \nabla_{\mathbf{z}_t} \mathcal{E}(\mathbf{c}, \mathbf{z}_t), \quad (4)$$

where $\tilde{\mathbf{z}}_{t-1} = (1 + \frac{1}{2}\beta_t)\mathbf{z}_t + \beta_t \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t) + \sqrt{\beta_t}\boldsymbol{\epsilon}$ is the original reverse process of unconditional diffusion models. In essence, λ_t is the guiding weight, which represents the learning rate of the correction. When there are multiple conditions in the reverse process of diffusion models, Eq. (4) is reformulated as:

$$\mathbf{z}_{t-1} = \tilde{\mathbf{z}}_{t-1} - \sum_{k=1}^K \lambda_t^k \nabla_{\mathbf{z}_t} \mathcal{E}(\mathbf{c}^k, \mathbf{z}_t), \quad (5)$$

where K represents the number of guidance terms.

In this work, the condition \mathbf{c} is the motion latent embeddings of local actions. The number K of guidance local actions is determined based on semantic parsing of the input motion description. To achieve the goal of the diffusion guidance, the energy function $\mathcal{E}(\mathbf{c}, \mathbf{z}_t)$ should meet all the following criteria: (i) if \mathbf{z}_t is a better match with \mathbf{c} , then $\mathcal{E}(\mathbf{c}, \mathbf{z}_t)$ is smaller; (ii) If \mathbf{z}_t perfectly conforms to the constraint set by \mathbf{c} , then $\mathcal{E}(\mathbf{c}, \mathbf{z}_t)$ should be zero.

Note that anything satisfying the above conditions can be employed as the energy function $\mathcal{E}(\mathbf{c}, \mathbf{z}_t)$, such as the Gram matrix [32] distance and the embedding distance. For simplicity in implementation, we utilize the ℓ_2 distance of latent embeddings as the energy function in practice.

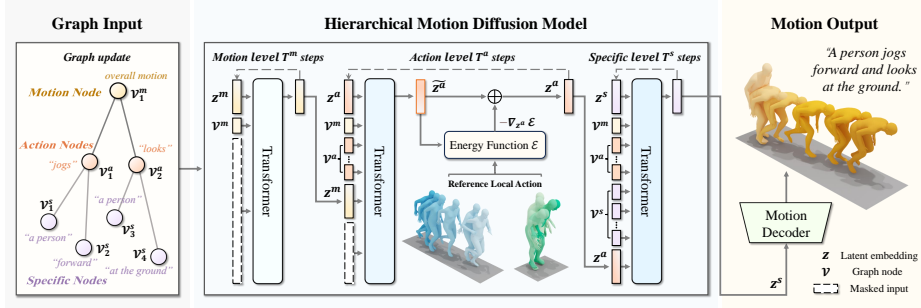


Fig. 3: The model architecture of the hierarchical motion diffusion model. Utilizing the semantic graph as input, the hierarchical diffusion model dissects the text-to-motion diffusion process into three semantic levels, which correspond to capturing the overall motion, local actions, and action specifics. To enhance generation stability, we exclusively implement local action guidance at the action level.

Guiding Weight Estimation. As illustrated in Fig. 3, the interactions among different levels in the semantic graph elucidate the characteristics of local actions and how these local actions contribute to the overall motion. Drawing inspiration from this insight, we employ graph attention networks [57] (GAT) to model the guiding weights in the local action-guided motion diffusion model.

We leverage the text encoder of CLIP [45] to initialize the representation of graph nodes. To represent the global motion node v^m , we utilize the [CLS] token to encapsulate the overall motion within the description. For the action node v^a , we adopt the token corresponding to the verb as the representation. In the case of the specific node v^s , we employ mean-pooling over tokens of each word in the attribute phrase to represent every action detail of the motion.

Given the initialized nodes $v = \{v^m, v^a, v^s\}$, we utilize a shared projection matrix $W \in \mathbb{R}^{D \times D}$, where D represents the dimension of node representation. This matrix transforms v into higher-level embeddings $h = \{h^m, h^a, h^s\}$. For each pair $\{h_i, h_j\}$ of connected nodes, we concatenate the node $h_i \in \mathbb{R}^D$ with its neighbor node $h_j \in \mathbb{R}^D$, creating the input data $\tilde{h}_{ij} = [h_i, h_j] \in \mathbb{R}^{2D}$ of the graph attention module, which is formulated as:

$$\tilde{h}_{ij} = [h_i, h_j] = [Wv_i, Wv_j]. \quad (6)$$

The semantic graph comprises multiple types of edges. To avoid over-fitting to infrequent edge types, we employ a shared transformation matrix $M \in \mathbb{R}^{2D \times 1}$ that applies to all edge types, and a relationship embedding matrix $M_r \in \mathbb{R}^{2D \times N}$ that is specific for different edges to represent multi-relational weights, where N denotes the number of edge types. The attention coefficient e_{ij} is formulated as:

$$e_{ij} = \sigma(M^\top \tilde{h}_{ij}) + \sigma(R_{ij} M_r^\top \tilde{h}_{ij}), \quad (7)$$

