

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

Xuanchen Wang¹ Heng Wang¹ Dongnan Liu¹ Weidong Cai¹

¹University of Sydney

xwan0579@uni.sydney.edu.au

{heng.wang, dongnan.liu, tom.cai}@sydney.edu.au

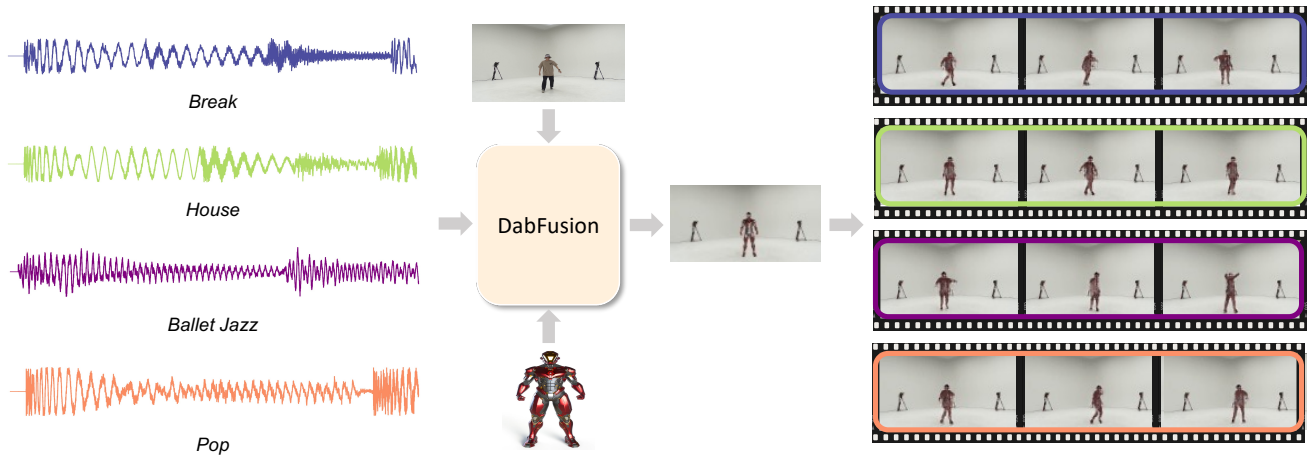


Figure 1. We introduce DabFusion, a diffusion-based framework designed to generate videos of individuals dancing, utilizing a music input and an initial reference image as the conditions for video generation.

Abstract

Generating dance from music is crucial for advancing automated choreography. Current methods typically produce skeleton keypoint sequences instead of dance videos and lack the capability to make specific individuals dance, which reduces their real-world applicability. These methods also require precise keypoint annotations, complicating data collection and limiting the use of self-collected video datasets. To overcome these challenges, we introduce a novel task: generating dance videos directly from images of individuals guided by music. This task enables the dance generation of specific individuals without requiring keypoint annotations, making it more versatile and applicable to various situations. Our solution, the Dance Any Beat Diffusion model (DabFusion), utilizes a reference image and a music piece to generate dance videos featuring various dance types and choreographies. The music is analyzed by our specially designed music encoder, which identifies essential features including dance style, movement, and rhythm. DabFusion excels in generating dance videos not only for individuals in the training dataset but also for any

previously unseen person. This versatility stems from its approach of generating latent optical flow, which contains all necessary motion information to animate any person in the image. We evaluate DabFusion’s performance using the AIST++ dataset, focusing on video quality, audio-video synchronization, and motion-music alignment. We propose a 2D Motion-Music Alignment Score (2D-MM Align), which builds on the Beat Alignment Score to more effectively evaluate motion-music alignment for this new task. Experiments show that our DabFusion establishes a solid baseline for this innovative task. Video results can be found on our project page: <https://DabFusion.github.io>.

