MotionLCM: Real-time Controllable Motion Generation via Latent Consistency Model

Supplementary Material

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

- Appendix A: Additional experiments.
- Appendix B: Supplementary quantitative results.
- Appendix C: Details of the evaluation metrics.

A Additional Experiments

A.1 Comparison to other ODE Solvers

To validate the effectiveness of latent consistency distillation, we compare three ODE solvers (DDIM [8], DPM [6], DPM++ [7]). The quantitative results shown in Tab. 1 demonstrate that our MotionLCM notably outperforms baseline methods. Moreover, unlike DDIM [8], DPM [6], and DPM++ [7], requiring more peak memory per sampling step when using CFG [5], MotionLCM only requires one forward pass, saving both time and memory costs.

Table 1: Quantitative results with the testing CFG scale $w = 7.5$. MotionLCM notably outperforms baseline methods [6–8] on HumanML3D [2] dataset, demonstrating the effectiveness of latent consistency distillation. **Bold** indicates the best result.

Methods	R-Precision (Top 3) ↑			FID ↓		
	1-Step	2-Step	4-Step	1-Step	2-Step	4-Step
DDIM [8]	$0.651^{\pm.003}$	$0.691^{\pm.002}$	$0.765^{\pm.003}$	$4.022^{\pm.043}$	$2.802^{\pm.038}$	$0.966^{\pm.018}$
DPM [6]	$0.651^{\pm.003}$	$0.691^{\pm.002}$	$0.777^{\pm.002}$	$4.022^{\pm.043}$	$2.798^{\pm.038}$	$0.727^{\pm.015}$
DPM++ [7]	$0.651^{\pm.003}$	$0.691^{\pm.002}$	$0.777^{\pm.002}$	$4.022^{\pm.043}$	$2.798^{\pm.038}$	$0.684^{\pm.015}$
MotionLCM	$\mathbf{0.803}^{\pm.002}$	$\mathbf{0.805}^{\pm.002}$	$\mathbf{0.798}^{\pm.002}$	$\mathbf{0.467}^{\pm.012}$	$\mathbf{0.368}^{\pm.011}$	$\mathbf{0.304}^{\pm.012}$

A.2 Impact of different testing CFGs

As shown in Fig. 1, we provide an extensive ablation study on the testing CFG [5]. It can be observed that, under different testing CFGs, increasing the number of inference steps continuously improves the performance. However, further increasing the inference steps results in comparable performance but significantly increases the time cost.

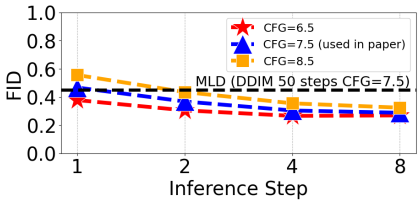

Fig. 1: Comparison of testing CFGs

B More Qualitative Results

In this section, we provide more qualitative results of our MotionLCM. Fig. 2 presents more generation results on the text-to-motion task. Fig. 3 displays additional visualization results on the motion control task. All videos shown in the figures can be found in the supplementary video (*i.e.*, *supp.mp4*).

