Act As You Wish: Fine-Grained Control of Motion Diffusion Model with Hierarchical Semantic Graphs Supplementary Material

Peng Jin[1,4] **Yang Wu**[3*] **Yanbo Fan**[3] **Zhongqian Sun**[3] **Yang Wei**[3] **Li Yuan**[1,2,4*]

[1] School of Electronic and Computer Engineering, Peking University, Shenzhen, China
[2] Peng Cheng Laboratory, Shenzhen, China [3] Tencent AI Lab, China
[4] AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, China

jp21@stu.pku.edu.cn dylan.yangwu@qq.com yuanli-ece@pku.edu.cn

Abstract This appendix provides additional discussions (Sec. A), implementation details (Sec. B), more qualitative results (Sec. C), several additional experiments (Sec. D), details of motion representations and metric definitions (Sec. E).

Code Code is available at https://github.com/jpthu17/GraphMotion. In this code, we provide the process of the training and evaluation of the proposed method, and the pre-trained weights.

A Additional Discussions

A.1 Potential Negative Societal Impacts

While our work effectively enhances the quality of human motion synthesis, there is a potential risk that it may be used for generating fake content, such as generating fake news, which can pose a threat to information security. Moreover, when factoring in energy consumption, there is a possibility that the widespread use of generative models for synthesizing human motions may contribute to increased carbon emissions and exacerbate global warming.

A.2 Limitations of our Work

Although our method makes some progress, there are still many limitations worth further study. (1) The proposed GraphMotion inherits the randomness of diffusion models. This property benefits diversity but may yield undesirable results sometimes. (2) The human motion synthesis capabilities of GraphMotion are limited by the performance of the pre-trained motion variational autoencoders, which we will discuss in experiments (Tab. D and Tab. E). This defect also exists in the existing state-of-the-art methods, such as MLD [3] and T2M-GPT [26], which also use motion variational autoencoder. (3) Though the proposed GraphMotion brings negligible extra cost on computations, it is still limited by the slow inference speed of existing diffusion models. We will discuss the inference time in experiments (Tab. C). This defect also exists in the existing state-of-the-art methods, such as MDM [25] and MLD [3], which also use diffusion models.

A.3 Future Work

In this paper, we focus on improving the controllability of text-driven human motion generation. Recently, large language models have made remarkable progress, making large language models a promising text extractor for human motion generation. However, despite their strengths in general reasoning and broad applicability, large language models may not be optimized for extracting subtle motion nuances. In future research, we will incorporate the features of large-scale languages into

*Corresponding author: Yang Wu, Li Yuan.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

our model, using hierarchical semantic graphs to give large language models the ability to extract fine-grained motion description structures. In addition, the application of hierarchical semantic graphs to other cross-modal tasks [10, 11, 15], such as cross-modal retrieval [9, 12] and visual question answering, is also a promising research direction.

B Implementation Details

B.1 Details of Hierarchical Semantic Graphs

To obtain actions, attributes of action as well as the semantic role of each attribute to the corresponding action, we implement a semantic parser of motion descriptions based on a semantic role parsing toolkit [21, 2]. Specifically, given the motion description, the parser extracts verbs that appeared in the sentence and attribute phrases corresponding verb, and the semantic role of each attribute phrase. The overall sentence is treated as the global motion node in the hierarchical graph. The verbs are considered as action nodes and connected to the motion node with direct edges, allowing for implicit learning of the temporal relationships among various actions during graph reasoning. The attribute phrases are specific nodes that are connected with action nodes. The edge type between action and specific nodes is determined by the semantic role of the specifics in relation to the action. As shown in Tab. A, we extract three types (motions, actions, and specifics) of nodes and twelve types of edges to represent various associations among the nodes.

Table A: **Node types and edge types in the parsed hierarchical semantic graph.** Each edge type corresponds to a type of semantic role.

