This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MikuDance: Animating Character Art with Mixed Motion Dynamics

Jiaxu Zhang[1,2] Xianfang Zeng[2†] Xin Chen[3] Wei Zuo[2] Gang Yu[2‡] Zhigang Tu[1‡]
[1]Wuhan University [2]StepFun [3]ByteDance

Project page: https://kebii.github.io/MikuDance

Abstract

We propose MikuDance, a diffusion-based pipeline incorporating mixed motion dynamics to animate stylized character art. MikuDance consists of two key techniques: Mixed Motion Modeling and Mixed-Control Diffusion, to address the challenges of high-dynamic motion and reference-guidance misalignment in character art animation. Specifically, a Scene Motion Tracking strategy is presented to explicitly model the dynamic camera in pixel-wise space, enabling unified character-scene motion modeling. Building on this, the Mixed-Control Diffusion implicitly aligns the scale and body shape of diverse characters with motion guidance, allowing flexible control of local character motion. Subsequently, a Motion-Adaptive Normalization module is incorporated to effectively inject global scene motion, paving the way for comprehensive character art animation. Through extensive experiments, we demonstrate the effectiveness and generalizability of MikuDance across various character art and motion guidance, consistently producing high-quality animations with remarkable motion dynamics.

1. Introduction

Character art plays a crucial role in the film, game, and digital design industries. Animating character art, which brings static character images to life, has been an increasingly prominent challenge in computer vision and graphics. Traditional animation software, such as MMD [11], and Live2D [22], requires professional skills, posing significant barriers for non-experts. Recently, image-to-video generation methods [7, 19, 21, 40, 50] have emerged as a promising solution for animation. However, these methods are primarily designed for animating real-world humans and cannot be directly applied to character art due to the following two key challenges.

The first challenge in animating character art arises from the *high-dynamic motion guidance* for both the complex foreground and background, making unified control and consistently temporal maintaining difficult. For instance, in the second drawing shown in Figure 1, the girl is portrayed in an elegant dress against an artistic background, driven by large-scale dance motion and camera movements. Existing image animation methods, such as Animate Anyone [14] and DISCO [36], are primarily limited to animating humans with a static camera and a clean background. In contrast, animating character art requires the model to handle large-scale motion within complex scenes. As a result, simultaneously modeling the high-dynamic motion of both characters and the backgrounds becomes a critical task.

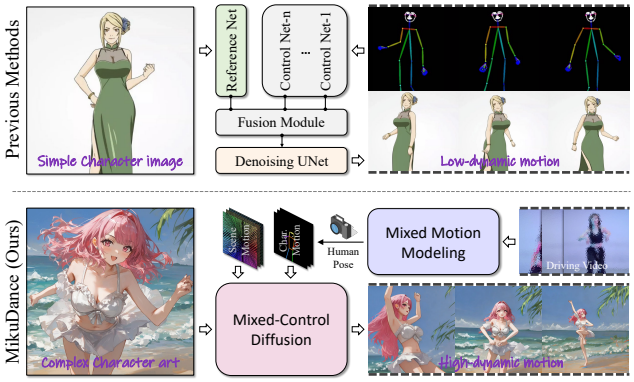

Figure 1. **We propose MikuDance**, a Diffusion-based pipeline for animating complex and stylized character art with high-dynamic motion guidance. The core insight of MikuDance lies in its *Mixed Motion Modeling* and *Mixed-Control Diffusion* capabilities.

The second challenge in animating character art stems from the *unique body shapes and diverse scales* of characters, which often misalign with motion guidance. For example, anime characters exhibit a large head-to-body ratio, exaggerated poses, and varied artistic styles. As shown in Figure 1, previous methods employ separate networks to process the reference image and motion guidance, oversimplifying the task by assuming a pre-aligned human body [30, 51], or performing alignment through pre-processing [37], which often leads to unnecessarily complex motion control architectures. However, since art images feature distinct characters, explicit alignment becomes impractical. Thus, implicitly aligning the reference and guidance within a unified structure presents a significant task.

†Xianfang Zeng is the project leader.
‡Corresponding authors: skicy@outlook.com, tuzhigang@whu.edu.cn

To address these challenges and leverage recent advancements in video generation for character art animation, we propose MikuDance. MikuDance animates in-the-wild character art by utilizing mixed character-scene motion guidance to generate videos with large-scale motion dynamics. It incorporates two key techniques: Mixed Motion Modeling and Mixed-Control Diffusion.

Mixed Motion Modeling explicitly represents character motions and 3D dynamic camera movements within a unified 2D space, enabling local and global motion guidance of both foreground and background in animations. Unlike previous methods that use camera parameters directly as the control signal [9, 39, 42], we propose a Scene Motion Tracking (SMT) strategy to model the global motion. SMT strategy projects the reference image to a scene point cloud and tracks corresponding points across consecutive camera frames, transforming camera poses into a pixel-wise scene motion representation. This scene motion, resembling keypoint-based character motion, establishes the foundational basis of our mixed motion control approach.

Mixed-Control Diffusion addresses misalignments in scales and body shapes of characters by integrating all reference and motion guidance into a unified Reference UNet [29]. This design is based on our observation that the mixed and implicit alignment approach outperforms other sophisticated control networks while preserving an elegant model architecture. Moreover, as scene motion guides the global dynamics of the animation, we carefully design a Motion-Adaptive Normalization (MAN) module to flexibly inject the scene motion into the Reference UNet, effectively integrating global dynamics and maintaining local consistency in character art animation.

By leveraging these two techniques and a mixed-source training approach, MikuDance animates diverse character art with mixed motion dynamics. We evaluate MikuDance using a range of reference characters and motion guidance. Both qualitative and quantitative results demonstrate that MikuDance can generate high-quality animation, particularly in maintaining consistency in character local motion and effectively handling large-scale scene motion.

Contributions of our MikuDance are listed below:

- Mixed Motion Modeling is proposed to explicitly model character and camera motions within a unified pixel-wise space, enabling the effective representation of high-dynamic motion.
- Mixed-Control Diffusion is exploited to implicitly align character shape, pose, and scale with the motion guidance, enabling cohesive motion control for character art animation.
- Extensive experiments demonstrate the effectiveness and generalizability of our MikuDance, achieving superior animation quality and high-dynamic motion control compared to state-of-the-art methods.

2. Related Work

Human image animation, which aims to generate human action videos from a reference image, has been studied extensively in recent years. Early methods, such as FOMM [31] and Liquid Warping GAN [20], warped the source human image using affine transformations guided by dense optical flow. Recently, diffusion-based approaches have gained traction due to their strong generalizability [3]. For instance, PIDM [2] introduced a texture diffusion module to model the correspondence between human appearance and poses. DreamPose [16] incorporated an adapter module to integrate both CLIP [27] and VAE [17] features of reference images. Animate Anyone [14] proposed a UNet-based ReferenceNet to extract appearance from reference images, along with a Pose Guider to encode the driving pose sequences, and a temporal module introduced by AnimateDiff [8] to enhance video consistency.

Following Animate Anyone, many diffusion-based models have been successively proposed, incorporating specific human motion guidance such as dense poses [43], 3D human models [51], and hand sequences [49]. Additionally, DISCO [36] proposed decoupling human subjects from backgrounds, and UniAnimate [37] employed a unified Denoising UNet to generate long-term video. While these methods introduce increasingly stringent human body priors, they do not account for the variable shapes and scales found in anime characters. Moreover, Animate-X [32] generalizes the pipeline to anthropomorphic characters but neglects the broader movements of the entire scene. In contrast, we design the Mixed-Control Diffusion to address the challenges of misalignment and comprehensive motion control in animating character art.

