STM2PE-Diff : Synthetically Trained Music-to-Pose Encoder Diffusion for Automated Choreography Generation

Anonymous CVPR submission

Paper ID *****

Abstract

Automated choreography generation, which aims to seamlessly harmonize human movements with music, is a multifaceted challenge demanding both technical precision and artistic expressiveness.

We present STM2PE-Diff, a novel framework for generating human dance videos conditioned on a refernce image and music sequence using a latent diffusion model. Our approach integrates a Music-to-Pose Encoder (M2PEnc), trained with a novel synthetic dataset generation pipeline (SDGPip), which maps audio features into structured 3D pose and shape parameters that capture human geometry and dynamic motion patterns synchronized with musical input. By combining these encoded parameters with a reference image through a multi-level attention mechanism within the latent diffusion framework, we synthesize visually coherent and rhythmically synchronized dance animations of individuals from the given reference image.

Experiments on benchmark datasets demonstrate that STM2PE-Diff achieves state-of-the-art performance, producing high-quality dance videos that accurately reflect pose diversity and temporal consistency. Additionally, our method exhibits robust generalization capabilities, validated by its strong performance on a newly introduced in-the-wild dataset.

1. Introduction

Recent advancements in human motion video generation, such as AnimageAnyone [9], AnimateAnything [5], Champ [15], MagicPose [3], MagicAnimate [24], MagicDance [2], and UniAnimate [22], have demonstrated the effectiveness of pose-conditioned approaches. These methods excel at translating pose sequences into realistic and visually compelling videos of human subjects in motion. **Pose guidance** plays a pivotal role in ensuring realism, consistency, and adaptability by bridging user-defined inputs (e.g., reference poses) with high-quality outputs suitable for diverse applications.

However, generating human motion videos from music presents a unique challenge: the inherently ambiguous and complex relationship between musical features and human poses. Unlike pose-conditioned methods that rely on explicit spatial inputs, music-to-motion generation requires capturing intricate temporal patterns and stylistic nuances embedded in audio signals. Current methods often fail to adequately model this relationship, resulting in outputs that lack rhythmic synchronization or stylistic coherence with the input music.

To address this gap, we propose **STM2PE-Diff**, a novel framework for generating dance videos conditioned on music and reference images. The first innovation of our approach is the Music-to-Pose Encoder (**M2PEnc**) that maps musical features into structured spatial pose representations, removing ambiguity in the music-to-pose relationship. These representations serve as pose guidance within a latent diffusion framework, enabling the synthesis of dance motions that are rhythmically aligned with input music and visually coherent to the reference image. By leveraging multi-level attention mechanisms (cross-attention, self-attention, temporal attention) within the diffusion process, our method ensures temporal consistency across frames while preserving stylistic fidelity to the reference image.

The second key innovation of our framework is the synthetic dataset generation pipeline (**SDGPip**) designed to address the scarcity of paired music-to-motion data. This pipeline combines established models such as EDGE [20], SMPL [12], DwPose [26], and CHAMP [15] to generate music-to-motion training data. The resulting dataset enables meaningful learning of diverse and natural pose features in the spatial domain while enhancing model robustness across various scenarios.

Notably, our SDGPip and M2PEnc are designed to efficiently utilize existing latent diffusion architectures, allowing for effective adaptation without requiring extensive retraining from scratch. This approach streamlines the synthesis process, making it more practical for real-world applications.

Our contributions are summarized as follows:

- **Novel Synthetic Dataset Pipeline**: We design a synthetic dataset generation pipeline that leverages CHAMP, EDGE, SMPL, and DwPose keypoints to produce paired music-pose data, ensuring diversity and naturalness of dance motions.
- **Music-to-Pose Encoder (M2PEnc)**: We introduce the M2PEnc, which directly maps musical features into structured spatial pose representations, thereby reducing ambiguity in the music-to-pose relationship.
- **Efficient Latent Diffusion Framework Adaptation**: We propose a novel method to adapt the latent diffusion framework with multi-level attention mechanisms for synchronized dance video synthesis, leveraging our synthetic dataset pipeline and M2PEnc to efficiently utilize existing architectures without requiring extensive retraining from scratch.
- **State-of-the-Art Performance**: We demonstrate state-of-the-art performance on benchmark datasets and robust generalization on a newly introduced in-the-wild dataset.

By bridging the gap between music and motion through our synthetic data-driven approach and efficient leveraging of diffusion modeling techniques, we present a robust solution for automated choreography generation that is scalable and adaptable to diverse applications.

2. Related Works

2.1. Music Feature Extraction

Music feature extraction plays a pivotal role in bridging audio signals to dance motion generation. Several approaches have been proposed for extracting features from music signals: Audio feature extraction methods such as Librosa [14] focus on extracting features like spectral centroid, spectral roll-off, and Mel-Frequency Cepstral Coefficients (MFCCs). Deep Learning-Based Methods such as MusicGen [4] and Whisper [16] employ encoder-decoder architectures to generate embeddings that capture complex musical patterns.

