Not Like Transformers: Drop the Beat Representation for Dance Generation with Mamba-Based Diffusion Model

Sangnjune Park[1] Inhyeok Choi[1] Donghyeon Soon[2] Youngwoo Jeon[1] Kyungdon Joo[1†]
[1]UNIST AIGS [2]DGIST CSE
{psj9116, inhyeok.choi, youngwoo.jeon, kyungdon}@unist.ac.kr
dhsoon@dgist.ac.kr

Abstract

Dance is a form of human motion characterized by emotional expression and communication, playing a role in various fields such as music, virtual reality, and content creation. Existing methods for dance generation often fail to adequately capture the inherently sequential, rhythmical, and music-synchronized characteristics of dance. In this paper, we propose a new dance generation approach that leverages a Mamba-based diffusion model. Mamba, specialized for handling long and autoregressive sequences, is integrated into our diffusion model as an alternative to the off-the-shelf Transformer. Additionally, considering the critical role of musical beats in dance choreography, we propose a Gaussian-based beat representation to explicitly guide the decoding of dance sequences. Experiments on AIST++ dataset show that our proposed method effectively reflects essential dance characteristics and advances performance compared to the state-of-the-art methods.

1. Introduction

Dance is a specified human motion that embodies rhythmic sequences of body poses, serving as an integral medium for social communication and artistic expression. Recently, generative AI models have enabled automated music-driven dance generation [9, 10, 10, 11, 13, 16–18]. These advances can significantly reduce the creative burden, benefiting various industries such as music, content creation, video games, and virtual reality applications.

Previous methods [9–11, 13, 16–18] in 3D dance generation primarily adopt Transformer-based architectures due to their effectiveness in modeling global contexts via attention mechanism. However, Transformers [19] often struggle with handling long and autoregressive sequences because of their limited inductive bias towards temporal causality. Recent studies [3, 20, 21] have explored integrating Mamba-based architectures into dance generation pipelines to better capture dance characteristics, yet these models still partially depend on attention modules through hybrid designs.

Moreover, existing approaches [10, 11, 13, 16–18] lack effective methods for injecting beat information. Generally, choreography or dance instructing involves segmenting movements according to beats, making beat information as crucial as music for the dance motions. Although music features contain low-dimensional beat information, this representation is insufficiently influential in guiding the generation process. Beat-It [8] introduces a relatively informative beat representation disentangled from music; however, it heavily relies on neural network embeddings and encoding processes, rather than directly modeling the fundamental impact of beats on dance movements.

In this paper, we introduce *MambaDance*, a Mamba-based architecture for 3D dance generation, which leverages Mamba's SSM modules which use strong inductive bias for modeling autoregressive temporal dynamics [1, 4]. We achieve this by proposing a Mamba-based diffusion model, whose dance decoder block consists of single-modal and cross-modal Mamba modules, interleaved with Adaptive Linear Modulation modules to effectively integrate beat and music information (see Fig. 1). Additionally, we propose a new explicit beat representation based on intuitive signal distributions, highlighting the crucial role of beats in structuring dance sequences and distinguishing them from normal human motion. Comprehensive experiments conducted on a 3D dance dataset demonstrate that our method effectively generates dance sequences faithfully reflecting rhythmic and structural characteristics, consistently achieving superior performance.

In summary, our contributions are as follows:

- We propose *MambaDance*, a Mamba-based diffusion model, tailored for generating autoregressive dance data.
- We propose a new beat representation as an essential conditional input, enabling the model to accurately capture dance-specific motion characteristics.
- We conduct experiments demonstrating improvements driven by specific modules and the beat representation.

