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

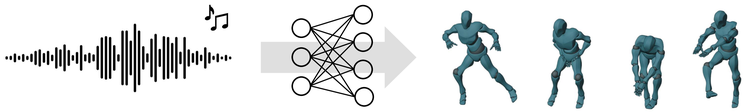
Sangjune Park[1], Inhyeok Choi[1], Donghyeon Soon[2], Youngwoo Jeon[1], Kyungdon Joo[1†]

[1]Artificial Intelligence Graduate School, UNIST [2]Department of Computer Science, DGIST

WACV 2026 TUCSON, AZ 3/6 - 3/10

Goal

☑ **Music-driven 3D dance generation task**



- Dance is a specific form of human motion along to music, which is expressive, symbolic and rhythmic
- Our goal is to generate plausible and diverse dance movements, conditioned by given music

Motivation

☑ **H1. Mamba[1] operates 3D dance data better than Transformer**

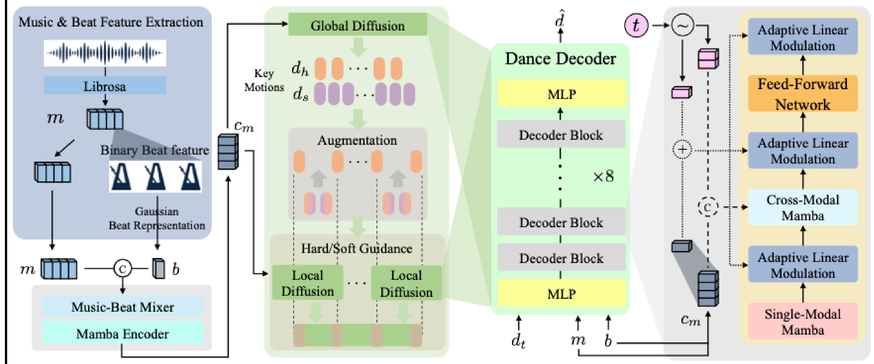
- Mamba's inductive bias helps modeling the autoregressive data
- Based on the previous Transformer-based two-stage diffusion framework[6], we leverage Mamba for dance decoder

☑ **H2. More expressive beat representation helps the generated dance better-aligned to the beat**

- Properties of beats: (i) frames closer to beats carry stronger signals, (ii) this strength decays rapidly yet smoothly with temporal distance.
- We formulate previous simple beat representation NBD[3], and use Gaussian decay function that can satisfy the properties

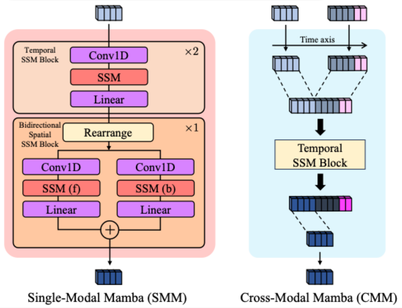
$$b(i) = \exp\left(-\frac{\mathrm{NBD}(i)^2}{2\,(\alpha \cdot l(i))^2}\right)$$

Method Overview



- Inspired by two-stage diffusion framework in Lodge[6],
- Mamba-based Diffusion: $d \approx \hat{d} = f_\theta(d_t, t, m, b)$
- Gaussian beat representation: $m \in \mathbb{R}^{L\times 35} \rightarrow b_{\text{raw}} \in \mathbb{R}^{L\times 1} \rightarrow b \in \mathbb{R}^{L\times 1}$

Mamba Modules



- **SMM** processes motion latent with SSM module of Mamba
- **CMM** processes concatenated sequence of motion latent, musical condition, and timestep embedding
- **AdaLM** is Adaptive Group Normalization (AdaGN[4]) for 1D sequences

Gaussian Beat Representation

- (a) **1D raw beat:** 1-channel binary signal from music feature
- (b) **NBD[3]:** each entry denotes the temporal distance to the nearest beat frame
- (c) **Gaussian Beat Rep.:** Gaussian decay function provides explicit and interpretable rhythmic cues by locally stressing beat frames

