

# Dance Any Beat: Blending Beats with Visuals in Dance Video Generation

Xuanchen Wang  
University of Sydney  
`xwan0579@uni.sydney.edu.au`

Heng Wang  
University of Sydney  
`heng.wang@sydney.edu.au`

Dongnan Liu  
University of Sydney  
`dongnan.liu@sydney.edu.au`

Weidong Cai  
University of Sydney  
`tom.cai@sydney.edu.au`

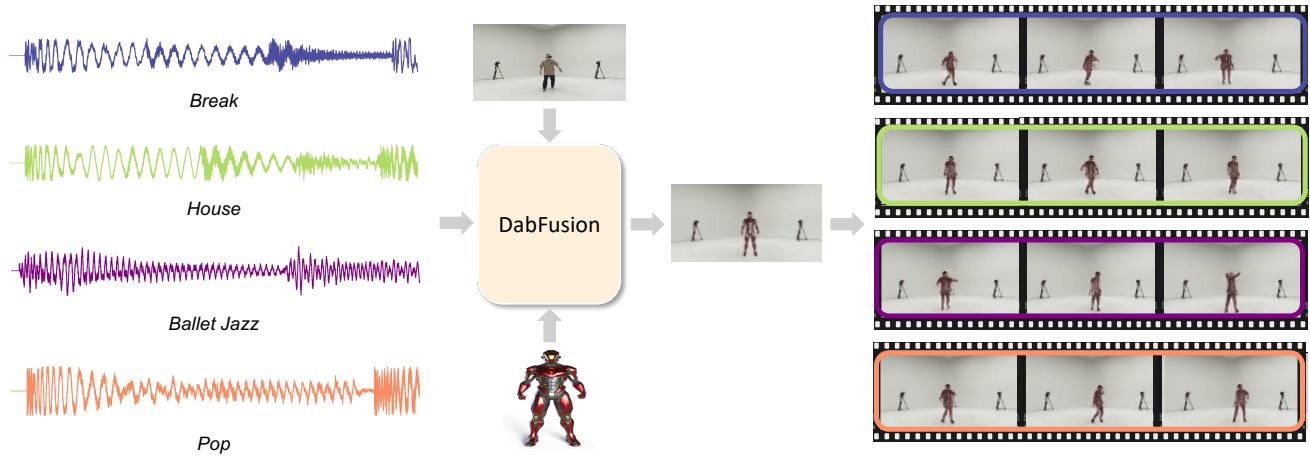


Figure 1: We present a diffusion-based framework DabFusion to dance any beat given any image. With various music styles of AIST++ dataset, our model predicts matching future video frames ( $128 \times 128$ ) starting from the same reference image.

## Abstract

The task of generating dance from music is crucial, yet current methods, which mainly produce joint sequences, lead to outputs that lack intuitiveness and complicate data collection due to the necessity for precise joint annotations. We introduce a Dance Any Beat Diffusion model, namely DabFusion, that employs music as a conditional input to directly create dance videos from still images, utilizing conditional image-to-video generation principles. This approach pioneers the use of music as a conditioning factor in image-to-video synthesis. Our method unfolds in two stages: training an auto-encoder to predict latent optical flow between reference and driving frames, eliminating the need for joint annotation, and training a U-Net-based diffusion model to produce these latent optical flows guided by music rhythm encoded by CLAP. Although capable of producing high-quality dance videos, the baseline model struggles with rhythm alignment. We enhance the model by adding beat

information, improving synchronization. We introduce a 2D motion-music alignment score (2D-MM Align) for quantitative assessment. Evaluated on the AIST++ dataset, our enhanced model shows marked improvements in 2D-MM Align score and established metrics. Video results can be found on our project page: <https://DabFusion.github.io>.

## 1. Introduction

Music-to-dance generation offers significant potential for automating choreography, benefiting various real-world applications. Existing methods [26, 27, 47, 35] focus on generating joint sequences, which, despite being effective, often produce outputs that are less intuitive for users and require precise joint annotations during data collection. This complicates the process. The field of conditional image-to-video (I2V) generation—where videos are produced from a single still image guided by textual descriptions [54, 39, 11], driving videos [6, 65, 55], or pose

sequences [38, 45, 44, 68]—offers a way to address these challenges. However, using music as a driving condition remains underexplored, presenting an opportunity for innovation. In this paper, we propose the Dance Any Beat Diffusion model (DabFusion), which generates dance videos from still images and music, enhancing intuitiveness and eliminating the need for joint annotations. DabFusion also pioneers the use of music as a condition in image-to-video synthesis.

DabFusion is highly versatile, capable of generating dance videos using the AIST++ dataset and animating individuals in new scenarios. As shown in Figure 1, we animate previously unseen individuals by merging their images with a background from the AIST++ dataset, bringing them to life through dance. This capability makes DabFusion practical for real-world applications, such as aiding dancers and choreographers in learning and practice, fostering engagement on social media, and enabling users without formal dance training to create personalized dance content. When generating dance videos using the AIST++ dataset, as illustrated in Figure 2, DabFusion excels in producing varied dance styles with different dancers from multiple perspectives and diverse initial poses. This flexibility enhances the intuitiveness and applicability of the generated content.

Inspired by recent I2V works [32, 44, 57], our model’s training process is structured into two phases. First, a latent flow auto-encoder is trained unsupervisedly to estimate the latent optical flow between reference and driving frames in a video, aiding in warping the reference frame to generate movement. Then, a U-Net based diffusion model generates latent flows guided by music and a starting image. For music information extraction, we use CLAP [59], the largest audio representation foundation model, to facilitate high-quality, music-aligned dance video generation.