Node type	Description
Motion	global motion description
Action	verb
Specific	attribute of action
Edge type	**Description**
ARG0	agent
ARG1	patient
ARG2	instrument, benefactive
ARG3	start point
ARG4	end point
ARGM-LOC	location (where)
ARGM-MNR	manner (how)
ARGM-TMP	time (when)
ARGM-DIR	direction (where to/from)
ARGM-ADV	miscellaneous
ARGM-MA	motion-action dependencies
OTHERS	other argument types, e.g., action

B.2 Classifier-free Diffusion Guidance

Following MLD [3], our denoiser network is learned with classifier-free diffusion guidance [8]. The classifier-free diffusion guidance improves the quality of samples by reducing diversity in conditional diffusion models. Concretely, it learns both the conditioned and the unconditioned distribution (10% dropout [23]) of the samples. Finally, we perform a linear combination in the following manner, which is formulated as:

$$\begin{aligned}
\widehat{\epsilon^m}_{scale} &= \alpha^{'}\phi_m(z^m, t^m, \mathcal{V}^m) + (1-\alpha^{'})\phi_m(z^m, t^m, \varnothing), \\
\widehat{\epsilon^a}_{scale} &= \alpha^{'}\phi_a(z^a, t^a, [\mathcal{V}^m, \mathcal{V}^a, z^m]) + (1-\alpha^{'})\phi_a(z^a, t^a, \varnothing), \\
\widehat{\epsilon^s}_{scale} &= \alpha^{'}\phi_s(z^s, t^s, [\mathcal{V}^m, \mathcal{V}^a, \mathcal{V}^s, z^a]) + (1-\alpha^{'})\phi_s(z^s, t^s, \varnothing),
\end{aligned} \tag{A}$$

Where $\alpha^{'}$ is the guidance scale and $\alpha^{'} > 1$ can strengthen the effect of guidance [3]. We set $\alpha^{'}$ to 7.5 in practice following MLD. Please refer to our code for more details.

B.3 Implementation Details for Different Datasets

Following MLD [3], we utilize a frozen text encoder of the CLIP-ViT-L-14 [19] model for text representation. The dimension of node representation D is set to 768. The dimension of latent embedding D' is set to 256. For the motion variational autoencoder, motion encoder $\mathcal{E}$ and decoder $\mathcal{D}$ all consist of 9 layers and 4 heads with skip connection [20]. We set the token sizes C^m to 2, C^a to 4, and C^s to 8. We set λ to 1e-4. All our models are trained with the AdamW [13, 17] optimizer using a fixed learning rate of 1e-4. We use 4 Tesla V100 GPUs for the training, and there are 128 samples on each GPU, so the total batch size is 512. The number of diffusion steps of each level is 1,000 during training, and the step sizes β_t are scaled linearly from 8.5$\times$1e-4 to 0.012. We keep running a similar number of iterations on different data sets. 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. For 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. Code is available at https://github.com/jpthu17/GraphMotion. In this code, we provide the process of the training and evaluation of the proposed method, and the pre-trained model.

