MuseDance: A Diffusion-based Music-Driven Image Animation System

Zhikang Dong[1*] Weituo Hao[2] Ju-Chiang Wang[2] Peng Zhang[3†] Pawel Polak[1]

[1]Stony Brook University [2]Bytedance [3]Apple

Abstract

Image animation is a rapidly developing area in multimodal research, with a focus on generating videos from reference images. While much of the work has emphasized generic video generation guided by text, music-driven dance image animation remains underexplored. In this paper, we introduce MuseDance, an end-to-end model that animates reference images using both music and text inputs. By integrating music as a conditioning modality, MuseDance generates personalized videos that not only adhere to textual descriptions but also synchronize character movements with the rhythm and dynamics of the music. Unlike existing methods, MuseDance eliminates the need for explicit motion guidance, such as pose sequences or depth maps, reducing the complexity of video generation while enhancing accessibility and flexibility. To support further research in this field, we present a new multimodal dataset comprising of 3,122 dance videos, each paired with the corresponding background music and text descriptions. Our approach leverages diffusion-based methods to achieve robust generalization, precise control, and temporal consistency, setting a new benchmark for the task of music-driven image animation. The dataset of this work is available at `https://github.com/Dongzhikang/musedance`.

1. Introduction

The field of Artificial Intelligence Generated Content (AIGC) has made significant strides in recent years [3, 10, 21, 22, 44, 48, 52, 61, 62, 75, 80]. In particular, image animation has advanced through the use of various guidance, such as motion [55, 56, 82] and text [15]. Current motion transfer models rely on motion guidance inputs to animate reference images. However, these methods usually fail to align with user preferences when generating a dance video that matches a given piece of music. Finding suitable dance motion guidance for specific music can be challenging. For example, no suitable guidance exists for Mozart or Beethoven. Such guidance is scarce and typically requires specialized domain knowledge that most users lack. Moreover, existing models focus mainly on human motion, limiting creativity, while any object, such as animated characters in Disney movies, could "dance".

To overcome these limitations, we propose a new task: music-driven dance generation, allowing users of all skill levels to create diverse dance videos directly from music, without pose guidance, featuring not only humans but also various objects. Additionally, users can enhance customization through text prompts, allowing them to generate unique and personalized dance videos. This novel approach has significant potential in film, social media, gaming, and education, where AI-generated dance animations can offer interactive and engaging experiences.

Despite its potential, music-driven image animation remains challenging and underexplored: (1) **Absence of explicit motion guidance**—Traditional image animation methods rely on structured motion inputs, such as human pose keypoints, depth maps, or skeleton sequences. However, music does not provide direct motion trajectories. It requires the model to infer realistic dance movements directly from audio cues, beat structure, and rhythm patterns; (2) **Synchronization complexity**—Aligning dance movements with music beats and genre-specific styles is challenging, as each dance form follows distinct movement patterns. For example, hip-hop features sharp, rhythmic motions, while ballet emphasizes fluid, graceful transitions. Ensuring motion remains temporally coherent while adhering to the music's beat structure remains a non-trivial problem; and (3) **Dataset limitations**—While datasets like AIST++ [42] provide dance motion data, they are restricted to predefined choreographies and a limited range of musical styles. In contrast, our dataset includes a wider variety of dance and music styles and features both human and non-human subjects, offering a more diverse and comprehensive resource for music-driven dance generation.

In this paper, we introduce MuseDance, a flexible end-to-end multimodal image animation framework that brings a static reference image to life using a music piece and a text prompt describing the desired motions. As shown in

*Work done during the internship at Bytedance.
†Work done at Bytedance.

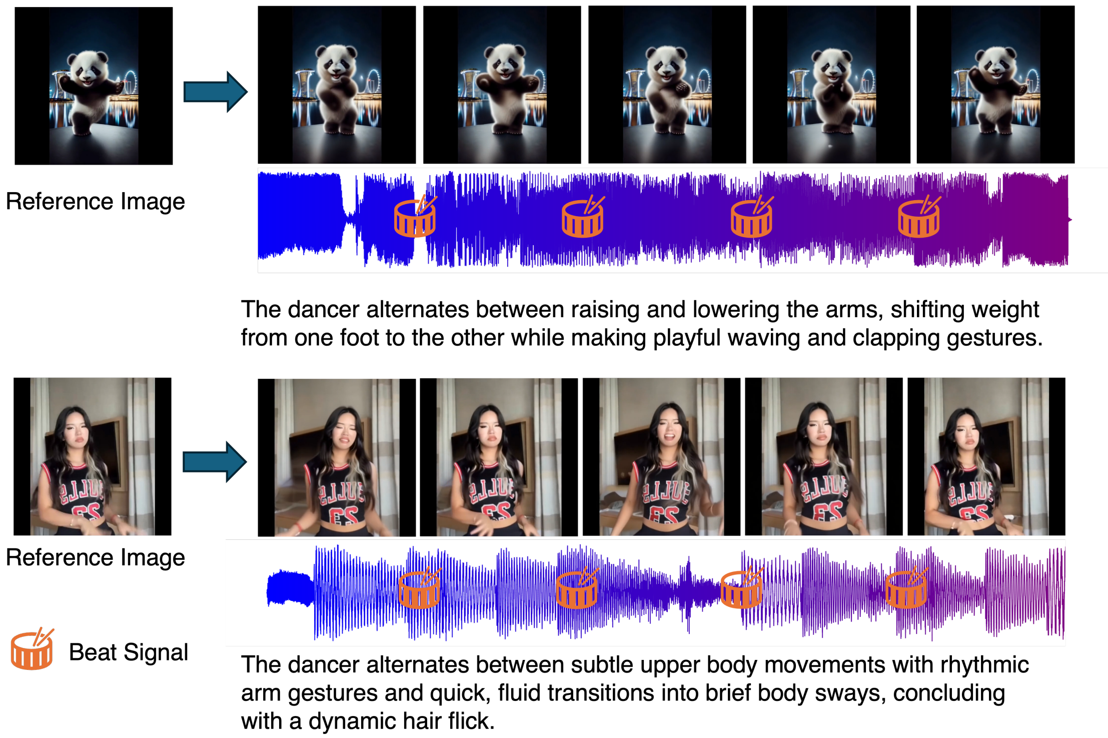

Figure 1. MuseDance generates a dancing video from a reference image, synchronizing movements to the provided music, aligning with the beats, and visually interpreting the guidance of a text prompt for a seamless, music-driven animation.

