Music-Driven Dance Video Generation via PCA-Whitened Latent Diffusion with Rotary Position Embedding

Nokap Tony Park

SK Telecom

Abstract. We present a music-driven dance video generation system that introduces stochastic choreographic diversity via a *cross-attention DDPM in a PCA-whitened guidance latent space*. PCA whitening compresses 320-channel spatial guidance to $\boldsymbol{z}_{\text{guid}} \in \mathbb{R}^{64}$ (97.6% variance), providing a well-conditioned isotropic Gaussian prior. Rotary Position Embedding (RoPE) with full bidirectional self-attention enables globally coherent motion without attention masks. On AIST++, our system achieves 2D-BA= **0.380**, surpassing the deterministic baseline (0.303) and GT (0.327), with Div= **0.208** and MulDiv= **0.203** — 85% and 83% of GT diversity respectively.

Keywords: Music-driven dance · Diffusion model · PCA whitening · RoPE

1 Introduction

Generating photorealistic dance videos synchronized with music requires temporally coherent body motion that captures both the rhythmic and stylistic character of a given track. M2PE-DIFF [6] established a strong deterministic baseline by mapping Jukebox [1] audio embeddings to a Pose Guidance tensor that conditions a pose-guided LDM [12] (2D-BA= 0.303). However, its one-to-one mapping produces the same choreography for every music input, limiting creative exploration.

We propose a system that introduces stochastic diversity by replacing the deterministic mapping with a *cross-attention DDPM in a PCA-whitened guidance latent space*, and adopts *Rotary Position Embedding (RoPE)* for full global self-attention across all frames, enabling temporally coherent diverse choreography from a single music input.

Our contributions are: (1) a **PCA-whitened guidance latent space** compressing 320-channel guidance to 64-dimensional isotropic vectors (97.6% variance) for well-conditioned diffusion; (2) a **cross-attention DDPM with RoPE full attention** jointly denoising guidance and keypoint latents for globally coherent motion; and (3) **inference-time structured initialization with SDEdit** that enables diverse choreography, achieving MulDiv= **0.203** (83% of GT) while maintaining 2D-BA= 0.380.

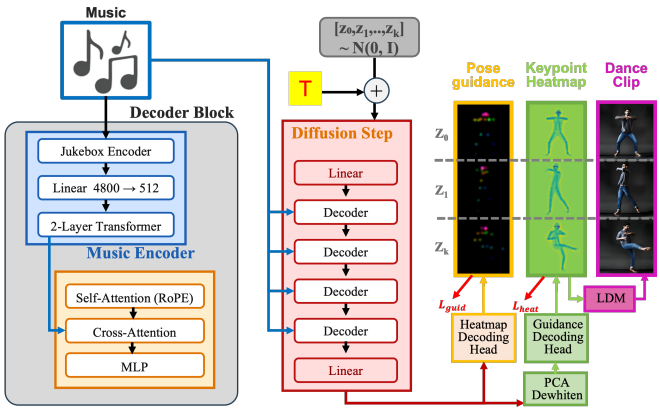

Fig. 1: Overview of the proposed M2DV-V1701 architecture. A frozen Jukebox encoder extracts music features, which are contextualized by a Transformer-based Music Encoder and used as cross-attention memory throughout the Diffusion Step module. A PCA-whitened joint latent $\boldsymbol{z}_{\text{total}} = [\boldsymbol{z}_{\text{guid}}; \boldsymbol{z}_{\text{kp}}]$ is denoised conditioned on the diffusion timestep t and music memory. The denoised latents are decoded into a Pose Guidance tensor and a Keypoint Heatmap, which together drive a pretrained LDM to synthesize the final dance video. Dashed boxes and red arrows indicate training-only components ($\mathcal{L}_{\text{guid}}$, $\mathcal{L}_{\text{heat}}$, DDPM Loss).

2 Related Work

Early methods map audio to 3D pose sequences [3, 7, 11], requiring separate rendering; M2PE-DIFF [6] instead generates the Pose Guidance tensor for a video LDM [12] but is deterministic. MDM [10] and FLAME [2] apply diffusion in 3D joint / VQ-VAE spaces; our PCA-whitened latent is derived analytically (no learned encoder), ensuring isotropic Gaussian coverage. RoPE [9] replaces hard attention masks with a soft distance bias; we adopt it for full bidirectional attention in the DDPM decoder, enabling global temporal coherence.

3 Method

The system takes Jukebox [1] music features $\boldsymbol{m} \in \mathbb{R}^{T\times 4800}$ and generates a Pose Guidance tensor $\hat{F}$ and keypoint heatmap $\hat{H}$ driving a fine-tuned LDM [6, 12]. As illustrated in Fig. 1, four components are trained jointly: a music encoder, a PCA whitening module, a cross-attention DDPM with RoPE, and a spatial decoder.

PCA-whitened guidance latent. Let $F_t \in \mathbb{R}^{320\times 64\times 64}$ be the Pose Guidance frame. Spatial pooling ($8\times$) and PCA whitening reduce each frame to

$$\boldsymbol{z}_{\text{guid}} = \frac{(\bar{F}_t - \boldsymbol{\mu})V}{\boldsymbol{\Lambda}^{1/2}} \in \mathbb{R}^{64}, \tag{1}$$

retaining 97.6% of training variance. Whitening ensures unit-variance per dimension, so $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is a faithful diffusion prior. Keypoints are linearly scaled to $\boldsymbol{z}_{\text{kp}} \in \mathbb{R}^{36}$; the joint latent $\boldsymbol{z}_{\text{total}} = [\boldsymbol{z}_{\text{guid}}; \boldsymbol{z}_{\text{kp}}] \in \mathbb{R}^{100}$ is the DDPM target.

Music encoder. Jukebox embeddings are projected and contextualized by a 4-layer Transformer encoder ($d = 512$, 8 heads) to yield per-frame music memory $\boldsymbol{e}_\mu \in \mathbb{R}^{T\times 512}$.

Cross-attention DDPM with RoPE. We adopt a cosine-schedule DDPM [5] ($T_{\text{diff}} = 1000$) with a 4-layer Transformer decoder. Each layer applies: (1) pre-norm RoPE self-attention over all T frames — no window masks — so nearby frames attend more strongly through cosine similarity of rotated Q/K vectors; (2) pre-norm cross-attention to $\boldsymbol{e}_\mu$; (3) pre-norm FFN. Inference uses DDIM [8] with 50 steps and $\eta = 0.0$; drawing K independent seeds yields K diverse choreographies.

Training objective. The loss combines noise prediction ($\mathcal{L}_{\text{ddpm}}$), guidance reconstruction ($\mathcal{L}_{\text{guid}}$), keypoint regression ($\mathcal{L}_{\text{kp}}$), velocity smoothness ($\mathcal{L}_{\text{vel}}$), foot-contact penalty ($\mathcal{L}_{\text{foot}}$), and latent smoothness ($\mathcal{L}_{\text{sm}}$). Music is randomly dropped ($p = 0.1$) for classifier-free guidance. We train on 96 A100 GPUs (batch 384, AdamW LR$= 1.5\times10^{-4}$, BF16) for 40K epochs.

Inference-time distribution matching. DDIM sampling under strong conditioning exhibits *posterior variance collapse*: per-dimension standard deviation of DDIM samples is $\sim$2.6$\times$ smaller than GT for $\boldsymbol{z}_{\text{guid}}$ and $\sim$3.6$\times$ for $\boldsymbol{z}_{\text{kp}}$. We correct this with a per-dimension affine transform calibrated on a held-out set:

$$\hat{z}_d = \frac{z_d - \mu_d^{\text{DDIM}}}{\sigma_d^{\text{DDIM}}} \cdot \sigma_d^{\text{GT}} + \mu_d^{\text{GT}}, \tag{2}$$

applied to both $\boldsymbol{z}_{\text{guid}}$ and $\boldsymbol{z}_{\text{kp}}$. This restores GT-level latent variance without retraining.

4 Experiments

Dataset and metrics. We evaluate on AIST++ [3] (1,020 sequences, 10 genres), reporting on 710 test clips of 5 s at 15 fps. *2D-BA*: Gaussian-weighted (σ = 1 frame) alignment between librosa beat times and kinetic-energy peaks over 18 OpenPose joints. *Div/MulDiv*: mean pairwise L2 across all / same-music clips in normalized 18-joint space.

Implementation. At inference, the guidance latent of a randomly selected different dancer is used as a structured initialization; SDEdit [4] then adds noise and restores music-conditioned structure via DDIM, enabling diverse choreography while preserving beat alignment.

Results. Our method achieves 2D-BA= **0.380**, exceeding both the deterministic baseline (+25%) and the GT reference (0.327), confirming strong music synchronization. Diversity reaches Div= **0.208** and MulDiv= **0.203**, corresponding to **85%** of GT diversity (0.245), demonstrating that the model generates varied choreographies from the same music input. Fig. 2 shows qualitative examples for a single music clip, confirming visually distinct choreographies across random z_{guid} initializations while maintaining coherent body motion.

Table 1: Results on AIST++ (710 test clips). GT 2D-BA: mean over test clips. GT Div: pairwise diversity of all GT clips.

Method	2D-BA↑	Div↑	MulDiv↑
GT (reference)	0.327	0.245	–
M2PE-DIFF [6] (det.)	0.303	0.000	0.000
Ours	**0.380**	**0.208**	**0.203**

Fig. 2: Qualitative results for a single music input (click for full videos). **GT**: ground-truth dancer sequence. **GT**$_z$: our model initialized with the GT dancer's guidance latent z_{guid} via SDEdit, preserving overall style while varying fine-grained motion. **Rand 1–4**: four diverse choreographies generated from different-music donor latents as random z_{guid} initializations, each music-synchronized yet kinematically distinct. Top row: rendered video frames; bottom row: pose-guidance heatmaps.

5 Conclusion

We presented a music-driven dance video generation system based on a cross-attention DDPM in a PCA-whitened guidance latent space with RoPE full bidirectional attention. PCA whitening provides a well-conditioned 64-dimensional isotropic space for diffusion; RoPE enables globally coherent motion without attention masks. On AIST++, our system surpasses the deterministic baseline and GT in beat alignment (2D-BA= **0.380**) while generating diverse choreographies from a single music input. Structured SDEdit initialization achieves Div= 0.208 and MulDiv= 0.203 (≈83% of GT), substantially closing the gap to human-performed choreography.

Limitations. The PCA bottleneck may lose fine spatial detail. Extending to in-the-wild data and learning disentangled latent axes (energy level, body-part emphasis) for user control are directions for future work.

Acknowledgements

This work was supported by a grant from the Korea Creative Content Agency (KOCCA), funded by the Ministry of Culture, Sports and Tourism (MCST), under Project Number RS-2024-00398536 (Contribution Rate: 100%).

References

1. Dhariwal, P., Jun, H., Payne, C., Kim, J.W., Ramesh, A., Chen, M.: Jukebox: A generative model for music. In: International Conference on Machine Learning (ICML) (2020)
2. Kim, J., Kim, J., Choi, S.: FLAME: Free-form language-based motion synthesis & editing. In: AAAI Conference on Artificial Intelligence (2023)
3. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: AI choreographer: Music conditioned 3d dance generation with AIST++. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
4. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. ICLR (2022)
5. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning (ICML) (2021)
6. Park, N.T.: M2PE-DIFF: Music-to-pose encoder for dance video generation leveraging latent diffusion framework. In: Proceedings of the 33rd ACM International Conference on Multimedia (MM '25) (2025). https://doi.org/10.1145/3746027.3754808
7. Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C.C., Liu, Z.: Bailando: 3d dance generation by actor-critic GPT with choreographic memory. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
8. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (ICLR) (2021)
9. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: RoFormer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
10. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2212.04048 (2022)
11. Tseng, J., Castellon, R., Liu, K.: EDGE: Editable dance generation from music. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
12. Zhu, S., Chen, J., Dai, Z., Dong, Y., Xu, X., Cao, L., Yao, Y., Zhu, H., Zhan, S.: Champ: Controllable and consistent human image animation with 3d parametric guidance. In: European Conference on Computer Vision (ECCV) (2024)