A critical aspect of dance video evaluation is synchronization between the dancer’s movements and the music’s beat, an area needing improvement in our initial model. To address this, we integrate Librosa [30], a tool for audio signal analysis and beat extraction, with CLAP’s output. This enhancement improves alignment between dance poses and music beats, offering a more coherent and engaging viewing experience. To quantitatively assess this alignment, we introduce the 2D motion-music alignment (2D-MM Align) score, inspired by the Beat Alignment Score [27] for 3D motion-music correlation. This metric evaluates synchronization between motion and music in 2D scenarios, demonstrating the effectiveness of our approach in producing rhythmically aligned dance videos.

Our contribution can be summarized as follows:

- We introduce a novel approach capable of directly generating dance videos from music, significantly enhancing the intuitiveness of the generated content and its applicability in real-world settings.

- Our research broadens the horizon of conditional image-to-video generation by exploring music as a novel conditional input. We investigate the extraction of meaningful information from music through the integration of large-scale foundation models and established signal-processing methodologies.
- We establish a baseline and propose a beat alignment module for this innovative task, supported by extensive evaluations and the introduction of a new evaluation metric to assess motion-music alignment.

## 2. Related Works

Generating dance sequences from music uniquely intersects motion synthesis [1, 4, 12, 14] and music interpretation [58, 30, 59], aiming to create choreographed movements synchronized with input music. This extends beyond traditional motion synthesis, as choreographed movements are complex to animate. Early research [43, 69, 9] focused on producing 2D dance sequences due to the availability of online dance videos and advances in 2D human pose estimation [5]. However, 2D predictions lack expressiveness and applicability, prompting a shift to 3D dance generation. Recent methods leverage LSTMs [49, 2, 20, 61], GANs [25, 48, 10, 21], and Transformers [17, 27, 47, 35, 64] for 3D motion generation. The AIST++ dataset [27], a rich compilation of 3D motion data, has significantly advanced the field. Notable models include FACT [27], a full-attention-based cross-modal transformer using sequence-to-sequence learning for lifelike 3D dance sequences, and Bailando [47], which combines a pose VQ-VAE with a Motion GPT for temporal coherence via actor-critic learning. While existing research focuses on generating motion sequences from music, Ren et al. [37] synthesize videos using these sequences. Our work bypasses intermediary joint sequence generation, directly creating dance videos from music.

Image-to-Video (I2V) generation [3, 8, 16, 29, 34, 54, 60, 62, 66, 55] transforms static images into dynamic video sequences. This domain relates closely to single-image video prediction [22, 28], which infuses still images with temporal continuity to create believable motion. A specialized subfield, conditional image-to-video generation [3, 8, 16, 29, 54, 62, 55], incorporates guiding conditions or cues into the video generation process, enriching the content with specific characteristics or actions. For example, Wang et al. [54] developed a conditional GAN architecture to create human videos from static images, conditioned on labels describing facial expressions or actions. Yang et al. [62] introduced a pose-guided approach that extracts poses from an image and uses GANs to generate a sequence of poses and then video frames, translating static postures into animated sequences. Building on these, Wang et al. [55] unveiled the Latent Image Animator (LIA), which animates

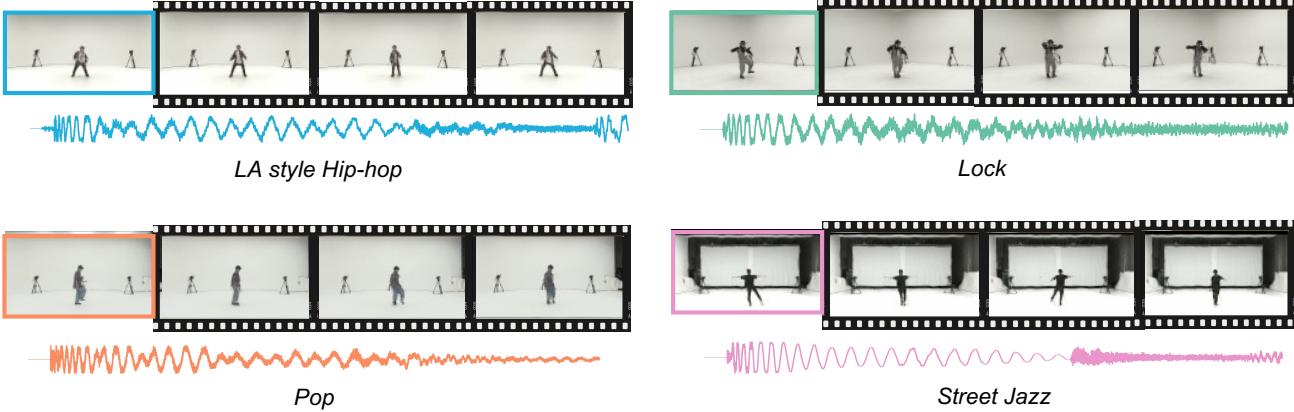


Figure 2: Exemplar videos generated from our DabFusion. Taking first image as starting frame and the unique music clip as guiding dance style, our framework is capable of generating varied styles of dance videos featuring different dancers from multiple perspectives with diverse initial poses and positions.

still images into videos under the guidance of an auxiliary video, leveraging a self-supervised auto-encoder to navigate latent space for animation. However, using audio, particularly music, as a conditioning factor in I2V remains largely unexplored. Introducing musical elements as conditions for video generation could open new research avenues and applications, suggesting an exciting direction for future work in I2V technology.

### 3. Method

Our methodology aims to generate latent optical flows guided by musical inputs. Recent works [31, 44, 52, 53, 32] in motion transfer have shown the effectiveness of using latent optical flow for warping images. This approach is more resource-efficient, requiring less computational power and time compared to high-dimensional pixel or latent feature spaces. Diffusion models are chosen for their exceptional quality and robust controllability. Our methodology begins by training an auto-encoder to discern optical flow between video frames. This trained auto-encoder then aids in training the diffusion model to generate latent flows. We extract musical information using CLAP to encode the music. Our enhanced model also incorporates beat information for better representation.

#### 3.1. Latent Flow Estimation

The objective of this phase is the training of an auto-encoder capable of accurately capturing and modeling motion between video frames within a latent space. As illustrated in Figure 3, the architectural components of the model comprise an image encoder, a flow predictor, and an image decoder. During the training process, two frames are randomly selected from a single video to serve as a reference frame  $x_{ref}$  and a driving frame  $x_{dri}$ , both sharing di-

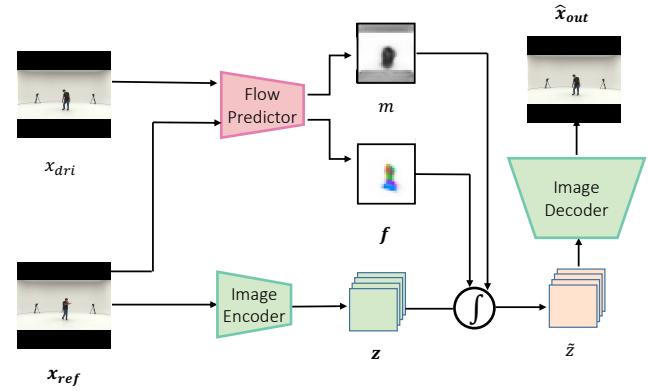


Figure 3: Training of latent flow auto-encoder. The flow predictor learns to estimate the latent flow  $f$  and occlusion map  $m$  between the reference frame  $x_{ref}$  and driving frame  $x_{dri}$ . The image encoder encodes  $x_{ref}$  into a latent representation  $z$ ,  $f$  and  $m$  are utilized to manipulate  $z$  into  $\tilde{z}$  which is then decoded by an image decoder to generate a output image  $\hat{x}_{out}$ . The objective of the training is to minimize the disparity between  $x_{dri}$  and  $\hat{x}_{out}$ .

mensions of  $H_x \times W_x \times 3$ . The image encoder processes  $x_{ref}$ , converting it into a compact latent representation,  $z$ , with dimensions  $H_z \times W_z \times C_z$ . Subsequently, similar to [44, 53], the flow predictor, receiving both  $x_{ref}$  and  $x_{dri}$  as inputs, computes the backward latent optical flow,  $f$ , and an occlusion map,  $m$ , to denote the transformation between these frames. The estimated latent flow,  $f$ , maintains the spatial dimensions of  $z$  and incorporates two channels to articulate the horizontal and vertical displacement across frames. Using backward flow estimation, the warping of  $f$  is realized through an efficient differentiable bilin-

ear sampling [18]. Concurrently, the occlusion map, sized  $H_z \times W_z \times 1$ , facilitates the reconstruction of areas within  $z$  that become obscured or revealed due to movement. Values within this map range from 0 to 1, where 1 signifies unoccluded regions, and 0 denotes complete occlusion. The warped latent map,  $\tilde{z}$ , is derived through the following equation:

$$\tilde{z} = m \odot f_w(z, f), \quad (1)$$

where  $f_w(\cdot, \cdot)$  denotes the back-warping operation and  $\odot$  is the element-wise multiplication. Finally, the image decoder takes  $\tilde{z}$  as input, reconstructing visible portions while simultaneously inpainting occluded regions to produce the output image  $\hat{x}_{out}$ . The training objective is to minimize a reconstruction loss that quantifies the difference between  $\hat{x}_{out}$  and  $x_{dri}$ . This loss employs the perceptual loss proposed by Johnson et al. [19], which utilizes features extracted by a pre-trained VGG-19 network [46]. Formally, the reconstruction loss is expressed as:

$$\mathcal{L}_{rec} = \sum_{i=1}^N |V_i(\hat{x}_{out}) - V_i(x_{dri})|, \quad (2)$$

where  $V_i(\cdot)$  extracts the  $i$ th channel features from a specified layer of VGG-19, and  $N$  is the total number of feature channels within that layer.

### 3.2. Latent Flow Generation

We use diffusion models to generate latent flow. These generative models reverse a diffusion process, which gradually turns data into Gaussian noise. The reverse process then restores this noise back to the original data distribution. Mathematically, the diffusion process is a Markov chain that adds noise to the data over  $T$  steps. The process starts with the original data  $x_0$  and ends with a sample  $x_T$  that resembles Gaussian noise. The transition from  $x_{t-1}$  to  $x_t$  can be defined by the following equation:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1 - \alpha_t} \epsilon_t, \quad (3)$$

where  $\alpha_t$  is a variance schedule that determines the amount of noise added at each step, and  $\epsilon_t$  is sampled from a standard Gaussian distribution  $\mathcal{N}(0, \mathcal{I})$ . The reverse diffusion process seeks to reconstruct the original data from noise by learning the conditional distribution  $p(x_{t-1}|x_t)$ . This is achieved by training a denoising model to estimate the parameters of the Gaussian distribution of  $x_{t-1}$  given  $x_t$ . The model outputs parameters  $\mu_\theta(x_t, t)$  and  $\sigma_\theta(x_t, t)$ , which define the Gaussian distribution from which  $x_{t-1}$  is sampled:

$$x_{t-1} = \mu_\theta(x_t, t) + \sigma_\theta(x_t, t) \cdot \epsilon, \quad (4)$$

where  $\epsilon$  is again sampled from a standard Gaussian distribution. The model is trained to minimize the difference between the generated and actual data distributions,

with an objective function often approximating the negative log-likelihood of the data under the model. A practical approach for training diffusion models is employing the Mean Squared Error (MSE) between the predicted and actual noise used in data corruption at each diffusion step. Given the diffusion process, the goal during training is to predict the noise  $\epsilon$  that was added to the original data at each step. The loss function for a single training step can be formulated as:

$$\mathcal{L} = \mathbb{E}_{t, x_0, \epsilon} [\|\epsilon - \epsilon_\theta(x_t, t)\|^2], \quad (5)$$

where  $\epsilon_\theta(x_t, t)$  is the predicted noise by the model given the noised data  $x_t$  at time step  $t$ .  $\epsilon$  is the actual noise added to the data  $x_0$  to obtain  $x_t$ .

Our approach utilizes a 3D U-Net [7] as the denoising mode. And the training of the model is in conjunction with the trained image encoder and flow predictor. Given an input video  $x_0^N = \{x_0, x_1, \dots, x_N\}$  and its corresponding music  $m$ . We first use the flow predictor to estimate latent flow sequence  $f_1^N = \{f_1, \dots, f_N\}$  and occlusion map sequence  $m_1^N = \{m_1, \dots, m_N\}$  between starting frame  $x_0$  and remaining frames  $\{x_1, \dots, x_N\}$ . The size of  $f_1^N$  is  $N \times H_z \times W_z \times 2$  and the size of  $m_1^N$  is  $N \times H_z \times W_z \times 1$ . Subsequently, we concatenate these two sequences along the last dimension to get  $a_0 = [f_1^N, m_1^N]$  which has the size  $N \times H_z \times W_z \times 3$ . Then, we perform diffusion process which gradually add 3D Gaussian noise to  $a_0$  to map it to a standard Gaussian noise. This is then reversed, using the image encoder to map the initial frame  $x_0$  to latent space  $z_0$  and incorporating music embeddings  $e$  for conditioning the denoising model. We will discuss how to encode music in Sec. 3.3. The updated loss function, accounting for the conditioning on  $z_0$  and  $e$ , is given by:

$$\mathcal{L} = \mathbb{E}_{t, a_0, \epsilon} [\|\epsilon - \epsilon_\theta(a_t, t, z_0, e)\|^2]. \quad (6)$$

The process of inference is shown in Figure 4, given an initial image  $x_0$  and a piece of music  $m$ , the image is first encoded into a latent space representation  $z_0$ , while the music is transformed into an embedding  $e$ . Subsequently, a volume of randomly sampled Gaussian noise, with dimensions  $N \times H_z \times W_z \times 3$ , undergoes a gradual denoising process by the U-Net. This process yields  $h_1^N$  which is the concatenation of a latent flow sequence  $f_1^N$  and corresponding occlusion map sequence  $m_1^N$ .  $z_0$  is then warped by each latent flow in the latent flow sequence and corresponding occlusion map in the occlusion map sequence to generate a new latent map sequence  $\tilde{z}_1^N$ . These maps are sequentially fed into the image decoder, which synthesizes the frames of the new video.

### 3.3. Music Encoding

Extracting meaningful information from music is crucial to our task. We use CLAP, developed by Wu et al. [59], for

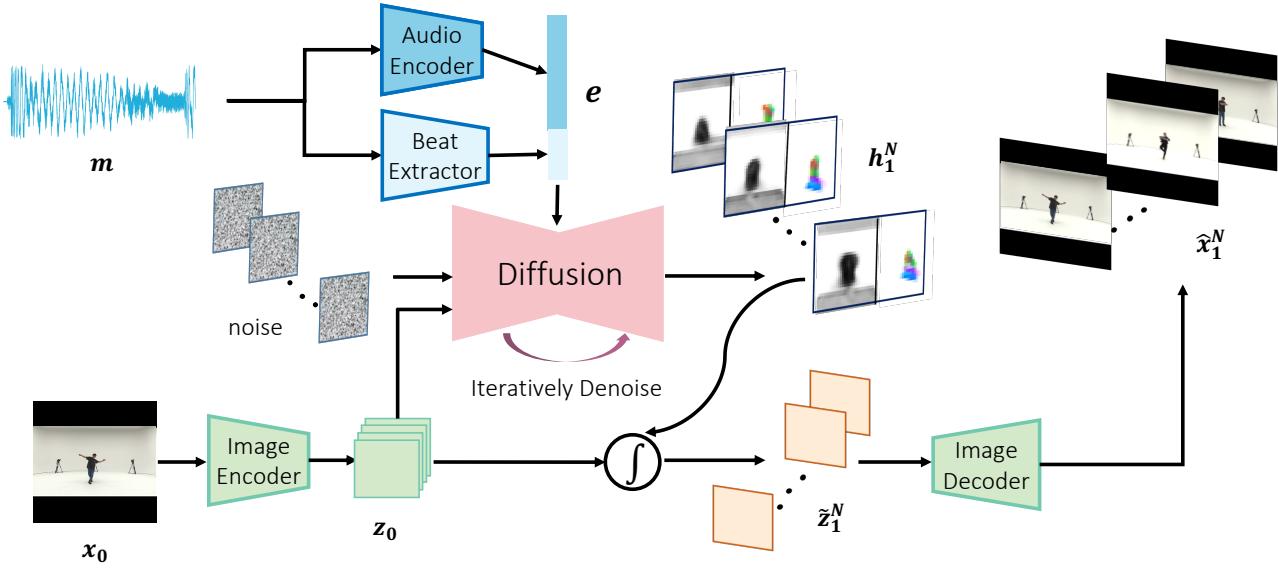


Figure 4: Inference process of DabFusion. DabFusion incorporates noise input along with image embedding  $z_0$  and music embedding  $e$  as conditions. Following the denoising stage of the diffusion model, we obtain  $h_1^N$ , comprising a concatenated sequence of latent flow and corresponding occlusion maps.  $h_1^N$  is utilized to transform  $z_0$  into a new sequence of latent maps, denoted as  $\tilde{z}_1^N$ , which is subsequently decoded to produce an image sequence.

music encoding. CLAP is a state-of-the-art audio representation model trained on a large dataset of text-audio pairs, learning a joint text-audio embedding space through contrastive learning. Our choice of CLAP is strategic, aiming to decode dance style information from musical cues using text as a bridge for music-to-dance style translation. While CLAP excels at encoding musical style, it struggles with detecting rhythmic patterns. To address this, we integrate beat information extracted via Librosa with CLAP’s output, creating a composite music embedding. This enhanced approach improves synchronization between motion and beat compared to our baseline model.

## 4. Experiments

### 4.1. Dataset and Implementation Details

Our model is trained and evaluated using the AIST++ [27] dataset, a leading resource in 3D human dance. AIST++ is recognized as the most comprehensive dataset in its domain, encompassing 10 dance genres. Each genre features 6 unique pieces of music, with each piece accompanying multiple videos, totaling 12,670 videos. These videos showcase a range of choreographies from basic to advanced, offering a wide spectrum of movements for detailed analysis. The pivotal advantage of AIST++ for our research is its meticulous synchronization of music and dance movements, which is crucial for music-driven dance generation and analysis.

We organize our dataset by music pieces, allocating

10,564 videos and 50 music tracks to the training set, and 2,106 videos with 10 tracks to the test set. Each video is segmented into frames, resized to  $128 \times 128$  pixels. In the latent flow estimation phase, we adopt the architecture from [19] for both our image encoder and decoder. The flow predictor follows [44]. The model is trained over 150 epochs with a batch size of 100, using the Adam optimizer [23], starting with a learning rate of  $2 \times 10^{-4}$ , decreased by a factor of 0.1 after epochs 60, 90, and 120. For the denoising model’s training, we employ a conditional 3D U-Net architecture from [13], featuring four down-sampling and up-sampling 3D convolutional blocks. Time step  $t$  is encoded by the sinusoidal position embedding [51]. The training spanned 250 epochs, starting with the same initial learning rate as the first stage, and adjusted at epochs 100, 150, and 200. We set  $T = 1000$  and apply a cosine noise schedule [33] and dynamic thresholding [41] at the 90% during sampling.

### 4.2. Evaluation Metrics

To assess the quality of generated videos, we use the Fréchet Video Distance (FVD) [50]. For quantitative assessment of image-level quality, we employ SSIM [56], PSNR [15], and LPIPS [67]. Additionally, we use the CLIP-Score (CS) [42] to measure the cosine similarity between CLIP embeddings of the dance video and Wav2CLIP [58] embeddings of the music, evaluating the coherence between the dance video and the music. To evaluate synchronization between the dancer’s movements and the music beat,

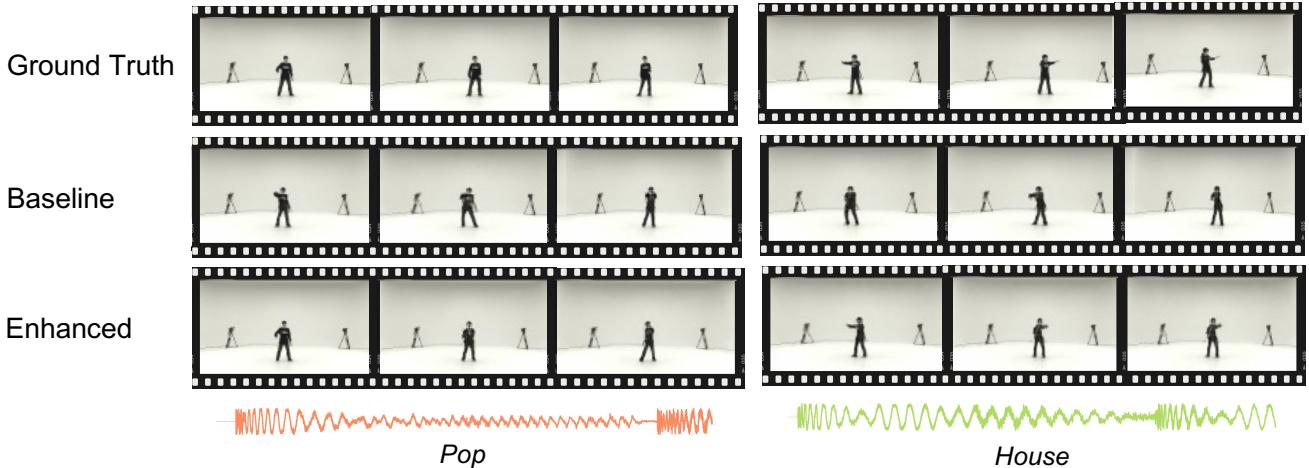


Figure 5: Video quality comparison between our models. We use the same starting image and music piece to generate videos with both our models, and select three frames from the same position.

we introduce the 2D motion-music alignment score (2D-MM Align). This metric, inspired by the Beat Alignment Score [27] for 3D scenes, adapts the concept for 2D environments. The alignment score is calculated as the average distance between each kinematic beat and its closest music beat. 2D-MM Align is defined as:

$$2\text{D-MM Align} = \frac{1}{n} \sum_{i=1}^n \exp \left( -\frac{\min_{\forall f_j^y \in F^y} \|f_i^x - f_j^y\|^2}{2\sigma^2} \right), \quad (7)$$

where  $F^y = \{f_j^y\}$  is the music beats,  $F^x = \{f_i^x\}$  is the kinematic beats and  $\sigma$  is a normalized parameter. Additionally, we incorporate the Audio-video Alignment Score (AV-Align) [63] to quantify audio-video synchronization. This metric analyzes optical flow in the video stream and audio peaks in the audio stream and assesses their temporal alignment within a three-frame window.

Table 1: Quantitative results among our models and MM-Diffusion. DabFusion<sub>b</sub> means our baseline model (without beat information) and DabFusion<sub>e</sub> means our enhanced model (with beat information).

Model	FVD ↓	LPIPS↓	PSNR↑	SSIM↑
MM-Diffusion	<b>180.31</b>	<b>0.025</b>	<b>34.81</b>	0.951
DabFusion <sub>b</sub> (ours)	194.44	0.032	26.17	0.957
DabFusion <sub>e</sub> (ours)	193.98	0.033	26.88	<b>0.959</b>

### 4.3. Result Analysis

**Video Quality Evaluation.** To evaluate the quality of the videos generated by our system, we use the music and the first frame from the test set, resulting in 2,106 generated

Table 2: Quantitative results across different angles. C01 is in front of the dancer, C02 is in the upper right of the dancer, C03 is in the right of the dancer, C09 is in front of the dancer but closer than C01.

Model	Camera	FVD ↓	LPIPS↓	PSNR↑	SSIM↑
MM-Diffusion	C01	176.93	0.019	35.52	0.949
DabFusion <sub>b</sub>	C01	<b>158.16</b>	<b>0.029</b>	<b>26.41</b>	<b>0.963</b>
	C02	197.01	0.039	25.031	0.952
	C03	158.56	0.036	26.30	0.959
	C09	195.83	0.037	25.09	0.953
DabFusion <sub>e</sub>	C01	158.02	<b>0.027</b>	<b>26.33</b>	<b>0.964</b>
	C02	198.17	0.040	25.23	0.954
	C03	<b>157.94</b>	0.036	26.29	0.960
	C09	195.17	0.035	25.28	0.956

videos. We visually present outputs from our baseline and enhanced models alongside ground truth in Figure 5. As shown in Figure 5, the quality of the videos generated by our model closely approximates that of real videos. Given the novelty of our task and the lack of directly comparable previous works in the I2V field, we benchmark our baseline model DabFusion<sub>b</sub> and enhanced model DabFusion<sub>e</sub> against MM-Diffusion [40], the leading model in unconditional video generation. We select MM-Diffusion because it is trained on a subset of the AIST++ dataset, closely aligning with our training dataset and demonstrating state-of-the-art video quality performance on AIST++. We randomly generate an equivalent number of videos with MM-Diffusion for comparison. A quantitative comparison between our models and MM-Diffusion is provided in Table 1. Analysis of Table 1 reveals that the quality of

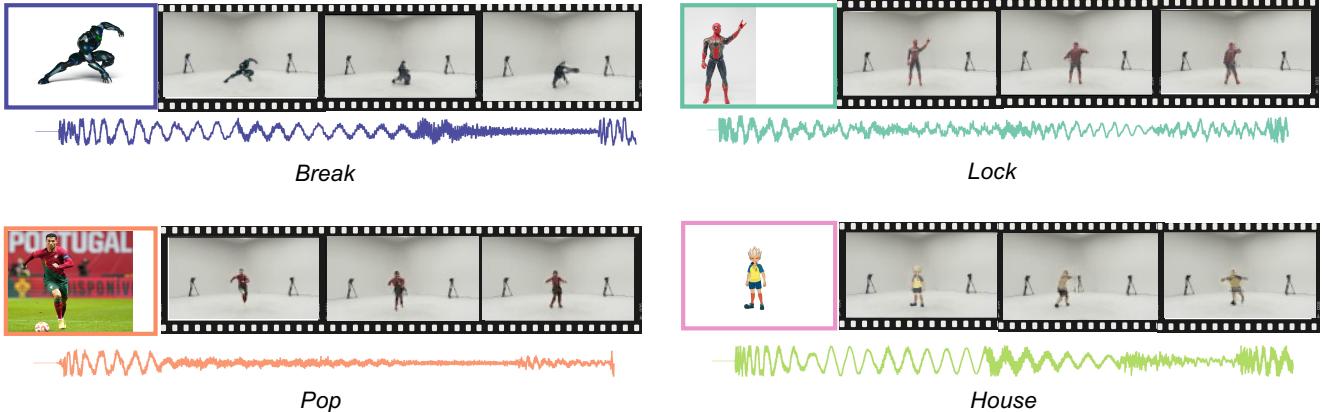


Figure 6: Exemplar videos generated using our DabFusion. Showcasing dance videos with unseen individuals in various poses. Following preprocessing of the input image, DabFusion demonstrates its ability to produce dance videos of high quality.

videos produced by our models closely matches that of MM-Diffusion. These comparisons underscore the competitive quality of videos generated by our approach. Furthermore, the comparison between our baseline and enhanced models suggests that incorporating beat information does not significantly impact the quality of the generated videos, as evidenced by their nearly identical evaluation scores.

A notable characteristic of the AIST++ dance dataset is its use of nine cameras C01-C09 to capture dance performances, offering multiple perspectives. We analyze four specific angles, summarized in Table 2, to assess their impact on video quality. MM-Diffusion, trained on a subset of the AIST++ dataset using only data from the C01 camera, is used to generate an equivalent number of videos for direct comparison. Our models outperform MM-Diffusion in terms of FVD and SSIM scores, especially for camera position C01, demonstrating superior video quality. Analysis of scores among different cameras indicates that direct camera angles, such as C01 and C03, where the cameras point straight at the dancer, result in higher-quality video generation compared to oblique angles like C02. Historically, I2V research has not emphasized the impact of camera positioning. Our findings reveal that camera placement significantly influences the quality of resulting videos, underscoring its importance in future I2V research. Additionally, we explored the impact of camera distance by comparing videos from C01 and C09, which share the same angle but differ in their proximity to the dancer. The findings suggest that cameras positioned further from the dancer tend to yield better results.

**Alignment Evaluation.** In Table 3, we assess the synchronicity between music-video and music-motion pairings. The results indicate that our enhanced model surpasses the baseline in all evaluation metrics, demon-

Table 3: Alignment results between our models: video-audio synchrony and motion-beat alignment.

	CS↑	2D-MM Align↑	AV Align↑
GT	0.153	0.223	0.152
DabFusion <sub>b</sub>	0.138	0.202	0.141
DabFusion <sub>e</sub>	<b>0.142</b>	<b>0.215</b>	<b>0.148</b>

Table 4: Alignment results based on dance styles.

Dance Style	Model	CS↑	2D-MM Align↑	AV Align↑
Break	GT	0.138	0.231	0.226
	DabFusion <sub>b</sub>	0.122	0.212	0.218
	DabFusion <sub>e</sub>	<b>0.131</b>	<b>0.223</b>	<b>0.229</b>
Ballet Jazz	GT	0.063	0.198	0.101
	DabFusion <sub>b</sub>	<b>0.061</b>	0.188	0.097
	DabFusion <sub>e</sub>	0.057	<b>0.193</b>	<b>0.098</b>
House	GT	0.121	0.218	0.134
	DabFusion <sub>b</sub>	0.109	0.198	0.121
	DabFusion <sub>e</sub>	<b>0.112</b>	<b>0.213</b>	<b>0.128</b>

ing the positive impact of beat information on both video-audio and motion-beat alignment. Figure 5 illustrates the differences, showing that frames generated by our enhanced model more closely resemble the ground truth videos. We also evaluate the results according to dance style categories corresponding to the music. These assessments, detailed in Table 4, exhibit significant disparities in alignment scores among various dance genres. Notably, Ballet Jazz yields the lowest alignment scores, attributed to its challenging choreography compared to other dance styles. Consequently, the alignment scores reflect the varying levels of difficulty inherent to each dance genre.

#### 4.4. Choreograph Anyone

DabFusion can choreograph individuals whether they belong to the AIST++ dataset or not, requiring several pre-processing steps. First, we use YOLO [36] for object detection and the Segment Anything Model [24] for segmentation to identify and segment the person from unseen scenarios. We then substitute the individual in the image from the AIST++ dataset with the newly segmented person, creating a combined image. This fused image is set as the starting frame to generate dance videos with musical conditions. We randomly select images featuring diverse individuals in various poses to create dance videos. As illustrated in Figure 6, preprocessing of the input image results in a new image compatible with DabFusion for dance video generation.

Table 5: Ablation study for result comparison between 40-frame video and 80-frame video. The results are from our enhanced model.

	CS↑	2D-MM Align↑	AV Align↑	FVD ↓
40 frames	0.142	0.215	0.148	<b>193.98</b>
80 frames	<b>0.153</b>	<b>0.219</b>	<b>0.179</b>	194.84

Table 6: Ablation study for result comparison between Music embedding and text embedding. The results are from our baseline model.

Embedding	CS↑	2D-MM Align↑	AV Align↑	FVD ↓
text	0.136	0.181	0.126	200.54
music	<b>0.138</b>	<b>0.202</b>	<b>0.141</b>	<b>194.44</b>

Table 7: Ablation study for result comparison about arbitrary length video generation. The results are from our enhanced model. We use the last frame of the first generated videos as the starting frame of next video to generate longer videos.

	CS↑	2D-MM Align↑	AV Align↑	FVD ↓
40 frames	0.142	<b>0.215</b>	<b>0.148</b>	<b>193.98</b>
80 frames	<b>0.145</b>	0.203	0.139	210.86
120 frames	0.131	0.197	0.121	256.32
160 frames	0.123	0.179	0.101	323.72

#### 4.5. Ablation Study

**Influence of Video Length.** We observe that video length impacts benchmark performance. Our models are trained to generate videos of 40 and 80 frames. Due to training time and computational resource considerations, we designate the 40-frame videos as the baseline. This comparison is detailed in Table 5. The results indicate comparable

video quality between the two lengths; however, the 80-frame videos demonstrate superior performance in alignment evaluations. This improvement is primarily attributed to the extended music duration, providing more information for enhancing alignment accuracy.

**Text Embedding vs. Music Embedding.** Given that CLAP is trained on text-audio pairs, we aim to uncover the type of information it extracts from music, hypothesizing that CLAP primarily identifies dance style cues. To test this, we conduct an experiment using our baseline model, focusing solely on the embedding from CLAP. We substitute the music embedding with one derived from CLAP using the text description of the dance style corresponding to the music. The findings, presented in Table 6, indicate that the impact of different embeddings on FVD and CS scores is minimal. However, videos generated with text-based embeddings exhibit inferior motion-beat alignment. This suggests that CLAP extracts more than just dance style information; it also obtains details crucial for enhancing motion-beat synchronization.

**Arbitrary-Length Video Generation.** Given the specialty of our task, our model theoretically can generate videos corresponding to any length of music. By using the final frame of one video sequence as the initial frame of the subsequent sequence, we can create videos of arbitrary length. We conducted experiments with four varying lengths of videos generated sequentially using our enhanced model; the results are presented in Table 7. As video length increases, quality tends to drop rapidly. This decline is primarily because the final frames generated by the model lack the fidelity of actual video frames, becoming more apparent with longer videos. Additionally, alignment scores diminish due to this quality degradation. For instance, employing a pose estimation algorithm to determine keypoints for calculating the 2D-MM Alignment score becomes less accurate with lower video quality, adversely affecting the alignment score.

## 5. Conclusion

In this paper, we explore the intersection of two distinct research areas: conditional image-to-video generation and music-to-dance generation. We introduce a novel approach, DabFusion, that directly generates dance videos from a single image and accompanying music. This research investigates the role of music as a conditional input in image-to-video synthesis and sets a foundational benchmark for direct dance video generation in response to musical cues. We further examine the extraction of crucial information for enhancing motion-music beat alignment, leveraging both the large-scale foundational model CLAP and the signal-processing tool Librosa. Through extensive experimentation, we have established a robust baseline for this novel task.