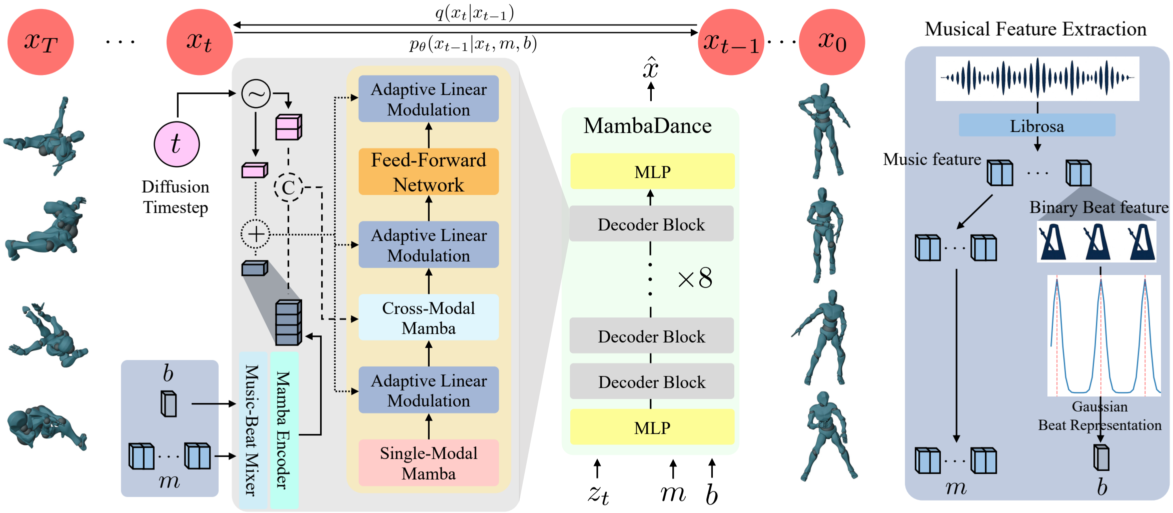

Figure 1. **The overall architecture of *MambaDance*.** We extract music feature $m \in \mathbb{R}^{L\times 35}$ with Librosa [14]. The last dimension corresponds to the binary mask of beat, subsequently represented as $b \in \mathbb{R}^{L\times 1}$ (blue box). In the dance decoder block, condition token c is constructed by concatenating entangled musical condition and timestep token. This token injects the musical information by Cross-Modal Mamba, along with the summation of mean-pooled conditional token and time step embedding by Adaptive Linear Modulation (gray box).

2. Method

2.1. Mamba for Dance Generation

Following EDGE [18], we define sliced dance motion as a sequence of L-length poses $x \in \mathbb{R}^{L\times 151}$, represented with SMPL [12] format. The music clip is processed using Librosa [14] to extract musical features $m \in \mathbb{R}^{L\times 35}$ as a condition for the diffusion model. The last dimension of m corresponds to a one-hot encoded beat, where we extract beat representations $b \in \mathbb{R}^{L\times 1}$ for the other condition. In the context of the diffusion-based dance generation and efficacy of Mamba for sequence modeling [1, 4], we propose a novel approach, referring to *MambaDance*, for music-to-3D dance generation. As in Fig. 1, we leverage Mamba for sampling $\hat{x}_\theta$, mainly focusing on the inductive bias to process complex sequential 3D dance data. Dance decoder (green box in Fig. 1) processes motion latent $x \in \mathbb{R}^{L\times E}$ produced by an MLP with 8 dance decoder layers, and outputs predicted motion sequences $\hat{x} \in \mathbb{R}^{L\times 151}$. Here, E stands for the latent dimension hyperparameter. Musical condition $c_m \in \mathbb{R}^{L\times E}$ is constructed by MLP-based Music-Beat Mixer and Mamba Encoder in each dance decoder layer, which consists of Single-Modal Mamba (SMM), Cross-Modal Mamba (CMM), Feed-Forward Network (FFN) and Adaptive Linear Modulation (AdaLM).

SMM transforms input noisy motion sequence using two Temporal SSM Block and bi-directional Spatial SSM Block. The Temporal SSM Block, following [4], transits arbitrary sequence $a \in \mathbb{R}^{L\times E}$ along L sequence length axis, and the Spatial SSM Block transits rearranged sequence $a' \in \mathbb{R}^{E\times L}$ along E spatial latent dimension bidirectionally. In contrast, CMM leverages the same Temporal SSM Block to get the information of conditions by concatenating all the input sequences including timestep embedding $e_t \in \mathbb{R}^{2\times E}$ as $[x, c_m, e_t]$, and outputs the motion part extracted after the transition (see Fig. 2).