In our work, we utilize the Jukebox encoder. The Jukebox encoder, developed by OpenAI [6], is a generative model designed to extract meaningful representations of musical features using VQ-VAE to compress audio signals into discrete latent spaces while preserving critical musical information such as rhythm, melody, and timbre. The encoder operates hierarchically at three levels of compression and generates 4800-dimensional feature vectors at a frequency of 345Hz.

Jukebox excels at capturing long-term dependencies in raw audio signals using its autoregressive Sparse Transformer architecture [6]. By integrating Jukebox embeddings, our model encodes rich musical features like rhythm and melody into structured representations, enabling the generation of dance motions that are rhythmically synchronized and stylistically coherent with the input music.

2.2. Music-to-Pose Generation

Recent advancements in music-to-pose generation have introduced innovative methods for synthesizing dance motions aligned with musical inputs. EDGE employs a transformer-based diffusion model conditioned on Jukebox embeddings, enabling fine-grained motion control through joint-wise conditioning and ensuring physically plausible movements with its Contact Consistency Loss. POPDG uses an iDDPM-based framework with a Space Augmentation Algorithm to enhance spatial body joint connections and an Alignment Module for improved temporal synchronization.

Non-diffusion-based methods, although less performing, also contribute valuable insights. FACT uses a cross-modal transformer to generate key poses from music and interpolate them into complete sequences, while Danceformer adopts a two-stage framework for pose generation and interpolation. Bailando combines VQ-VAEs for body segmentation with a motion GPT to map music features into dance sequences.

Our work leverages EDGE as a teacher diffusion model to augment (music, dance pose sequence) pairs, introducing semantic diversity and enhancing model robustness through diffusion-driven data augmentation [7, 11, 19]. This approach generates diverse and high-quality synthetic data, enriching the training dataset and allowing our 3D transformer model to capture the intricate links between music and dance movements more effectively.

2.3. Latent Diffusion Framework and Conditioned Human Animation Generation

Current progress in latent diffusion models has significantly improved the generation of realistic human motion videos conditioned on diverse attributes such as pose and identity. **Pose conditioning** provides explicit spatial and temporal constraints, enabling precise control over the motion in generated videos. Various methods employ different pose representations for conditioning. For instance, MagicPose [3] utilizes DensePose [8] for dense motion guidance, AnimateAnything [5] leverages optical flow to guide frame-by-frame animation, and CHAMP [15] incorporates 3D SMPL parameters for multi-hypothesis pose generation.

Identity conditioning, on the other hand, encodes appearance and background information into the generated videos to ensure visual consistency. While most approaches use a separate U-Net for appearance encoding, UniAnimate [22] adopts a unified encoder that maps pose, reference images, and noise into a shared feature space.

To enhance **temporal consistency** across video frames, temporal attention blocks are integrated into diffusion net-works. Additionally, many frameworks adopt multi-stage training strategies to disentangle multimodal features effectively, such as appearance and motion. For example, MagicDance [2] fine-tunes separate modules for appearance and pose control before integrating them into a unified framework.

Our work builds upon recent advancements in diffusion-based generative models by integrating appearance encoding, multi-level attention mechanisms, and a novel Music-to-Pose Encoding mechanism.

3. Methods

Figure 1 provides an overview of the architecture of our proposed framework, which generates temporally coherent dance motion videos conditioned on music input sequence and a reference image. Figure 1a illustrates the synthetic dataset generation pipeline, as detailed in Section 3.1, which produces paired music-to-motion data to enable effective training. Figure 1b depicts the Music-to-Pose Encoder (M2PEnc), discussed in Section 3.2, which maps musical features into structured spatial pose representations used as pose guidance. Figure 1c shows how we adapt a latent diffusion framework by integrating M2PEnc with a denoising 3D U-Net, leveraging multi-level attention mechanisms (cross-attention, self-attention, and temporal attention) for synchronized dance video synthesis, as explained in Section 3.3. Finally, Section 3.4 describes the training and inference methods used in our work.

3.1. Synthetic Dataset Generation

Bridging Music and Dance. Transforming music into dance motion is inherently challenging due to the ambiguous relationship between musical features and human poses. A single piece of music can correspond to multiple dance movements, making it difficult for the Music-to-Pose Encoder to learn consistent latent representations during training. Incorrect or inconsistent representations can lead to anatomically inaccurate or stylistically misaligned dance motions when combined with the downstream Denoising U-Net.