$$\tilde{e}_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathbb{N}_i} \exp(e_{ik})},$$

where σ refers to a nonlinear function, *i.e.* LeakyReLU [39] with a negative input slope set to 0.2. $\mathbf{R}_{ij} \in \mathbb{R}^{1 \times N}$ denotes a one-hot vector denoting the type of edge between node i and j . \mathbb{N}_i is the set of neighborhood nodes of node i . Finally, we use the attention coefficient \tilde{e} as the guiding weight λ , which is formulated as:

$$\lambda_t^k = \rho_t \tilde{e}^k, \quad \tilde{e}^k \in \{(\mathbf{u}, \mathbf{v}^m) | \mathbf{u} \in \{\mathbf{v}_k^a\}_{k=1}^K\}, \quad (8)$$

where ρ_t is a predefined parameter used to amplify or reduce the guiding strength. \tilde{e}^k is the attention coefficient corresponding to the k_{th} reference local action.

3.3 Hierarchical Motion Diffusion Model

To enhance generation stability, we decompose the diffusion process into three semantic levels and build three transformer-based denoising networks, which correspond to motions, actions, and specifics. The motion level provides a good initial value for the subsequent semantic levels. Subsequently, we exclusively implement local action guidance at the action level. Finally, at the specific level, we further refine the generated results to match the original motion description, rather than solely adhering to a reference local action.

Motion Variational Autoencoder. Following previous works [42, 67], we encode the motion into the latent space using a motion variational autoencoder [35] (VAE). Specifically, we construct the motion encoder and decoder based on the transformer [41, 56]. For the motion encoder, we utilize Q learnable query tokens along with the motion sequence $\mathbf{x}^{1:L} = \{x^i\}_{i=1}^L$ as inputs to generate motion latent embeddings $\mathbf{z} \in \mathbb{R}^{Q \times D'}$, where D' is the dimension of latent representation. For the motion decoder, we input the latent embeddings $\mathbf{z} \in \mathbb{R}^{Q \times D'}$ and the motion query tokens to generate a human motion sequence.

Corresponding to the three-level structure of the hierarchical motion diffusion model, we encode human motion sequences independently into three latent representation spaces: $\mathbf{z}^m \in \mathbb{R}^{Q^m \times D'}$, $\mathbf{z}^a \in \mathbb{R}^{Q^a \times D'}$, and $\mathbf{z}^s \in \mathbb{R}^{Q^s \times D'}$. To generate motion progressively from coarse to fine, we gradually increase the number of learnable query tokens, *i.e.*, $Q^m \leq Q^a \leq Q^s$.

Hierarchical Motion Diffusion. We utilize the semantic graph as the input for the hierarchical diffusion model. The node embeddings \mathcal{V} are formulated as:

$$\mathcal{V}_i = \sigma' \left(\sum_{j \in \mathbb{N}_i} \tilde{e}_{ij} \mathbf{h}_j \right) + \mathbf{v}_i, \quad (9)$$

where σ' is a nonlinear function. Following graph attention networks [57] (GAT), we adopt ELU [12] as the nonlinear function σ' and apply skip connection [21, 46] to mitigate over-smoothing [62] in graph networks.

To improve generation stability, we partition the diffusion process into three semantic levels, aligning with motions, actions, and specifics. In the motion level model ϕ_m , We employ the global motion node \mathcal{V}^m as input to predict the noise component ϵ^m . The training objective for the motion level is formulated as:

$$\mathcal{L}_M = \mathbb{E}_{\mathbf{z}, \epsilon, t} \left[\|\epsilon^m - \phi_m(\mathbf{z}^m, t^m, \mathcal{V}^m)\|_2^2 \right]. \quad (10)$$

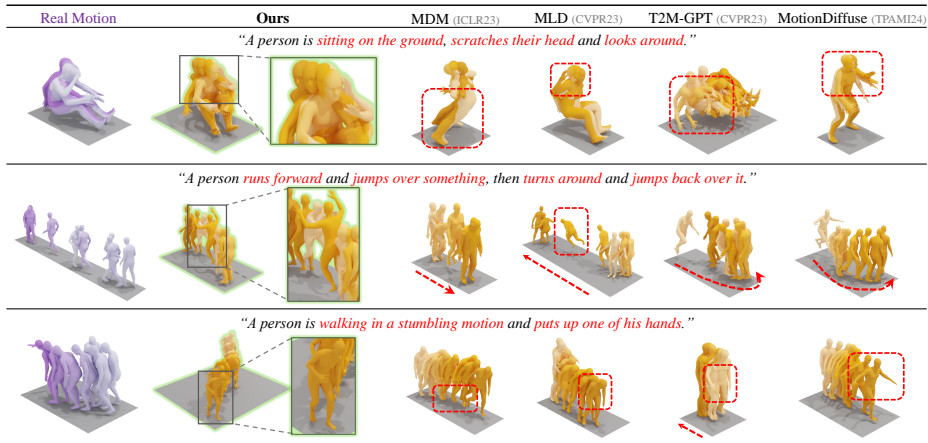


Fig. 4: Qualitative comparisons. The darker colors indicate the later in time. The motions generated by our method closely align with the descriptions, outperforming others that exhibit degraded motions or improper semantics.