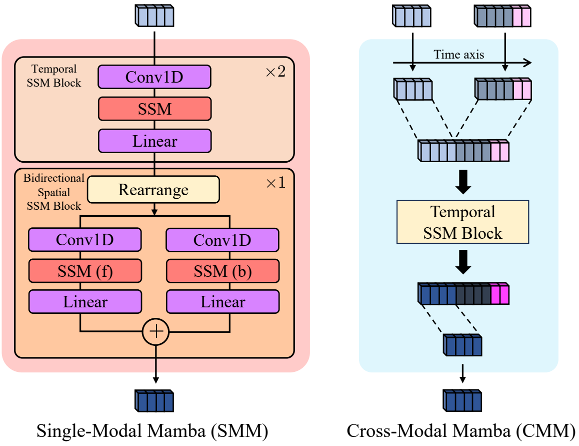

Figure 2. **Single-Modal Mamba (left) and Cross-Modal Mamba (right).** Light blue, dark blue, and pink blocks correspond to motion, condition, and timestep tokens, respectively.

Additionally, we introduce Adaptive Linear Modulation (AdaLM), a normalization-based modulation technique for 1-dimensional input sequences. Previous adaptive normalization techniques [2, 7] apply conditioned scaling and shifting to the calculated mean and standard deviation for the input images. Similarly, we scale and shift the group-normed 1D motion input with the given summation of mean-pooled musical and timestep conditions $c_{\text{mod}} \in \mathbb{R}^E$:

$$\text{AdaLM}(x, \gamma, \beta) = (1 + \gamma) * \text{GroupNorm}(x) + \beta, \quad (1)$$

where γ and β are from the projected conditioning input c_{mod} by internal activation network.

2.2. Beat Representation

Previous studies have identified a strong consistency between musical beats and the dance motion beats [8, 10, 16, 18]. Capturing this rhythmic correspondence is essential for generating coherent and expressive dance sequences. In many existing approaches, beat information is included in the music features. For instance, Librosa [14] is commonly used to extract a 1-dimensional binary beat signal, which is included as a part of a 35-dimensional music feature vector m. Beat-It [8] addresses sparsity and less informativeness of the binary representation by introducing a beat representation in the form of a vector, where each entry denotes the temporal distance to the nearest beat frame. However, the resulting signal remains monotonic and fails to explicitly reflect the decreasing influence of frames further from beats.

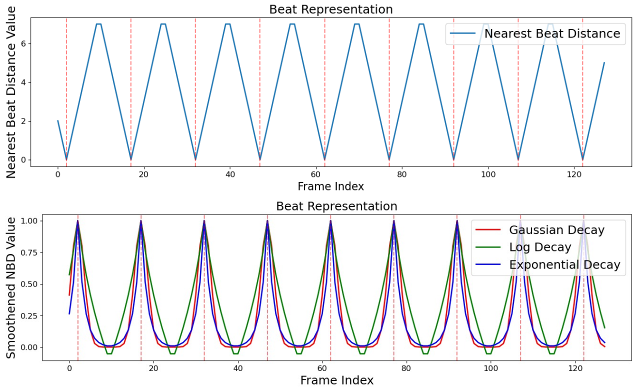

Figure 3. **Visualizations of the nearest beat distance [8] (top) and our beat representation (bottom) for a sliced music clip.**

In practical dance choreography, beats serve as natural segmentation points that structure and differentiate motion phrases. Therefore, an effective beat representation should satisfy two key properties: *frames closer to beats should exhibit higher signal strength, and this strength should decay rapidly yet smoothly as temporal distance increases*. To meet these criteria, we propose a beat representation $b \in \mathbb{R}^{L \times 1}$ using a Gaussian decay function, intuitively modeling signal attenuation from the beat frames.

As illustrated in Fig. 3, we first compute nearest beat distance d for each frame as the minimum distance to the closest preceding or following beat frame. Given a scale factor α and the interval l between adjacent beats, we define:

$$b = \exp\left(-\frac{d^2}{2\sigma^2}\right). \quad (2)$$

Among various decay functions (Gaussian, logarithmic, and exponential), we adopt the Gaussian form due to its advantageous properties. It naturally produces values in the range $[0, 1]$, eliminating the need for further normalization, and its bell-shaped curve provides a smooth and localized emphasis around the beat frames. The resulting signal offers an explicit and interpretable rhythmic cue that aligns well with the expressive structure of dance motion.

