

Motion Mamba Supplementary

1 Implementation Details

Motion Mamba operates within the latent spaces, leveraging the capabilities of the Motion Variational AutoEncoder (VAE) $\mathcal{V} = \{\mathcal{E}, \mathcal{D}\}$, as proposed in the seminal work by Chen et al. [6]. For the configuration of the Motion Mamba denoiser ϵ_θ , we have opted for an architecture comprising 11 layers ($N = 11$), with the latent dimensionality set to $z \in \mathbb{R}^{2,d}$. The Hierarchical Temporal Mamba (HTM) modules are arranged in a scan pattern of $\{S^{2N_n-1}, \dots, S^1\}$, while the Bidirectional Spatial (BSH) modules incorporate a block-level bidirectional scan policy. Additionally, we utilize a pretrained *CLIP-VIT-L-14* model in a frozen state to derive text embeddings $\tau_\theta^w(w^{1:N}) \in \mathbb{R}^{1 \times d}$.

All models under the Motion Mamba framework are meticulously trained using the AdamW Optimizer, with the learning rate steadfastly maintained at 10^{-4} . We have standardized our global batch size at 512, which is judiciously distributed across 4 GPUs to facilitate data-parallel training. The training regime is extended over 2,000 epochs to ensure convergence to an optimal set of parameters. For the diffusion sampling process, we maintain the number of steps at 1,000 and 50 during the training and inference phases, respectively. The entire training procedure is executed on a single-node GPU server, outfitted with 4 NVIDIA A100 GPUs, spanning approximately 4 hours. Inference speed evaluations of our Motion Mamba models are conducted on a single NVIDIA V100 GPU for fair comparison, while module development and additional inference tasks are performed on a single NVIDIA GeForce RTX 3090/4090 GPU.

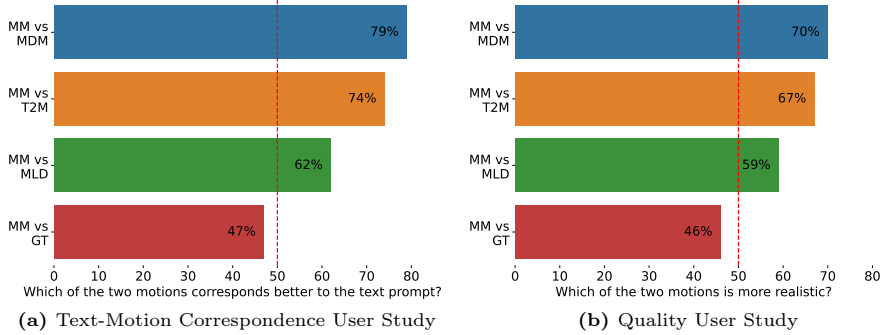


Fig. 1: User Study in two aspects including text-motion correspondence and quality, we compare *Motion Mamba* (MM) with previous methods including MDM [49], T2M [17], MLD [6] and ground truth.

2 User Study

In this work, we undertake a comprehensive evaluation of *Motion Mamba*’s performance, encompassing both qualitative analyses across various datasets and a user study to assess its real-world applicability. A diverse collection of 20 motion sequence sets, prompted randomly and extracted from the HumanML3D [17] test set, were generated utilizing three distinct methodologies—MDM [49], T2M [17], MLD [6]—alongside *Motion Mamba* and a baseline of ground truth motions. Subsequently, 50 participants were randomly selected to evaluate the motion sequences generated by these methods.

The user study was administered through a Google Forms interface, as depicted in Fig. 4, ensuring that motion sequences were presented anonymously without revealing their generative model origins. Our analysis focused on two critical dimensions: the fidelity of text-to-motion correspondence and the overall quality of the generated motions.

Empirical results, illustrated in Fig. 1a and Fig. 1b, unequivocally demonstrate *Motion Mamba*’s superior performance relative to the benchmark methods in terms of both text-motion alignment and motion quality. Specifically, *Motion Mamba* achieved significant margins over MDM [49], T2M [17], and MLD [6] by 79%, 74%, and 62% in text-motion correspondence, respectively, as highlighted in Fig. 1a. When juxtaposed with ground truth data—meticulously captured with state-of-the-art, noise-free devices—*Motion Mamba*’s generated sequences exhibited a remarkably close adherence to the intended text descriptions, underscoring its proficiency in aligning textual prompts with motion sequences.

Further reinforcing these findings, *Motion Mamba*’s generated motions were also found to surpass the aforementioned methods by substantial margins of 70%, 67%, and 59%, respectively, in terms of quality, as reported in Fig. 1b. This underscores *Motion Mamba*’s ability to not only closely match the text-motion correspondence of high-fidelity ground truth data but also to produce high-quality motion sequences that resonate well with real user experiences.

3 Visualization

Our study delves into the visualization of motion generation by capturing intricate motion sequences, utilizing prompts and their variations derived from HumanML3D [17]. We meticulously compare our proposed Motion Mamba methodology with established state-of-the-art techniques, namely MotionDiffuse [54], MDM [49], and MLD [6]. Presenting three distinct motion sequences, we meticulously analyze and visualize each, offering a comprehensive assessment of our approach’s efficacy.