In the action level model ϕ_a , we concatenate the action node \mathcal{V}^a , the motion node \mathcal{V}^m , and the generated result \mathbf{z}^m from the motion level as the input. The training objective for the action level is formulated as:

$$\mathcal{L}_A = \mathbb{E}_{\mathbf{z}, \epsilon, t} \left[\|\epsilon^a - \phi_a(\mathbf{z}^a, t^a, [\mathcal{V}^m, \mathcal{V}^a, \mathbf{z}^m])\|_2^2 \right]. \quad (11)$$

In the specific level model ϕ_s , we utilize the results generated by the action level and nodes across all semantic levels to predict the noise component. The training objective for the specific level is formulated as:

$$\mathcal{L}_S = \mathbb{E}_{\mathbf{z}, \epsilon, t} \left[\|\epsilon^s - \phi_s(\mathbf{z}^s, t^s, [\mathcal{V}^m, \mathcal{V}^a, \mathcal{V}^s, \mathbf{z}^a])\|_2^2 \right]. \quad (12)$$

Finally, the total training objective is denoted as $\mathcal{L} = \mathcal{L}_M + \mathcal{L}_A + \mathcal{L}_S$. The output at the specific level is considered the final result, and the motion decoder is employed to decode the latent representation into the motion sequence.

4 Experiments

Experimental Settings. *Datasets.* We compare the proposed method with other methods on two commonly used public benchmarks: HumanML3D [17] and KIT [43]. **HumanML3D** [17] originates from and textually reannotates the HumanAct12 [19] and AMASS [40] datasets. HumanML3D comprises 14,616 human motions and 44,970 text descriptions. **KIT** [43] contains 3,911 human motion sequences and 6,278 textual annotations.

Metrics. Following previous works, we use the following five metrics to measure the performance of the model. (1) **R-Precision.** In the feature space of the

Table 1: Comparisons to current state-of-the-art methods on the HumanML3D test set. “ \uparrow ” denotes that higher is better. “ \downarrow ” denotes that lower is better. “ \rightarrow ” denotes that results are better if the metric is closer to the real motion. We repeat all the evaluations 20 times and report the average with a 95% confidence interval. **Bold** and underlined indicate the best and second-best results, respectively.

Methods	R-Precision \uparrow			FID \downarrow	MM-Dist \downarrow	Diversity \rightarrow	MModality \uparrow
	Top-1	Top-2	Top-3				
Real Motion	0.511 \pm .003	0.703 \pm .003	0.797 \pm .002	0.002 \pm .000	2.974 \pm .008	9.503 \pm .065	-
Hier [15] _{ICCV21}	0.301 \pm .002	0.425 \pm .002	0.552 \pm .004	6.532 \pm .024	5.012 \pm .018	8.332 \pm .042	-
TEMOS [42] _{ECCV22}	0.424 \pm .002	0.612 \pm .002	0.722 \pm .002	3.734 \pm .028	3.703 \pm .008	8.973 \pm .071	0.368 \pm .018
TM2T [18] _{ECCV22}	0.424 \pm .003	0.618 \pm .003	0.729 \pm .002	1.501 \pm .017	3.467 \pm .011	8.589 \pm .076	2.424 \pm .093
T2M [17] _{CVPR22}	0.457 \pm .002	0.639 \pm .003	0.740 \pm .003	1.067 \pm .002	3.340 \pm .008	9.188 \pm .002	2.090 \pm .083
MotionDiffuse [68] _{TPAMI24}	0.491 \pm .001	0.681 \pm .001	0.782 \pm .001	0.630 \pm .001	3.113 \pm .001	9.410 \pm .049	1.553 \pm .042
MDM [53] _{ICLR23}	0.320 \pm .005	0.498 \pm .004	0.611 \pm .007	0.544 \pm .044	5.566 \pm .027	<u>9.559\pm.086</u>	2.799\pm.072
MLD [10] _{CVPR23}	0.481 \pm .003	0.673 \pm .003	0.772 \pm .002	0.473 \pm .013	3.196 \pm .010	9.724 \pm .082	2.413 \pm .079
Fg-T2M [58] _{ICCV23}	0.492 \pm .002	0.683 \pm .003	0.783 \pm .002	0.243 \pm .019	3.109 \pm .007	9.278 \pm .072	1.614 \pm .049
MotionGPT [27] _{NeurIPS23}	0.492 \pm .003	0.681 \pm .003	0.778 \pm .002	0.232 \pm .008	3.096 \pm .008	9.528\pm.071	2.008 \pm .084
T2M-GPT [67] _{CVPR23}	0.491 \pm .003	0.680 \pm .003	0.775 \pm .002	0.116 \pm .004	3.118 \pm .011	9.761 \pm .081	1.856 \pm .011
GraphMotion [31] _{NeurIPS23}	<u>0.504\pm.003</u>	0.699\pm.002	0.785 \pm .002	0.116 \pm .007	3.070 \pm .008	9.692 \pm .067	<u>2.766\pm.096</u>
ReMoDiffuse [69] _{ICCV23}	0.510\pm.005	<u>0.698\pm.006</u>	0.795\pm.004	<u>0.103\pm.004</u>	2.974\pm.016	9.018 \pm .075	1.795 \pm .043
GuidedMotion (Ours)	0.503 \pm .002	0.691 \pm .002	<u>0.788\pm.002</u>	0.057\pm.006	<u>3.040\pm.012</u>	9.864 \pm .077	2.473 \pm .096

