

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:

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

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):

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

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 $\{z_t\}_{t=1}^T$:

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

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:

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

6. Loss function

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

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

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}_{pos} measuring the similarity of joint positions, a velocity loss \mathcal{L}_{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:

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

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) pro-

posed 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) acceleration along both horizontal plane and vertical axis. 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 , 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 [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.