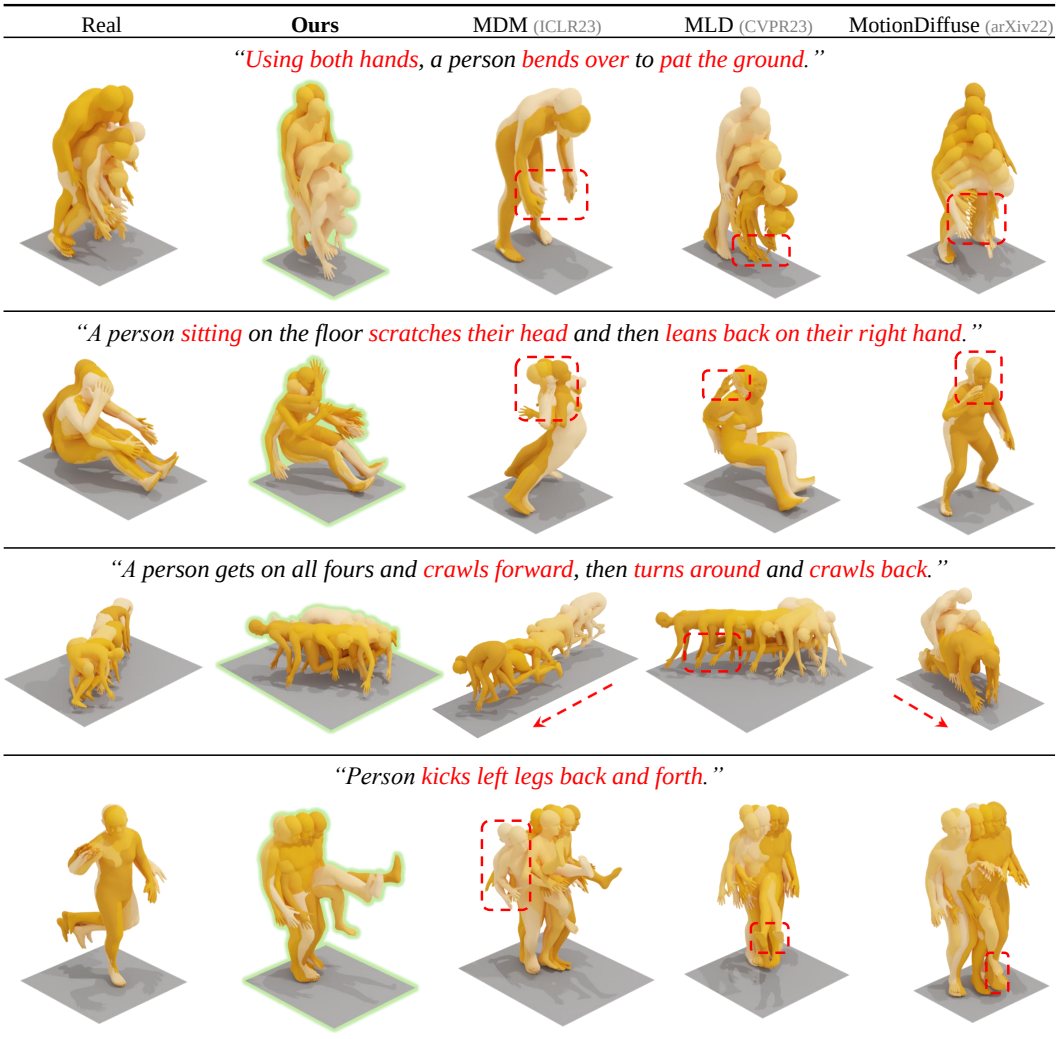

Figure A: **Qualitative comparison of the existing methods.** The darker colors indicate the later in time. The generated results of our method better match the descriptions, while others have downgraded motions or improper semantics, demonstrating that our method achieves superior controllability compared to well-designed baseline models. We have provided a supplemental video in our supplementary material. In the supplemental video, we show comparisons of text-driven motion generation. We suggest the reader watch this video for dynamic motion results.

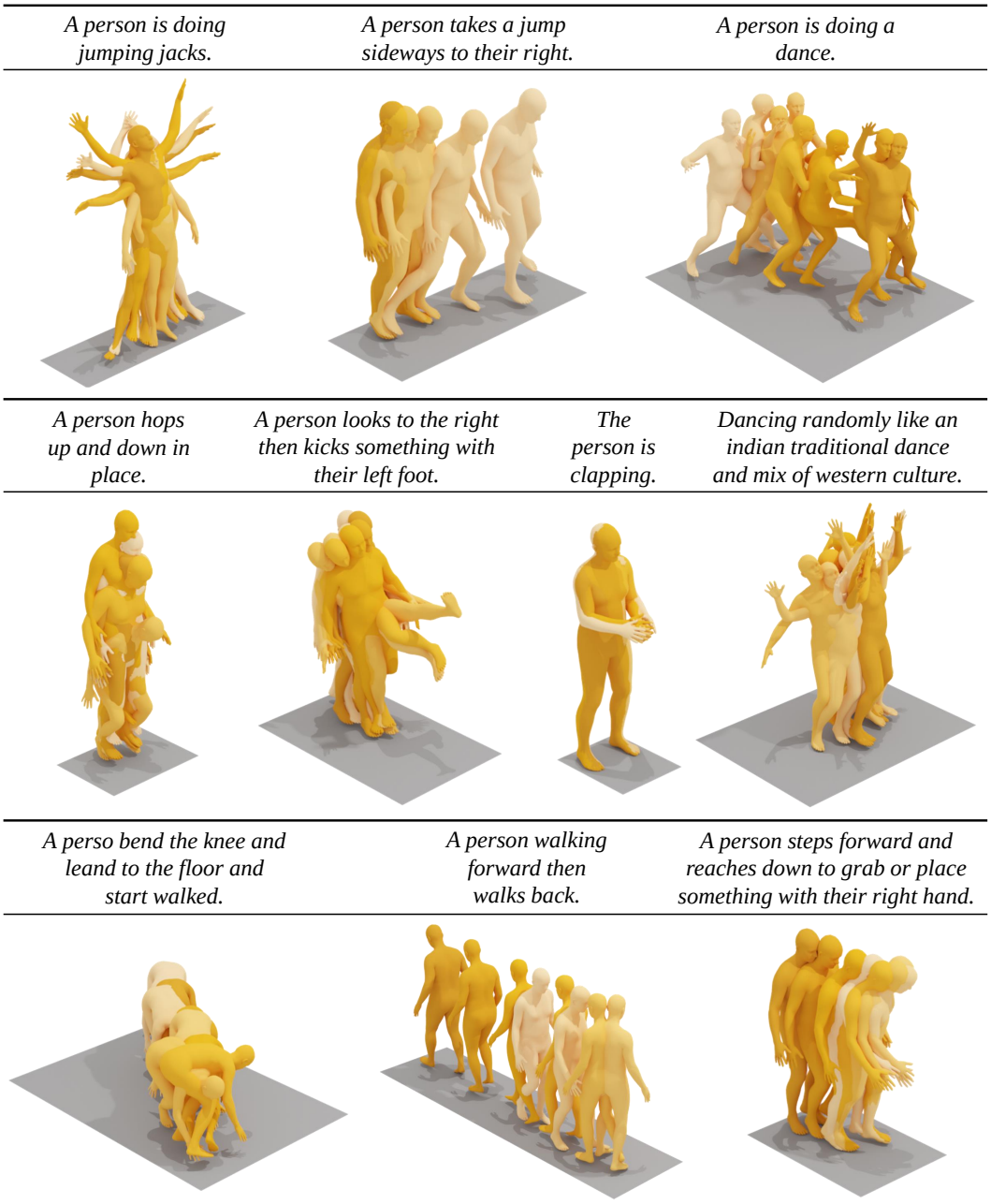

Figure B: **Additional qualitative motion results are generated with text prompts of the HumanML3D test set.** The darker colors indicate the later in time. These results demonstrate that our method can generate diverse and accurate motion sequences.

C Additional Qualitative Analysis

C.1 Qualitative Comparison of the Existing Methods

We provide additional qualitative motion results in Fig. A. Compared to other methods, our method generates motions that match the text descriptions better, indicating that our method is more sensitive to subtle differences in texts. The generated results demonstrate that our method achieves superior controllability compared to well-designed baseline models.

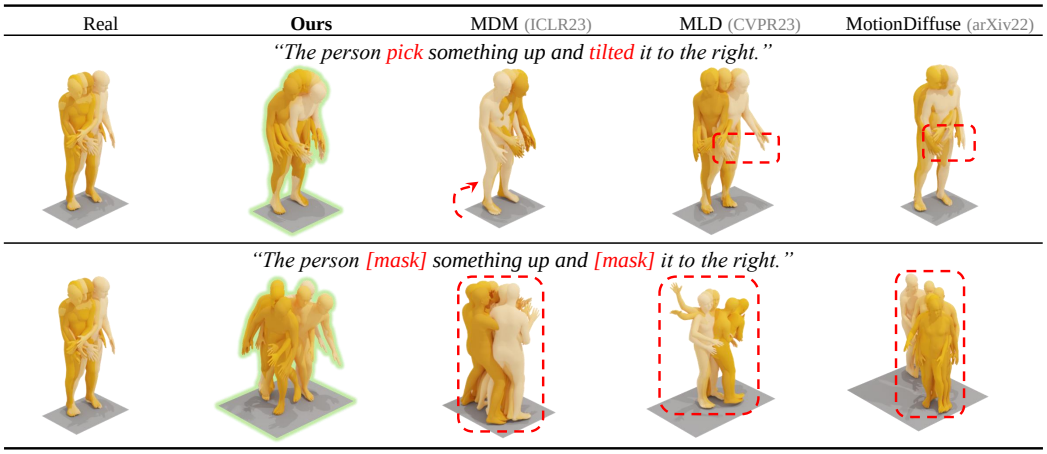

Figure C: **Qualitative analysis on the imbalance problem.** The darker colors indicate the later in time. When the verbs and action names are masked, existing models tend to generate motion randomly. In contrast, our method can generate motion based solely on the action specifics. These results show that our method is not overly focused on the verbs and action names.