3. Experiments

3.1. Experiment Design

We evaluate *MambaDance* on the AIST++ [10] dataset sampled at 30 FPS. AIST++ consists of 1,408 high-quality short dance sequences, performed by professional dancers across 10 genres. We set the length L of training motion and music sequences to 128.

We compare our method with the following baselines, which show recent advances by leveraging Transformer-based diffusion architectures in music-conditioned 3D dance generation.

- **EDGE** [18]: The first approach to use Transformer-based diffusion model for 3D dance generation with rich music representation from Jukebox.
- **POPDG** [13]: A follow-up method utilizing an improved diffusion model (iDDPM [15]) based on additional alignment module and space augmentation algorithm.

3.2. Evaluation Metrics

We evaluate the generated dance motions using standard metrics. To measure motion realism, we use the Fréchet Inception Distance (FID), reporting both kinematic (FID_k) and geometric (FID_g) variants. Physical plausibility is assessed using the Physical Foot Contact (PFC) score and the Physical Body Contact (PBC) score. The Beat Alignment Score (BAS) measures how well the generated motion beats align with the music beats. Motion diversity is quantified by the Diversity metric (Div), which computes the average pairwise distance of kinematic and geometric features. Since the metrics do not exactly represent human evaluation, we conduct a user study and report the win rate (Wins) of ours over the baselines.

Unlike prior works [13, 18], we calculate all the metrics on full-length dance sequences instead of sliced motions to better reflect global temporal coherence, aligned with the goal of generating complete dances for entire music. In addition, we report the mean and standard deviation across 10 independent results to ensure statistical reliability. Further details regarding the evaluation metrics and user study are provided in Section 7.

Model	Fidelity				Beat	Diversity		Wins (↑)
	FID_k (↓)	FID_g (↓)	PFC (↓)	PBC (→)	BAS (↑)	Div_k (→)	Div_g (→)	
GT	-	-	1.25	8.35	-	9.61	7.78	-
EDGE	$\underline{125.99}^{\pm 128.69}$	$\underline{28.72}^{\pm 4.29}$	$\underline{3.1883}^{\pm 0.5318}$	$\underline{5.6923}^{\pm 0.4121}$	$\mathbf{0.2572}^{\pm 0.0112}$	$\mathbf{11.45}^{\pm 3.25}$	$\underline{4.91}^{\pm 0.56}$	82.5%
POPDG	$777.32^{\pm 711.65}$	$60.08^{\pm 5.98}$	$4.8615^{\pm 0.6010}$	$5.9301^{\pm 0.6438}$	$0.2318^{\pm 0.0129}$	$24.08^{\pm 7.40}$	$\mathbf{7.87}^{\pm 0.59}$	88.5%
Ours	$\mathbf{33.60}^{\pm 7.04}$	$\mathbf{28.05}^{\pm 0.74}$	$\mathbf{1.5297}^{\pm 0.1743}$	$\mathbf{6.1807}^{\pm 0.6495}$	$\underline{0.2518}^{\pm 0.0172}$	$\underline{6.17}^{\pm 0.96}$	$3.59^{\pm 0.20}$	-

Table 1. **Quantitative results on the AIST++ dataset.** GT motion is used as the reference. For each metric, ↓ indicates lower is better, and → indicates closer to the real motion is better. The best and second-best results are highlighted in bold and underline, respectively.

3.3. Comparisons

As shown in Table 1, our model consistently outperforms baselines in metrics related to fidelity, such as FID_k, FID_g, PFC, and PBC. These results indicate that *MambaDance* produces more natural and physically plausible 3D dance motions. Furthermore, as reflected in the Wins metric from our user study, dance sequences generated by our model are more frequently preferred by human evaluators compared to those produced by baselines. In contrast to EDGE and POPDG, which often suffer from static or repetitive motion patterns, *MambaDance* generates more expressive movements that are better aligned with the musical rhythm and beat. In particular, our model achieves improved coordination between upper and lower body motions, resulting in smoother full-body dynamics and a notable reduction in artifacts such as foot sliding. These qualitative improvements are illustrated in Fig. 4 and the supplementary video.