Controllable video generation builds upon the success of image generation by integrating additional spatial and temporal control signals. For example, VideoCrafter [4] and DynamiCrafter [41] control the first frame of the generated video through an image injection module. DragNUWA [46] integrates motion trajectories to control object movements. Animate Anything [5] introduces area guidance and motion strength guidance to achieve fine-grained motion control. In this work, the central focus is on the unified guidance of both character motion and camera movements.

Moving forward, existing works on camera control in video generation [9, 42, 45] often use Plücker coordinates [15] as an embedding of camera poses. Recent works such as Human4DiT [30] and HumanVid [39] consider camera movement in human videos by using an independent camera encoder and directly adding this embedding into the Denoising UNet. However, this camera embedding has a significant domain gap compared to the image pixel-wise guidance of character motion, making it difficult to fuse and maintain consistency in animation. In this work, we

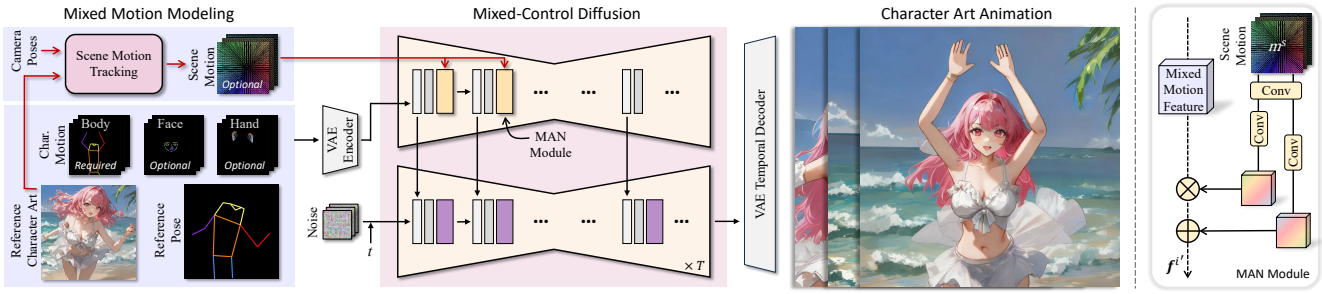

Figure 2. **Illustration of our MikuDance pipeline.** Given a reference character art and a driving video, the pixel-wise scene motion is predicted using the Scene Motion Tracking (SMT) strategy, which is combined with the character poses to form the character-scene mixed motion guidance. The Mixed-Control Diffusion subsequently generates the animation in a latent space, guided by the character poses and the scene motion injected through the Motion-Adaptive Normalization (MAN) module.

explicitly track camera movement in pixel space and integrate it with character motion guidance using the Motion-Adaptive Normalization module, enabling high-dynamic motion modeling in character art animation.

3. Method

As illustrated in Figure 2, given a character art $\mathcal{I}$ and a driving video $\mathcal{V}$, the purpose of our MikuDance is to animate the image $\mathcal{I}$ with reference to the human and camera motion in video $\mathcal{V}$. Specifically, we utilize Xpose [44] to separately extract pose sequences of the human body, face, and hand, and employ DROID-SLAM [33] to extract the camera poses $\{\boldsymbol{p}_l^c\}_{l=1}^L$, $\boldsymbol{p}^c \in \mathbb{R}^{L\times 7}$ from $\mathcal{V}$. L indicates the sequence length. The character's initial body pose, which exhibits significant scale and pose differences compared to the driving video, is also extracted from $\mathcal{I}$. Next, the image $\mathcal{I}$ and all the reference and driving pose images are encoded into the latent space through a VAE Encoder. The camera poses $\boldsymbol{p}^c$ are processed through the Scene Motion Tracking strategy to obtain the pixel-wise scene motion guidance. Then, the Mixed-Control Diffusion is used to animate $\mathcal{I}$ guided by the mixed motion guidance of the character poses and scene motion in the latent space. Finally, the latent output is decoded through the VAE Temporal Decoder to produce the character art animation.

3.1. Preliminaries on Stable Diffusion

Stable Diffusion (SD) [28] is a popular Latent Diffusion Model for text-to-image generation. SD consists of a VAE [17] for auto-encoding the images, and a UNet [29] for noise estimation to iteratively transform a noise image into a latent image by the reverse diffusion process [12]. Given an input data distribution $\boldsymbol{x}_0$, the forward process applies a Markov noising process of T steps on $\boldsymbol{x}_0$ to obtain $\{\boldsymbol{x}_t\}_{t=0}^T$:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\left(\sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\boldsymbol{I}\right), \tag{1}$$

where $\alpha_t \in (0,1)$ are constant hyper-parameters. When α_t is small enough, $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The reverse process takes a noisier data distribution $\boldsymbol{x}_t$ and generates a less noisy distribution $\boldsymbol{x}_{t-1}$ using an UNet, which is trained with the simple loss function:

$$\mathcal{L}_{simple} := \mathbb{E}_{\boldsymbol{\epsilon},t,c}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, c)\|_2^2\right], \tag{2}$$

where $\boldsymbol{\epsilon}$ is the Gaussian noise. c is the text condition. $\boldsymbol{\epsilon}_\theta(\cdot)$ is the trainable noise predictor. In this work, inspired by Animate Anyone [14], we utilize the pre-trained SD-1.5 as our base model to develop our animation pipeline, MikuDance.

3.2. Mixed Motion Modeling

Following existing human image animation methods [3, 14, 36], we use image-based pose sequences as motion guidance for characters. Unlike previous approaches that directly extract the character's whole-body poses, we separately extract poses for the body, face, and hands, allowing the face and hands to be optional and enabling more flexible motion control. However, character animation often involves high-dynamic motion throughout the entire scene to enhance the visual impact of storytelling. Traditional pose sequences offer only character motion guidance, lacking representations for background dynamics. To address this, we introduce the Scene Motion Tracking strategy.

Scene Motion Tracking (SMT). As illustrated in Figure 3, given a camera pose $\boldsymbol{p}_l^c$ from the driving video at the l-th frame, a scene point cloud $\boldsymbol{\phi}^l \in \mathbb{R}^{N\times 3}$ of the character art $\mathcal{I}$ is constructed in the $\boldsymbol{p}_l^c$ coordinate system using the depth map of $\mathcal{I}$. N is the number of points, which depends on the size of the image. Next, $\boldsymbol{\phi}^l$ is transferred into the world coordinate system through the camera-to-world matrix $\mathcal{T}^l \in \mathbb{R}^{N\times 4\times 4}$ of $\boldsymbol{p}_l^c$, resulting in the point cloud $\boldsymbol{\phi}^w$. Subsequently, by applying the world-to-camera matrix $\mathcal{Y}^{l+1} \in \mathbb{R}^{N\times 4\times 4}$ of the camera pose $\boldsymbol{p}_{l+1}^c$, we obtain the point cloud $\boldsymbol{\phi}^{l+1}$ for the next frame. Finally, we project $\boldsymbol{\phi}^l$ and $\boldsymbol{\phi}^{l+1}$ into the image coordinate system using the intrinsic matrices $\mathcal{K} \in \mathbb{R}^{N\times 3\times 4}$ of $\boldsymbol{p}_l^c$ and $\boldsymbol{p}_{l+1}^c$, respectively. The projected images are denoted as $\mathcal{I}^l$ and $\mathcal{I}^{l+1}$. Since these two projected images are rendered from the same scene point cloud in the world coordinate system, we can calculate the scene motion $\boldsymbol{m}^s \in \mathbb{R}^{N\times 2}$ in the image space based on the correspondence of the points. This