1. Introduction

Music-to-dance generation not only revolutionizes choreography by automating complex routines but also enhances performance training, enriches interactive gaming experiences, and opens new avenues for virtual reality entertainment, offering profound benefits across multiple real-world applications. Existing methods [21, 22, 27, 39] focus on generating keypoint sequences, which, despite be-

ing effective, often produce outputs that are less intuitive and do not typically support the animation of specific individuals, which limits their practical utility in personalized applications. Ren et al. [29] attempt to advance the field, but their approach still begins by generating keypoint sequences, which are then used to synthesize videos of specific individuals. This method requires an additional motion transfer model and a video of the target person, complicating the pipeline and limiting practical applications, as obtaining a video of the target person can be challenging in real-world scenarios. Another challenge with existing techniques is their dependency on precisely annotated keypoint data. This requirement not only complicates the data collection process but also restricts the usability of these methods with self-collected or less structured video datasets. To address these limitations, we introduce a new and innovative task: generating dance videos directly from images of individuals, guided by accompanying music. By utilizing images of individuals, this new task facilitates the generation of personalized dance sequences that can feature specific people. Furthermore, it eliminates the need of keypoint annotations, thereby increasing the versatility and applicability of the task across a variety of scenarios. Our solution, named the Dance Any Beat Diffusion model (DabFusion), takes a reference image paired with a selected piece of music as input to create dance videos featuring various dance types and choreographies. For extracting information from music, we develop a novel music encoder that captures essential features including dance style, movement, and rhythm, which are critical for this task. To accurately extract dance styles, we fine-tune the CLAP [51] model, a text-audio pretraining foundation model, on the AIST++ [22] dataset, which serves as the training dataset for our task. For movement information, we fine-tune the Wav2CLIP [50] model, a visual-audio pretraining foundation model, on the AIST++ dataset. The synchronization of dance movements with the music’s rhythm is essential for realistic dance generation. To achieve this, we employ Librosa [23], a tool for audio signal analysis, to extract beat information from the music. By integrating these elements, our music encoder can generate a comprehensive representation of the music.

DabFusion effectively generates dance videos not just for individuals included in the AIST++ dataset, but also for people it has never encountered before. As shown in Fig. 1, we animate previously unseen individuals by merging their images with a background from the AIST++ dataset, bringing them to life through dance. DabFusion can choreograph dance sequences for previously unseen individuals to various music pieces, producing distinctly different videos. This diversity in output highlights the robust information extraction capability of our music encoder. When generating dance videos using the AIST++ dataset, as illustrated in

Fig. 2, DabFusion excels in producing varied dance styles with different dancers from multiple perspectives and diverse initial poses. The capability of DabFusion to generate various dance forms enables it to serve a broad audience, ranging from professional choreographers seeking inspiration to casual users wanting to see themselves or others dance to their favorite tunes. The versatility of DabFusion arises from its method of generating latent optical flow, which captures all essential motion information needed to animate any person in the image. This technique is widely adopted in the field of image-to-video (I2V) generation [4, 8, 30, 31, 37, 46, 47, 56]. Inspired by recent I2V works [25, 36, 49], 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. We conduct a thorough evaluation of DabFusion’s performance using the AIST++ dataset, with particular focus on video quality, audio-video synchronization, and the alignment between motion and music. To more effectively assess motion-music alignment—a crucial aspect of this new task—we introduce a 2D motion-music alignment (2D-MM Align) score, inspired by the Beat Alignment Score [22] 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 experimental results demonstrate that DabFusion establishes a robust baseline for this innovative task, indicating its potential as a pioneering tool in the field of dance video synthesis.

Our contributions can be summarized as follows:

- We introduce a novel task that directly generates dance videos from music, significantly enhancing the intuitiveness of the generated content and its applicability in real-world settings.
- We propose DabFusion, a solution for the new task. DabFusion is capable of creating dance sequences not only for individuals present in the training dataset but also for person it has never previously encountered.
- We propose a novel music encoder designed to generate a comprehensive representation of music, capturing key elements including dance style, movement, and rhythm.
- We establish a solid baseline for the new task, supported by extensive evaluations on the AIST++ dataset.