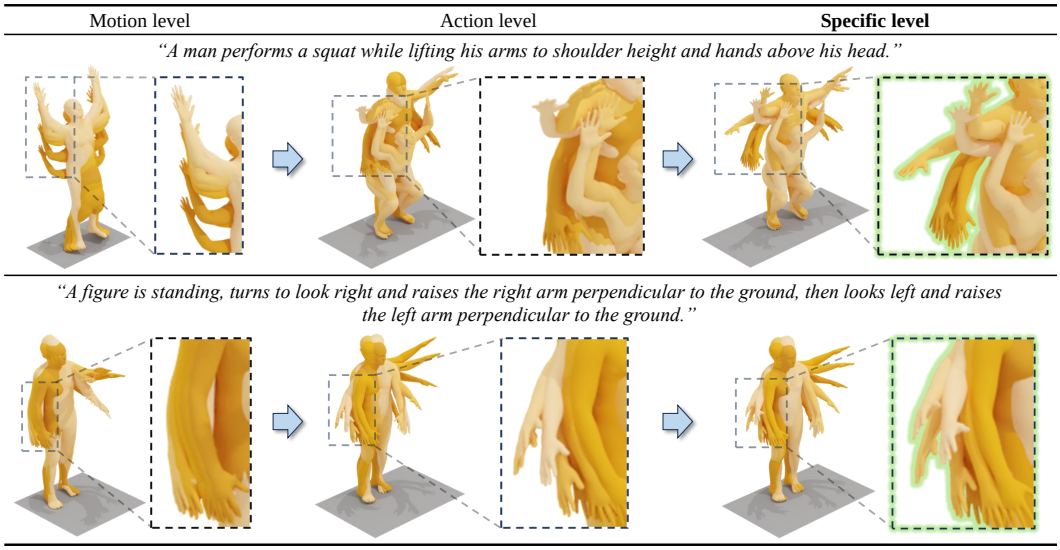

Figure D: **Qualitative comparison of different hierarchies.** The output at the higher level (e.g., specific level) has more action details. Specifically, the motion level generates only coarse-grained overall motion. The action level generates local actions better than the motion level but lacks action specifics. The specific level generates more action specifics than the action level.

C.2 Additional Visualization Results

In Fig. B, we provide additional qualitative motion results which are generated with text prompts of the HumanML3D test set. These results demonstrate that our method can generate diverse and accurate motion sequences from a variety of motion descriptions.

C.3 Qualitative Analysis on the Imbalance Problem

To demonstrate the imbalance problem of other methods and prove that our method does not have this problem, we mask the verbs and action names in the motion description to force the model to generate motion only from action specifics. As shown in Fig. C, when the verbs and action names are masked, existing models tend to generate motion randomly. In contrast, our method can generate

Table B: **Ablation study about the total number of diffusion steps on the HumanML3D test set.** "↑" denotes that higher is better. "↓" denotes that lower is better. We repeat all the evaluations 20 times and report the average with a 95% confidence interval. "✘" denotes that this method does not apply this parameter. To speed up the sampling process, we use DDIM in practice following MLD.

Methods	Diffusion Steps			R-Precision ↑			FID ↓
	Motion T^m	Action T^a	Specific T^s	Top-1	Top-2	Top-3	
The total number of diffusion steps is 1000 with DDPM [7]							
MDM [25]	1000	✘	✘	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$
MotionDiffuse [27]	1000	✘	✘	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$0.630^{\pm.001}$
The total number of diffusion steps is 50 with DDIM [22]							
MLD [3]	50	✘	✘	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$
GraphMotion (Ours)	20	15	15	$0.489^{\pm.003}$	$0.676^{\pm.002}$	$0.771^{\pm.002}$	$0.131^{\pm.007}$
GraphMotion (Ours)	15	15	20	$\mathbf{0.496^{\pm.003}}$	$\mathbf{0.686^{\pm.003}}$	$\mathbf{0.778^{\pm.002}}$	$\mathbf{0.118^{\pm.008}}$
The total number of diffusion steps is 150 with DDIM [22]							
MLD [3]	150	✘	✘	$0.461^{\pm.002}$	$0.649^{\pm.003}$	$0.797^{\pm.002}$	$0.457^{\pm.011}$
GraphMotion (Ours)	50	50	50	$\mathbf{0.504^{\pm.003}}$	$\mathbf{0.699^{\pm.002}}$	$\mathbf{0.785^{\pm.002}}$	$\mathbf{0.116^{\pm.007}}$
The total number of diffusion steps is 300 with DDIM [22]							
MLD [3]	300	✘	✘	$0.473^{\pm.002}$	$0.664^{\pm.003}$	$0.765^{\pm.002}$	$0.403^{\pm.011}$
GraphMotion (Ours)	100	100	100	$\mathbf{0.486^{\pm.003}}$	$\mathbf{0.671^{\pm.004}}$	$\mathbf{0.767^{\pm.003}}$	$\mathbf{0.096^{\pm.008}}$
The total number of diffusion steps is 1000 with DDIM [22]							
MLD [3]	1000	✘	✘	$0.452^{\pm.002}$	$0.639^{\pm.003}$	$0.751^{\pm.002}$	$0.460^{\pm.013}$
GraphMotion (Ours)	400	300	300	$0.475^{\pm.003}$	$0.659^{\pm.003}$	$0.756^{\pm.003}$	$0.136^{\pm.007}$
GraphMotion (Ours)	300	300	400	$\mathbf{0.484^{\pm.003}}$	$\mathbf{0.694^{\pm.003}}$	$\mathbf{0.787^{\pm.003}}$	$\mathbf{0.132^{\pm.008}}$

Table C: **Evaluation of Inference time costs on the HumanML3D test set.** We evaluate the average time per sample with different diffusion schedules and FID. "↓" denotes that lower is better. Please note the bad FID of MDM with DDIM is mentioned in their GitHub issues #76. "✘" denotes that this method does not apply this parameter. We use DDIM in practice following MLD.

Methods	Reference	Diffusion Steps			Average time per sample (s) ↓	FID ↓
		Motion T^m	Action T^a	Specific T^s		
The total number of diffusion steps is 1000 with DDPM [7]						
MDM [25]	ICLR 2023	1000	✘	✘	178.7699	**0.544**
MLD [3]	CVPR 2023	1000	✘	✘	5.5045	0.568
The total number of diffusion steps is 50 with DDIM [22]						
MDM [25]	ICLR 2023	50	✘	✘	20.5678	7.334
MLD [3]	CVPR 2023	50	✘	✘	0.9349	0.473
GraphMotion	Ours	20	15	15	0.9094	0.131
GraphMotion	Ours	15	15	20	0.7758	**0.118**
The total number of diffusion steps is 150 with DDIM [22]						
MLD [3]	CVPR 2023	150	✘	✘	2.4998	0.457
GraphMotion	Ours	50	50	50	2.5518	**0.116**
The total number of diffusion steps is 1000 with DDIM [22]						
MLD [3]	CVPR 2023	1000	✘	✘	16.6654	0.460
GraphMotion	Ours	400	300	300	22.1238	0.136
GraphMotion	Ours	300	300	400	17.0912	**0.132**

motion that matches the description based solely on the action specifics. These results show that our method is not overly focused on the verbs and action names.

C.4 Qualitative Comparison of Different Hierarchies

We provide different levels of qualitative comparison in Fig. D. The results show that the output at the higher level (e.g., specific level) has more action details. Specifically, the motion level generates only coarse-grained overall motion. The action level generates local actions better than the motion level but lacks action specifics. The specific level generates more action specifics than the action level.

D Additional Experiments

D.1 Analysis of the Diffusion Steps

In Tab. B, we show the ablation study of the total number of diffusion steps on the HumanML3D test set. Following MLD [3], we adopt the denoising diffusion implicit models [22] (DDIM) during interference. As shown in Tab. B, our method consistently outperforms the existing state-of-the-art methods with the same total number of diffusion steps, which demonstrates the efficiency of our method. We find that the number of diffusion steps at the higher level (e.g., specific level) has a greater impact on the result. Therefore, in scenarios requiring high efficiency, we recommend allocating more diffusion steps to the higher level. Moreover, with the increase of the total diffusion steps, the performance of our method is further improved, while the performance of MLD saturates. These results further prove the superiority of our design.

D.2 Analysis of the Inference Time

In Tab. C, we provide the evaluation of inference time costs. Our method is as efficient as the one-stage diffusion methods during the inference stage, even though we decompose the diffusion process into three parts. This is because we can control the total number $T^m + T^a + T^s$ of iterations by restricting it to be the same as those of the one-stage diffusion methods. As shown in Tab. C, the inference speed of our method is comparable to that of the existing state-of-the-art methods with the same total number of diffusion steps, which demonstrates the efficiency of our method.