pre-trained network introduced by T2M [17], motion-retrieval precision is determined by the matching accuracy of the top 1/2/3 text descriptions with a motion sequence and 32 text descriptions. (2) **Fréchet Inception Distance (FID)**. We measure the distribution distance between generated and real motion using FID [22] on the extracted motion features [17]. (3) **Multimodal Distance (MM-Dist)**. We calculate the average Euclidean distances between each text feature and the corresponding generated motion feature. (4) **Diversity**. All generated motions are randomly sampled into two equal-sized subsets. Motion features [17] are then extracted, and the average Euclidean distances between the two subsets represent diversity. (5) **Multimodality (MModality)**. For each text description, we generate 20 motion sequences, creating 10 pairs of motions. The average Euclidean distance between motion features is calculated for each pair. The result is the average across all text descriptions.

Implementation details. For text representation, we employ a frozen text encoder from the CLIP-ViT-L-14 [45] model. The dimension of node representation D is set to 768. The dimension of latent embedding D' is set to 256. We set the token sizes Q^m to 2, Q^a to 4, and Q^s to 8. The predefined parameter ρ in Eq. (8) is set to 0.01. All our models are trained using the AdamW [34, 38] optimizer with a fixed learning rate of 1e-4. Training is performed on 4 Tesla V100 GPUs, with 128 samples on each GPU, resulting in a total batch size of 512. We keep running a similar number of iterations on different datasets. For the HumanML3D dataset, the model is trained for 6,000 epochs during the motion variational autoencoder stage and 3,000 epochs during the diffusion stage. In the case of the KIT dataset, the model is trained for 30,000 epochs during the motion variational autoencoder stage and 15,000 epochs during the diffusion stage.

Table 2: Comparisons to other methods on the KIT test set. We repeat all the evaluations 20 times and report the average with a 95% confidence interval. **Bold** and underlined indicate the best and second-best results, respectively.

Methods	R-Precision \uparrow			FID \downarrow	MM-Dist \downarrow	Diversity \rightarrow	MModality \uparrow
	Top-1	Top-2	Top-3				
Real Motion	0.424 \pm .005	0.649 \pm .006	0.779 \pm .006	0.031 \pm .004	2.788 \pm .012	11.08 \pm .097	-
Hier [15] <small>ICCV21</small>	0.255 \pm .006	0.432 \pm .007	0.531 \pm .007	5.203 \pm .107	4.986 \pm .027	9.563 \pm .072	2.090 \pm .083
TEMOS [42] <small>ECCV22</small>	0.353 \pm .006	0.561 \pm .007	0.687 \pm .005	3.717 \pm .051	3.417 \pm .019	10.84 \pm .100	0.532 \pm .034
TM2T [18] <small>ECCV22</small>	0.280 \pm .005	0.463 \pm .006	0.587 \pm .005	3.599 \pm .153	4.591 \pm .026	9.473 \pm .117	3.292 \pm .081
T2M [17] <small>CVPR22</small>	0.370 \pm .005	0.569 \pm .007	0.693 \pm .007	2.770 \pm .109	3.401 \pm .008	10.91 \pm .119	1.482 \pm .065
MotionDiffuse [68] <small>TPAMI24</small>	0.417 \pm .004	0.621 \pm .004	0.739 \pm .004	1.954 \pm .062	<u>2.958\pm.005</u>	11.10\pm.143	0.730 \pm .013
Fg-T2M [58] <small>ICCV23</small>	0.418 \pm .005	0.626 \pm .004	0.745 \pm .004	0.571 \pm .047	3.114 \pm .015	10.93 \pm .083	1.019 \pm .029
T2M-GPT [67] <small>CVPR23</small>	0.416 \pm .006	0.627 \pm .006	0.745 \pm .006	0.514 \pm .029	3.007 \pm .023	10.92 \pm .108	1.570 \pm .039
MDM [53] <small>ICLR23</small>	0.164 \pm .004	0.291 \pm .004	0.396 \pm .004	0.497 \pm .021	9.191 \pm .022	10.85 \pm .109	1.907 \pm .214
MLD [10] <small>CVPR23</small>	0.390 \pm .008	0.609 \pm .008	0.734 \pm .007	0.404 \pm .027	3.204 \pm .027	10.80 \pm .117	2.192 \pm .071
GraphMotion [31] <small>NeurIPS23</small>	0.429 \pm .007	0.648 \pm .006	0.769\pm.006	0.313 \pm .013	3.076 \pm .022	<u>11.12\pm.135</u>	3.627 \pm .113
ReMoDiffuse [69] <small>ICCV23</small>	0.427 \pm .014	0.641 \pm .004	0.765 \pm .055	0.155\pm.006	2.814\pm.012	10.80 \pm .105	1.239 \pm .028
GuidedMotion (Ours)	0.430\pm.006	0.652\pm.005	<u>0.768\pm.005</u>	<u>0.213\pm.017</u>	3.034 \pm .021	10.99 \pm .101	4.138\pm.145

Table 3: Comparisons to other methods on the complex motion subset. We filter the HumanML3D test set containing at least 3 local actions and 150 frames or more in length as a new test set to verify the ability to generate complex motions.