## References

- [1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7144–7153, 2019.
- [2] Omid Alemi, Jules Fran oise, and Philippe Pasquier. Groovenet: Real-time music-driven dance movement generation using artificial neural networks. *networks*, 8(17):26, 2017.
- [3] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Bjorn Ommer. Understanding object dynamics for interactive image-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5171–5181, 2021.
- [4] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6158–6166, 2017.
- [5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.
- [6] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [7]  zg n  cek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II* 19, pages 424–432. Springer, 2016.
- [8] Michael Dorkenwald, Timo Milbich, Andreas Blattmann, Robin Rombach, Konstantinos G Derpanis, and Bjorn Ommer. Stochastic image-to-video synthesis using cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3742–3753, 2021.
- [9] Joao P Ferreira, Thiago M Coutinho, Thiago L Gomes, Jos  F Neto, Rafael Azevedo, Renato Martins, and Erickson R Nascimento. Learning to dance: A graph convolutional adversarial network to generate realistic dance motions from audio. *Computers & Graphics*, 94:11–21, 2021.
- [10] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019.
- [11] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3615–3625, 2022.
- [12] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal in-
- painting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7134–7143, 2019.
- [13] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.
- [14] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016.
- [15] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.
- [16] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18219–18228, 2022.
- [17] Ruozhi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Dixin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. *arXiv preprint arXiv:2006.06119*, 2020.
- [18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.
- [19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II* 14, pages 694–711. Springer, 2016.
- [20] Hsuan-Kai Kao and Li Su. Temporally guided music-to-body-movement generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 147–155, 2020.
- [21] Jinwoo Kim, Heeseok Oh, Seongjean Kim, Hoseok Tong, and Sanghoon Lee. A brand new dance partner: Music-conditioned pluralistic dancing controlled by multiple dance genres. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3490–3500, 2022.
- [22] Yunji Kim, Seonghyeon Nam, In Cho, and Seon Joo Kim. Unsupervised keypoint learning for guiding class-conditional video prediction. *Advances in neural information processing systems*, 32, 2019.
- [23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.
- [25] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. *Advances in neural information processing systems*, 32, 2019.
- [26] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance

- motions with transformer. *arXiv preprint arXiv:2008.08171*, 2020.
- [27] 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.
- [28] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Flow-grounded spatial-temporal video prediction from still images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 600–615, 2018.
- [29] Aniruddha Mahapatra and Kuldeep Kulkarni. Controllable animation of fluid elements in still images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3667–3676, 2022.
- [30] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *SciPy*, pages 18–24, 2015.
- [31] Haomiao Ni, Yihao Liu, Sharon X Huang, and Yuan Xue. Cross-identity video motion retargeting with joint transformation and synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 412–422, 2023.
- [32] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18444–18455, 2023.
- [33] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.
- [34] Junting Pan, Chengyu Wang, Xu Jia, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Video generation from single semantic label map. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2019.
- [35] Qiaosong Qi, Le Zhuo, Aixi Zhang, Yue Liao, Fei Fang, Si Liu, and Shuicheng Yan. Diffdance: Cascaded human motion diffusion model for dance generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1374–1382, 2023.
- [36] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [37] Xuanchi Ren, Haoran Li, Zijian Huang, and Qifeng Chen. Self-supervised dance video synthesis conditioned on music. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 46–54, 2020.
- [38] Yurui Ren, Ge Li, Shan Liu, and Thomas H Li. Deep spatial transformation for pose-guided person image generation and animation. *IEEE Transactions on Image Processing*, 29:8622–8635, 2020.
- [39] 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.
- [40] 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.
- [41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.
- [42] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [43] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to body dynamics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7574–7583, 2018.
- [44] 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.
- [45] 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.
- [46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [47] 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.
- [48] Guofei Sun, Yongkang Wong, Zhiyong Cheng, Mohan S Kankanhalli, Weidong Geng, and Xiangdong Li. Deepdance: music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia*, 23:497–509, 2020.
- [49] Taoran Tang, Jia Jia, and Hanyang Mao. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1598–1606, 2018.
- [50] 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.
- [51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

- [52] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Bryan Catanzaro, and Jan Kautz. Few-shot video-to-video synthesis. *Advances in Neural Information Processing Systems*, 32, 2019.
- [53] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1152–1164, 2018.
- [54] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal gan for video generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1160–1169, 2020.
- [55] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022.
- [56] 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.
- [57] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018.
- [58] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE, 2022.
- [59] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [60] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2018.
- [61] Nelson Yalta, Shinji Watanabe, Kazuhiro Nakadai, and Tetsuya Ogata. Weakly-supervised deep recurrent neural networks for basic dance step generation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.
- [62] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahu Lin. Pose guided human video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.
- [63] 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*, volume 38, pages 6639–6647, 2024.
- [64] Zijie Ye, Haozhe Wu, Jia Jia, Yaohua Bu, Wei Chen, Fanbo Meng, and Yanfeng Wang. Choreonet: Towards music to dance synthesis with choreographic action unit. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 744–752, 2020.
- [65] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [66] Jiangning Zhang, Chao Xu, Liang Liu, Mengmeng Wang, Xia Wu, Yong Liu, and Yunliang Jiang. Dtvnet: Dynamic time-lapse video generation via single still image. In *Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 300–315. Springer, 2020.
- [67] 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.
- [68] 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.
- [69] Wenlin Zhuang, Congyi Wang, Jinxiang Chai, Yangang Wang, Ming Shao, and Siyu Xia. Music2dance: Dancenet for music-driven dance generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2):1–21, 2022.