To address this ambiguity, we constrain the Music-to-Pose Encoder to learn from synthetic datasets that pair music sequences with latent representations of 3D pose and shape guidance sequences. By anchoring musical features to structured spatial representations, this approach removes variability in music-to-motion mapping, enabling synchronized, visually coherent, and anatomically plausible dance motion generation.

Source Dataset. The AIST [21] dance dataset consists of 13,940 videos captured from nine multi-view cameras (C01 to C09), spanning ten music genres, each paired with six unique tracks. The dataset features diverse choreographies ranging from basic to advanced movements, enabling models to learn a wide variety of dance motions.

AIST++ extends AIST by providing 1,408 sequences of 3D human SMPL motions, represented as joint rotations and root trajectories. Its precise synchronization between music and dance is critical for learning meaningful correlations between musical features and human poses.

We follow the recommended AIST++ dataset split: 980 videos for training, 20 for validation, and 20 for testing. The dataset spans approximately 5.2 hours (560,000 frames). For training, we crop each frame from camera C09 into a 640×640 sub-image centered on the dancer. These cropped images, along with corresponding music tracks and SMPL motion data, are used both as ground truth for MSE loss computation and for generating synthetic datasets as described in the subsequent section.

To expand diversity beyond the predefined motions in AIST++, we utilize EDGE [20] to generate pseudo-motions. Specifically, we select 300 new music tracks spanning 10 genres and partition them into 5-second slices. Corresponding SMPL motions are then generated for each slice. These pseudo-motions enable us to explore a broader range of music-to-dance mappings, extending beyond the explicit pairings available in the original dataset.

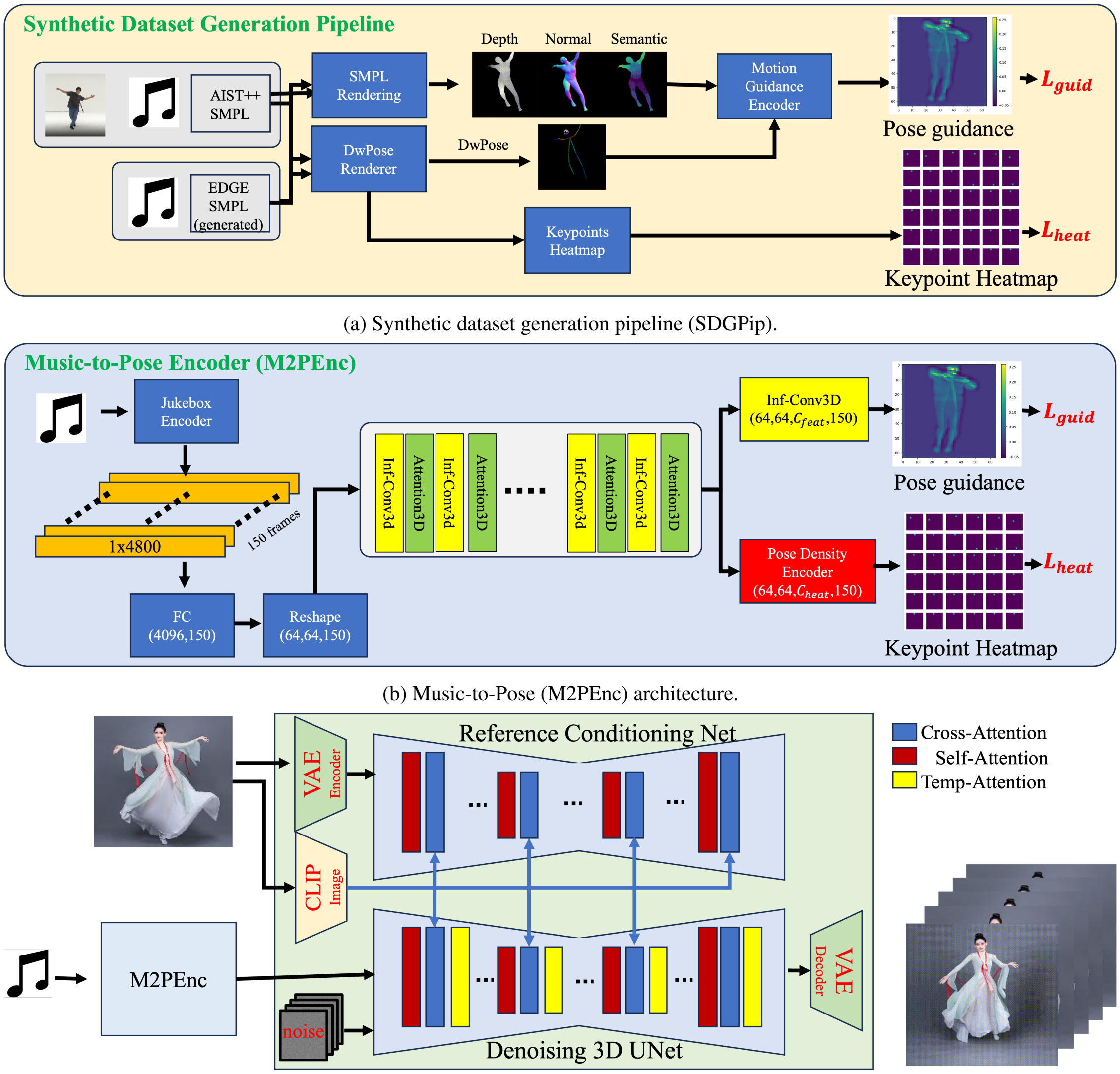
SDGPip Architecture. The pipeline, illustrated in Figure 1 (a), leverages advanced deep learning techniques inspired by Jukebox [6], SMPL [12], DwPose [26], and CHAMP [15] to generate structured spatial representations of human motion. It begins with the SMPL rendering module, which generates three types of maps to represent human body geometry:

- **Depth**: Encodes the distance between body surfaces and the camera.
- **Normal**: Captures surface orientation to model spatial dynamics.
- **Semantic**: Segments body parts into distinct regions for detailed analysis.

Simultaneously, skeletal keypoints are extracted by leveraging DwPose. These keypoints are processed by the DwPose Renderer to produce:

- **Skeleton keypoint map**: Visualizes the skeletal structure, highlighting joint connections.
- **Keypoint Heatmaps**: Represents the spatial probability of anatomical landmarks, enabling precise localization of keypoints.

Next, Pose Guidance is computed by combining Depth, Normal, Semantic maps, and skeleton keypoint maps through CHAMP's multi-level Guidance Encoder. This structured spatial representation serves as a critical input to the downstream U-Net during training, ensuring anatomically accurate and rhythmically synchronized dance motion generation.



(a) Synthetic dataset generation pipeline (SDGPip).

(b) Music-to-Pose (M2PEnc) architecture.

(c) Video diffusion model architecture.

Figure 1. **Overview of our proposed framework.** The system integrates a Music-to-Pose Encoder (M2PEnc) with a denoising 3D U-Net architecture, leveraging multi-level attention mechanisms (cross-attention, self-attention, and temporal attention) to generate dance videos synchronized with music input. M2PEnc maps musical features into structured spatial pose representations, which serve as pose guidance within the U-Net. This framework is trained on a synthetic dataset generated leveraging CHAMP, EDGE, SMPL, and DwPose. The synthetic dataset generation pipeline (SDGPip) addresses the scarcity of paired music-to-motion data, enabling STM2PE-Diff to produce high-quality, rhythmically aligned, and visually coherent dance animations.

3.2. Music-to-Pose Guidance Encoder

Dance and music are intrinsically connected, with music shaping elements such as rhythm, tempo, mood, emotion, structure, and phrasing, while dance translates these characteristics into physical movements, transitions, and expressions. The M2PEnc aims to computationally model this relationship by bridging musical inputs with spatial pose representations.

3.2.1. M2PEnc Architecture

The M2PEnc transforms music into dance feature sequences that condition the downstream U-Net. Specifically,

- The Jukebox Encoder produces 4800-dimensional feature vectors for each video frame that encode rhythm, melody, and dynamics of the input music. For each batch, N frames ($N = D \times F$) of music are processed, where D represents the duration (in seconds) and F is the frame rate (fps).

- The fully-connected layer reduces the dimensionality of the feature to $(N, 4096)$.
- A reshape operation reshapes the feature into an image-like spatial shape of $(N, 64, 64)$ for further processing.
- L alternating layers of Inflated 3D Convolutions (Inf-Conv3D) and 3D attention modules capture the temporal dependency in the feature. Specifically, Inf-Conv3D captures the spatio-temporal dependencies in music-driven motion by expanding 2D convolutions into the temporal dimension. 3D Attention modules enhance long-range temporal relationships across frames.

Throughout the processing, the feature maps maintain a fixed spatial resolution of $(64, 64)$. Upon completion, the encoder produces two outputs through two additional layers of Inf-Conv3D:

- **Pose Guidance**: High-level motion features of shape $(N, C_{\text{guid}}, 64, 64)$, which serve as conditioning inputs to the downstream U-Net. These features encapsulate temporal and spatial information derived from the music input, guiding the generation of coherent and rhythmically aligned dance motions.
- **Keypoint Heatmaps**: Detailed spatial distributions of keypoint locations, represented as heatmaps of shape $(N, C_{\text{heat}}, 64, 64)$. These maps are used to enhance model performance by minimizing a loss function, such as the mean squared error (MSE), which encourages precise alignment with ground truth keypoints. The use of keypoint density maps during training forces the model to focus on learning meaningful spatial distributions of keypoints, thereby improving the accuracy and expressiveness of generated dance motions.