Figure 1, MuseDance builds on the Diffusion model with pretrained Stable Diffusion weights [6] and incorporates modifications to ReferenceNet [29] to capture spatiotemporal information. Additionally, we design a hierarchical structure that includes a modality fusion mechanism for injecting music embeddings into video sequences, a beat encoder to explicitly align generated videos with the music beat, and motion alignment modules to ensure frame consistency and adherence to input constraints. The model trains in two stages: in the first stage, the model learns to generate single frames by capturing the reference image's appearance, guided by its DensePose [28] representation, and disentangles appearance and motion using text prompts. In the second stage, it generates video sequences guided by music, beat, and motion while keeping the first-stage modules frozen to preserve appearance quality. The music, beat, and motion modules enable semantic alignment of music with video sequences, synchronization with beats, and temporal consistency across frames, resulting in more natural and dynamic animations.

In summary, our contributions are threefold: (1) We propose MuseDance, a novel end-to-end diffusion-based method that uses music and text as driving dynamics to animate the reference image in a way that aligns with the semantic meaning of the input. (2) We introduce a novel music-dance video generation dataset, where each sample includes a dance video, the corresponding background music track, and the textual description of the motion in the video. This dataset trains models to animate reference images by learning motion dynamics from music and text. (3) Our model is able to generate temporally consistent dance videos with multi-modal control across diverse objects ranging from real-human dancers to cartoon figures, despite their vastly different characteristics.

2. Related Work

Video Generation Diffusion Model. Video generation is a very important task in AIGC. Methods like variational RNNs [2, 12, 19, 40] and GANs [43, 47, 54, 59, 67, 68, 78] have been explored to tackle this problem. However, most of those works are limited to low-resolution, the lack of large scale high-quality datasets or loose control ability. Diffusion models are proposed to solve this problem. [7] introduces temporal dimension to the latent diffusion image generation model. Make-A-Video [57] enhances DALL·E2 [51], a text-to-image model, by using joint text-image priors and super-resolution strategies to produce high-quality videos. Stable Video Diffusion [6] presents a large-scale text-to-video foundation model, which also supports various downstream tasks like image-to-video generation, camera motion adaptability and multi-view objects synthesis. In addition to open source models, closed-source video generative models, in particular GEN-2 [25], PikaLabs [39], Sora [9] and Kling [38] provide state-of-the-art video generation capabilities for general use.

Music-guided Dance Generation. Music-guided dance generation in 3D sequences has been explored in recent works. Bailando [58] proposes a pose Q-VAE with a motion GPT to predict future pose tokens given music. EDGE [63] presents a physics-constrained transformer-based diffusion to generate more realistic 3D dance sequences. M2C [45] introduces a music code extractor to replace existing music feature processor to enhance music's role in 3D dance motion generation. LM2D [77] integrates lyrics information to enable the generation of more diverse 3D dance sequences. 3D dance sequence generation models produce only skeleton keypoints, rather than full dance videos, limiting their practical application. Music-guided dance generation in 2D videos is still largely unexplored. [70] utilizes a diffusion model to generate optical flow, which is then combined with a reference image for animation. MusicInfuser [31] takes music and text as input and generate dancing video without reference image. [16] leverages 2D pose and music to animate reference image.

Human Motion Transfer. Earlier works [4, 8, 17, 24, 72] on human body motion transfer demonstrate lower accuracy and require significant human intervention. Recently, deep learning techniques have enabled more realistic motion transfer with highly automated training pipelines. MoCoGAN [65] introduces an unsupervised adversarial training method for transferring motion and facial expressions onto target subjects. [1] extends the StyleGAN [36] generator to learn the warped local features. [13] utilizes a video-to-video synthesis method to generate new motions by giving a 2D video and 2D skeleton sequences. Dreampose [35] proposes a diffusion model to animate a reference human image using pose sequences and fabric textures. [33] utilizes a lightweight pose guider to enable controllable continuous character movement across various downstream tasks.

In human dance transfer domain, DISCO [69] generates human dance videos from dance skeleton sequences and a reference human image. Their method generalizes to unseen human references, backgrounds, and poses. MagicAnimate [74] combines a video encoder with an appearance encoder to generate temporally consistent dance videos from a reference image. MagicPose [14] incorporates facial keypoints with body skeletons as guidance to generate realistic human dance videos. However, these approaches still require pose guidance to generate dance videos, and such pose sequences are not always available, limiting the generalization capability of these models.

3. Method

We propose a two-step training framework to animate images in dancing based on music and text input. In the first step, the model is trained on individual frames from the target video to learn visual features and acquire prior knowledge. In the second step, we introduce music and text as triggers to generate animated frames that align with these inputs. The process is illustrated below.

3.1. Preliminaries

Latent Diffusion Models denote a class of diffusion models that operate within the encoded latent space produced by an autoencoder, represented as $\mathcal{D}(\mathcal{E}(\cdot))$. One of the most widely used models in this category is Stable Diffusion [52], which incorporates a VQ-VAE and a time-conditioned U-Net architecture. Additionally, Stable Diffusion utilizes a text encoder from the CLIP [50] model to encode text prompts into embeddings. Given an image $I \in \mathbb{R}^{H_I \times W_I \times 3}$ and its corresponding text embedding $c_{\text{text}} \in \mathbb{R}^{D_c}$, we obtain the latent representation $z_0 = \mathcal{E}(I) \in \mathbb{R}^{H_z \times W_z \times D_z}$ and apply it to a predefined diffusion process across T timesteps, modeled as a Gaussian process. This process approximates a standard Gaussian distribution, $z_T \sim \mathcal{N}(0, I)$. The training objective in Stable Diffusion is to iteratively denoise z_T back to the original latent representation z_0.

During inference, the original latent z_t is reconstructed using sampling methods, such as denoising diffusion implicit models [60]. Then, the latent z_t is decoded by the decoder $\mathcal{D}$ to generate the final, clear image. Latent Diffusion Models can produce high-fidelity images and align the generated images with the text-conditioned prompt.

Cross Attention Mechanism is a key component in the U-Net of latent diffusion models. This mechanism integrates information from the latent representation and the conditioning embedding, enabling latent diffusion models to generate images that semantically align with the given condition. More generally, the conditioning modality can be text, motion flows, audio, etc., providing semantic guidance for the generation process.

3.2. Appearance Pretraining

As shown in Figure 2, the first training stage aims to generate motion consistent with the reference image while preserving appearance. From a video $V = I_1, \ldots, I_N$, we sample an input frame I_i within $(w, N - w)$ and a target frame I_j within $(i - w, i + w)$, with $j \neq i$. To emphasize the foreground object, we use DensePose [28] as a segmentation mask rather than motion guidance. Unlike pose-based methods that rely on skeletal keypoints, DensePose highlights the object against the background in the reference image, guiding the model to learn structure and appearance without predefined motion constraints. We obtain the DensePose mask of the reference image D_i and encode

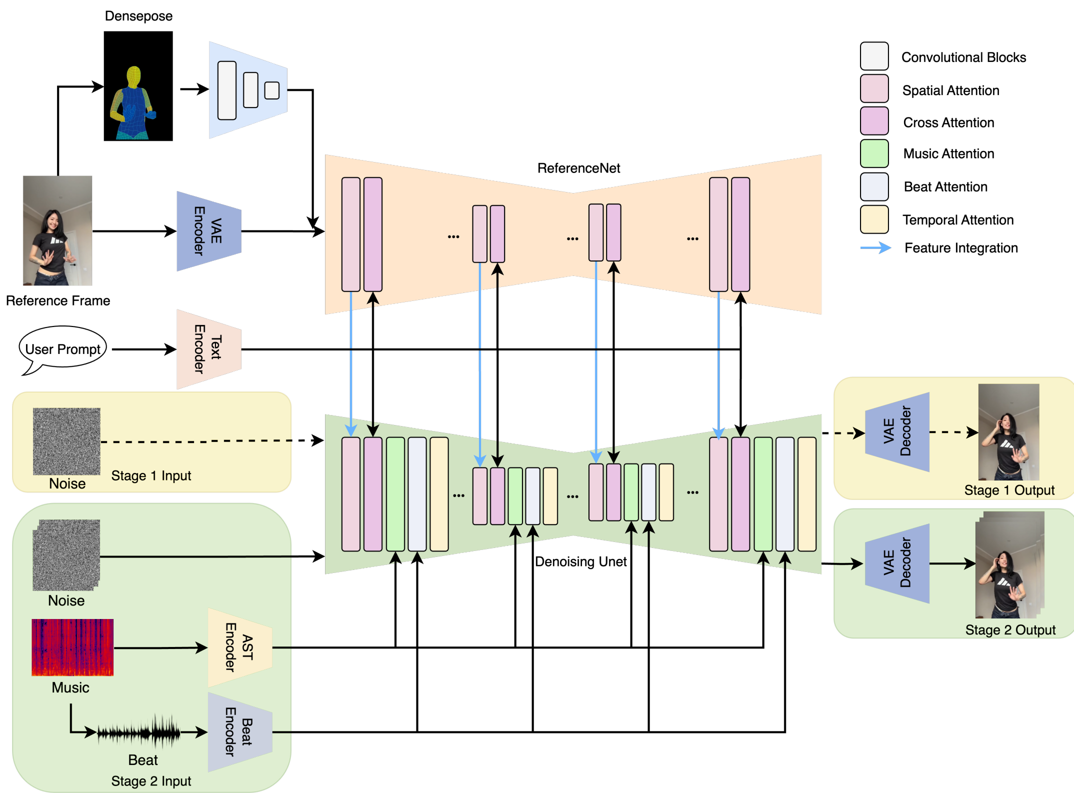

Figure 2. In the first training stage, we train the model to capture spatial information by generating individual frames, with reference and target frames randomly sampled from a short time window. DensePose is used to help the model focus on the object, while text prompts assist in understanding motion. In the second training stage, we freeze the spatial attention blocks to preserve the model's frame generation ability and introduce music, beat, and motion modules to incorporate music dynamics, align with the beat, and improve frame-to-frame consistency.

it with F_m, a convolutional network that downscales and extracts mask features, producing $m_i = F_m(D_i)$. As in ControlNet [79], we add these features residually to sharpen object focus while maintaining flexibility. This use of DensePose keeps the model motion-guidance-free, avoiding constraints from explicit poses or keypoint trajectories.

$$z_0 = \mathcal{E}(I_i) + \text{Conv}(D_i), \tag{1}$$

Following [33, 73], we use ReferenceNet—a U-Net-based Stable Diffusion model with the same layers as our backbone—to extract visual features from the reference image R. Let $x_d \in \mathbb{R}^{H_z \times W_z \times d}$ be features from the denoising U-Net and $x_R \in \mathbb{R}^{H_z \times W_z \times d}$ from ReferenceNet. We concatenate them along the W dimension, apply spatial self-attention, and retain the first half as output. This output is then used in cross-attention with CLIP text features, $\mathrm{CrossAttn}(z_t, c_{\text{text}})$, enabling the model to capture pose and appearance with semantic guidance.

Because ReferenceNet shares the structure of the denoising U-Net, its feature maps integrate seamlessly to enrich generated frames with detailed foreground and background information. Unlike ControlNet, which enforces alignment between reference and target frames, our reference and target are separate images from a short time window, meaning they share only spatially related features rather than alignment. Hence, ControlNet is unsuitable for our setting.

3.3. Dynamic Trigger Video Generation

Figure 2 illustrates the second stage of our training process. Here, the model learns to generate dance videos based on the reference image, music input, and text guidance. To preserve the visual generation ability from the appearance pretraining stage, we freeze the spatial attention blocks.

To achieve temporal alignment in driving the reference image, we add three new modules to the denoising U-Net: a music understanding module, a beat alignment module, and a motion alignment module.

Music Understanding Module. This module aims to extract musicality information from the music and use it to control frame generation. We use the Audio Spectrogram Transformer (AST) [26] to obtain music embeddings, as it effectively captures higher-level musical semantics such as genre, style, and mood [11, 23, 27, 41]. Given the hidden states from the previous module, $z_t \in \mathbb{R}^{K\times(H_zW_z)\times d}$, where K is the number of generated frames, and the music embedding $c_{\text{music}} \in \mathbb{R}^{L\times d}$, where L is the sequence length of the music embedding, we apply a cross-attention mechanism between the music embeddings and frames to facilitate information flow across these two modalities, allowing the music dynamics to control frame generation. To further improve temporal alignment, we reshape the hidden states into $z_t \in \mathbb{R}^{(H_zW_z)\times K\times d}$ and compute self-attention along the temporal dimension.

Beat Alignment Module. We observe that, in most music dance videos, the music beat serves as a strong signal, often marking the start, stop, or change in dance style. To capture this pattern, we incorporate beat information into the denoising U-Net. We use Librosa to identify beat locations in the music, converting them into a one-hot encoded format. This produces a binary vector $b_{\text{binary}} \in \mathbb{R}^{K}$, where frames with a beat are assigned a value of 1, and others a value of 0. We align the beat information with video sequences, inspired by token processing in NLP tasks, and apply a lookup embedding layer to transform the discrete embedding into a continuous dense embedding $b_{\text{dense}} \in \mathbb{R}^{K\times d}$. We then apply the cross-attention mechanism to help the hidden states learn the beat information. Similar to the music understanding module, we reshape the hidden states and apply temporal attention layers to ensure temporal continuity.

Motion Alignment Module. In video generation, maintaining content continuity across frames is crucial, especially for generating coherent dance motions. In addition to the temporal attention layers in the music understanding and beat alignment modules, we employ a motion alignment module to capture temporal dependencies across frames. Inspired by [29, 73], we use several previously generated frames as guidance, concatenating them with the current hidden states and performing self-attention across the temporal sequence dimension. Specifically, we form a concatenated hidden state $z_{\text{motion}} = \text{concat}(z'_t, z_t)$, where $z'_t \in \mathbb{R}^{(H_zW_z)\times M\times d}$ represents the hidden states from the previous M generated frames. By applying self-attention on z_{motion} across the temporal axis, we select the last K hidden states as the current generated frames.

Instead of simply concatenating embeddings, we sequentially integrate them to ensure coherence between the music, beat, and motion modules. First, the music understanding module encodes high-level semantic information, which implicitly captures rhythm and maintains temporal continuity. Building on this, the beat alignment module refines the temporal structure by anchoring motion changes to beat locations, ensuring precise rhythm synchronization. Finally, the motion alignment module enhances temporal consistency between frames, aligning motion with both the music semantics and beat locations to smooth transitions and maintain natural motion flow.

4. Experimental Results

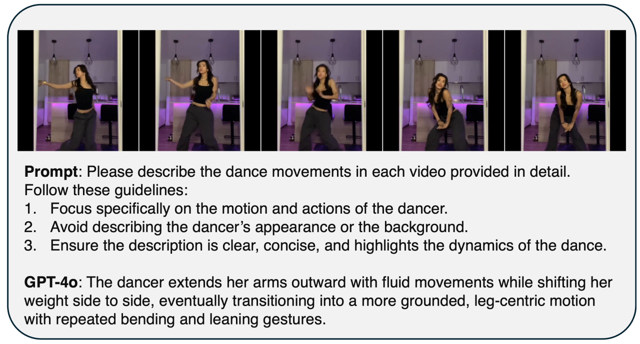

Figure 3. An example of textual data generation, we provide a series of frames and a detailed prompt to instruct GPT-4o to generate motion captions.

4.1. Music Dance Dataset

In this paper, we introduce the first music-dance video dataset. The raw videos are collected from YouTube, totaling 304 videos. These videos feature a diverse range of dance genres, from popular styles found on short video platforms to traditional Chinese dance. We manually collect these videos to ensure that each one features a single object-centric dance, with a clear screen, minimal special effects, and a stable camera angle that is generally front-facing but diverse in perspective. Since all data is manually collected, we make sure that the videos are of high quality and free from harmful content. The videos also provide diversity in dancer appearance and background settings. Given that dancers often perform to the same music, some music tracks overlap across videos. Human-object videos are primarily from TikTok dance collections or dance challenge series, while non-human object videos consist of synthetic animations of animals dancing. We split each video into multiple clips. All these sub-videos are in a vertical, object-centered format, paired with background music, and vary in length from 10 to 30 seconds. We manually edit each video to remove irrelevant segments, such as intros, outros, and conversational parts, retaining only the music-driven dance sections. To ensure consistency, we include only videos with a single dancer, with plans to expand the dataset to include multiple-dancer videos in the future. Following these preprocessing steps, our dataset comprises 3,122 videos, each paired with background music and lasting approximately 4 seconds. The dataset includes a total of 489 unique music tracks. All dance videos undergo manual review to ensure that any harmful or inappropriate content is excluded.

We also include a text description of motion for each pair of music and video. Figure 3 illustrates our process for gen-

Dataset	Videos	Music	Text
Everybody Dance Now [13]	105 Videos	✗	✗
TikTok [34]	350 Videos	✗	✗
AIST++[1] [42]	1408 Videos	60 Songs	✗
MuseDance (ours)	3122 Videos	489 Songs	3122 Captions

Table 1. Current Music Dance Dataset Comparison.

erating these descriptions. We leverage OpenAI's GPT-4o API to generate video captions, sampling each video every 10 frames and combining these samples with a text prompt for GPT-4o. To separate motion from appearance, we instruct GPT-4o to focus only on motions and actions, ignoring the dancer's appearance and background. Under this setup, each sample in our dataset includes a triplet: a short dancing video, background music, and a motion description text.

In Table 1, we compare existing music-dance datasets with our own. The Everybody Dance Now [13] and TikTok [34] datasets contain only the video modality and have significantly smaller dataset sizes. The AIST++ [42] dataset, a subset of the AIST [64] dataset, includes both video and music modalities. However, it is limited to 10 dance genres and 60 music tracks, resulting in less diversity in music and videos compared to our dataset. Its backgrounds are simple and clear, lacking the richness and realism needed for diverse dance videos. Using an LLM to generate text descriptions won't resolve the lack of dataset diversity. LLM-generated captions remain limited, as these datasets lack the dance styles and real-world variations needed to enhance motion diversity. Our dataset includes a wider range of dance styles—such as street dance, traditional dance, and hand gesture dance—and features both professional and non-professional dancers, making it more representative. Additionally, while AIST++ mostly features simple studio backgrounds and Everybody Dance Now and TikTok Dance don't have natural music-video pairings, our dataset unifies diverse dance motions, synchronized music, and real-world settings, making it a richer resource for music-driven dance generation.

4.2. Implement Details

In our experiments, both training and inference processes are conducted on a computational platform with 32 NVIDIA A100 GPUs, each has 80 GB memory. The training framework consists of two stages, each comprising 30,000 steps. The batch size is set to 12 in the first stage and 2 in the second stage for each GPU, with video dimensions maintained at 640×640 pixels. During the second stage of training, each instance generates a 4-second video at a frame rate of 12 FPS. The music has a sample rate of 16,000 Hz and is in mono. To ensure consistency in the generated content, the hidden states of the last two generated frames are utilized within the motion module. A learning rate of 1e-5 is applied across both training stages, and the Adam [37] optimizer is employed for parameter updates. The ReferenceNet and Denoising U-Net are initialized based on `stable-diffusion-v1-5`, while the motion module is initialized with weights derived from Animatediff [29]. To enhance video generation quality, a dropout rate of 0.05 is applied. Additionally, we use DDIM to sample the generated frames.

4.3. Quantitative Results

Similar to the approach in [14, 69, 74], we randomly select 228 videos as the test split, including various figures, such as human and non-human ones. We evaluate the quality of our generated dancing videos using several metrics. For single-frame quality, we employ SSIM [71], LPIPS [81], and PSNR [32]. To assess overall video quality, we use Fréchet Video Distance (FVD) [66].

Due to the lack of open-source code and publicly available datasets for music-driven video generation, direct comparison remains challenging. To address this, we evaluate baselines from two complementary perspectives: (1) Two-stage baselines. We use EDGE [63] to generate motion sequences from music. These sequences are projected into 2D and then rendered into videos using DISCO [69]. We further adapt recent 3D pose–driven renderers, Champ [83] and MIMO [46], which synthesize videos from generated 3D pose. While these pipelines approximate music-to-dance generation, their decoupled design prevents joint optimization of motion and appearance. (2) One-stage baselines. We evaluate MM-Diffusion [53] and MusicInfuser [31], fine-tuning both on our MuseDance dataset for fairness. While MM-Diffusion is a multimodal model for audio-conditioned video, MusicInfuser generates video from text and music without using a reference image, so it is related but not directly comparable to our setting. We run each test three times with different seeds and report mean$\pm$std for all metrics. The comparison results are presented in Table 2. Our findings indicate that MuseDance outperforms all baselines across both music–video alignment and video generation quality. Among one-stage baselines, MM-Diffusion underperforms relative to all approaches, reflecting its lack of a dedicated design for dance video synthesis and difficulty handling complex appearance and motion. MusicInfuser performs better on some image-level aspects but still lags behind on temporal consistency and alignment compared to MuseDance, as it does not use a reference image, lacks any explicit beat modeling, and was developed for a broader video generation task rather than identity-preserving, rhythm-synchronized image animation. Turning to two-stage pipelines, EDGE+MIMO

[1]For a fair comparison, we only consider videos filmed from a front-facing camera perspective with a single dancer.

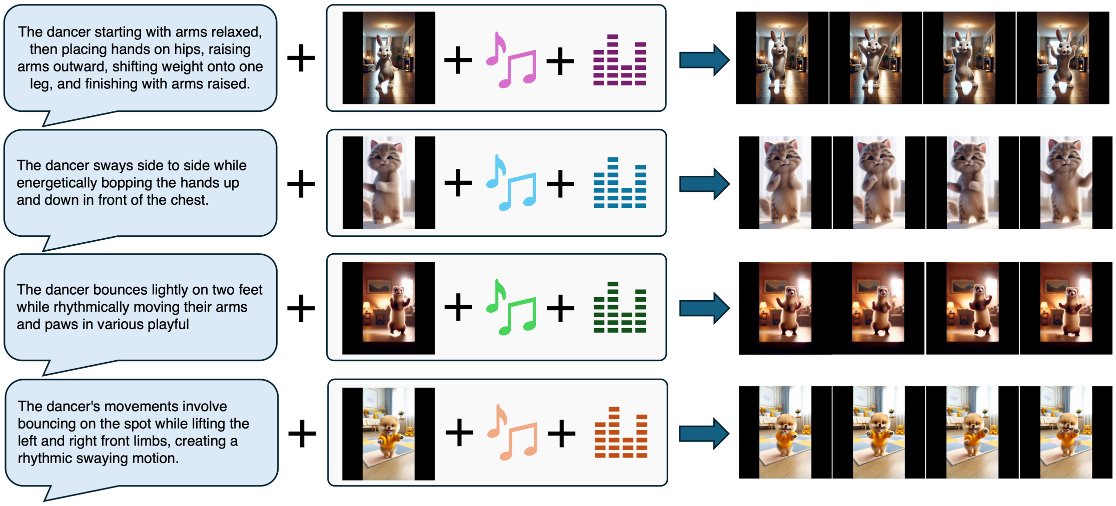

Figure 4. Music driven dancing video generation on non-human objects.

and EDGE+Champ improve over EDGE+DISCO in image quality, yet all remain constrained by their reliance on independently trained components, where motion generation and video rendering cannot be jointly optimized. By contrast, MuseDance unifies music, beat, and motion cues within a single diffusion framework, leading to higher fidelity, stronger temporal coherence, and superior music–motion synchronization.

Method	Image			Video
	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
EDGE + DISCO	27.62±0.12	0.601±0.011	0.276±0.008	401.95±8.21
EDGE + MIMO	28.95±0.09	0.651±0.010	**0.261±0.009**	356.31±7.36
EDGE + Champ	28.43±0.08	0.629±0.012	0.269±0.010	372.42±9.42
MM-Diffusion	27.79±0.11	0.469±0.013	0.299±0.010	576.95±11.34
MusicInfuser	28.82±0.10	0.658±0.009	0.265±0.008	352.18±7.88
MuseDance (ours)	**29.57±0.07**	**0.675±0.008**	0.266±0.007	**315.76±6.57**

Table 2. Model performance on MuseDance (mean±std over 3 runs). Higher PSNR/SSIM and lower LPIPS/FVD are better.

To assess the alignment between our generated videos and the background music, we evaluate the synchronization between kinematic beats extracted from the videos and the amplitude envelope of the accompanying music. Following the method proposed in [18], we extract 2D body keypoints from all frames and compute kinematic beats based on keypoint movement. We use three metrics for evaluation: (1) Mean Euclidean distance error between kinematic and music beats, (2) Median Euclidean distance error, and (3) the Audio-Video Alignment Score (AV Align) proposed in [76]. The results, summarized in Table 3, demonstrate that our method outperforms all baselines. This can be attributed to our explicit integration of beat and rhythm information as temporal guidance during training. Notably, MM-Diffusion achieves the lowest alignment performance among all approaches, as it primarily captures high-level semantic information from audio while neglecting beat and rhythm details. MusicInfuser performs well on alignment metrics due to its joint music–text conditioning, which helps capture rhythm in an end-to-end framework. However, it lacks explicit beat information, leading to weaker performance compared to our method. To further demonstrate alignment, especially for non-human objects, we use MemFlow [20] to extract optical flow and visualize the correlation between motion magnitude and the audio signal, with examples in the supplementary material.

	Mean Distance Error ↓	Median Distance Error ↓	AV Align ↑
Ground Truth	0.218±0.003	0.170±0.001	0.188±0.002
EDGE+DISCO	0.287±0.009	0.204±0.006	0.154±0.005
EDGE+MIMO	0.274±0.008	0.196±0.006	0.162±0.005
EDGE+Champ	0.283±0.010	0.201±0.007	0.158±0.006
MM-Diffusion	0.538±0.012	0.334±0.010	0.099±0.007
MusicInfuser	0.252±0.007	0.182±0.006	0.171±0.005
MuseDance (ours)	**0.234±0.007**	**0.173±0.005**	**0.179±0.004**

Table 3. Music–video alignment metrics (mean±std over 3 runs). Lower distance errors and higher AV Align indicate better alignment.

4.4. Ablation Studies

To illustrate the effectiveness of each module in the second training stage, we conduct ablation studies by removing the music, motion, or beat module and analyzing their impact on output quality. As shown in Table 4 , when all three modules are removed, the model produces the lowest-quality outputs, with weak perceptual similarity and poor temporal consistency. Introducing the music module alone provides valuable semantic guidance, enriching the generation process by aligning motion with musical content. However, since the AST encoder does not explicitly encode temporal music dynamics, the music module alone does not significantly enhance temporal coherence. Combining the motion module with the music module leads to a substantial improvement in both visual quality and temporal smoothness, as the motion module structures movement over time while the music module provides high-level semantic information. Additionally, the temporal attention layers in the motion module help identify temporal patterns within the music embeddings, allowing the model to better capture rhythmically coherent motion. The best results are achieved when all three modules—music, motion, and beat—are included, with the beat module further refining synchronization to ensure that generated movements align with rhythm while preserving expressive variation. These findings highlight the modules' complementary roles: the music module shapes content, the motion module ensures temporal structure, and the beat module refines rhythmic alignment, creating more natural and dynamic dance generation.

Module			Metrics			
Music	Motion	Beat	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
✗	✗	✗	24.49±0.11	0.608±0.010	0.295±0.009	729.65±10.24
✓	✗	✗	24.54±0.10	0.623±0.009	0.284±0.008	608.77±7.57
✓	✓	✗	28.89±0.09	0.671±0.008	0.273±0.007	391.82±6.92
✓	✓	✓	**29.57±0.07**	**0.675±0.008**	**0.266±0.007**	**315.76±6.57**

Table 4. Ablation results on removing modules in the second training stage (mean±std over 3 runs).

4.5. Qualitative Results

Non-human Object Generation. Unlike existing works, our model has the capability to generate dancing videos of non-human objects. As shown in Figure 4, our model produces realistic dancing motions for non-human objects based on the music input and tempo. We observe that regardless of whether the text description provides detailed motion instructions or just generalized guidance, the model still performs well. This demonstrates the strong language understanding ability of our model.

Text Semantic Preservation. We evaluate our model's semantic consistency by controlling the input text prompt. Figure 5 shows animations of different reference images using the same text guidance but varying music dynamics. Our results show that the model accurately follows text guidance and adapts flexibly to music, generating coherent dance videos.

Human Evaluation. We conduct a human evaluation with 23 participants to assess our generated videos on three aspects: Quality, which measures clarity and visual fidelity; Consistency, which reflects motion smoothness and coherence; and Alignment, which evaluates synchronization with background music. Each aspect is rated on a 1–5 Mean Opinion Score (MOS) scale, with detailed instructions provided to participants. Table 5 shows that MuseDance outperforms the baselines in all three aspects. For baselines, we remove input modalities during inference, excluding music and beat embeddings. Music input improves alignment

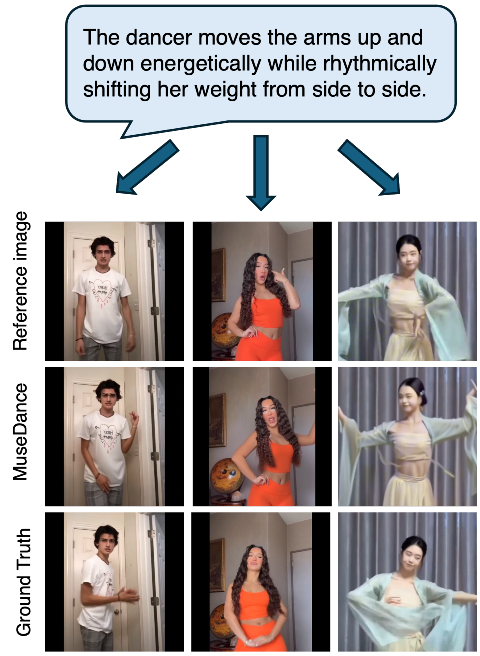

Figure 5. Dance video generations with the same text prompt but different reference images and music. Frames are shown at matching time points from both the generated videos and ground truth.

through rhythmic cues, while text input enhances consistency by guiding motion. Rating criteria are provided in the supplementary material.

Method	Quality	Consistency	Alignment
EDGE + DISCO [63, 69]	3.28	2.79	3.33
EDGE + MIMO [46]	3.87	3.02	3.39
EDGE + Champ [83]	3.59	2.95	3.48
MM-Diffusion [53]	3.68	3.31	3.19
MusicInfuser [31]	3.73	3.34	3.28
MuseDance w/o Music & Text	3.39	2.56	3.02
MuseDance w/o Music	3.49	3.17	3.31
MuseDance w/o Text	3.74	3.28	4.06
MuseDance (ours)	**3.92**	**3.84**	**4.18**

Table 5. Human evaluation results for three perspectives: Quality, Consistency, and Alignment, scored on a 1-5 MOS scale.

5. Conclusions

This work presents an end-to-end framework for animating static images into dance videos using only music dynamics and text guidance. We introduce the first music-driven dance video dataset from public YouTube videos and propose a diffusion-based model that fuses visual, auditory, and textual inputs. Our model generates realistic, synchronized dance videos and shows strong performance on this music driven image animation task. Future work will enhance temporal guidance by adding precise timeline annotations to improve motion coherence in longer sequences.

References

[1] Badour Albahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Transactions on Graphics (TOG)*, 40(6):1–11, 2021. 3

[2] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017. 2

[3] Apoorva Beedu, Zhikang Dong, Jason Sheinkopf, and Irfan Essa. Mamba fusion: Learning actions through questioning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 1

[4] Thaddeus Beier and Shawn Neely. Feature-based image metamorphosis. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 529–536. 2023. 3

[5] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011. 1

[6] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[7] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2

[8] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 715–722. 2023. 3

[9] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. https://openai.com/research/video_generation_models_as_world_simulators, 2024. Accessed: 2024-08-22. 3

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[11] Fred Bruford, Frederik Blang, and Shahan Nercessian. Synthesizer sound matching using audio spectrogram transformers. *arXiv preprint arXiv:2407.16643*, 2024. 5

[12] Lluis Castrejon, Nicolas Ballas, and Aaron Courville. Improved conditional vrnns for video prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7608–7617, 2019. 2

[13] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5933–5942, 2019. 3, 6

[14] 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 *Forty-first International Conference on Machine Learning*, 2023. 3, 6

[15] Xi Chen, Zhiheng Liu, Mengting Chen, Yutong Feng, Yu Liu, Yujun Shen, and Hengshuang Zhao. Livephoto: Real image animation with text-guided motion control. In *European Conference on Computer Vision*, pages 475–491. Springer, 2025. 1

[16] Zeyuan Chen, Hongyi Xu, Guoxian Song, You Xie, Chenxu Zhang, Xin Chen, Chao Wang, Di Chang, and Linjie Luo. X-dancer: Expressive music to human dance video generation. *arXiv preprint arXiv:2502.17414*, 2025. 3

[17] German KM Cheung, Simon Baker, Jessica Hodgins, and Takeo Kanade. Markerless human motion transfer. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, pages 373–378. IEEE, 2004. 3

[18] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 7

[19] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International conference on machine learning*, pages 1174–1183. PMLR, 2018. 2

[20] Qiaole Dong and Yanwei Fu. Memflow: Optical flow estimation and prediction with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19068–19078, 2024. 7, 2

[21] Zhikang Dong and Paweł Polak. Cp-pinns: Data-driven changepoints detection in pdes using online optimized physics-informed neural networks. In *2024 Conference on AI, Science, Engineering, and Technology (AIxSET)*, pages 90–97. IEEE, 2024. 1

[22] Zhikang Dong, Juni Kim, and Paweł Polak. Mapping the invisible: Face-gps for facial muscle dynamics in videos. In *2024 IEEE First International Conference on Artificial Intelligence for Medicine, Health and Care (AIMHC)*, pages 209–213. IEEE, 2024. 1

[23] Zhikang Dong, Xiulong Liu, Bin Chen, Pawel Polak, and Peng Zhang. Musechat: A conversational music recommendation system for videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12775–12785, 2024. 5

[24] Efros, Berg, Mori, and Malik. Recognizing action at a distance. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 726–733. IEEE, 2003. 3

[25] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 3

[26] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. 4

[27] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10699–10709, 2022. 5

[28] Rıza 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*, pages 7297–7306, 2018. 2, 3

[29] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2, 5, 6

[30] Umut Güçlü, Jordy Thielen, Michael Hanke, and Marcel A. J. van Gerven. Brains on beats, 2016. 1

[31] Susung Hong, Ira Kemelmacher-Shlizerman, Brian Curless, and Steven M Seitz. Musicinfuser: Making video diffusion listen and dance. *arXiv preprint arXiv:2503.14505*, 2025. 3, 6, 8

[32] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 6

[33] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 3, 4

[34] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762, 2021. 6

[35] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22623–22633. IEEE, 2023. 3

[36] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[37] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[38] Kuaishou. Kling video model. `https://kling.kuaishou.com/en`, 2024. 3

[39] Pika Labs. `https://www.pika.art/`. 3

[40] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018. 2

[41] Kang Li, Yan Song, Li-Rong Dai, Ian McLoughlin, Xin Fang, and Lin Liu. Ast-sed: An effective sound event detection method based on audio spectrogram transformer. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 5

[42] 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*, pages 13401–13412, 2021. 1, 6

[43] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[44] Xiulong Liu, Zhikang Dong, and Peng Zhang. Tackling data bias in music-avqa: Crafting a balanced dataset for unbiased question-answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4478–4487, 2024. 1

[45] Matthew Marchellus and In Kyu Park. M2c: Concise music representation for 3d dance generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3126–3135, 2023. 3

[46] Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Mimo: Controllable character video synthesis with spatial decomposed modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21181–21191, 2025. 6, 8

[47] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1789–1798, 2017. 2

[48] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[49] Jordi Pons and Xavier Serra. musicnn: Pre-trained convolutional neural networks for music audio tagging. *arXiv preprint arXiv:1909.06654*, 2019. 1

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[51] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3

[53] 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*, pages 10219–10228, 2023. 6, 8

[54] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping.

In *Proceedings of the IEEE international conference on computer vision*, pages 2830–2839, 2017. 2

[55] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 1

[56] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021. 1

[57] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[58] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2022. 3

[59] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3626–3636, 2022. 2

[60] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[61] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022. 1

[62] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

[63] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 3, 6, 8

[64] 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 *ISMIR*, page 6, 2019. 6

[65] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018. 3

[66] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

[67] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033*, 2017. 2

[68] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016. 2

[69] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9326–9336, 2024. 3, 6, 8

[70] Xuanchen Wang, Heng Wang, Dongnan Liu, and Weidong Cai. Dance any beat: Blending beats with visuals in dance video generation. *arXiv preprint arXiv:2405.09266*, 2024. 3

[71] 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

[72] Feng Xu, Yebin Liu, Carsten Stoll, James Tompkin, Gaurav Bharaj, Qionghai Dai, Hans-Peter Seidel, Jan Kautz, and Christian Theobalt. Video-based characters: creating new human performances from a multi-view video database. In *ACM SIGGRAPH 2011 papers*, pages 1–10. 2011. 3

[73] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Luc Van Gool, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. 4, 5

[74] 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*, pages 1481–1490, 2024. 3, 6

[75] Chenxi Yao, Qishi Zhan, Zeyu Cao, Danping Li, Yangfan Lin, Yuanxun Shao, Lin Wang, Zhao Wang, Jiaqing Zhang, Yunfei Zhang, et al. Generative ai for simulating real world dynamics applications and challenges. *Authorea Preprints*, 2025. 1

[76] 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*, pages 6639–6647, 2024. 7

[77] Wenjie Yin, Xuejiao Zhao, Yi Yu, Hang Yin, Danica Kragic, and Mårten Björkman. Lm2d: Lyrics-and music-driven dance synthesis. *arXiv preprint arXiv:2403.09407*, 2024. 3

[78] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. *arXiv preprint arXiv:2202.10571*, 2022. 2

[79] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 4

[80] Peng Zhang, Zhikang Dong, Bin Chen, and Xiulong Liu. Implementing dialog-based music recommendations for videos, 2025. US Patent App. 18/368,253. 1

[81] 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 conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[82] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022. 1

[83] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162. Springer, 2024. 6, 8