Methods	R-Precision \uparrow			FID \downarrow	MM-Dist \downarrow	Diversity \rightarrow	MModality \uparrow
	Top-1	Top-2	Top-3				
Real Motion	0.456 \pm .002	0.640 \pm .002	0.740 \pm .002	0.002 \pm .000	3.245 \pm .009	8.738 \pm .064	-
MDM [53] <small>ICLR23</small>	0.300 \pm .008	0.473 \pm .006	0.581 \pm .011	0.579 \pm .057	5.437 \pm .041	8.987\pm.101	2.808\pm.063
MLD [10] <small>CVPR23</small>	0.417 \pm .006	0.603 \pm .005	0.710 \pm .006	0.783 \pm .069	3.243\pm.013	9.235 \pm .164	<u>2.642\pm.118</u>
T2M-GPT [67] <small>CVPR23</small>	0.431 \pm .003	<u>0.612\pm.003</u>	<u>0.712\pm.002</u>	<u>0.314\pm.004</u>	3.448 \pm .011	9.277 \pm .081	2.125 \pm .011
GuidedMotion (Ours)	0.451\pm.003	0.635\pm.003	0.732\pm.002	0.144\pm.008	<u>3.447\pm.011</u>	9.284 \pm .057	2.503 \pm .113

Comparisons to State-of-the-Art. We provide qualitative motion results in Fig. 4. Compared to other methods, our method generates motions that match the text descriptions better and are more realistic. Moreover, we compare the proposed GuidedMotion with other methods on two benchmarks. Tab. 1 shows the results on the HumanML3D test set. Tab. 2 presents the results on the KIT test set. Across both benchmarks, the proposed GuidedMotion, which allows for continuous refinement of motion results, achieves performance comparable to existing state-of-the-art methods that lack fine-grained control.

Evaluation on Complex Motion Generation. We analyze the distribution of the number of local actions in each motion in Fig. 5. As shown in Fig. 5, motions typically consist of multiple local actions. However, generating complex motions, characterized by many local actions, poses a challenge. Compared with the direct generation of complex motion, our local-to-global paradigm, utilizing local actions as a prior, simplifies the intricacies involved in generating complex motions. To demonstrate the advantages of the proposed local-to-global generation paradigm in generating complex motions, we create a new test set from the HumanML3D test set, consisting of motions with at least 3 local actions and

Table 4: Ablation study of each part on the HumanML3D test set.

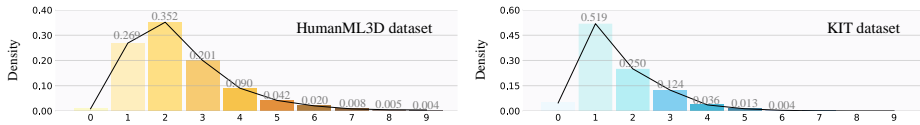
Motion Level	Action Level	Specific Level	Local Action Guidance	R-Precision Top-3 \uparrow	FID \downarrow
✓				0.760 \pm .003	0.186 \pm .011
✓	✓			0.771 \pm .003	0.133 \pm .009
✓	✓		✓	0.778 \pm .002	0.119 \pm .009
✓	✓	✓		0.769 \pm .004	0.107 \pm .009
✓	✓	✓	✓	0.788\pm.002	0.057\pm.006

Table 5: Evaluation of the motion VAE models on the motion part on the HumanML3D test set.

Methods	Token Size	R-Precision Top-3 \uparrow	FID \downarrow
Real Motion	-	0.797 \pm .002	0.002 \pm .000
Motion Level	2	0.791 \pm .003	1.906 \pm .003
Action Level	4	0.793 \pm .003	0.068 \pm .002
Specific Level	8	0.800\pm.004	0.019\pm.003

Table 6: Effect of diffusion steps on the HumanML3D test set. We use DDIM in practice and set T^m , T^a , and T^s to 50 for optimal performance.