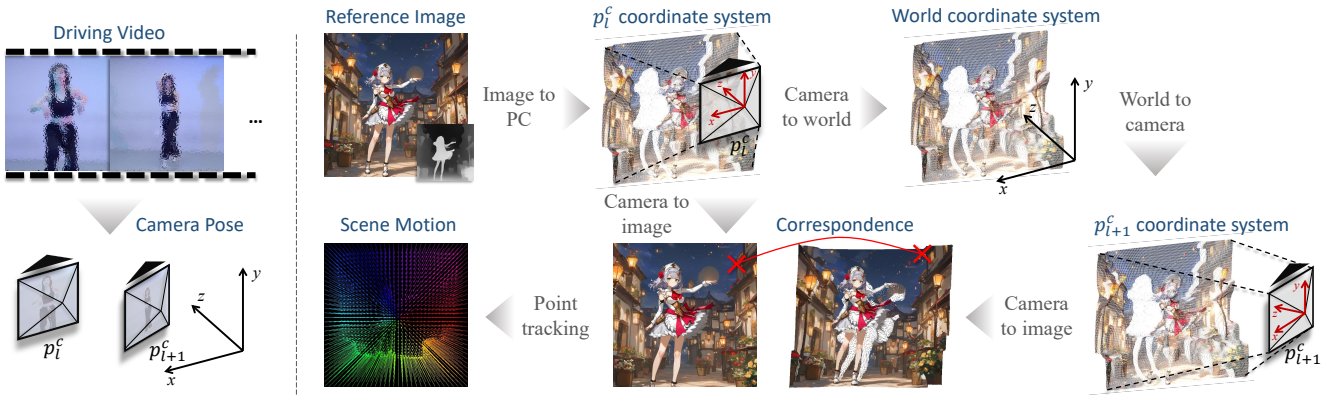

Figure 3. **Illustration of the Scene Motion Tracking strategy.** To effectively guide global background motion, 3D camera poses extracted from the driving video are transformed into a pixel-wise 2D space through the projection of the scene's point cloud (PC).

process is formulated as:

$$(z^l - z^{l+1}) \begin{bmatrix} m^s \\ 1 \end{bmatrix} = \mathcal{K}^l \begin{bmatrix} \phi^l \\ 1 \end{bmatrix} - \mathcal{K}^{l+1} \mathcal{Y}^{l+1} \mathcal{T}^l \begin{bmatrix} \phi^l \\ 1 \end{bmatrix}, \quad (3)$$

where z represents the projected Z-axis coordinates, serving as a scale factor for the scene motion m^s.

Notably, our SMT strategy diverges from the optical flow commonly used in video generation methods [6, 23] in two key respects: first, the scene motion extracted by SMT is independent of the driving video's content, whereas optical flow is content-dependent. Second, SMT tracks 3D points from the point cloud, while optical flow tracks pixel movements in the image domain, without considering the actual 3D scene. Consequently, our SMT strategy provides decoupled camera dynamic information, which is crucial for continuous background motion in character art animation.

In the proposed SMT process, we assume that the character and scene are static and standardized in the first camera. However, in real applications of character art animation, the reference scene often misaligns with the camera scale of the driving video, and the character pose varies in each frame. This ambiguity cannot be explicitly eliminated and necessitates the model to be implicitly perception guided by the character pose and the art image. Therefore, we propose Mixed-Control Diffusion in the next section.

3.3. Mixed-Control Diffusion

The concept of our Mixed-Control Diffusion is to mix and fuse all motion guidance for the character and scene within a unified reference space, thereby achieving aligned motion control over the animation.

As illustrated in Figure 2, drawing inspiration from Animate Anyone [14], we utilized the pre-trained SD-1.5 as the base Denoising UNet and a copy of it as the Reference UNet to achieve controllable image-to-video generation. Distinct from Animate Anyone and other related work, we eliminate separate encoders for motion guidance and simultaneously encode the reference character art, the reference pose, and all character pose guidance using the VAE Encoder, embedding them into the same latent space. Next, all embedded guidance is concatenated along the channel dimension to serve as the input for the Mixed-Control Reference UNet. To accommodate this mixed input, we expand the channel of the input convolution layer in the Reference UNet, initializing the added parameters with zero convolution weights [47]. Additionally, the reference image is embedded using the CLIP image encoder [27] and serves as the key features in the cross-attention operations of both the Denoising UNet and the Reference UNet. This process is commonly used in existing work and is therefore omitted from Figure 2.