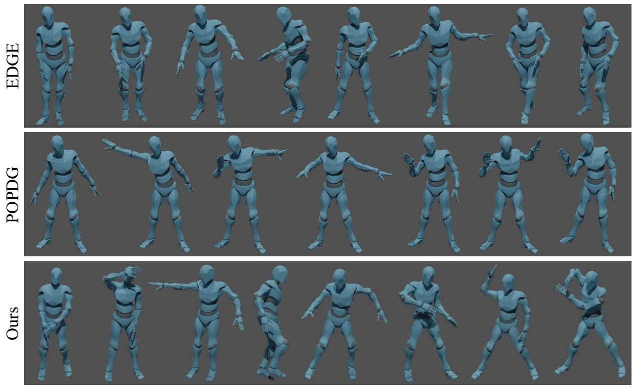

Figure 4. **Qualitative comparison.** Each row shows a set of sampled frames captured at consistent intervals from the full motion sequence.

3.4. Ablation Studies

We analyze the effects of the proposed beat representation and AdaLM through ablation experiments on the AIST++ dataset. As shown in Table 2, the physical plausibility of the generated motion improves, leading to a lower FID_k while the diversity of movements increases under varying musical conditions, reflected in a higher Div_k. In addition, when the Gaussian-based beat representation is applied, the results show a clear improvement in BAS and further increases in Div_k due to more structurally diverse motion patterns. Overall, both modules contribute meaningfully to the improvement of key aspects in dance generation, including realism, diversity, and rhythm synchronization.

Ablations		Metrics		
Beat	AdaLM	FID_k (↓)	BAS (↑)	Div_k (→)
		$47.10^{\pm 6.71}$	$0.2390^{\pm 0.0141}$	$4.08^{\pm 0.36}$
	✓	$43.30^{\pm 2.50}$	$0.2436^{\pm 0.0187}$	$4.36^{\pm 0.25}$
✓	✓	$\mathbf{33.60}^{\pm 7.04}$	$\mathbf{0.2518}^{\pm 0.0172}$	$\mathbf{6.17}^{\pm 0.96}$

Table 2. **Ablation study on beat representation and AdaLM.**

4. Conclusion

In this paper, we have proposed *MambaDance*, a novel approach for music-conditioned 3D dance generation. The proposed method fully substitutes self-attention and cross-attention for Single-Modal Mamba and Cross-Modal Mamba, respectively, which effectively capture long-range dependencies with linear complexity. Furthermore, we introduce an informative beat representation based on Gaussian decay, considering the role of beat and emphasizing the significance of beat information during dance decoding. Experimental results demonstrate the superiority of our approach over baselines across fidelity, beat alignment, and diversity. By addressing the challenges of 3D dance generation, our study highlights the potential for advancements in applications such as AI-driven choreography, creative content creation, and virtual performance systems. Nonetheless, our method occasionally exhibits motion glitches, a common limitation in non-autoregressive generation pipelines. Future work may mitigate this issue, for example, by incorporating an improved motion stitching algorithm during inference.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-II220612, Geometric and Physical Commonsense Reasoning based Behavior Intelligence for Embodied AI, and No.RS-2020-II201336, Articial Intelligence Graduate School Program (UNIST)).

References

[1] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, pages 10041–10071, 2024. 1, 2

[2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794, 2021. 2

[3] Congyi Fan, Jian Guan, Xuanjia Zhao, Dongli Xu, Youtian Lin, Tong Ye, Pengming Feng, and Haiwei Pan. Align your rhythm: Generating highly aligned dance poses with gating-enhanced rhythm-aware feature representation, 2025. arXiv preprint arXiv:2503.17340. 1

[4] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2023. arXiv preprint arXiv:2312.00752. 1, 2

[5] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *Proceedings of the International Conference on Learning Representations*, 2022. 1

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 1

[7] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 2

[8] Zikai Huang, Xuemiao Xu, Cheng Xu, Huaidong Zhang, Chenxi Zheng, Jing Qin, and Shengfeng He. Beat-it: Beat-synchronized multi-condition 3d dance generation. In *European Conference on Computer Vision*, page 273–290, 2024. 1, 3

