– Supplementary Materials –
FineMoGen: Fine-Grained Spatio-Temporal Motion Generation and Editing

Mingyuan Zhang[1] **Huirong Li**[1] **Zhongang Cai**[1,2] **Jiawei Ren**[1] **Lei Yang**[2] **Ziwei Liu**[1]

[1] S-Lab, Nanyang Technological University
[2] Sensetime, China

Contents

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

A Evaluation Metrics

The following quantitative assessment metrics, utilized in MotionDiffuse [3], are adopted again for evaluation: Frechet Inception Distance (FID), R-Precision, Diversity, Multimodality, and Multi-Modal Distance.

(1) FID serves as a quantitative metric to compute the distance between feature representations of real and generated motion sequences, effectively measuring generation quality.

(2) R-Precision examines the correspondence between text descriptions and generated motion sequences, indicating the likelihood of the actual text appearing within the top k rankings after ordering. For each generated motion, its ground-truth text description is combined with 31 randomly chosen mismatched descriptions from the test set, creating a description pool. The Euclidean distances between the motion feature and text feature for each description in the pool are then calculated and ranked, with average accuracy determined at top-1, top-2, and top-3 positions.

(3) Diversity assesses the variability and richness of generated action sequences by comparing two randomly sampled subsets of equal size S_d, conditioned on different descriptions, which can be defined as follows [2]:

$$Diversity = \frac{1}{S_d}\sum_{i=1}^{S_d} \left\| \mathbf{v}_i - \mathbf{v'}_i \right\|, \tag{1}$$

where $\mathbf{v}_i$ and $\mathbf{v'}_i$, $i = 1, 2, \ldots, S_d$, are corresponding motion features of these two subsets.

(4) Multimodality gauges the mean fluctuation of generated motion sequences in response to a single text description. Given a set of motions belonging to C descriptions, two subsets of equal size S_m are randomly sampled for each description, and multimodality is defined accordingly [2]:

$$Multimodality = \frac{1}{S_m \times C}\sum_{c=1}^{C}\sum_{i=1}^{S_m} \left\| \mathbf{v}_{c,i} - \mathbf{v'}_{c,i} \right\|, \tag{2}$$

where $\mathbf{v}_{c,i}$ and $\mathbf{v'}_{c,i}$, $i = 1, 2, \ldots, S_m$, are corresponding motion features of these two subsets.

(5) Multi-Modal Distance (MM Dist) quantifies the average Euclidean distance between motion feature representations and their corresponding text description features.

B HuMMan-MoGen dataset

B.1 demographic composition

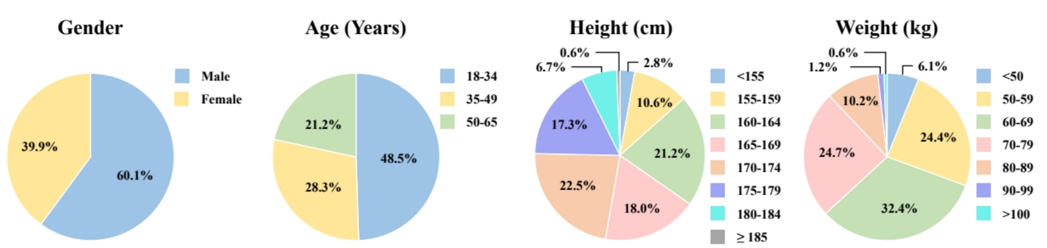

Figure 1: **The distribution of participants in HuMMan-MoGen**. This statistical chart is referenced from the HuMMan dataset [1].

Figure 1 shows the distribution of participants in our HuMMan-MoGen. The diversity inherent in the HuMMan dataset alleviates inductive bias.

B.2 License

Copyright 2022 S-Lab

Redistribution and use for non-commercial purpose in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
3. Neither the name of the copyright holder nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.
4. In the event that redistribution and/or use for commercial purpose in source or binary forms, with or without modification is required, please contact the contributor(s) of the work.
5. Prohibition of creating videos containing potentially offensive content, such as violence, pornography, defamation, and the like; Prohibition of generating fraudulent videos; Prohibition of inferring biometric information of participants.
6. If you use our data, you are required to agree: 'When a participant who provided data wishes to have data related to themselves removed, we will email you with the corresponding action numbers. You need to delete the relevant content and refrain from redistributing this data to others.'

B.3 Discussion of Potential Bias

The raw motion data used in our proposed HuMMan-MoGen originates from HuMMan and primarily consists of fitness-related actions. It inherently excludes violent and pornographic motions. However, due to the nature of fine-grained motion generation, these actions, when combined at a granular level, might inadvertently generate unexpected and undesirable sequences, posing a risk of misuse. We forbid this kind of abuse condition in the license above.

B.4 Discussion of Misuse

The first scenario involves utilizing the generated motion sequences to produce realistic-style videos. Since our algorithm only provides human skeletal motion data, achieving this effect would require integration with other image or video generation techniques. The second scenario entails combining high-precision 3D human models to render lifelike videos. However, our algorithm does not offer motion sequences for hand movements or changes in facial expressions. Creating such videos would necessitate the assistance of other algorithms. These two methods of generating fabricated videos are currently not fully developed, but there is still a certain risk of misuse. We forbid these kind of abuse condition in the license above.

C Quantitative Results of Spatial Composition

To measure the consistency between the generated motion sequence and the descriptions of each body part, we follow the approach proposed by Guo *et al*. [2]. We trained a separate comparison model for each body part.

Table 1: **Ablation study on HuMMan-MoGen test set.** All methods use zero-shot setting, it means that they are not trained on the spatial composition data. Here we report the average score from individual ones of seven different body parts.

Methods	Spatial Independence	Temporal Independence	MoE	R Precision↑	FID↓	Diversity →	MultiModality↑
Real motions	-	-	-	0.61	0.003	5.94	1.68
Baseline	-	-	-	0.43	2.87	5.85	5.39
	✓	-	-	0.49	2.04	5.75	5.25
	-	✓	-	0.41	3.56	**5.91**	**5.58**
	-	-	✓	0.45	2.41	5.81	5.32
FineMoGen	✓	✓	✓	**0.51**	**1.09**	5.71	5.17

Table 1 show the quantitative results of spatial composition. Spatial independence contribute a lot to the performance while temporal independence reduce it. This phenomenon is similar to the experimental results in temporal composition.

D Motion Editing

We use ChatGPT-4 to accept the natural instructions from users and edit the fine-grained descriptions accordingly. To create fine-grained spatio-temporal descriptions from users, the instruction is:

> *Can you help me create some motion sequences. I will give you a sentence about what I want to do. You should help me divide this action into several different stages. For each stage, you should tell me: 1) the specific action; 2) 7 detailed description about head, spine, left upper limb, right upper limb, left lower limb, right lower limb, and trajectory; 3) the lasting frames (30 frames per second)*

After the initialization, the users can edit it with natural instructions. We decorate their commands by:

> *Based on current description you provided, I want to modify it by the command "**#users' command#**". Please give me the modified description.*

References

[1] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. HuMMan: Multi-modal 4d human dataset for versatile sensing and modeling. In *17th European Conference on Computer Vision, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 557–577. Springer, 2022.

[2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022.

[3] 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.