3.3. Leveraging the Latent Diffusion Framework

Figure 1c illustrates our approach to leveraging the latent diffusion framework by combining a Music-to-Pose Encoder (M2PEnc) with a denoising 3D U-Net. The M2PEnc maps the input music sequence into a sequence of poses, which serves as conditioning input to the U-Net. This pose sequence is designed to match the input feature shape of the denoising 3D U-Net, allowing for seamless integration by summing the pose guidance with the latent features of the U-Net. This alignment ensures that the motion guidance, derived from the pose sequence, effectively captures the rhythmic and structural characteristics of the music.

In parallel, a reference human image is processed using a frozen Variational Autoencoder (VAE) and CLIP image encoder to extract latent embeddings. These embeddings are then processed by the reference U-Net, which encodes appearance and background information. The weights of the reference U-Net are connected to the denoising 3D U-Net, serving as additional conditioning to ensure visual fidelity of appearance and background throughout the synthesis process of the dance video.

The denoising 3D U-Net integrates both the pose sequence guidance and reference embeddings through multi-level attention mechanisms:

- **Cross-attention**: Aligns the pose sequence guidance with reference embeddings, ensuring that pose features correspond accurately to visual attributes derived from the reference image.
- **Self-attention**: Refines spatial coherence within each frame, maintaining local consistency in pose and appearance.
- **Temporal attention**: Enforces smooth transitions across video frames, preserving temporal coherence throughout the sequence.

Finally, the denoised latent representations are passed through the frozen VAE decoder to reconstruct high-quality video frames.

In our work, we leverage the methods proposed in CHAMP [15]. By reusing CHAMP's pretrained weights alongside the stable diffusion's denoising 3D U-Net architecture, we can maintain high-quality synthesis results without extensive retraining, thereby streamlining the process and enhancing practicality for real-world applications.

3.4. Training and Inference

The training process for our model is divided into three distinct stages, each designed to progressively refine the system's ability to generate high-quality dance motions synchronized with music. Notably, the first stage introduces the method to train M2PEnc proposed in this paper, while the second and third stages leverage established methods and weights with some minor changes for computational efficiency.

In the first stage, M2PEnc is pretrained using synthetic datasets to establish its ability to map musical features into structured spatial representations. The training objective minimizes two key loss functions: the guidance loss, L_{guid}, which quantifies the difference between the predicted and ground truth pose guidance features,denoted as $\hat{f}$ and f respectively, and the heatmap loss, L_{heat}, which measures the alignment between predicted keypoint heatmaps $\hat{k}$ and their corresponding ground truth k. These losses are designed to ensure that M2PEnc effectively captures the relationship between music and pose structure:

$$L_{\text{guid}} = \frac{1}{HWC_{\text{guid}}} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C_{\text{guid}}} |f_{h,w,c} - \hat{f}_{h,w,c}| \tag{392}$$

393

$$L_{\text{heat}} = \frac{\exp(\hat{k}_{h,w,c})}{HWC_{\text{heat}}} - k_{h,w,c}. \tag{394}$$

In the second stage, training is performed on individual video frames. The weights of the M2PEnc remain frozen, and the temporal layer of the Denoising U-Net is temporarily excluded. For each image in the synthetic dataset, there exists a corresponding Jukebox frame. The U-Net and Reference Encoder are optimized to predict the selected video frame computed from a corresponding Jukebox frame by minimizing a pixel-wise Mean Squared Error (MSE) loss between the predicted frame x and a randomly selected image y. This loss is defined as:

$$L_{\text{pix}} = \text{MSE}(x, y). \tag{1}$$

The Denoising U-Net and ReferenceNet are initialized with pre-trained weights from CHAMP [15], while the weights of the VAE's Encoder and Decoder, the CLIP image encoder, as well as the temporal layer remain fixed throughout this stage.

Finally, in the third stage, we fine-tune the temporal layer and M2PEnc jointly by combining the pixel-wise loss from the U-Net (L_{pix}) and the guidance loss from M2PEnc (L_{guid}), as defined below:

$$L_{\text{total}} = 0.5 \cdot L_{\text{pix}} + 0.5 \cdot L_{\text{guid}}.$$

The weights of the remaining network components are kept fixed during this stage. This joint optimization ensures that the temporal layer and M2PEnc work cohesively, enabling the model to generate rhythmically aligned and visually coherent dance motions.

During inference, due to memory constraints, we generate dance videos in multiple short segments of 5 seconds each. To ensure temporal continuity across segments, we adopt UNIAnimate's noise-conditioning method [22], which effectively merges these segments into seamless full-length videos.

4. Experiments and Results

4.1. Implementation Details

STM2PE-Diff was trained using eight A100 GPUs over multiple stages. Initially, the M2PEnc module was trained for over 20,000 epochs with a batch size of 24. Each batch consisted of a sequence of 75 frames, which corresponds to a duration of $D = 5$ seconds at a frame rate of $F = 15$ fps. The synthetic dataset described in Section 3.1 was utilized for training this module.