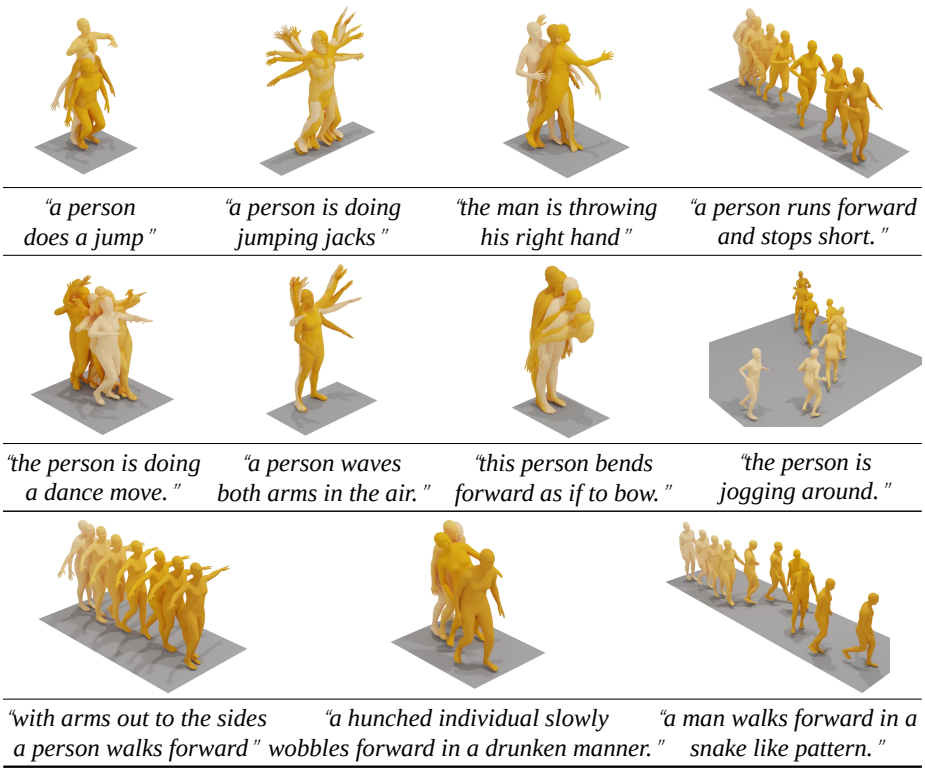

Fig. 2: More qualitative results of MotionLCM on the text-to-motion task.

C Metric Definitions

Time cost: To assess the inference efficiency of models, we follow [1] to report the Average Inference Time per Sentence (AITS) measured in seconds. We calculate AITS on the test set of HumanML3D [2] by setting the batch size to 1 and excluding the time cost for model and dataset loading parts.
Motion quality: Frechet Inception Distance (FID) measures the distributional difference between the generated and real motions, calculated using the feature extractor associated with a specific dataset, *e.g.*, HumanML3D [2].