Method	The person walks up the corner stairs.	The person first walks forward then walks backward.	The person is walking in a semi-circle and in clockwise direction.
Ours			
MLD			
MDM			
MD			

Fig. 2: We compared the proposed Motion Mamba with well-established state-of-the-art methods such as MotionDiffuse [54], MDM [49], and MLD [6]. We presented three distinct motion prompts and visualized them in the form of motion sequence. The results demonstrated our superior performance compared to existing methods.

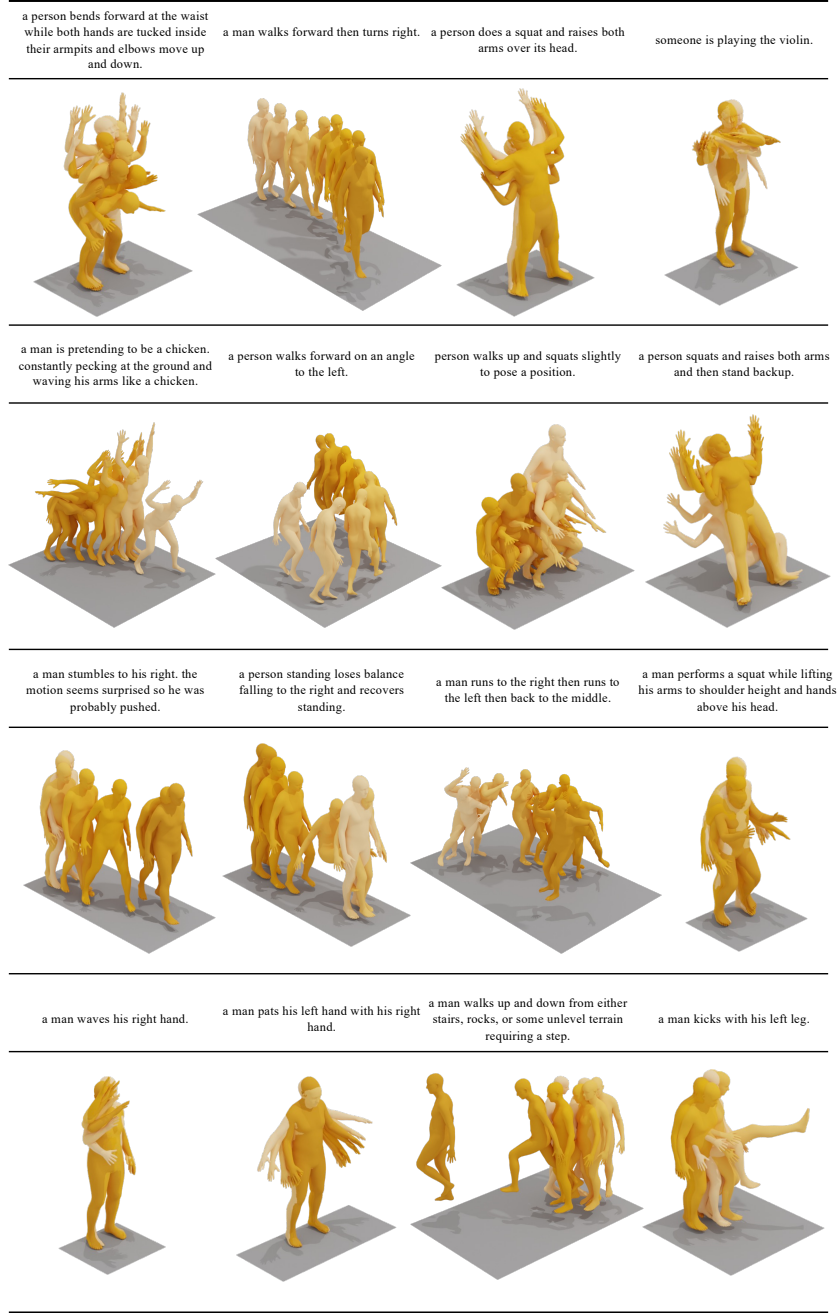


Fig. 3: We have included extra examples to showcase the proposed Motion Mamba model. These examples feature randomly selected prompts sourced from HumanML3D [17], providing additional visualizations of the model’s capabilities.

Motion Mamba User Study

* Indicates required question

Video A

mm_sample_2

Text Prompt: The character punches with his right hand and kicks with his left foot simultaneously.

Video B

mm_example_1

Text Prompt: The character punches with his right hand and kicks with his left foot simultaneously.

which of the two motions corresponds better to the text prompt? *

☐ A

☐ B

Which of the two motions is more realistic? *

☐ A

☐ B

Submit [Clear form](#)

This content is neither created nor endorsed by Google. [Report Abuse](#) [Terms of Service](#) [Privacy Policy](#)

Google Forms

Fig. 4: This figure presents the User Interface (UI) deployed for our User Study, wherein participants are presented with two videos, labeled as Video A and Video B, respectively. These videos are selected randomly from a pool consisting of outputs generated by three distinct methods, in addition to the Ground Truth (GT) for comparison. Participants are posed with two types of evaluative questions to gauge the effectiveness of the generated motions. The first question, "Which of the two motions is more realistic?", aims to assess the overall quality and realism of the motion capture. The second question, "Which of the two motions corresponds more accurately to the text prompt?", is designed to evaluate the congruence between the generated motion and the provided text prompt. This dual-question approach facilitates a comprehensive assessment of both the quality of the motion generation and its fidelity to the specified text prompts.

