

Supplementary Material

No MoCap Needed: Post-Training Motion Diffusion Models with Reinforcement Learning using Only Textual Prompts

Anonymous WACV Algorithms Track submission

Paper ID 2161

1. Datasets details for the Leave-one-out experiments

To create our held-out category datasets, we performed keyword-based filtering of the HumanML3D dataset using carefully selected terms that capture the semantic essence of each motion category. Our approach involved scanning all textual descriptions in the dataset and including any text-motion pair, where the description contained at least one keyword from the target category list.

For the *Object Manipulation* category, we used the following keywords:

pick, place, lift, carry, push, pull, throw, catch, press, wipe, clean, stack

For the *Posture and Balance* category, we used:

t-pose, sit, stand, squat, kneel, bend, stretch, lean, balance, stumble, trip, fall, handstand, bridge

These keywords were selected by analyzing all textual descriptions in the HumanML3D dataset [1], ensuring comprehensive coverage of the target motion categories, while minimizing overlap between them. This keyword-based approach enables automated dataset construction and captures all relevant motions, including those described with varied linguistic expressions of the same underlying action.

The filtering process resulted in 3,194 text-motion pairs for *Object Manipulation* (approximately 23% of HumanML3D) and 4,384 pairs for *Posture and Balance* (approximately 31% of HumanML3D). Some examples of the included prompts are shown in tables 1 and 2.

Table 1. Example prompts included in our *Object Manipulation* dataset. These examples demonstrate the diversity of object interaction scenarios captured in this split of the dataset.

Example Prompts from the *Object Manipulation* Dataset

Prompt 1: “A person holds something with their hand and then they throw it.”
 Prompt 2: “A person is catching something and then throwing it back with his left hand.”
 Prompt 3: “The person lifted a weight with his right hand.”
 Prompt 4: “A person throws something with his left hand and receives something with his two hands.”

Table 2. Example prompts included in our *Posture and Balance* dataset. These examples showcase the range of static postures and balance-related movements captured in this category.

Example Prompts from the *Posture and Balance* Dataset

Prompt 1: “A person readies himself to do a handstand and fails.”
 Prompt 2: “A person stumbles forward and back almost falling over.”
 Prompt 3: “A person, slightly squatting with arms stretched out to the sides, lowers their arms, pauses and then raises their arms to the same stretched out position.”
 Prompt 4: “A person steps to left sits down and then stands up to return to first position.”

Table 3. Ablation study on forgetting after fine-tuning. We evaluate models pretrained on KIT-ML and fine-tuned on HumanML3D, reporting results on the KIT-ML test set to assess the impact of fine-tuning on the original distribution. Results demonstrate consistent preservation of performance on the original dataset, with some metrics showing improvements.

Method	R@1 ↑	R@2 ↑	R@3 ↑	FID ↓	MMDist ↓	Diversity →	MModality ↑
Ground Truth	0.401	0.601	0.730	0	2.636	9.103	-
MoMask	0.433	0.656	0.781	0.204	2.779	-	1.131
MotionGPT	0.366	0.558	0.680	0.510	3.096	10.35	2.328
StableMoFusion	0.384	0.589	0.707	0.724	3.021	8.666	0.852
MDM-SMPL	0.215	0.375	0.467	0.593	3.339	8.832	1.269
StableMoFusion (Ours)	0.378	0.589	0.713	0.666	3.010	8.798	0.844
MDM-SMPL (Ours)	0.221	0.378	0.471	0.554	3.324	8.654	1.242

2. Additional Forgetting Analysis

To provide a comprehensive assessment of our method’s impact on the original data distribution, we present additional forgetting analysis for the Kit-to-Human experimental setting. Table 3 reports results for models pretrained on KIT-ML [6] and fine-tuned on HumanML3D, then evaluated on the original KIT-ML test set.

The results in Table 3 corroborate our findings from the Human-to-Kit setting, demonstrating that our RL fine-tuning approach does not cause catastrophic forgetting. For StableMoFusion [4], we observe stable performance across all metrics, with R@3 showing a slight improvement (0.707 → 0.713) and FID improving from 0.724 to 0.666 (-8.0%). Similarly, MDM-SMPL maintains comparable retrieval performance while achieving better FID (0.593 → 0.554, -6.6%).

Comparing both forgetting experiments, we observe a consistent pattern: our method not only preserves the original model capabilities but often leads to modest improvements in generation quality. This suggests that the exposure to different motion styles and textual descriptions during RL fine-tuning acts as a form of regularization, enhancing the model’s understanding of motion-text relationships across domains.

The absence of catastrophic forgetting is particularly noteworthy given that our fine-tuning uses only textual prompts without access to ground-truth motions from the target domain. This property makes our approach highly practical for real-world deployment, where maintaining performance on existing capabilities while adding new ones is crucial for system reliability.