Experiments

☑ **Quantitative Comparison**

Dataset	Model	Fidelity			Beat	Diversity		Wins (↑)
		FID_k (↓)	FID_g (↓)	PFC (↓)	BAS (↑)	Div_k (→)	Div_g (→)	
FineDance[8]	GT	-	-	0.1852	-	10.9924	7.7424	-
	EDGE[2]	$179.01^{\pm 3.10}$	$1234.28^{\pm 236.95}$	$0.3994^{\pm 0.0177}$	$0.2261^{\pm 0.0013}$	$\mathbf{10.34}^{\pm 0.36}$	$31.74^{\pm 2.76}$	3.5%
	POPDG[5]	$190.23^{\pm 1.18}$	$1479.14^{\pm 6.44}$	$0.3765^{\pm 0.0124}$	$0.2361^{\pm 0.0033}$	$7.22^{\pm 0.14}$	$14.53^{\pm 0.12}$	5.5%
	Lodge[6]	$\underline{84.99}^{\pm 2.07}$	$\underline{64.57}^{\pm 10.74}$	$\underline{0.0585}^{\pm 0.014}$	$\underline{0.2410}^{\pm 0.0063}$	$\underline{7.98}^{\pm 0.20}$	$\mathbf{7.67}^{\pm 0.68}$	$\underline{35.0\%}$
	Ours	$\mathbf{51.36}^{\pm 0.67}$	$\mathbf{43.11}^{\pm 10.54}$	$\mathbf{0.0119}^{\pm 0.0008}$	$\mathbf{0.2441}^{\pm 0.0044}$	$6.38^{\pm 0.17}$	$\underline{6.44}^{\pm 0.88}$	**56.0%**
AIST++[7]	GT	-	-	1.2544	-	9.61	7.78	-
	EDGE[2]	$125.99^{\pm 128.69}$	$\underline{28.72}^{\pm 4.29}$	$3.1883^{\pm 0.5318}$	$\underline{0.2572}^{\pm 0.0112}$	$\mathbf{11.45}^{\pm 3.25}$	$4.91^{\pm 0.56}$	10.5%
	POPDG[5]	$777.32^{\pm 711.65}$	$60.08^{\pm 5.98}$	$4.8615^{\pm 0.6010}$	$0.2318^{\pm 0.0129}$	$24.08^{\pm 7.40}$	$\mathbf{7.87}^{\pm 0.59}$	9.0%
	Lodge[6]	$\underline{67.13}^{\pm 2.79}$	$28.93^{\pm 0.47}$	$\underline{1.4087}^{\pm 0.1296}$	$0.2397^{\pm 0.0158}$	$3.34^{\pm 0.32}$	$3.54^{\pm 0.14}$	$\underline{32.0\%}$
	Ours	$\mathbf{65.86}^{\pm 3.11}$	$\mathbf{26.58}^{\pm 1.02}$	$\mathbf{1.0622}^{\pm 0.2343}$	$\mathbf{0.2701}^{\pm 0.0116}$	$\underline{3.57}^{\pm 0.37}$	$\underline{4.98}^{\pm 0.43}$	**48.5%**

☑ **Qualitative Result (Details in VIDEO!!!)**

References

[1] Albert Gu and Tri Dao, "Mamba: Linear-time Sequence Modeling with Selective State Spaces." ArXiv 2023.
[2] Jonathan Tseng, et al., "EDGE: Editable Dance Generation from Music." CVPR 2023.
[3] Zikai Huang, et al., "Beat-It: Beat-Synchronized Multi-Condition 3D Dance Generation." ECCV 2024.
[4] Prafulla Dhariwal and Alexander Nichol, "Diffusion Models Beat GANs on Image Synthesis." NeurIPS 2021.
[5] Zhenye Luo, et al., "POPDG: Popular 3D Dance Generation with PopDanceSet." CVPR 2024.
[6] Ronghui Li, et al., "Lodge: A Coarse to Fine Diffusion Network for Long Dance Generation Guided by the Characteristic Dance Primitives." CVPR 2024.
[7] Ruilong Li, et al., "AI Choreographer: Music Conditioned 3D Dance Generation with AIST++." ICCV 2021.
[8] Ronghui Li, et al., "FineDance: A Fine-grained Choreography Dataset for 3D Full Body Dance Generation." ICCV 2023.

ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL