

MotionHiFlow: Text-to-Motion via Hierarchical Flow Matching

Supplementary Material

In this supplementary material, we include additional experimental results and training details: 1) We validate the necessity of designed cross-scale transition in Section A; 2) We conduct ablation on TMDiT in Section B; 3) We include additional implementation details and comparisons of our topology-aware Motion VAE in Section C; 4) We present the study of inference hyper-parameters in Section D; Finally, 5) we provide more visualization results in Section F.

A. Necessity of Designed Cross-Scale Transition

In our framework, we design the cross-scale transition as a *denoise-upsample-renoise* cascade. To validate the effectiveness of this design, we conduct quantitative comparisons against two alternatives: 1) direct linear interpolation for up-sampling (*upsample*), and 2) linear interpolation for up-sampling, and followed by a re-noising step as in [5] (*upsample-renoise*). As shown in Table A1, our denoise-upsample-renoise transition, which preserves noise consistency and ensures smooth transitions across scales, consistently outperforms both alternatives in FID (generation quality) and MM-Dist (text-alignment) metrics.

Table A1. Ablation on the cross-scale transition.

Method	FID↓	MM-Dist↓
upsample	0.304	3.033
upsample-renoise	0.057	2.723
denoise-upsample-renoise (Ours)	0.032	2.691

B. Ablation on TMDiT

Our Text-Motion Diffusion Transformer (TMDiT), with the word-level text encoding and non-shared parameter strategy, significantly improves the generation performance of the Baseline method. In this section, we start from the Baseline and examine the effectiveness of each element. By leveraging word-level text encoding, the text-alignment (MM-Dist) performance improves from 3.043 to 2.849 and generation quality (FID) performance improves from 0.074 to 0.066. To further investigate the effect of the non-shared parameter strategy, we conduct experiments on the number of shared/non-shared layers, and report the results in Table A2. We observe that: 1) Our strategy outperforms the shared strategy, with 6 shared blocks achieving the best system performance; 2) when the number of shared layers is too large, the model struggles to simultaneously capture modality-specific features, due to the significant cross-modality gap between text and motion data; 3) when the number of shared layers is too small, the model fails to learn the shared representations

Table A2. Ablation on the non-shared parameter strategy in TMDiT. The x - d - y s represents x double-stream blocks with unshared parameters, followed by y single-stream blocks with shared parameters in the architecture.

Parameter Strategy	Setting	FID ↓	MM-Dist ↓
shared	0d-9s	0.066	2.849
non-shared	1d-8s	0.064	2.786
	2d-7s	0.051	2.741
	3d-6s	0.045	2.738
	4d-5s	0.047	2.739

necessary for effective cross-modal alignment, leading to suboptimal generation performance.

C. Analysis on topology-aware Motion VAE

In this section, we provide additional training details for our topology-aware motion VAE and compare it with existing motion tokenizers. We train our topology-aware Motion VAE using the same loss functions and hyperparameter settings as in [2, 6, 7]. Specifically, the loss function of our motion VAE is formulated as: $\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{aug}}\mathcal{L}_{\text{aug}}$. Here, \mathcal{L}_{aug} is defined in Equation (11). Following MoMask [2], the reconstruction loss is defined as: $\mathcal{L}_{\text{rec}} = |M - \hat{M}|_1 + \lambda_{\text{vel}}|M_{\text{vel}} - \hat{M}_{\text{vel}}|_1 + \lambda_{\text{pos}}|M_{\text{pos}} - \hat{M}_{\text{pos}}|_1$, where $\hat{M} = \mathcal{D}(\mathcal{E}(M))$ represents the reconstructed human motion, M_{vel} and M_{pos} denote the velocity and position components of the motion, respectively. The weighting λ_{KL} , λ_{aug} , λ_{vel} and λ_{pos} are empirically set to 0.02, 0.5, 0.5, and 0.5, respectively.

Quantitative Comparisons. We further compare our topology-aware Motion VAE with existing motion tokenizers, including T2M-GPT [8], MotionGPT [4] and Momask [2]. Specifically, we replace our proposed topology-aware Motion VAE with theirs, and report the reconstruction (in the r-FID and MPJPE metrics) and generation (in the g-FID and MM-Dist metric) performance in Table A3. As shown, our proposed Motion VAE consistently outperforms all other methods in terms of both reconstruction and generation.