Methods	Diffusion Steps			FID \downarrow
	T^m	T^a	T^s	
<i>1000 diffusion steps with DDPM [23]</i>				
MDM [53] <small>ICLR23</small>	1000	✗	✗	0.544 \pm .044
MotionDiffuse [68] <small>TPAMI24</small>	1000	✗	✗	0.630 \pm .001
<i>50 diffusion steps with DDIM [50]</i>				
MLD [10] <small>CVPR23</small>	50	✗	✗	0.473 \pm .013
GuidedMotion (Ours)	20	15	15	0.136 \pm .007
GuidedMotion (Ours)	15	20	15	0.120 \pm .006
GuidedMotion (Ours)	15	15	20	0.117\pm.006
<i>150 diffusion steps with DDIM [50]</i>				
MLD [10] <small>CVPR23</small>	150	✗	✗	0.457 \pm .011
GuidedMotion (Ours)	50	50	50	0.057\pm.006
<i>300 diffusion steps with DDIM [50]</i>				
MLD [10] <small>CVPR23</small>	300	✗	✗	0.403 \pm .011
GuidedMotion (Ours)	100	100	100	0.062\pm.007

**Fig. 5: The distribution of the number of local actions in each motion.** Motions typically consist of multiple local actions rather than just one local action.

lasting 150 frames or more. As shown in Tab. 3, our method maintains generation quality even for complex motions and is superior to other methods on most metrics. These results demonstrate the benefits of the proposed local-to-global paradigm in generating complex motions comprising multiple local actions.

Ablative Analysis. *Analysis of each part of our method.* To explore the impact of each part of our method, we provide the ablation results in Tab. 4. In the proposed hierarchical motion diffusion model, the high semantic layer generates results based on the results from the low semantic layer. As shown in Tab. 4, higher semantic levels, such as the specific level, exhibit superior performance. Moreover, the proposed local action guidance significantly enhances the quality of the generated motion, providing conclusive evidence for the effectiveness of the proposed method. We observe that the performance in the ‘‘R-Precision Top-3’’ metric at the specific level, without local action guidance, is lower compared to the action level. This is likely due to the specific level refining results based on the action details. When two motion descriptions share overlapping action details, the model may produce similar features in the generated motions, thus adversely affecting R-Precision. Despite this, since the specific level can enhance the quality (FID) of motion generation, we still recommend its utilization.

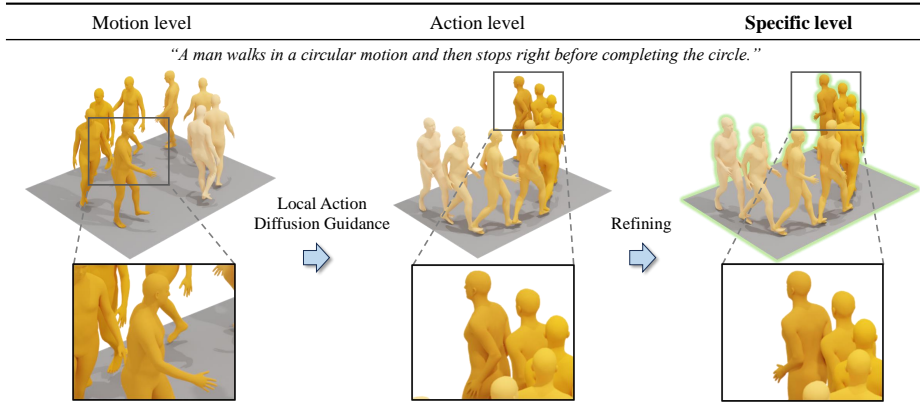


Fig. 6: Qualitative comparison of different hierarchical levels. The output at the higher level (*e.g.*, specific level) contains more action details.

Analysis of the motion VAE models. In Tab. 5, we show the evaluation of the motion VAE models on the HumanML3D test set. Among the three levels, the performance of the specific level is the best, which indicates that increasing the token size enhances the reconstruction ability of the motion VAE models. Therefore, we take the output at the specific level as the final result and use the motion decoder to decode the latent representation into the motion sequence.

Analysis of the diffusion steps. In Tab. 6, we provide the ablation results of the total number of diffusion steps on the HumanML3D test set. We observe that the number of diffusion steps at the higher level, such as the specific level, has a more pronounced impact on quantitative results. Simultaneously, the number of diffusion steps at the action level determines the control ability of the local action guidance to the global motion. Therefore, we recommend allocating a sufficient number of diffusion steps to each level. As illustrated in Tab. 6, the performance is similar when the total number of diffusion steps is set to 150 and 300, prompting us to adopt a setting of 150 steps in practice.

Qualitative Analysis. *Visualization of different hierarchies.* The results in Fig. 6 show that the output at the higher level (*e.g.*, specific level) contains more action details. Specifically, the motion level generates only coarse-grained overall motion. The action level generates local actions with guidance but lacks action specifics. The specific level generates more action specifics than the action level.

