Dance Video Generation using Music-to-Pose Encoder Trained on Synthetic Dataset Generation Pipeline leveraging Latent Diffusion Framework

Nokap Tony Park
SK Telecom
Seoul, Korea
tony.nokap.park@sk.com

We present a framework for generating human dance videos conditioned on music and a reference image. A Music-to-Pose Encoder (M2PEnc), trained with a Synthetic Dataset Generation Pipeline (SDGPip), maps musical features into structured 3D pose parameters, ensuring precise rhythm–motion alignment. These pose sequences condition a latent diffusion model (LDM) with multi-level attention to synthesize motions that are rhythmically synchronized, visually coherent, and faithful to the reference subject. Extensive benchmark evaluations demonstrate state-of-the-art performance and strong generalization across subjects, styles, and genres. Comprehensive ablation studies confirm the contributions of each component, and a user study verifies the naturalness and expressiveness of the generated dances. Together, these results underscore the robustness and effectiveness of the proposed approach.

1. Introduction

Recent progress in human motion video generation highlights the effectiveness of pose-conditioned latent diffusion methods such as CHAMP [12], MagicPose [2], and Uni-Animate [20]. However, music-driven video generation remains challenging due to the complex mapping between audio features and human poses, requiring modeling of subtle temporal patterns and stylistic nuances.

We address this challenge with a framework for generating dance videos conditioned on music and a reference image, introducing:

- **Music-to-Pose Encoder (M2PEnc):** Translates musical features into structured spatial pose representations, reducing ambiguity in music-to-pose mapping.
- **Synthetic Dataset Generation Pipeline (SDGPip):** Integrates EDGE [17], SMPL [8], DwPose [23], and CHAMP [12] to produce diverse music-motion training data for M2PEnc.

Leveraging multi-level attention within a latent diffusion model, our method generates motions that are rhythmically synchronized with music and visually coherent with the reference. Experiments on benchmark and in-the-wild datasets (4.3) show state-of-the-art performance and strong generalization.

2. Related Works

2.1. Music Feature Extraction

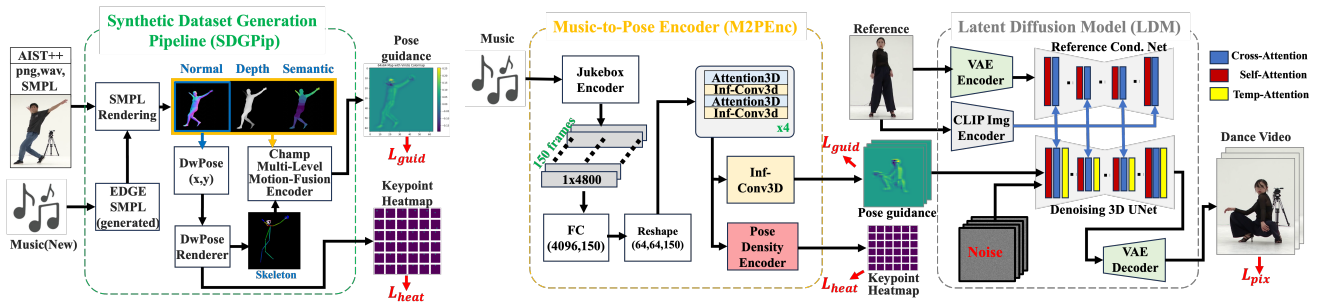
Extracting meaningful features from music is essential for music-to-motion generation. Traditional methods such as Librosa [11] extract low-level features like MFCCs, while deep learning-based approaches like MusicGen [3] and Whisper [13] capture complex musical patterns through embeddings. We use OpenAI's Jukebox encoder [4], which compresses audio signals into latent spaces using hierarchical VQ-VAE [19], preserving critical musical information such as rhythm and melody.

2.2. Music-to-Pose Generation

Music-to-pose methods synthesize 3D dance motions (SMPL[8]) aligned with musical inputs. Diffusion-based approaches like EDGE [17] use transformers for fine-grained motion control, while POPDG [10] enhances synchronization through spatial augmentation. Non-diffusion methods such as FACT [6] interpolate key poses, and Bailando [15] combines VQ-VAE[19] with motion GPT for music-driven sequences. Our SDGPip leverages EDGE for augmentation of 3D dance motions to improve robustness and diversity in music-to-motion mapping.

2.3. Latent Diffusion Framework

Pose conditioned latent diffusion models integrate pose and identity conditioning for human motion video generation. Techniques like MagicPose [2] utilize DensePose for motion guidance, while CHAMP [12] incorporates SMPL parameters for multi-class pose feature generation. Our method proposes M2PEnc to produce multi-class pose features used as input to Latent diffusion model to help generate synchronized and visually coherent dance videos conditioned on music and reference images.



(a) Synthetic dataset generation pipeline (SDGPip). (b) Music-to-Pose (M2PEnc) architecture(left) and Latent diffusion framework(right).

Figure 1. **Overview of our proposed framework.** The M2PEnc maps music to spatial pose features and leverages a denoising 3D U-Net to generate music-synchronized dance videos. The SDGPip creates synthetic pairs of music and pose features training data used by M2PEnc.

3. Methods

Figure 1 overviews our framework for generating dance videos from music and reference images. Figure 1a (Section 3.1) illustrates the synthetic data pipeline, while Figure 1b (Sections 3.2–3.3) details the Music-to-Pose Encoder and its integration with the latent diffusion model. Training and inference methods are described in Section 3.4.

3.1. Synthetic Dataset Generation

Transforming music into dance motion faces a core challenge: the subjective mapping between musical elements and human movement. To resolve this, we introduce the Synthetic Dataset Generation Pipeline (SDGPip), which creates synchronized pairs of musical features and anatomically structured 3D poses. This dataset bridges the gap between sound and motion, enabling the proposed model to learn rhythmically aligned and realistic dance generation

3.1.1. Dataset and Pseudo-Motions

We utilize AIST++ [6], which provides 1,020 SMPL [8] motion sequences synchronized with music across ten genres. The dataset includes 980 training, 20 validation, and 20 test videos, spanning approximately 5.2 hours of dance sequences. For each (motion, music) frame pair, we use the corresponding AIST[18] video frame from camera C09 for the training of latent diffusion model. We cropped the center 640×640 box of each frame to focus on the dancer.

To enhance diversity beyond predefined motions, we generate pseudo-motions using EDGE [17]. We introduce a new audio dataset comprising 300 music tracks across ten genres, each randomly cropped into a 2-minute segment, resulting in a total of 10 hours of music. By slicing them into 5-second segments and synthesizing corresponding SMPL motions, we expand the range of music-to-dance mappings.

3.1.2. SDGPip Architecture

SDGPip generates structured spatial representations using SMPL [8], DwPose [23], and CHAMP [12], including:

- **Depth Maps:** Encode body surface distances.
- **Normal Maps:** Capture surface orientation.
- **Semantic Maps:** Segment body parts into regions.
- **Skeleton Keypoints and Heatmaps:** Highlight joint connections for precise localization.

These representations are combined via CHAMP's multi-level Guidance Encoder[12] to produce pose guidance that conditions the Denoising U-Net to produce dance videos, ensuring rhythmically aligned and anatomically plausible dance motions.

3.2. Music-to-Pose Guidance Encoder

The Music-to-Pose Encoder (M2PEnc) translates musical features into pose guidance for dance synthesis. Jukebox encodes music into 4800D feature vectors per frame, capturing rhythm, melody, and dynamics. These features are reduced to $(N, 4096)$ (where $N = D \times F$, duration $\times$ frame rate) and reshaped to $(N, 64, 64)$ for spatio-temporal processing. Alternating layers of Inflated 3D Convolutions (Inf-Conv3D) and 3D attention modules model temporal dependencies, with Inf-Conv3D capturing spatio-temporal relationships and attention enhancing long-range frame coherence.

The M2PEnc outputs two components: (1) pose guidance features $(N, C_{guid}, 64, 64)$ that condition the U-Net for rhythmically aligned synthesis, and (2) keypoint heatmaps $(N, C_{heat}, 64, 64)$ optimized via MSE loss to ensure precise anatomical alignment with ground truth. This dual-output design ensures both high-level motion coherence and detailed pose accuracy.

3.3. Leveraging the Latent Diffusion Framework

Our framework integrates the Music-to-Pose Encoder (M2PEnc) with a denoising 3D U-Net within the latent diffusion framework, enabling rhythmic alignment and visual fidelity in music-driven dance generation. The M2PEnc-generated pose sequences are summed with the U-Net's latent features to synchronize music and motion, while reference images are encoded using a frozen Variational Autoencoder (VAE) and CLIP encoder to extract latent em-

beddings that preserve the subject's appearance and background. Three attention mechanisms—cross-attention for aligning pose guidance with reference embeddings, self-attention for maintaining spatial coherence within frames, and temporal attention for ensuring smooth transitions across frames—jointly refine synthesis quality. Leveraging CHAMP's pretrained stable diffusion v1.5 3D U-Net weights[12], our model achieves high-quality latent motion and reference feature synthesis without extensive retraining, with the final video frames decoded through the VAE.

3.4. Training and Inference

The training process consists of three stages. In the first stage, the Music-to-Pose Encoder (M2PEnc) is pretrained using synthetic datasets to map musical features into structured spatial pose representations. Two loss functions are minimized: the guidance loss,

$$L_{\text{guid}} = \frac{1}{HWC_{\text{guid}}} \sum_{h,w,c} |f_{h,w,c} - \hat{f}_{h,w,c}|, \tag{1}$$

which aligns predicted ($\hat{f}$) and ground truth (f) pose guidances, and the heatmap loss,

$$L_{\text{heat}} = \frac{\exp(\hat{k}_{h,w,c})}{HWC_{\text{heat}}} - k_{h,w,c}, \tag{2}$$

which refines predicted ($\hat{k}$) and ground truth (k) keypoint heatmaps. In the second stage, M2PEnc weights are frozen, and the Denoising U-Net is trained to minimize the pixel-wise Mean Squared Error (MSE) loss between predicted(x) and ground truth(y) video frame:

$$L_{\text{pix}} = \text{MSE}(x, y). \tag{3}$$

The network is initialized with CHAMP's pretrained weights. In the third stage, the temporal layer is enabled fine-tuned jointly with M2PEnc using a combined loss

$$L_{\text{total}} = 0.5 \cdot L_{\text{pix}} + 0.5 \cdot L_{\text{guid}}, \tag{4}$$

ensuring cohesive integration of temporal dynamics with pose guidance.

During inference, dance videos are generated in 5-second segments, which are seamlessly merged using UNI-Animate's noise-conditioning method [20], ensuring temporal continuity across full-length videos.

4. Experiments and Results

4.1. Implementation Details

The proposed framework was trained in three stages on eight A100 GPUs using the AdamW optimizer (learning rate 1×10^{-5}). First, M2PEnc was pretrained for 20K steps (batch size 24) on 75-frame (music, pose guidance, keypoint heatmap) sequences. Next, the denoising U-Net was trained for 60K steps (batch size 32) on (music, reference frame, target frame) pairs, with frames cropped to the human bounding box and resized to 640×640. Finally, the temporal layer was fine-tuned for 20 steps (batch size 8) on 24-frame (music, reference frame, video sequence, pose guidance) pairs. The AdamW optimizer [9] was used throughout training with a learning rate of 1×10^{-5}.

4.2. Evaluation Metrics

The quality of generated dance videos was assessed using several metrics. Fréchet Video Distance (FVD) [16], computed with an I3D classifier pre-trained on Kinetics-400 [1], evaluates video realism and temporal coherence. Structural similarity (SSIM) [21], Peak Signal-to-Noise Ratio (PSNR) [5], and Learned Perceptual Image Patch Similarity (LPIPS) [25] measure image-level structural consistency, pixel fidelity, and perceptual similarity. Synchronization metrics include the 2D-BeatAlign score [6], which assesses alignment between dance movements (with joints computed using DwPose [23]) and musical beats extracted using Librosa [11]. AV-Align [24] evaluates temporal synchronization between audio and video features. These metrics comprehensively evaluate our model's ability to produce rhythmically synchronized and visually coherent dance videos.