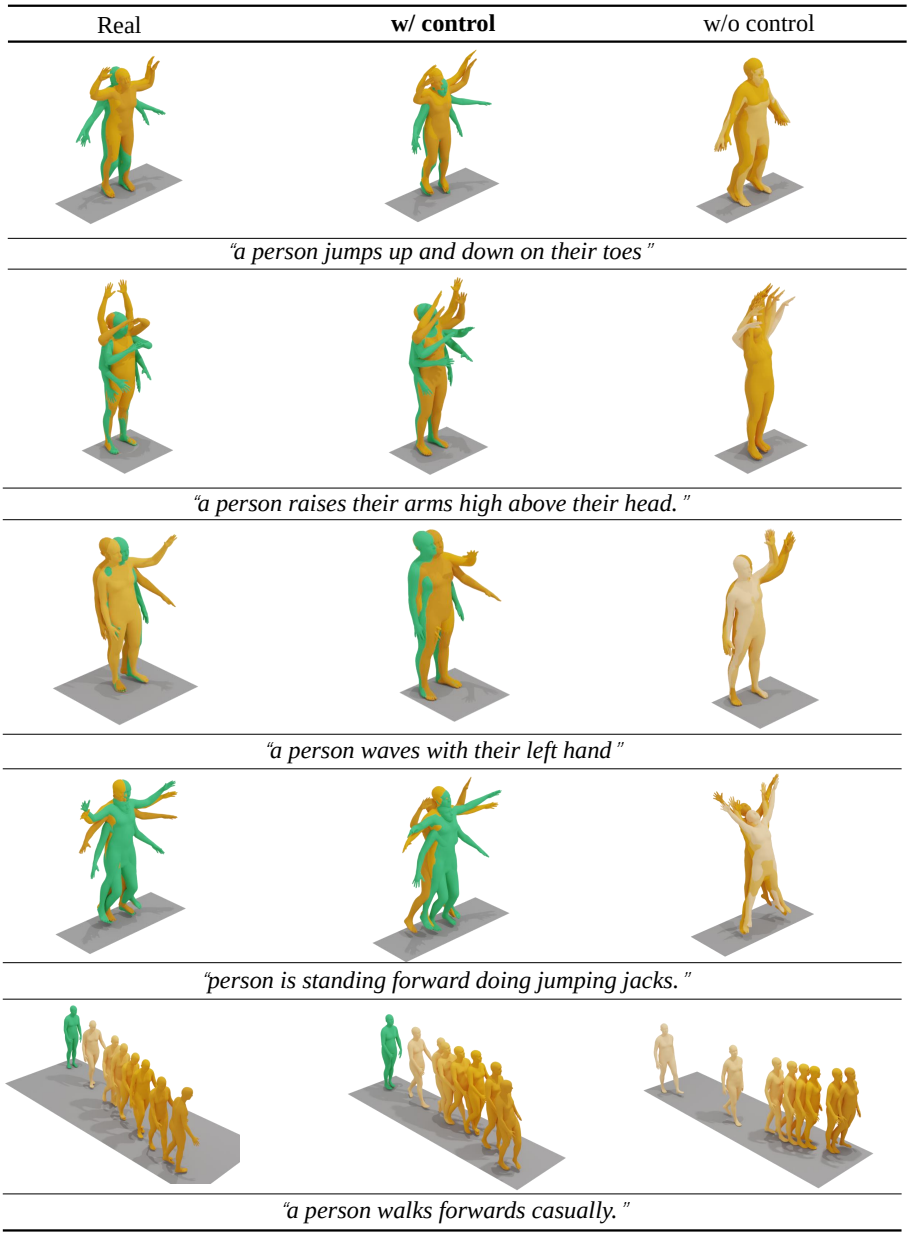

Fig. 3: More qualitative results of MotionLCM on the motion control task.

Motion diversity: Following [3, 4], we report Diversity and MultiModality to evaluate the generated motion diversity. Diversity measures the variance of the generated motions across the whole set. Specifically, two subsets of the same size

S_d are randomly sampled from all generated motions with their extracted motion feature vectors $\{\mathbf{v}_1, ..., \mathbf{v}_{S_d}\}$ and $\{\mathbf{v}_1^{'}, ..., \mathbf{v}_{S_d}^{'}\}$. Diversity is defined as follows,

$$\text{Diversity} = \frac{1}{S_d} \sum_{i=1}^{S_d} ||\mathbf{v}_i - \mathbf{v}_i^{'}||_2. \tag{1}$$

Different from Diversity, MultiModality (MModality) measures how much the generated motions diversify within each textual description. Specifically, a set of textual descriptions with size C is randomly sampled from all descriptions. Then we randomly sample two subsets with the same size I from all generated motions conditioned by the c-th textual description, with extracted feature vectors $\{\mathbf{v}_{c,1}, ..., \mathbf{v}_{c,I}\}$ and $\{\mathbf{v}_{c,1}^{'}, ..., \mathbf{v}_{c,I}^{'}\}$. MModality is formalized as follows,

$$\text{MModality} = \frac{1}{C \times I} \sum_{c=1}^{C} \sum_{i=1}^{I} ||\mathbf{v}_{c,i} - \mathbf{v}_{c,i}^{'}||_2. \tag{2}$$

Condition matching: [2] provides motion/text feature extractors to generate geometrically closed features for matched text-motion pairs and vice versa. Under this feature space, evaluating motion-retrieval precision (R-Precision) involves mixing the generated motion with 31 mismatched motions and then calculating the text-motion Top-1/2/3 matching accuracy. Multimodal Distance (MM Dist) calculates the mean distance between the generated motions and texts.
Control error: Following [9], we report Trajectory error, Location error, and Average error to assess the motion control performance. Trajectory error (Traj. err.) is defined as the proportion of unsuccessful trajectories, *i.e.*, if a control joint in the generated motion exceeds a certain distance threshold from the corresponding joint in the given control trajectory, it is considered a failed trajectory. Similar to the Trajectory error, Location error (Loc. err.) is defined as the ratio of unsuccessful joints. In our experiments, we adopt 50cm as the distance threshold to calculate the Trajectory error and Location error. Average error (Avg. err.) denotes the mean distance between the control joint positions in the generated motion and those in the given control trajectory.

References

1. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: CVPR. pp. 18000–18010 (2023)
2. 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)
3. 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)
4. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: ACMMM. pp. 2021–2029 (2020)

5. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
6. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. NeurIPS pp. 5775–5787 (2022)
7. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. arXiv preprint arXiv:2211.01095 (2022)
8. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
9. Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation. In: ICLR (2024)