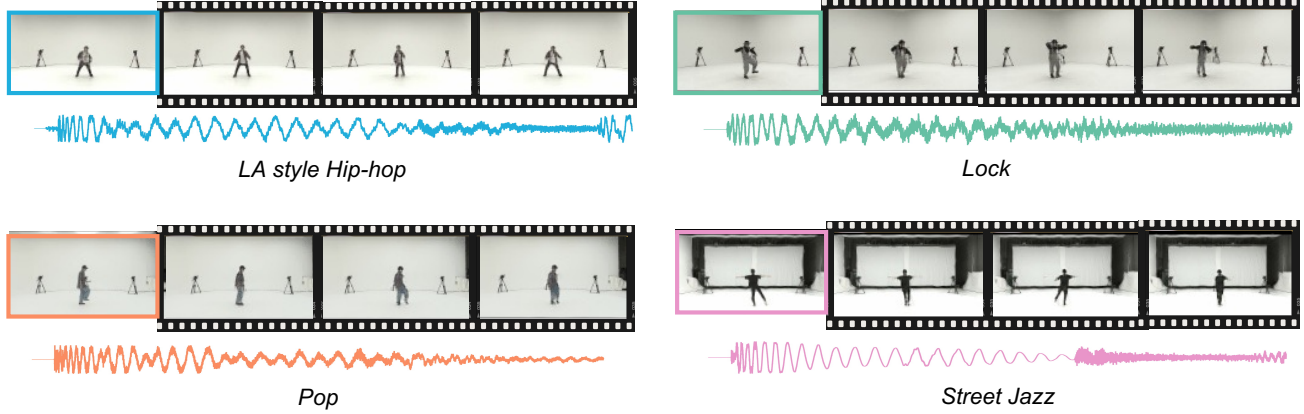


Figure 2. Exemplar videos generated from our DabFusion. Taking first image as starting frame and the unique music clip as guiding dance style, our framework is capable of generating varied styles of dance videos featuring different dancers from multiple perspectives with diverse initial poses and positions.

2. Related Works

Generating dance sequences from music uniquely intersects motion synthesis [1, 2, 9, 11] and music interpretation [23, 50, 51], aiming to create choreographed movements synchronized with input music. This extends beyond traditional motion synthesis, as choreographed movements are complex to animate. Early research [6, 35, 57] focused on producing 2D dance sequences due to the availability of online dance videos and advances in 2D human pose estimation [3]. However, 2D predictions lack expressiveness and applicability, prompting a shift to 3D dance generation. Recent methods leverage LSTMs [16, 41, 52], GANs [7, 17, 20, 40], and Transformers [13, 22, 27, 39, 54] for 3D motion generation. The AIST++ dataset [22], a rich compilation of 3D motion data, has significantly advanced the field. Notable models include FACT [22], a full-attention-based cross-modal transformer using sequence-to-sequence learning for lifelike 3D dance sequences, and Bailando [39], 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. [29] synthesized videos using these sequences. Our work bypasses intermediary joint sequence generation, directly creating dance videos from music.

3. Method

3.1. Overview

Our methodology generates latent optical flows driven by musical inputs, with each music piece encoded by our specially designed music encoder. Music encoding is detailed in Sec. 3.2. Recent works [24, 25, 36, 44, 45] 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. Our methodology begins by training an auto-encoder to discern optical flow between video frames, we discuss this in details in Sec. 3.3. This trained auto-encoder then aids in training the diffusion model to generate latent flows. The training of the diffusion model is detailed in Sec. 3.4. The overview of DabFusion is shown in Fig. 3, 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.