4.3. Result Analysis of Video Quality Evaluation

4.3.1. Qualitative Evaluation

We sliced 5-second clips from AIST[18] test set and used music and the first frame as reference, for the generation of dance videos. We compared against MM-Diffusion[14] by generating equivalent number of video clips using their public code base. As shown in Figure 2, our model generates dance sequences with superior image quality and rhythmic alignment, whereas MM-Diffusion struggles with fine-grained motion details. DabFusion[22] was excluded from comparison due to inaccessible code and inferior sample quality shown on their website.

Our model also generalizes robustly to unseen individuals, synthesizing diverse dance genres (Break, Hip-Hop, Waack, etc.) as shown in Figure 3. Consistent seed usage across music tracks demonstrates reliable rendering of stylistically accurate motions regardless of the subject's appearance.

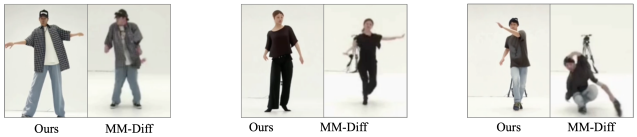


Figure 2. **Qualitative evaluation on AIST test set.** Example dance videos generated by the proposed framework. Click on the figure to view the results.

4.3.2. Quantitative Evaluation

For quantitative evaluation, we use 100 5-second clips sliced from AIST++ test set. For MM-Diffusion[14], we

Figure 3. **Qualitative evaluation using in-the-wild images.** The model generalizes well across individuals and dance genres. Click on the figure to view the results.

generated equivalent number of video clips using their public code base as mentioned earlier. We post-processed MM-Diffusion's 256×256 outputs to match to our models result by cropping to the dancer's bounding box and resizing to 640×640, ensuring metrics (PSNR, SSIM, LPIPS) reflect dance quality rather than inflated background uniformity. Upscaling low-resolution outputs may seem unfair, but it enables a balanced comparison of each model's strengths. Increasing resolution without architectural changes doesn't guarantee better performance, so upscaling helps evaluate their true capability at higher resolutions. Table 1 highlights the superiority of the proposed framework over MM-Diffusion in all metrics.

Since DabFusion's code is unavailable, we used its reported metrics from Table 1 of DabFusion [22] paper for reference. Our model was evaluated on the AIST test set; DabFusion's results are only directly comparable if they also used test clips, otherwise, they serve as a reference.

Alignment scores further validate our approach (Table 2): Our work achieves near-ground-truth 2D-BeatAlign (0.303 vs. 0.307) and outperforms MM-Diffusion in AV-Align (0.421 vs. 0.370). These results underscore the efficacy of our Music-to-Pose Encoder and latent diffusion framework in bridging music and motion.

4.4. Ablation Study

Need for M2PEnc. To assess the impact of the Music-to-Pose Encoder (M2PEnc), we ablated by removing the Pose Density Encoder and skipping the training stage 1, reducing it to a simple audio encoder. This led to noticeable drops in both video quality and alignment metrics (Table 1 and Table 2), confirming the need for M2PEnc design. We ran 20K more steps in training stage 2 and L_{guid} was ignored in the third stage.

The effect of pose guidance setups. We also ablate different pose guidance setups: DNS (Depth-Normal-Semantic maps), P (Pose skeleton), and PDNS (Pose skeleton + DNS) as shown in (Table 3). DNS shows moderate performance. P significantly improves metrics due to explicit skeletal guidance. PDNS achieves optimal results demonstrating that combining pose skeletons with geometry maps yields superior motion alignment and visual quality.