D.3 Analysis of the motion VAE models

We provide the evaluation of the motion VAE models. In Tab. D, we show the results on the HumanML3D test set. Tab. E shows the results on the KIT test set. Among the three levels, the performance of the specific level is the best, which indicates that increasing the token size can improve the reconstruction ability of the motion VAE models.

E Motion Representations and Metric Definitions

E.1 Motion Representations

Motion representation can be summarized into the following four categories, and we follow the previous work of representing motion in latent space.

Latent Format. Following previous works [18, 3, 26], we encode the motion into the latent space with a motion variational autoencoder [14]. The latent representation is formulated as:

$$\hat{x}^{1:L} = \mathcal{D}(z), \quad z = \mathcal{E}(x^{1:L}). \tag{B}$$

HumanML3D Format. HumanML3D [5] proposes a motion representation $x^{1:L}$ inspired by motion features in character control. This motion representation is well-suited for neural networks. To be specific, the i_{th} pose x^i is defined by a tuple consisting of the root angular velocity $r^a \in \mathbb{R}$ along the Y-axis, root linear velocities $(r^x, r^z \in \mathbb{R})$ on the XZ-plane, root height $r^y \in \mathbb{R}$, local joints positions $\boldsymbol{j^p} \in \mathbb{R}^{3N_j}$, velocities $\boldsymbol{j^v} \in \mathbb{R}^{3N_j}$, and rotations $\boldsymbol{j^r} \in \mathbb{R}^{6N_j}$ in root space, and binary foot-ground contact features $\boldsymbol{c^f} \in \mathbb{R}^4$ obtained by thresholding the heel and toe joint velocities. Here, N_j denotes the joint number. Finally, the HumanML3D format can be defined as:

$$x^i = \{r^a, r^x, r^z, r^y, \boldsymbol{j^p}, \boldsymbol{j^v}, \boldsymbol{j^r}, \boldsymbol{c^f}\}. \tag{C}$$

SMPL-based Format. SMPL [16] is one of the most widely used parametric human models. SMPL and its variants propose motion parameters θ and shape parameters β. $\theta \in \mathbb{R}^{3\times 23+3}$ is rotation vectors for 23 joints and a root, while β represents the weights for linear blended shapes. The global translation r is also incorporated to formulate the representation as follows:

$$x^i = \{r, \theta, \beta\}. \tag{D}$$

MMM Format. Master Motor Map [24] (MMM) representations propose joint angle parameters based on a uniform skeleton structure with 50 degrees of freedom (DoFs). In text-to-motion tasks,

Table D: **Evaluation of the VAE models on the motion part of the HumanML3D test set.** "↑" denotes that higher is better. "↓" denotes that lower is better. "→" denotes that results are better if the metric is closer to the real motion. The performance of the specific level is the best.

Methods	Token Size	R-Precision ↑			FID ↓	Diversity →
		Top-1	Top-2	Top-3		
Real motion	-	0.511	0.703	0.797	0.002	9.503
Motion Level	2	0.498	0.692	0.791	1.906	9.675
Action Level	4	0.514	0.703	0.793	0.068	**9.610**
Specific Level	8	**0.525**	**0.708**	**0.800**	**0.019**	9.863

Table E: **Evaluation of the VAE models on the motion part of the KIT test set.** "↑" denotes that higher is better. "↓" denotes that lower is better. "→" denotes that results are better if the metric is closer to the real motion. The performance of the specific level is the best.

Methods	Token Size	R-Precision ↑			FID ↓	Diversity →
		Top-1	Top-2	Top-3		
Real motion	-	0.424	0.649	0.779	0.031	11.08
Motion Level	2	**0.431**	0.623	0.745	1.196	10.66
Action Level	4	0.413	**0.644**	**0.770**	0.396	10.85
Specific Level	8	0.414	0.640	0.760	**0.361**	**10.86**

recent methods [1, 4, 18] converts joint rotation angles into $J = 21$ joint XYZ coordinates. Given the global trajectory t_{root} and $p_m \in \mathbb{R}^{3J}$, the preprocessed representation is formulated as:

$$x^i = \{p_m, t_{root}\}. \tag{E}$$

E.2 Metric Definitions

Following previous works, we use the following five metrics to measure the performance of the model. Note that global representations of motion and text descriptions are first extracted with the pre-trained network in [5].