3.2. Music Encoding

Extracting meaningful information from music is essential for generating dance videos, where dance type, dancer movement, and rhythm play critical roles. For dance type information, we utilize the CLAP model, a pipeline for contrastive language-audio pretraining that develops audio representations by integrating audio data with natural language descriptions. We fine-tune CLAP to enhance its capture of dance style information. Specifically, we freeze the weights of the audio and text encoders in CLAP and introduce multi-layer perceptrons as adapter layers. These layers are trained using data from the AIST++ dataset. For movement infor-

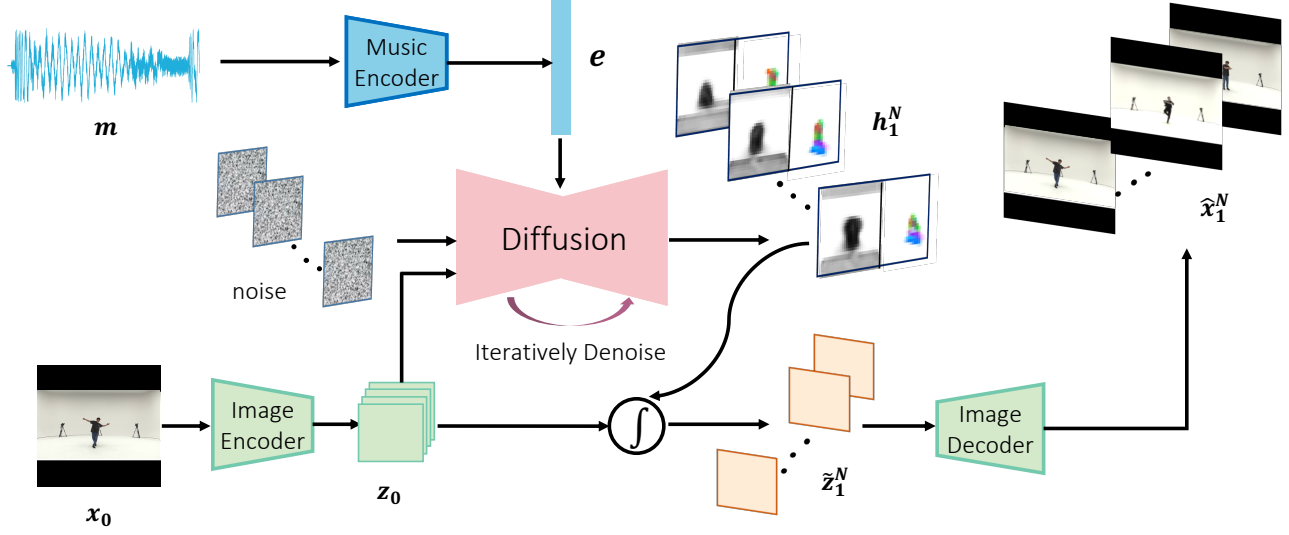


Figure 3. Overview of DabFusion. Given a reference image x_0 with dimensions $H_x \times W_x \times 3$ and a piece of music m . DabFusion incorporates noise input along with image embedding z_0 which has dimensions $H_z \times W_z \times C_z$ and music embedding e as conditions. Following the denoising stage of the diffusion model, we obtain h_1^N which has dimensions $H_z \times W_z \times 3 \times N$, 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.

mation, we employ Wav2CLIP, initially trained on audio-visual datasets. We apply a similar fine-tuning approach as with CLAP, using the AIST++ dataset to adapt Wav2CLIP for our specific requirements. Regarding rhythm, which is crucial for aligning dance movements with the music beat, we use Librosa, a tool for audio signal analysis, to extract beat information. Given a piece of music m , we pass it through fine-tuned CLAP, fine-tuned Wav2CLIP, and Librosa to extract different features. Specifically, we obtain the music representation e_c from CLAP, e_w from Wav2CLIP, and beat information e_b from Librosa. We then concatenate e_c , e_w , and e_b to form a comprehensive musical representation e , which is specifically tailored for generating dance videos.