Table A3. Quantitative comparison with existing motion tokenizers.

Method	r-FID↓	MPJPE↓	g-FID↓	MMDist↓
T2M-GPT [8]	0.070	58.0	0.141	3.121
MotionGPT [4]	0.067	55.8	0.232	3.096
Momask [2]	0.019	29.5	0.045	2.958
Ours ($d = 8$)	0.022	33.2	0.083	2.688
Ours ($d = 16$)	0.007	18.3	0.032	2.691

D. Study on Inference Hyper-parameters

During inference, two critical hyperparameters significantly impact performance: the Classifier-Free Guidance (CFG) [3] scale and the number of inference steps. As in Figure A1, system performance consistently improves as the number of inference steps increases, and stabilizes around 12 steps. For the CFG scale, as shown in Figure A2, the system achieves the best performance when CFG scale is set to 4.5. Either too small or too large scale leads a performance degradation.

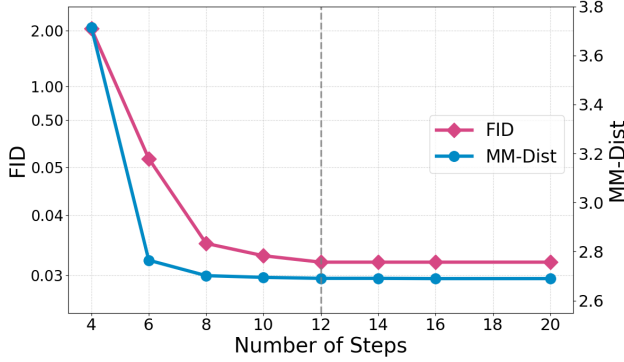


Figure A1. Analysis on different numbers of inference steps. Here, FID scores are reported in log scale.

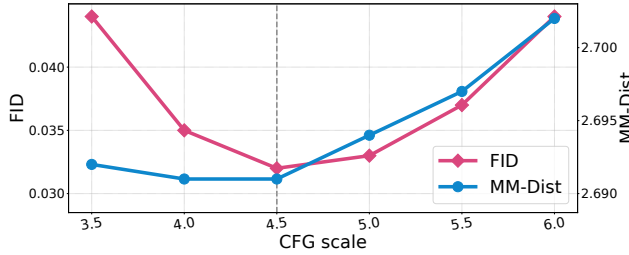


Figure A2. Analysis of Classifier-Free Guidance scales on the HumanML3D test set. The system achieves the best performance when guidance scale is set to 4.5.

E. Efficiency of inference

To assess computational efficiency, we compare the inference time of our MotionHiFlow against various methods on the HumanML3D [1] test set. As demonstrated in Table A4, our architecture exhibits superior efficiency compared to the inefficient spatial encoding method (MoGenTS) and autoregressive method (BAMM). By integrating hierarchical flow matching and performing inference at reduced scales in the early stages (using multi-scale scales $\{r_k\} = [1/3, 2/3, 1.0]$), our approach further achieves the best results while keeping lowest computational time costs.

Table A4. Comparison of Average Inference Time per Sentence (AITS, in seconds) across various models on the HumanML3D test set. Our proposed method achieves the best performance while maintaining competitive computational efficiency.

Method	AITS (second)	FID ↓	MM-Dist ↓
MoGenTS	0.75	0.033	2.867
BAMM	1.32	0.055	2.919
Ours (w/o hierarchical)	0.40	0.051	2.723
Ours (w/ hierarchical)	0.31	0.032	2.691

F. More visualizations

In Figure A3, we provide some qualitative results of our MotionHiFlow on the HumanML3D dataset [1]. As shown, our method can effectively produce text-aligned and realistic motions. For example, in the 2nd column of the 1st row, our model accurately generates actions "jump", "slow down", and "walk", while maintaining temporal coherence across the entire motion sequence. In the 1st column of the 4th row, our model also captures a variety of complex motions (e.g., squat, pick with both hands) and generates smooth and realistic human motion. These demonstrate the effectiveness of our MotionHiFlow framework visually. For more visualization videos, please refer to [visualize.html](#).