Subsequently, in the second stage, video frames were sampled, resized, and center-cropped to a resolution of 640×640. For each frame, a corresponding Jukebox feature frame was used as input to the M2PEnc module. This phase involved training for 60,000 steps with a batch size of 32.

To enhance temporal coherence, during the third stage, the temporal layer was trained using 24-frame video sequences with a batch size of 8 for 20,000 steps. Throughout all training processes, the AdamW optimizer [13] was employed with a learning rate of 1×10^{-5} to optimize the network.

4.2. Evaluation Metrics

To evaluate the quality of generated videos, we utilize the Fréchet Video Distance (FVD) [18] with the I3D classifier pre-trained on Kinetics-400 [1], which quantifies the discrepancy between the generated and the real video distribution. For assessing image-level quality, we use widely-used image metrics, namely SSIM [23], PSNR [10], and LPIPS [27].

Table 1. Quantitative evaluation results between STM2PE-Diff to existing methods.

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
STM2PE-Diff	**213.289**	**0.102**	**21.11**	**0.908**

4.3. Result Analysis Video Quality Evaluation.

To assess the effectiveness of our proposed method, we perform comparisons with existing approaches using both quantitative and qualitative metrics. Furthermore, we conduct a human evaluation to validate its efficacy.

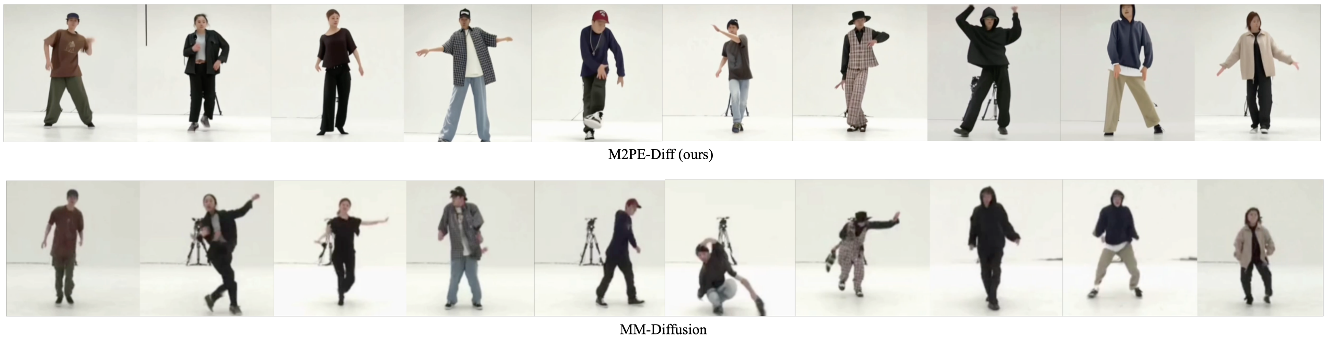
4.3.1. Qualitative Evaluation

AIST test set inputs. For the evaluation of the generated dance video quality, we use audio from 20 video clips and the first frame of each video as the reference image. The reference image is cropped to a box of size 640× 640 centered around the dancer. The predicted videos are presented in Figure 2.

Music-to-dance generation is a relatively new field with limited works for comparison. We compare STM2PE-Diff against MM-Diffusion [17]. MM-Diffusion results, obtained from their project page, demonstrate competence but lack the fine-grained dance realism achieved by our approach. DabFusion [25] samples are of lower quality and were excluded from the qualitative comparison.

Unseen individual inputs. STM2PE-Diff demonstrates exceptional generalization by animating individuals who were not part of the full dataset. Using an in-the-wild reference image as the initial frame, the model generates motion sequences that align seamlessly with the rhythm and style of the input music. As illustrated in Figure 3, the model successfully animates diverse individuals across various dance genres, including Break, House, Ballet Jazz, Street Jazz, LA Hip-Hop, Lock, Krump, Pop, Middle Hip-Hop, and Waack. This evaluation emphasizes how STM2PE-Diff choreographs realistic and stylistically accurate dance movements using only a reference image and music input.

As shown in Figure 2 and 3, STM2PE-Diff outperforms existing methods in video quality, movement realism, and alignment with musical cues. By producing visually appeal-



M2PE-Diff (ours)

MM-Diffusion

Figure 2. **Qualitative evaluation of AIST dataset.** Example dance videos generated by STM2PE-Diff using audio from the AIST test set and the first frame as the reference image. The model showcases its effectiveness in music-to-dance generation by producing high-quality, rhythmically synchronized dance animations across various genres. Click on the figure to view the results.

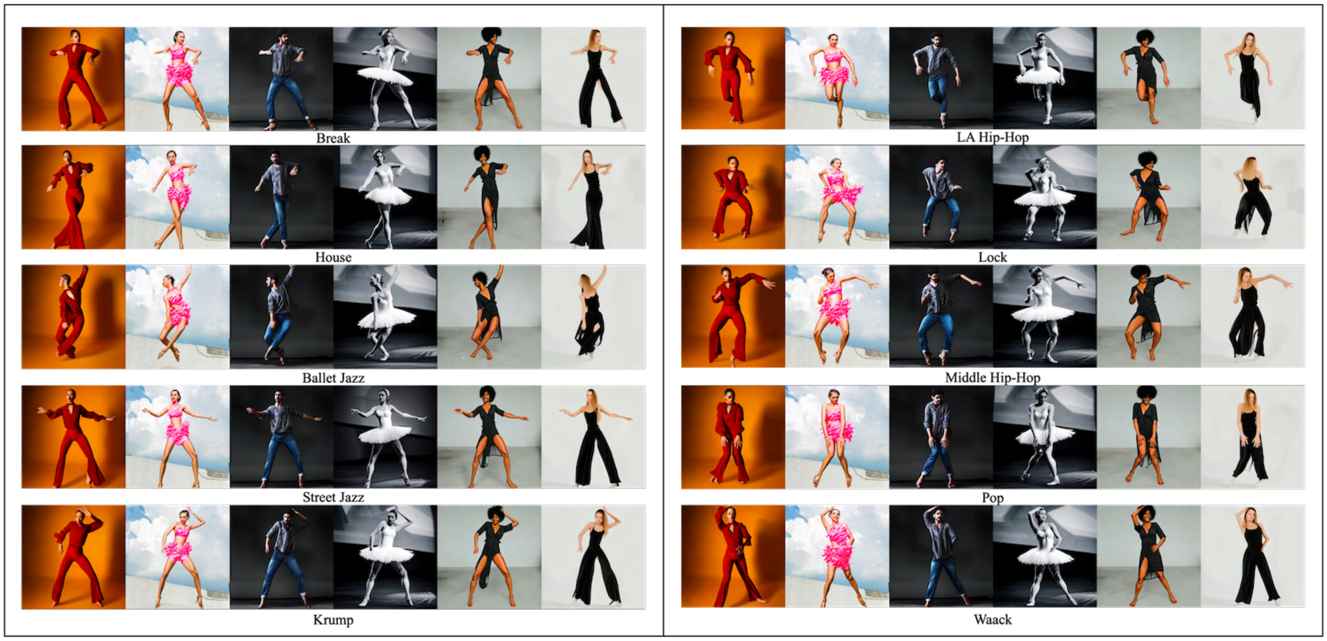


Figure 3. **Qualitative evaluation using in-the-wild images.** Examples of dance videos generated by STM2PE-Diff using in-the-wild reference images. The model demonstrates its ability to generalize well across different individuals and dance genres, producing realistic and rhythmically synchronized dance animations. Click on the figure to view the results.

ing and musically synchronized dance videos, our model sets a new benchmark in music-to-dance generation.

4.3.2. Quantitative Evaluation

The results of the quantitative evaluation are presented in Table 1. For comparison, we randomly generated an equivalent number of frames using MM-Diffusion by providing audio from the AIST test set as input.

Both MM-Diffusion and DabFusion models are trained to generate small-sized video clips; for instance, MM-Diffusion generates 256 × 256 clips with the application of a super-resolution module, while DabFusion generates 128 × 128 clips. In contrast, our model directly generates 640 × 640 videos, matching the size of the reference image.

Additionally, the cropping strategies employed by MM-Diffusion and DabFusion include large portions of the background. In our approach, we focus more on the dancer by cropping a box around the dancer's area and resizing it to 640 × 640 for comparison with our results.

Since the DabFusion codebase is not publicly available, instead of computing its results directly, we scaled its reported metrics based on Table 1 of the DabFusion paper [25].

Our results in Table 1 demonstrate that STM2PE-Diff significantly outperforms both MM-Diffusion and DabFusion across all metrics. The lower FVD and LPIPS values, combined with higher PSNR and SSIM values, indicate that STM2PE-Diff generates more realistic, visually coherent, and high-quality dance videos. This superior performance validates the effectiveness of our Music-to-Pose Encoder and the coupled U-Net architecture in capturing the intricate relationship between music and dance motions.

4.4. User Study

To evaluate the subjective quality of the generated dance videos, we conducted a user study comparing our proposed STM2PE-Diff framework with MM-Diffusion [17]. Participants were asked to assess two key aspects: rhythmic synchronization and visual consistency. Rhythmic synchronization measured how well the dance movements aligned with the rhythm of the input music, while visual consistency evaluated the realism and coherence of the generated videos.