Visualization of adjusting the guiding weight of each local action. The strength of our local action-guided motion generation lies in its capacity to fine-tune the generation process of motion diffusion models. In contrast to existing methods confined to producing a singular style of motion, our method offers flexibility in adjusting the guiding weight λ of action guidance. This affords precise control over how each local action influences the overall motion, catering to diverse user preferences. As illustrated in Fig. 7, we can manipulate the movement trajectories by varying the guiding weight of the local action. For example, increasing

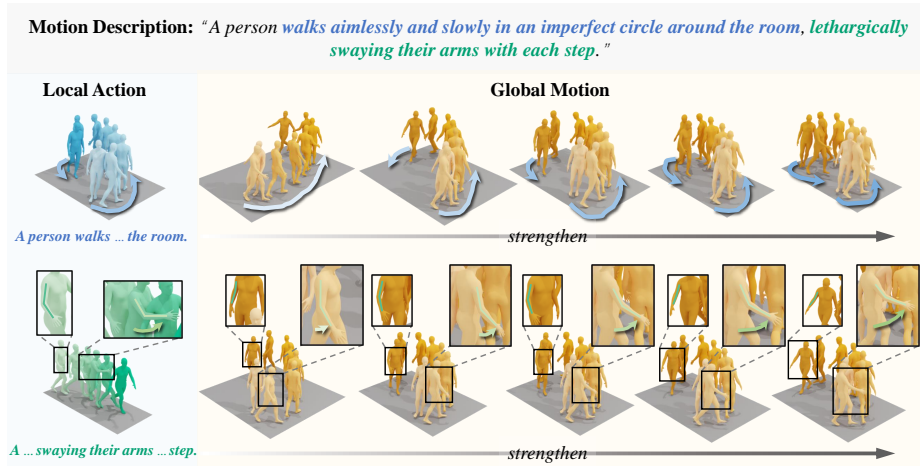


Fig. 7: The proposed GuidedMotion controls the generation process of motion diffusion models. Our method provides flexibility in adjusting the guiding weight λ of each local action, enabling fine-grained control over global motion.

the guiding weight of “walks aimlessly and slowly in an imperfect circle around the room” results in the human body walking in a tighter circle. Furthermore, we can refine the human body postures throughout the motion. For instance, by amplifying the guiding weight of “lethargically swaying their arms with each step,” the body exhibits more pronounced arm movements.

5 Conclusion

In this paper, we introduce GuidedMotion, a local action-guided motion diffusion model designed to enhance the controllability of text-driven human motion generation by employing local actions as fine-grained control signals. Our method empowers users to combine preferred local actions freely, generating motions that align with their mental imagery. Extensive experiments demonstrate that our method achieves superior controllability than the existing state-of-the-art methods. Furthermore, our method supports continuous guiding weight adjustment, allowing for the refinement of the motion results to align with user preferences.

Acknowledgements

This work was supported in part by the National Key R&D Program of China (No. 2022ZD0118101), Natural Science Foundation of China (No. 61972217, 32071459, 62176249, 62006133, 62271465, 62332002, 62202014), the Shenzhen Medical Research Funds in China (No. B2302037), and AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, China.

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. Ahuja, C., Morency, L.P.: Language2pose: Natural language grounded pose forecasting. In: 3DV. pp. 719–728 (2019)
3. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., van den Berg, R.: Structured denoising diffusion models in discrete state-spaces. In: NeurIPS. pp. 17981–17993 (2021)
4. Badler, N.I., Phillips, C.B., Webber, B.L.: Simulating humans: computer graphics animation and control. Oxford University Press (1993)
5. Barquero, G., Escalera, S., Palmero, C.: Belfusion: Latent diffusion for behavior-driven human motion prediction. In: ICCV. pp. 2317–2327 (2023)
6. Bhattacharya, U., Rewkowski, N., Banerjee, A., Guhan, P., Bera, A., Manocha, D.: Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In: VR. pp. 1–10 (2021)
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. In: NeurIPS. pp. 1877–1901 (2020)
8. Chen, L.H., Zhang, J., Li, Y., Pang, Y., Xia, X., Liu, T.: Humanmac: Masked motion completion for human motion prediction. arXiv preprint arXiv:2302.03665 (2023)
9. Chen, S., Zhao, Y., Jin, Q., Wu, Q.: Fine-grained video-text retrieval with hierarchical graph reasoning. In: CVPR. pp. 10638–10647 (2020)
10. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, J., Yu, G.: Executing your commands via motion diffusion in latent space. In: CVPR. pp. 18000–18010 (2023)
11. Cheng, Z., Li, K., Jin, P., Ji, X., Yuan, L., Liu, C., Chen, J.: Parallel vertex diffusion for unified visual grounding. arXiv preprint arXiv:2303.07216 (2023)
12. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 (2015)
13. Delmas, G., Weinzaepfel, P., Lucas, T., Moreno-Noguer, F., Rogez, G.: Posescript: 3D human poses from natural language. In: ECCV. pp. 346–362 (2022)
14. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: NeurIPS. pp. 8780–8794 (2021)
15. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. In: ICCV. pp. 1396–1406 (2021)
16. Gong, S., Li, M., Feng, J., Wu, Z., Kong, L.: DIFFUSEQ: Sequence to sequence text generation with diffusion models. arXiv preprint arXiv:2210.08933 (2022)
17. 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)
18. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3D human motions and texts. In: ECCV. pp. 580–597 (2022)
19. 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)
20. He, C., Saito, J., Zachary, J., Rushmeier, H., Zhou, Y.: NeMF: Neural motion fields for kinematic animation. In: NeurIPS. pp. 4244–4256 (2022)