References

1. Ahuja, C., Morency, L.P.: Language2pose: Natural language grounded pose forecasting. In: 2019 International Conference on 3D Vision (3DV). pp. 719–728. IEEE (2019)
2. Bao, F., Li, C., Cao, Y., Zhu, J.: All are worth words: a vit backbone for score-based diffusion models. arXiv preprint arXiv:2209.12152 (2022)
3. Baron, E., Zimerman, I., Wolf, L.: 2-d ssm: A general spatial layer for visual transformers. arXiv preprint arXiv:2306.06635 (2023)
4. Barsoum, E., Kender, J., Liu, Z.: Hp-gan: Probabilistic 3d human motion prediction via gan. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 1418–1427 (2018)
5. Bhattacharya, U., Rewkowski, N., Banerjee, A., Guhan, P., Bera, A., Manocha, D.: Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In: 2021 IEEE virtual reality and 3D user interfaces (VR). pp. 1–10. IEEE (2021)
6. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18000–18010 (2023)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems* **34**, 8780–8794 (2021)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
9. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1396–1406 (2021)
10. Gong, K., Lian, D., Chang, H., Guo, C., Jiang, Z., Zuo, X., Mi, M.B., Wang, X.: Tm2d: Bimodality driven 3d dance generation via music-text integration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9942–9952 (2023)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. *Advances in neural information processing systems* **27** (2014)
12. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
13. Gu, A., Goel, K., Gupta, A., Ré, C.: On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems* **35**, 35971–35983 (2022)
14. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)
15. Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., Ré, C.: Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems* **34**, 572–585 (2021)
16. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. *Advances in neural information processing systems* **30** (2017)
17. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (2022)

18. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020)
19. Gupta, A., Gu, A., Berant, J.: Diagonal state spaces are as effective as structured state spaces. *Advances in Neural Information Processing Systems* **35**, 22982–22994 (2022)
20. Harvey, F.G., Yurick, M., Nowrouzezahrai, D., Pal, C.: Robust motion in-betweening. *ACM Transactions on Graphics (TOG)* **39**(4), 60–1 (2020)
21. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)
22. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* **33**, 6840–6851 (2020)
23. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences* **79**(8), 2554–2558 (1982)
24. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems* **36** (2024)
25. Kalman, R.E.: A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering* **82**(1), 35–45 (1960)
26. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. *stat* **1050**, 1 (2014)
27. Kramer, M.A.: Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal* **37**(2), 233–243 (1991)
28. Li, S., Singh, H., Grover, A.: Mamba-nd: Selective state space modeling for multi-dimensional data. *arXiv preprint arXiv:2402.05892* (2024)
29. Li, Y., Cai, T., Zhang, Y., Chen, D., Dey, D.: What makes convolutional models great on long sequence modeling? In: The Eleventh International Conference on Learning Representations (2022)
30. Lin, X., Amer, M.R.: Human motion modeling using dvkans. *arXiv preprint arXiv:1804.10652* (2018)
31. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166* (2024)
32. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
33. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5442–5451 (2019)
34. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019)
35. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3D human motion synthesis with transformer VAE. In: International Conference on Computer Vision (ICCV) (2021)
36. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision. pp. 480–497. Springer (2022)

37. Petrovich, M., Black, M.J., Varol, G.: TEMOS: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision (ECCV) (2022)
38. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. *Big data* **4**(4), 236–252 (2016)
39. Plappert, M., Mandery, C., Asfour, T.: Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems* **109**, 13–26 (2018)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
41. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
42. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
43. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. *Parallel Distributed Processing* pp. 318–362 (1986)
44. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022)
45. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems* **28** (2015)
46. Smith, J., De Mello, S., Kautz, J., Linderman, S., Byeon, W.: Convolutional state space models for long-range spatiotemporal modeling. *Advances in Neural Information Processing Systems* **36** (2024)
47. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
48. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: European Conference on Computer Vision. pp. 358–374. Springer (2022)
49. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2022)
50. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. *Advances in neural information processing systems* **30** (2017)
51. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
52. Wang, J., Yan, J.N., Gu, A., Rush, A.M.: Pretraining without attention. *arXiv preprint arXiv:2212.10544* (2022)
53. Yan, J.N., Gu, J., Rush, A.M.: Diffusion models without attention. *arXiv preprint arXiv:2311.18257* (2023)
54. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024)

- 55. Zhang, Y., Black, M.J., Tang, S.: We are more than our joints: Predicting how 3d bodies move. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3372–3382 (2021)
- 56. Zhong, C., Hu, L., Zhang, Z., Xia, S.: Att2m: Text-driven human motion generation with multi-perspective attention mechanism. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 509–519 (2023)
- 57. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)