[9] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer, 2020. arXiv preprint arXiv:2008.08171. 1

[10] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 13401–13412, 2021. 1, 3

[11] Ronghui Li, YuXiang Zhang, Yachao Zhang, Hongwen Zhang, Jie Guo, Yan Zhang, Yebin Liu, and Xiu Li. Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1524–1534, 2024. 1

[12] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2

[13] Zhenye Luo, Min Ren, Xuecai Hu, Yongzhen Huang, and Li Yao. Popdg: Popular 3d dance generation with popdanceset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26984–26993, 2024. 1, 3, 2

[14] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceeding of the 14th Python in Science Conference*, pages 18–24, 2015. 2, 3

[15] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3

[16] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation via actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3, 2

[17] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando++: 3d dance gpt with choreographic memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45: 14192–14207, 2023.

[18] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 1, 2, 3

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[20] Kaixing Yang, Xulong Tang, Yuxuan Hu, Jiahao Yang, Hongyan Liu, Qinnan Zhang, Jun He, and Zhaoxin Fan. Matchdance: Collaborative mamba-transformer architecture matching for high-quality 3d dance synthesis. *arXiv preprint arXiv:2505.14222*, 2025. 1

[21] Kaixing Yang, Xulong Tang, Ziqiao Peng, Yuxuan Hu, Jun He, and Hongyan Liu. Megadance: Mixture-of-experts architecture for genre-aware 3d dance generation. *arXiv preprint arXiv:2505.17543*, 2025. 1

Not Like Transformers: Drop the Beat Representation for Dance Generation with Mamba-Based Diffusion Model

Supplementary Material

5. Preliminaries

Selective State Space Model. State Space Models (SSMs), particularly Structured State Space Models (S4 [5]) and Mamba [1, 4], have shown superior capabilities of modeling long-range dependencies of sequential data. These models map an input sequence $x_t \in \mathbb{R}^T$ to an transited output sequence $y_t \in \mathbb{R}^T$ through a hidden state $h_t \in \mathbb{R}^N$. SSM can be discretized with step size Δ as follows:

$$\begin{aligned} h_t &= Ah_{t-1} + Bx_t \\ y_t &= C^\top h_t, \end{aligned} \tag{3}$$

where $A \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{N\times 1}$, and $C \in \mathbb{R}^{N\times 1}$ are state matrix, input matrix, and output matrix, defined by state dimension N, respectively. This system can be expressed using a global convolution with a structured convolutional kernel $\bar{K}$ (note that x denotes general sequential input here):

$$\begin{aligned} \bar{K} &= (C^\top \bar{B}, C^\top \bar{A}\bar{B}, \ldots, C\bar{A}^{L-1}\bar{B}) \\ y &= x * \bar{K}. \end{aligned} \tag{4}$$

To deviate from linear time-invariance (LTI), Mamba1 [4] introduces selective scanning with time-varying parameters, overcoming computational challenges with associative scans. Mamba2 [1] further enhances the efficiency by conceptually connecting SSM and attention mechanism, enabling faster computations while maintaining competitive performance against Transformers [19].

Diffusion Model. We adopt DDPM [6] formulation, defined by a forward noising process of latents $\{z_t\}_{t=1}^T$:

$$q(z_t|x) \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}x, (1-\bar{\alpha}_t)\mathbf{I}), \tag{5}$$

where $x \sim p(x)$, and $\bar{\alpha}_t \in (0,1)$ are constants which follow a monotonically decreasing schedule. Given musical condition c_m from music feature m and beat representation b, the diffusion model reverses the forward diffusion process to estimate $\hat{x}_\theta(z_t, t, m, b) \approx x$ for all timestep t, where θ denotes the model parameters.