21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
22. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. pp. 6840–6851 (2020)
24. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
25. Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3D scenes. In: CVPR. pp. 16750–16761 (2023)
26. Jeong, H., Kwon, G., Ye, J.C.: Zero-shot generation of coherent storybook from plain text story using diffusion models. arXiv preprint arXiv:2302.03900 (2023)
27. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. In: NeurIPS (2023)
28. Jin, P., Huang, J., Liu, F., Wu, X., Ge, S., Song, G., Clifton, D., Chen, J.: Expectation-maximization contrastive learning for compact video-and-language representations. In: NeurIPS. pp. 30291–30306 (2022)
29. Jin, P., Huang, J., Xiong, P., Tian, S., Liu, C., Ji, X., Yuan, L., Chen, J.: Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In: CVPR. pp. 2472–2482 (2023)
30. Jin, P., Li, H., Cheng, Z., Li, K., Ji, X., Liu, C., Yuan, L., Chen, J.: Diffusion-ret: Generative text-video retrieval with diffusion model. In: ICCV. pp. 2470–2481 (2023)
31. Jin, P., Wu, Y., Fan, Y., Sun, Z., Wei, Y., Yuan, L.: Act as you wish: Fine-grained control of motion diffusion model with hierarchical semantic graphs. In: NeurIPS (2023)
32. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV. pp. 694–711 (2016)
33. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: ICCV. pp. 2151–2162 (2023)
34. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
35. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
36. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F.: A tutorial on energy-based learning. Predicting structured data **1**(0) (2006)
37. 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. In: NeurIPS (2023)
38. Loshchilov, I., Hutter, F., et al.: Fixing weight decay regularization in adam. arXiv preprint arXiv:1711.05101 (2017)
39. Maas, A.L., Hannun, A.Y., Ng, A.Y., et al.: Rectifier nonlinearities improve neural network acoustic models. In: ICML (2013)
40. 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)
41. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3D human motion synthesis with transformer vae. In: ICCV. pp. 10985–10995 (2021)
42. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: ECCV. pp. 480–497 (2022)

43. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. *Big data* **4**(4), 236–252 (2016)
44. Plappert, M., Mandery, C., Asfour, T.: Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems* **109**, 13–26 (2018)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: *ICML*. pp. 8748–8763 (2021)
46. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: *MICCAI*. pp. 234–241 (2015)
47. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418* (2023)
48. Shi, P., Lin, J.: Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255* (2019)
49. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: *ICML*. pp. 2256–2265 (2015)
50. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020)
51. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: *NeurIPS* (2019)
52. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: *ICLR* (2021)
53. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: *ICLR* (2023)
54. 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)
55. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023)
56. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: *NeurIPS* (2017)
57. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. In: *ICLR* (2018)
58. Wang, Y., Leng, Z., Li, F.W., Wu, S.C., Liang, X.: Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In: *ICCV*. pp. 22035–22044 (2023)
59. Wang, Y., Yu, J., Zhang, J.: Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490* (2022)
60. Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation. *arXiv preprint arXiv:2310.08580* (2023)
61. Xu, S., Li, Z., Wang, Y.X., Gui, L.Y.: Interdiff: Generating 3D human-object interactions with physics-informed diffusion. In: *ICCV*. pp. 14928–14940 (2023)
62. Yang, C., Wang, R., Yao, S., Liu, S., Abdelzaher, T.: Revisiting over-smoothing in deep gcn. *arXiv preprint arXiv:2003.13663* (2020)
63. Yu, H., Zhang, D., Xie, P., Zhang, T.: Point-based radiance fields for controllable human motion synthesis. *arXiv preprint arXiv:2310.03375* (2023)
64. Yu, J., Wang, Y., Zhao, C., Ghanem, B., Zhang, J.: Freedom: Training-free energy-guided conditional diffusion model. *arXiv preprint arXiv:2303.09833* (2023)

65. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. In: ICCV. pp. 16010–16021 (2023)
66. Zhai, Y., Huang, M., Luan, T., Dong, L., Nwogu, I., Lyu, S., Doermann, D., Yuan, J.: Language-guided human motion synthesis with atomic actions. In: ACM MM. pp. 5262–5271 (2023)
67. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. In: CVPR (2023)
68. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. TPAMI (2024)
69. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: ICCV (2023)
70. Zhang, Q., Song, J., Huang, X., Chen, Y., Liu, M.Y.: Diffcollage: Parallel generation of large content with diffusion models. arXiv preprint arXiv:2303.17076 (2023)
71. Zhang, Y., Huang, D., Liu, B., Tang, S., Lu, Y., Chen, L., Bai, L., Chu, Q., Yu, N., Ouyang, W.: Motiongpt: Finetuned llms are general-purpose motion generators. arXiv preprint arXiv:2306.10900 (2023)
72. Zhao, M., Bao, F., Li, C., Zhu, J.: Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. In: NeurIPS. pp. 3609–3623 (2022)
73. Zhu, W., Ma, X., Ro, D., Ci, H., Zhang, J., Shi, J., Gao, F., Tian, Q., Wang, Y.: Human motion generation: A survey. arXiv preprint arXiv:2307.10894 (2023)