3.3. 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 Fig. 4, 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 dimensions 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 [36, 45], the flow predictor, receiving both x_{ref} and x_{dri} as inputs, computes the backward latent optical

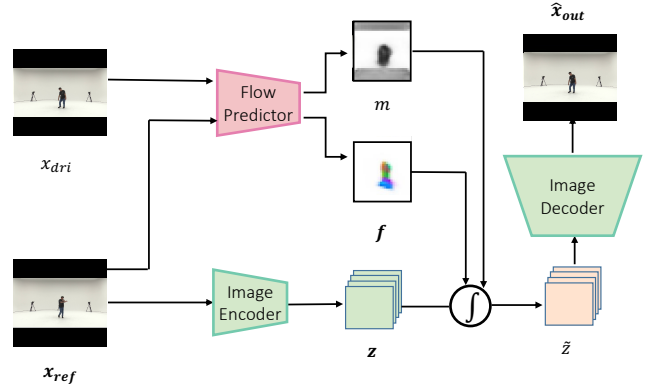


Figure 4. Training of latent flow auto-encoder. The flow predictor learns to estimate the latent flow f and occlusion map m between the reference frame x_{ref} and the driving frame x_{dri} . The image encoder encodes x_{ref} into a latent representation z , f and m are utilized to manipulate z into $\tilde{z}$ which is then decoded by an image decoder to generate an output image $\hat{x}_{out}$. The objective of the training is to minimize the disparity between x_{dri} and $\hat{x}_{out}$.

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 z using f is realized through an efficient differentiable bilinear sampling [14]. Concurrently, the occlu-

sion 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. [15], which utilizes features extracted by a pre-trained VGG-19 network [38]. 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.

3.4. 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 α_t is a variance schedule that determines the amount of noise added at each step, and ϵ_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 ϵ 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 ϵ 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 . ϵ is the actual noise added to the data x_0 to obtain x_t .

Our approach utilizes a 3D U-Net [5] 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 adds 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. The updated loss function, accounting for the conditioning on z_0 and e , is given by:

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

4. Experiments and Results

4.1. Dataset and Implementation Details

Our model is trained and evaluated using the AIST++ [22] 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.

DabFusion was trained using two A6000 GPUs. We organize our dataset by music pieces, allocating 10,564 videos and 50 music tracks to the training set, and 2,106 videos

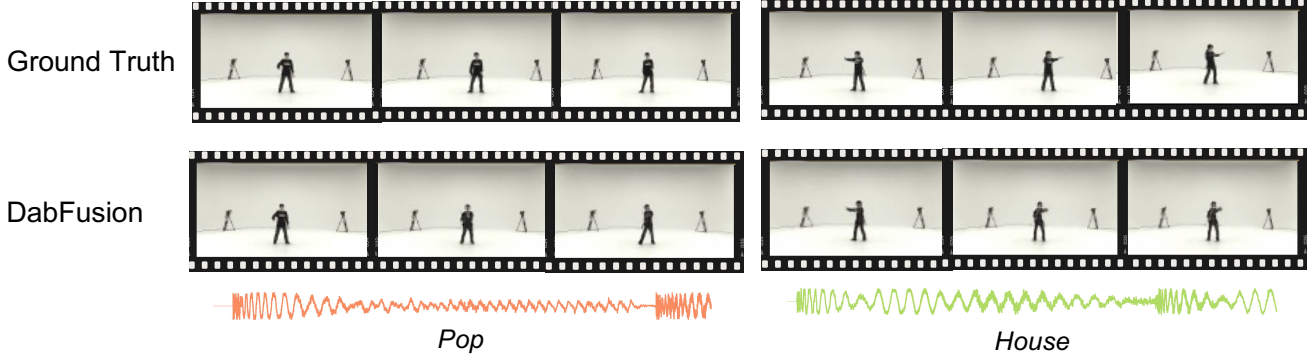


Figure 5. Comparison of video quality between ground-truth videos and those generated by DabFusion. We use the same starting image and music piece to generate videos with our model, and select three frames from the same position.

with 10 tracks to the test set. Each video is segmented into frames, resized to 128×128 pixels. When fine-tuning CLAP and Wav2CLIP, the adapter layers are composed of two MLP layers, each with a hidden size of 512. These adapter layers are trained over 100 epochs. In the latent flow estimation phase, we adopt the architecture from [15] for both our image encoder and decoder. The flow predictor follows [36]. The model is trained over 150 epochs with a batch size of 100, using the Adam optimizer [18], starting with a learning rate of 2×10^{-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 [10], featuring four down-sampling and up-sampling 3D convolutional blocks. The time step t is encoded by the sinusoidal position embedding [43]. 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 [26] and dynamic thresholding [33] at the 90% during sampling.

Table 1. Quantitative results between DabFusion and MM-Diffusion.

Model	FVD ↓	LPIPS ↓	PSNR ↑	SSIM ↑
MM-Diffusion	180.31	0.025	34.81	0.951
DabFusion (ours)	193.98	0.033	26.88	0.959

4.2. Evaluation Metrics

To assess the quality of generated videos, we use the Fréchet Video Distance (FVD) [42]. For quantitative assessment of image-level quality, we employ SSIM [48], PSNR [12], and LPIPS [55]. Additionally, we use the CLIP-Score (CS) [34] to measure the cosine similarity between CLIP embeddings of the dance video and Wav2CLIP embeddings of the music, evaluating the coherence between the dance video and the music. To evaluate synchroniza-

tion between the dancer’s movements and the music beat, we introduce the 2D motion-music alignment score (2D-MM Align). This metric, inspired by the Beat Alignment Score [22] 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:

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

where $F^y = \{f_j^y\}$ is the music beats, $F^x = \{f_i^x\}$ is the kinematic beats and σ is a normalized parameter. Additionally, we incorporate the Audio-video Alignment Score (AV-Align) [53] 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.

4.3. Result Analysis

Video Quality Evaluation. To evaluate the quality of the videos generated by our DabFusion, we use the music and the first frame from the test set, resulting in 2,106 generated videos. We visually present outputs from our model alongside ground truth in Fig. 5. As shown in Fig. 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, we benchmark our DabFusion against MM-Diffusion [32], the leading model that in unconditional video generation. The AIST++ dataset uniquely utilizes nine cameras ranging from C01 to C09 to record dances from multiple perspectives. DabFusion is trained on the whole dataset and MM-Diffusion is trained on data from only the C01 camera which is in front of the dancer. We randomly generate an equivalent number of videos with MM-Diffusion for comparison. A quantitative comparison between our models

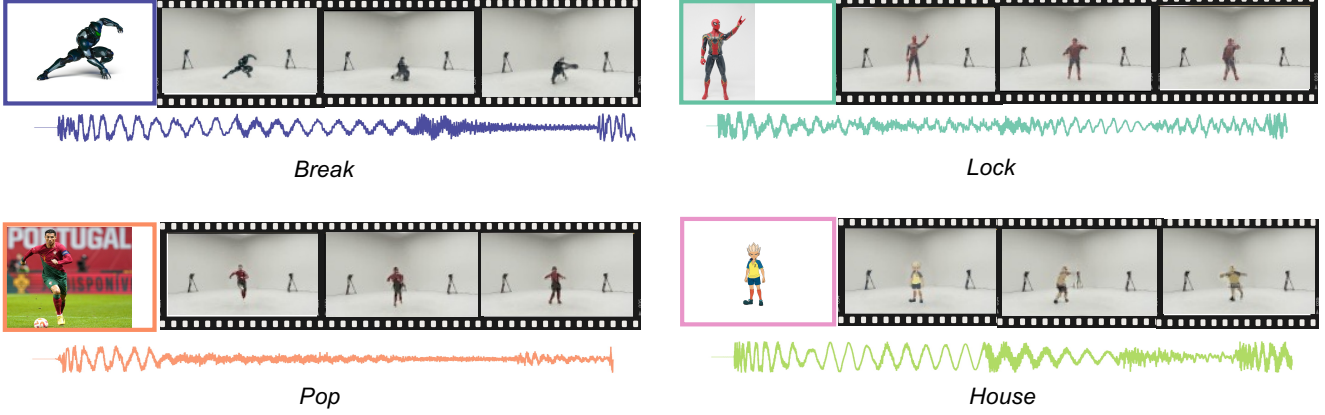


Figure 6. Exemplar videos generated using our DabFusion, showcasing dance videos with unseen individuals in various poses.