We adopt a standard reconstruction loss of the diffusion models, defined as:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x,t}\left[\|x - \hat{x}_\theta(z_t, t, m, b)\|_2^2\right]. \tag{6}$$

6. Loss function

Additionally, following EDGE [18], the auxiliary losses can be formulated as:

$$\begin{aligned} \mathcal{L}_{\text{pos}} &= \frac{1}{L}\sum_{i=1}^{L}\left\|\text{FK}(x^{(i)}) - \text{FK}(\hat{x}^{(i)})\right\|_2^2 \\ \mathcal{L}_{\text{vel}} &= \frac{1}{L-1}\sum_{i=1}^{L-1}\left\|(x^{(i+1)} - x^{(i)}) - (\hat{x}^{(i+1)} - \hat{x}^{(i)})\right\|_2^2 \\ \mathcal{L}_{\text{foot}} &= \frac{1}{L-1}\sum_{i=1}^{L-1}\left\|(\text{FK}'(\hat{x}^{(i+1)}) - \text{FK}'(\hat{x}^{(i)}))\cdot \hat{y}^{(i)}\right\|_2^2, \end{aligned} \tag{7}$$

where $\text{FK}(\cdot)$ and $\text{FK}'(\cdot)$ denote the forward kinematic function which convert joint angles into joint positions for all joints and foot joints, respectively. L indicates the number of frames and the index is denoted as superscript i. Also, $\hat{y}$ stands for the predicted binary foot contact label. A position loss $\mathcal{L}_{\text{pos}}$ measuring the similarity of joint positions, a velocity loss $\mathcal{L}_{\text{vel}}$ assessing the similarity of joint velocities, and a contact consistency loss $\mathcal{L}_{\text{foot}}$ ensuring accurate foot-ground contacts.

The total loss function for training *MambaDance* combines these terms as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{simple}} + \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{foot}}\mathcal{L}_{\text{foot}}, \tag{8}$$

7. Evaluation Metrics

To quantitatively evaluate the quality of the generated dance motions, we adopt several commonly used metrics from prior works. We used a sequence length of 128, which slightly differs from the original baseline setting of 150, and calculated all metrics for whole integrated dance, so the metric values may differ from those reported in prior works.

Motion Quality. To evaluate the quality of generated motions, we compute the Fréchet Inception Distance (FID) between motion features of generated and ground truth motion sequences. For each motion, we extract kinematic and geometric features, which respectively capture physical naturalness and overall dance choreography.

Physical Foot Contact Score. To evaluate the physical plausibility of foot movements in response to dance motion, we adopt the Physical Foot Contact Score (PFC) proposed in EDGE [18]. This physically-inspired metric assesses whether foot-ground interactions are realistic or not without requiring explicit physical modeling. It evaluates the center of mass (COM) acceleartion along both horizontal plane and vertical axiz. Lower PFC scores indicate more physically plausible motions.

Physical Body Contact Score. Inspired by POPDG [13], PBC measures the overall physical feasibility of full-body movements by analyzing inter-limb and upper-body contacts to identify implausible interpenetrations or unnatural poses.

Motion Diversity. To assess the diversity of the generated motions, we compute the average feature distance of generated motions and ground truth motions. Following Bailando [16], we consider both kinematic and geometric features, denoted as Div_k and Div_g, repectively. Higher values indicate greater variability in motion patterns.

Beat Alignment Score. To evaluate the beat consistency between the generated dance and the music, we follow Bailando [16] and compute the average temporal distance between each music beat and its nearest motion beat. A higher BAS value indicates better synchronization between the motion and the rhythm of the music.

User Study (Wins). For the user study, we gather 20 participants and each of them watches 10 pairs of dance videos, with each pair corresponding to one of the 10 music tracks in the test set. Every pair consists of two dance sequences generated for the same music–one by *MambaDance* and the other by either EDGE [18] or POPDG [13]. Evaluators are asked to choose which video performed better according to specific criteria. Two separate surveys are conducted, one comparing ours with EDGE and the other with POPDG. The criteria for ”better performance” are clearly defined as follows:

- Which one demonstrates more natural dance movements?
- Which one aligns better with the music in terms of beat and rhythm synchronization?
- Which one exhibits more diverse and dynamic movements?

To prevent positional bias, the order of the videos within each pair is randomized. For fair comparisons against both baselines, we generate two different dance sequences per music track, ensuring a balanced and unbiased evaluation for each baseline. The videos used for user study are included in the supplementary materials.