R-Precision. Under the feature space of the pre-trained network in [5], given one motion sequence and 32 text descriptions (1 ground-truth and 31 randomly selected mismatched descriptions), motion-retrieval precision calculates the text and motion Top 1/2/3 matching accuracy.

Frechet Inception Distance (FID). We measure the distribution distance between the generated and real motion using FID [6] on the extracted motion features [5]. The FID is calculated as:

$$\text{FID} = \|\mu_{gt} - \mu_{pred}\|^2 - \text{Tr}(\Sigma_{gt} + \Sigma_{pred} - 2(\Sigma_{gt}\Sigma_{pred})^{\frac{1}{2}}), \tag{F}$$

where Σ is the covariance matrix. Tr denotes the trace of a matrix. μ_{gt} and μ_{pred} are the mean of ground-truth motion features and generated motion features.

Multimodal Distance (MM-Dist). Given N randomly generated samples, we calculate the average Euclidean distances between each text feature f_t and the generated motion feature f_m from that text. The multimodal distance is calculated as:

$$\text{MM-Dist} = \frac{1}{N}\sum_{i=1}^{N} \|f_{t,i} - f_{m,i}\|, \tag{G}$$

where $f_{t,i}$ and $f_{m,i}$ are the features of the i_{th} text-motion pair.

Diversity. All generated motions are randomly sampled to two subsets ($\{x_1, x_2, ..., x_{X_d}\}$ and $\{x_1^{'}, x_2^{'}, ..., x_{X_d}^{'}\}$) of the same size X_d. Then, we extract motion features [5] and compute the average Euclidean distances between the two subsets:

$$\text{Diversity} = \frac{1}{X_d}\sum_{i=1}^{X_d} \|x_i - x_i^{'}\|. \tag{H}$$

Multimodality (MModality). We randomly sample a set of text descriptions with size J_m from all descriptions. For each text description, we generate $2 \times X_m$ motion sequences, forming X_m pairs of motions. We extract motion features and calculate the average Euclidean distance between each pair. We report the average of all text descriptions. We define features of the j_{th} pair of the i_{th} text description as $(x_{j,i}, x^{'}_{j,i})$. The multimodality is calculated as:

$$\text{MModality} = \frac{1}{J_m \times X_m} \sum_{j=1}^{J_m} \sum_{i=1}^{X_m} \|x_{j,i} - x^{'}_{j,i}\|. \tag{I}$$

References

[1] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *3DV*, pages 719–728, 2019.

[2] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, pages 10638–10647, 2020.

[3] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, pages 18000–18010, 2023.

[4] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *ICCV*, pages 1396–1406, 2021.

[5] 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.

[6] 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. In *NeurIPS*, 2017.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020.

[8] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[9] Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-language representations. In *NeurIPS*, pages 30291–30306, 2022.

[10] Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *CVPR*, pages 2472–2482, 2023.

[11] Peng Jin, Hao Li, Zesen Cheng, Jinfa Huang, Zhennan Wang, Li Yuan, Chang Liu, and Jie Chen. Text-video retrieval with disentangled conceptualization and set-to-set alignment. In *IJCAI*, pages 938–946, 8 2023.

[12] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, and Jie Chen. Diffusionret: Generative text-video retrieval with diffusion model. In *ICCV*, pages 2470–2481, 2023.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[15] Hao Li, Peng Jin, Zesen Cheng, Songyang Zhang, Kai Chen, Zhennan Wang, Chang Liu, and Jie Chen. Tg-vqa: Ternary game of video question answering. *arXiv preprint arXiv:2305.10049*, 2023.

[16] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015.

[17] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2017.

[18] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *ECCV*, pages 480–497, 2022.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[21] Peng Shi and Jimmy Lin. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*, 2019.

[22] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[23] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014.

[24] Ömer Terlemez, Stefan Ulbrich, Christian Mandery, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. Master motor map (mmm)—framework and toolkit for capturing, representing, and reproducing human motion on humanoid robots. In *2014 IEEE-RAS International Conference on Humanoid Robots*, pages 894–901, 2014.

[25] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In *ICLR*, 2023.

[26] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023.

[27] 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.