and MM-Diffusion is provided in Tab. 1. The results reveal that the quality of videos produced by our models closely matches that of MM-Diffusion. Given that our model can generate videos from various angles, producing videos from specific angles such as C02, which is situated in the upper right of the dance space, proves more challenging than from C01. Despite this, DabFusion demonstrates exceptional performance in the quality of the generated videos. To further assess the impact of camera angles on video qual-

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 (ours)	C01	158.02	0.027	26.33	0.964
	C02	198.17	0.040	25.23	0.954
	C03	157.94	0.036	26.29	0.960
	C09	195.17	0.035	25.28	0.956

ity, we analyze four specific angles. These are summarized in Tab. 2, illustrating the quality of videos generated from different perspectives. Our models show better performance in terms of FVD and SSIM scores, particularly with the C01 camera position, indicating higher video quality. Our analysis reveals that videos generated from cameras directly facing the dancer, such as C01 and C03, are generally of higher quality compared to those from angled cameras like C02. We also examine the effect of camera distance by comparing videos from C01 and C09, which have the same angle but differ in distance from the dancer. Results indicate that cameras positioned further away tend to produce better outcomes.

Alignment Evaluation. In Tab. 3, we assess the synchronicity between music-video and music-motion pair-

Table 3. Alignment results for DabFusion.

	CS ↑	2D-MM Align ↑	AV Align ↑
Ground Truth	0.153	0.223	0.152
DabFusion (ours)	0.142	0.215	0.148

Table 4. Alignment results based on dance styles.

Dance Style	Model	CS ↑	2D-MM Align ↑	AV Align ↑
Break	Ground Truth	0.138	0.231	0.226
	DabFusion (ours)	0.131	0.223	0.229
Ballet Jazz	Ground Truth	0.063	0.198	0.101
	DabFusion (ours)	0.057	0.193	0.098
House	Ground Truth	0.121	0.218	0.134
	DabFusion (ours)	0.112	0.213	0.128

ings. The results indicate that videos generated by our DabFusion closely match the ground-truth videos in terms of alignment metrics. We also evaluate the results according to dance style categories corresponding to the music. These assessments, detailed in Tab. 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 [28] for object detection and the Segment Anything Model [19] 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

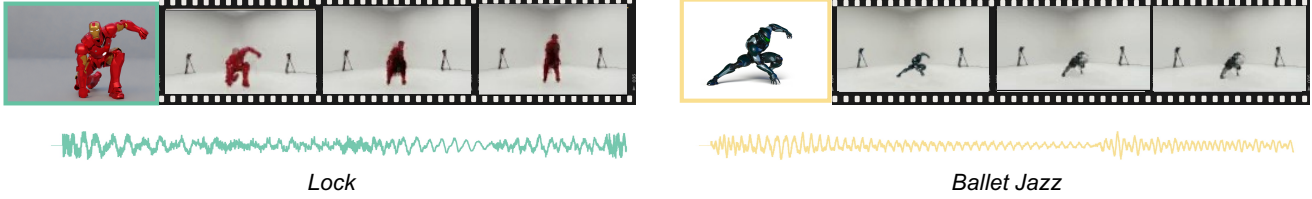


Figure 7. Failure cases for DabFusion in different scenarios. On the left, a person with unclear joint delineations leads DabFusion to produce low-quality video output. On the right, the initial crouching pose, uncommon in Ballet Jazz dance, results in repetitive frames.

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 Fig. 6, DabFusion is capable of generating high-quality dance videos featuring previously unseen individuals in a variety of contexts and poses.

Table 5. Ablation study for result comparison between 40-frame video and 80-frame video.

	CS $\uparrow$	2D-MM Align $\uparrow$	AV Align $\uparrow$	FVD $\downarrow$
40 frames	0.142	0.215	0.148	193.98
80 frames	0.153	0.219	0.179	194.84

Table 6. Ablation study for result comparison between DabFusion with and without beat information.

