

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

Supplementary Material

1. Preliminaries

Selective State Space Model. State Space Models (SSMs), particularly Structured State Space Models (S4 [3]) and Mamba [1, 2], 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 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:

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

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):

$$\begin{aligned} \bar{K} &= (C^\top \bar{B}, C^\top \bar{A} \bar{B}, \dots, C^\top \bar{A}^{L-1} \bar{B}) \\ y &= x * \bar{K}. \end{aligned} \quad (2)$$

To deviate from linear time-invariance (LTI), Mamba1 [2] 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 [10].

Diffusion Model. We adopt DDPM [4] formulation, defined by a forward noising process of latents $\{z_t\}_{t=1}^T$:

$$q(z_t|x) \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}x, (1 - \bar{\alpha}_t)\mathbf{I}), \quad (3)$$

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:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x,t} \left[\|x - \hat{x}_\theta(z_t, t, m, b)\|_2^2 \right]. \quad (4)$$

2. Implementation Details

We report hyperparameters as (FineDance, AIST++). The global sequence length is $N = (1024, 128)$, and the local window size is $n = (256, 64)$; both music and motion are processed at 30FPS. The global diffusion stage outputs 13 key motions m_{key} per sequence—5 hard ques d_h and 8 soft ques d_s . The key-motion length is $L_{\text{key}} = (8, 4)$, chosen relative to N . After choreographic augmentation, each d_s is mirrored to produce 16 soft-cue instances and placed at beat-aligned locations. Both global and local models are optimized with Adan [11] at learning rate of 2×10^{-4} . At inference, we use DDIM [8] sampling with 50 number of inference steps. The loss weights of each terms are as: $\lambda_{\text{pos}} = (1, 0.636)$, $\lambda_{\text{vel}} = (2.964, 2.964)$, $\lambda_{\text{acc}} = (2.964, 2.964)$, $\lambda_{\text{foot}} = (20, 10.942)$, and $\lambda_{\text{trans}} = (0.5, 0.5)$.

On the other hand, two diffusion models have same structure of dance decoder, except for the detail of SMM. We omit the Spatial SSM block of global diffusion, because the input sequence length and output sequence length are different (recall that Spatial SSM block processes $a' \in \mathbb{R}^{l \times E}$).

3. 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 [9]. 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) acceleration along both horizontal plane and vertical axis. Lower PFC scores indicate more physically plausible motions.

Physical Body Contact Score. Inspired by POPDG [6], 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 [7], we consider both kinematic and geometric features, denoted as Div_k and Div_g , respectively. 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 [7] 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 gathered 20 participants. Each participant evaluated dance videos generated from 2 datasets, 10 music tracks, and 4 models (*MambaDance*, EDGE [9], POPDG [6], and Lodge [5]). In total, every participant watched $2 \times 10 \times 4$ dance videos, where each set consisted of four sequences generated for the same music by the four models.

Participants were asked to select the best video in each set according to the following criteria:

- Which video demonstrates the most natural dance movements?
- Which video aligns best with the music in terms of beat and rhythm synchronization?
- Which video exhibits the most diverse and dynamic movements?

To mitigate positional bias, the order of the four videos within each set was randomized.

References

- [1] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *ICML*, 2024. 1
- [2] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2023. arXiv preprint arXiv:2312.00752. 1
- [3] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *ICLR*, 2022. 1
- [4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1
- [5] 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 *CVPR*, 2024. 2
- [6] Zhenye Luo, Min Ren, Xuecai Hu, Yongzhen Huang, and Li Yao. Popdg: Popular 3d dance generation with popdanceset. In *CVPR*, 2024. 2
- [7] 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 *CVPR*, 2022. 2
- [8] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1
- [9] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *CVPR*, 2023. 1, 2
- [10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 1
- [11] Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng YAN. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *IEEE TPAMI*, 2024. 1