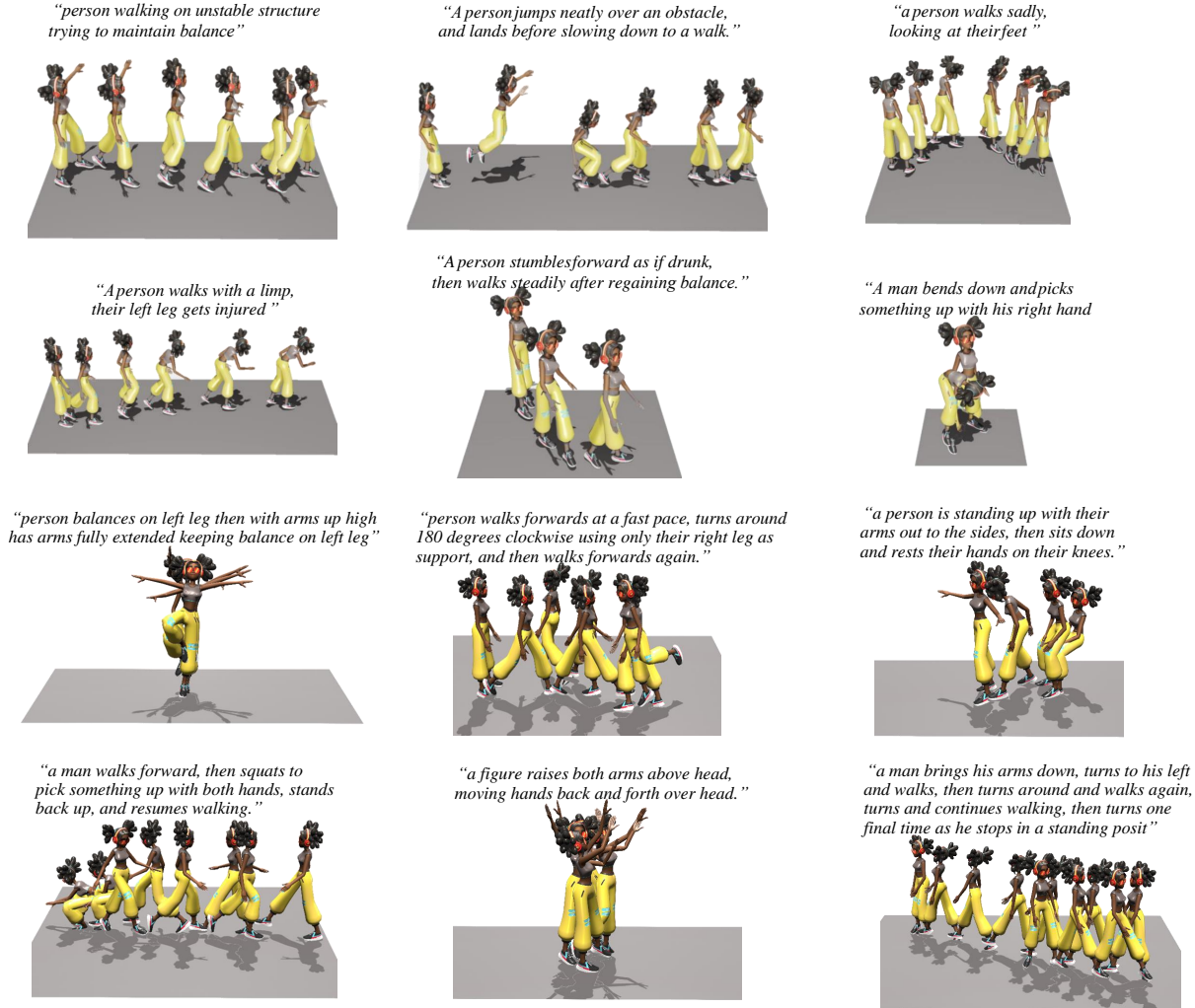


Figure A3. Visualization of human motion generated by MotionHiFlow. Our MotionHiFlow can generate text-aligned and realistic human motion.

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 *CVPR*, pages 5152–5161, 2022. 2
- [2] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 1900–1910. IEEE, 2024. 1
- [3] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. 2
- [4] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *NeurIPS*, 36:20067–20079, 2023. 1
- [5] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong MU, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. 1
- [6] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. BAMB: bidirectional autoregressive motion model. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XV*, pages 172–190. Springer, 2024. 1
- [7] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. MMM: generative masked motion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 1546–1555. IEEE, 2024. 1
- [8] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Y ong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *CVPR*, pages 14730–14740, 2023. 1