The results of the user study are summarized in Table 2. STM2PE-Diff achieved an 85% preference score for rhythmic synchronization, slightly surpassing MM-Diffusion's 83%. For visual consistency, STM2PE-Diff demonstrated a significant improvement, scoring 80% compared to MM-Diffusion's 55%. These findings highlight the ability of STM2PE-Diff to produce dance motions that are both rhythmically aligned and visually coherent, validating its effectiveness in music-to-dance generation.

Table 2. User Study Results

Model	Rhy. Sync	Vis. Cons.
MM-Diffusion	83%	55%
STM2PE-Diff	**85%**	**80%**

5. Ablation

To evaluate the impact of different pose guidance setups on the performance of STM2PE-Diff, we conducted an ablation study using three configurations: DNS (Depth-Normal-Semantic maps), P (Pose skeleton), and PDNS (Pose skeleton combined with Depth-Normal-Semantic maps). The results are summarized in Table 3.

The DNS setup provides spatial information about body geometry but it lacks explicit skeletal constraints, leading to moderate performance across metrics. The P setup, which uses skeleton keypoints as pose guidance, improves upon DNS in all metrics. This suggests that skeletal constraints contribute significantly to the alignment and coherence of generated poses. Finally, the PDNS setup combines skeleton keypoints with depth-normal-semantic maps, yielding the best results across all metrics.

The ablation study highlights the importance of combining skeletal pose guidance with detailed spatial representations for optimal performance. Future work could further explore advanced fusion strategies for pose guidance inputs to improve robustness and generalization further.

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(STM2PE-Diff)	213.289	0.102	21.11	0.908

6. Limitations

While STM2PE-Diff outperforms existing methods, it still encounters several limitations. Notably, capturing realistic and fine-grained details in facial and hand regions remains a challenge. Furthermore, in certain failure cases, the generated pose guidance does not align with the input reference image, resulting in deformations of the generated individual. Additionally, the background in the generated images often becomes distorted, leading to inconsistencies and artifacts. To address these limitations, future work will focus on synthesizing high-quality, high-definition video datasets and refining the model to improve shape parameter prediction. Moreover, we plan to explore the use of background masking techniques to maintain a consistent and stable background across frames, thereby enhancing the overall coherence and realism of dance animation results.

7. Conclusion

This study presents STM2PE-Diff, a cutting-edge framework for generating dance videos conditioned on music and reference images. By integrating a Music-to-Pose Encoder with a latent diffusion model, we address the inherent challenges of music-to-motion mapping, ensuring rhythmic synchronization and stylistic fidelity. Our synthetic dataset pipeline enhances the model's robustness by removing the ambiguity between music to dance mapping and providing diversity in the training data. The strong quantitative and qualitative results validate STM2PE-Diff's ability to produce high-quality dance motions, setting a new benchmark in music-driven choreography generation. Future research directions could include incorporating higher-resolution outputs, enhancing the accuracy of human shape representation, and improving background rendering. STM2PE-Diff represents a significant advancement in integrating AI with artistic expression, unlocking new avenues for creative innovation.

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)*, pages 6299–6308, 2017. 6

[2] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu,

Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicdance: Realistic human dance video generation with motions & facial expressions transfer. *arXiv preprint arXiv:2311.12052*, 2023. 1, 3

[3] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion, 2024. 1, 2

[4] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *NeurIPS 2023*, 2024. Last Modified: January 29, 2024. 2

[5] Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Animateanything: Fine-grained open domain image animation with motion guidance, 2023. 1, 2

[6] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 2, 3

[7] Haoyang Fang, Boran Han, Shuai Zhang, Su Zhou, Cuixiong Hu, and Wen-Ming Ye. Data augmentation for object detection via controllable diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1257–1266, 2024. 2

[8] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7297–7306, 2018. 2

[9] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 1

[10] National Instruments. Peak signal-to-noise ratio as an image quality metric. *NI Technical Documentation*, 2011. 6

[11] Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, and Karthik Nandakumar. Diffusemix: Label-preserving data augmentation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27621–27630, 2024. 2

[12] 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):248:1–248:16, 2015. 1, 3

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 6

[14] Brian McFee et al. librosa: Python library for audio and music analysis, 2025. Version 0.11.0. 2

[15] 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, 5, 6

[16] Alec Radford et al. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2503.09905*, 2025. 2

[17] 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 *CVPR*, 2023. 6, 8

[18] 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. 6

[19] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. 2

[20] 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)*, pages 448–457, 2023. 1, 3

[21] 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. 3

[22] 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 preprint arXiv:2406.01188*, 2024. 1, 2, 6

[23] 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):600–612, 2004. 6

[24] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[25] 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. 6, 7

[26] 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)*, pages 4212–4222, 2023. 1, 3

[27] 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. 6