Model	Beat	CS $\uparrow$	2D-MM Align $\uparrow$	AV Align $\uparrow$	FVD $\downarrow$
DabFusion	$\times$	0.138	0.202	0.141	194.44
	$\checkmark$	0.142	0.215	0.148	193.98

4.5. Ablation Study

Influence of Video Length. We observe that video length impacts benchmark performance. Our model is 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 Tab. 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.

Influence of Beat Information. With the rise of large models that generate music representations, many current methods rely solely on these models, often neglecting traditional audio analysis tools. In our experiment, we test the impact of including beat information in our dance video generation. The findings, presented in Tab. 6, reveal that while beat information does not greatly change the overall video

quality, it improves synchronization between video and audio, and alignment between motion and beat. Specifically, our model that incorporates beat information performs better across all metrics, highlighting the value of including beat information.

5. Discussion

Our DabFusion model effectively handles most of the dance generation scenarios; however, it does encounter some limitations in specific situations, as illustrated in Fig. 7. The first limitation occurs when a dancer’s skeleton cannot be clearly distinguished, such as in the left case of Fig. 7. Additionally, challenges arise when the dancer’s starting pose is atypical for the dance style being simulated, as shown in the right case of Fig. 7. This will lead DabFusion to generate repetitive frames. Moreover, DabFusion is currently limited to creating videos of a fixed length. While chaining sequences together using the final frame of one video as the starting frame for the next can theoretically extend video length, our tests show that this degrades video quality. This deterioration is due to the final frames lacking the fidelity of typical video frames. Future research could improve the quality of longer videos and enable choreography of new dancers without pre-segmentation, making the process simpler and more versatile.

6. Conclusion

We present the novel challenge of generating dance videos from music using images of specific individuals, tackled by our Dance Any Beat Diffusion model (DabFusion). It can animate any individual from a single reference image and synchronize their movements to diverse musical styles, demonstrating its versatility and adaptability. Our innovative music encoder is designed to pinpoint critical features such as dance style, movement, and rhythm, which are crucial for dynamic choreography. Employing the AIST++ dataset, we validate DabFusion’s ability to create high-quality dance videos that achieve excellent audio-video synchronization and precise motion-music alignment. This research establishes a solid baseline for future explorations in this intriguing field of dance video generation.

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. 3
- [2] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellström. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6158–6166, 2017. 3
- [3] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017. 3
- [4] 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. 2
- [5] Özgün Çiçek, 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. 5
- [6] Joao P Ferreira, Thiago M Coutinho, Thiago L Gomes, José 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. 3
- [7] 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. 3
- [8] 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. 2
- [9] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7134–7143, 2019. 3
- [10] 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. 6
- [11] 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. 3
- [12] 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
- [13] Ruozhi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. *arXiv preprint arXiv:2006.06119*, 2020. 3
- [14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in Neural Information Processing Systems*, 28, 2015. 4
- [15] 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. 5, 6
- [16] 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. 3
- [17] 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. 3
- [18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [19] 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. 7
- [20] 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. 3
- [21] 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. 1
- [22] 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, 2, 3, 5, 6
- [23] 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. 2, 3
- [24] 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. 3
- [25] 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. 2, 3

- [26] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 6
- [27] 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. 1, 3
- [28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 779–788, 2016. 7
- [29] 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. 2, 3
- [30] 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. 2
- [31] 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. 2
- [32] 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
- [33] 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. 6
- [34] 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. 6
- [35] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to body dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7574–7583, 2018. 3
- [36] 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. 2, 3, 4, 6
- [37] 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. 2
- [38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5
- [39] 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. 1, 3
- [40] 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. 3
- [41] 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. 3
- [42] 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
- [43] 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. 6
- [44] 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. 3
- [45] 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. 3, 4
- [46] 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. 2
- [47] 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. 2
- [48] 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
- [49] 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. 2
- [50] 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. 2, 3

- [51] 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. 2, 3
- [52] 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. 3
- [53] Guy Yariv, Itai Gat, Sagie Benaïm, 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. 6
- [54] 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. 3
- [55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 586–595, 2018. 6
- [56] 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. 2
- [57] 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. 3