3. Additional Qualitative Comparison

Figure 1 shows additional qualitative results comparing generations from our model against the baseline methods MDM-SMPL [5], MoMask [2], MotionGPT [3], and StableMoFusion [4].

References

- [1] 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 (CVPR)*, pages 5152–5161, 2022. 1
- [2] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions, 2023. 2
- [3] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language, 2023. 2
- [4] Sanghoon Kim, Jaehun Park, Hyunwoo Kim, and Junho Cho. Stablemotionfusion: Text-driven human motion synthesis via diffusion models. In *Computer Graphics Forum*, 2022. 2
- [5] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J. Black, Gül Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *CVPR Workshop on Human Motion Generation*, 2024. 2
- [6] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, 2016. 2

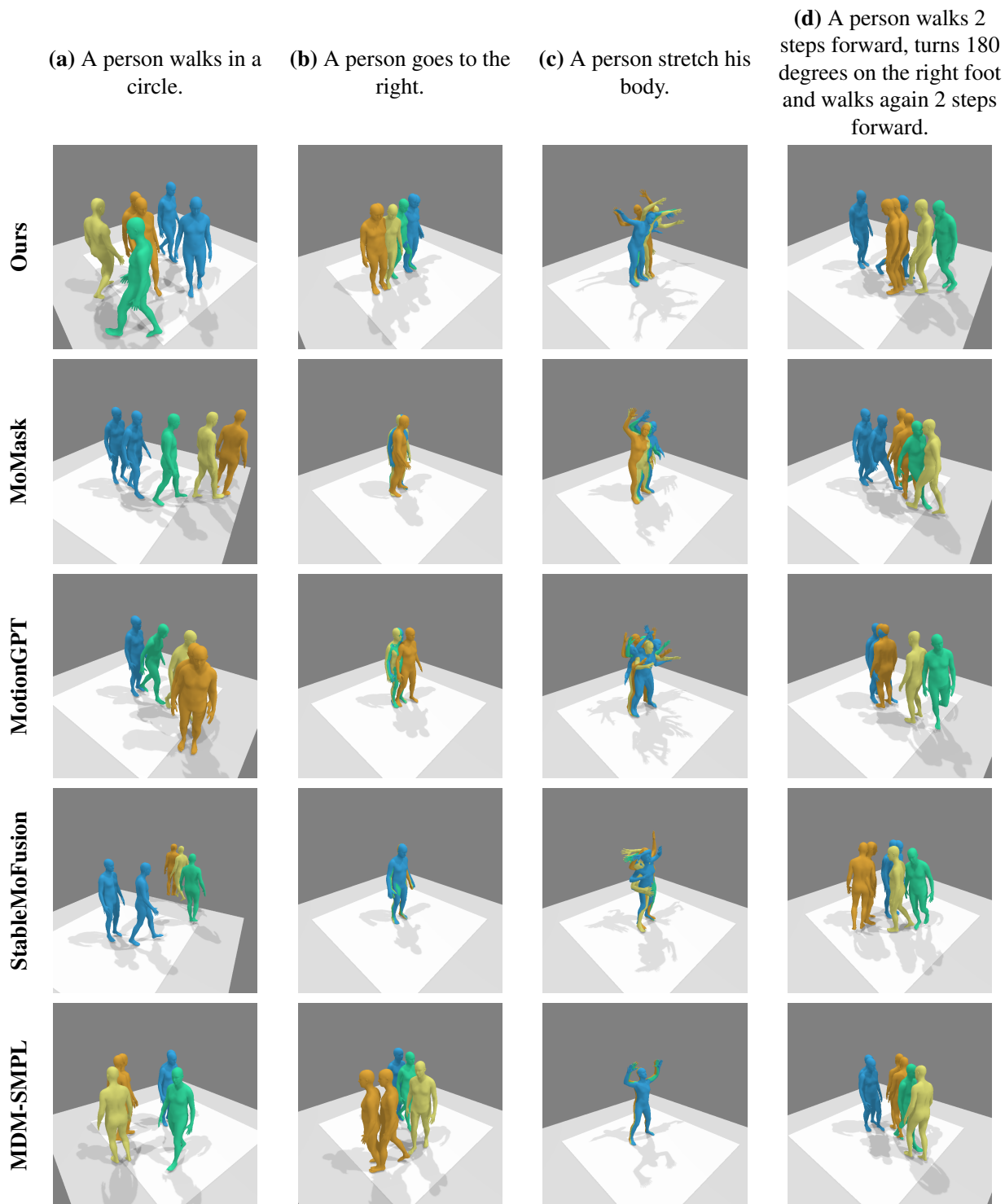


Figure 1. Example of generation of our fine-tuned model compared against other baselines.