Impact of the EDGE Method. Table 1 and Table 2 show that incorporating new pose information generated with EDGE [17] leads to improved alignment metrics, while the video quality metrics remain largely unchanged.

Table 1. Quantitative evaluation results.

Model	FVD ↓	LPIPS ↓	PSNR ↑	SSIM ↑
MM-Diffusion	1338.57	0.425	11.04	0.770
DabFusion(scaled)	1440.05	0.561	8.525	0.776
Ours	**213.289**	**0.102**	**21.11**	**0.908**
Simple Audio Encoder	725.10	0.294	16.48	0.812
Ours w/o EDGE [17]	212.98	0.108	21.14	0.906

Table 2. Alignment results.

	2D-BeatAlign↑	AV Align↑
Ground-Truth	0.307	0.448
MM-Diffusion	0.287	0.370
Ours	**0.303**	**0.421**
Simple Audio Encoder	0.261	0.350
Ours w/o EDGE [17]	0.298	0.407

Table 3. Ablation Study results on Pose Guidance Setups

Guidance Setup	FVD ↓	LPIPS ↓	PSNR ↑	SSIM ↑
DNS	257.196	0.194	19.95	0.812
P	239.724	0.156	20.50	0.857
PDNS(Ours)	**213.289**	**0.102**	**21.11**	**0.908**

4.5. User Study

A user study with 20 participants (5 choreographers, 15 engineers) evaluated 50 randomized 5-second dance clips each, using a 10-point Likert [7] scale for rhythmic synchronization and visual consistency, totaling 1,000 ratings. STM2PE-Diff achieved 85.2% rhythmic synchronization (vs. MM-Diffusion's 83.3%) and 80.4% visual consistency (vs. 55.4%), indicating superior musical alignment and more realistic, coherent video generation.

Table 4. User Study Results

Model	Rhy. Sync	Vis. Cons.
MM-Diffusion	83.3%	55.4%
Ours	**85.1%**	**80.4%**

5. Conclusion

We introduced a music-conditioned dance video generation framework that integrates a Music-to-Pose Encoder with a latent diffusion model to achieve rhythmic synchronization and stylistic fidelity. A pipeline for generating synthetic datasets addresses data scarcity and ambiguity, improving diversity in music-to-motion mapping. Evaluations, ablation studies, and a user study consistently demonstrate superior quality, alignment, and expressiveness over prior methods, confirming the robustness and effectiveness of the proposed approach across diverse subjects, styles, and genres.

Acknowledgements

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

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[2] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: realistic human poses and facial expressions retargeting with identity-aware diffusion. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. JMLR.org, 2024. 1

[3] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. Last Modified: January 29, 2024. 1

[4] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music, 2020. arXiv:2005.00341. 1

[5] National Instruments. Peak signal-to-noise ratio as an image quality metric. *NI Technical Documentation*, 2011. 3

[6] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3

[7] Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55, 1932. 4

[8] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 34(6), 2015. 1, 2

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 3

[10] 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 (CVPR)*. Computer Vision Foundation, 2024. 1

[11] Brian McFee et al. librosa: Python library for audio and music analysis, 2025. Version 0.11.0. 1, 3

[12] Author names not specified in the search results. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 3

[13] Alec Radford et al. Robust speech recognition via large-scale weak supervision, 2025. arXiv:2503.09905. 1

[14] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[15] 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 (CVPR)*, 2022. 1

[16] Gaurav Parmar Jun-Yan Zhu Jia-Bin Huang Songwei Ge, Aniruddha Mahapatra. On the content bias in fréchet video distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[17] Jonathan Tseng, Rodrigo Castellon, and C. Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 4

[18] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, Delft, Netherlands, 2019. 2, 3

[19] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017. 1

[20] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *arXiv:2406.01188*, 2024. 1, 3

[21] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 2004. 3

[22] Wang Xuanchen, Wang Heng, Liu Dongnan, and Weidong Cai. Dance any beat: Blending beats with visuals in dance video generation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. 3, 4

[23] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023. 1, 2, 3

[24] Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 3

[25] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3