In each denoising step t of the Mixed-Control Diffusion, the self-attention features from the Reference UNet are injected into the Denoising UNet through an addition operation. The Reference UNet requires inference only once, while the denoising process is repeated T times. The experiments demonstrate that our mixed control approach outperforms other control encoding and fusion methods by effectively addressing the misalignments between the reference image and the motion guidance. Furthermore, since scene motion exerts a global influence on the animation frames, it is intuitive to integrate it with character motion using an adaptive normalization method, as introduced below.

Motion-Adaptive Normalization (MAN) is designed to effectively mix the extracted pixel-wise scene motion, m^s, and enhance the temporal consistency of both foreground and background animations.

Inspired by SPADE [24], which employs a spatial-aware normalization method to capture semantics for image synthesis, we propose to implement spatial-aware normalization adapted by scene motion, as shown in the right part of Figure 2. Let the mixed motion feature of the i-th block in the Reference UNet be denoted as f^i. We first normalize it using the Instance Normalization operation. Then, the scene motion m^s is processed through three convolutional layers to obtain the motion-adapted standard deviation $\gamma^i \in \mathbb{R}^{C \times H \times W}$ and the mean $\beta^i \in \mathbb{R}^{C \times H \times W}$. Here, C, W, H represent the channel, height, and width of the feature, respectively. It is important to note that both β^i and γ^i have spatial dimensions, enabling pixel-wise guidance

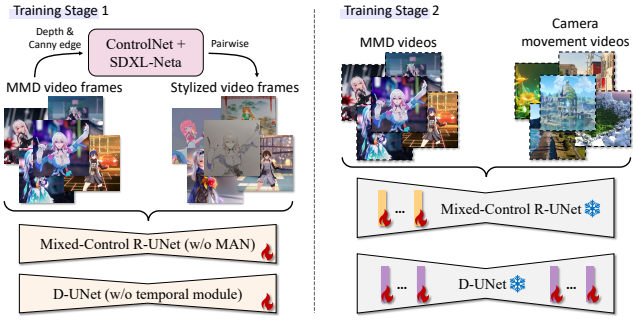

Figure 4. **The mixed-source training approach.** We utilize synthetic stylized video frames and non-character videos in the two training stages, respectively, to enhance generalizability.

of the entire scene motion. This process is formulated as:

$$\boldsymbol{f}^{i'} = \boldsymbol{\gamma}^{i}_{C,H,W}\left(\boldsymbol{m}^{s}\right)\frac{\boldsymbol{f}^{i}_{C,H,W} - \boldsymbol{\mu}^{i}_{C}}{\boldsymbol{\sigma}^{i}_{C}} + \boldsymbol{\beta}^{i}_{C,H,W}\left(\boldsymbol{m}^{s}\right), \quad (4)$$

where $\boldsymbol{\mu}^{i}_{C}$ and $\boldsymbol{\sigma}^{i}_{C}$ are the mean and standard deviation of $\boldsymbol{f}^{i}$ along the channel dimension.

With the proposed Mixed Motion Modeling and Mixed-Control Diffusion, along with the incorporation of the MAN module after each down-sampling block in the Reference UNet, we outline the complete pipeline of our MikuDance. Furthermore, to enhance MikuDance's ability to accommodate various styles of character art and the dynamics of large-scale camera movements, we present a mixed-source training approach in the next section.

3.4. Mixed-Source Training Approach

Considering that image animation is a data-intensive task, proposing an effective data and training pipeline is as crucial as the model itself. In our MikuDance, as illustrated in Figure 4, we adopt a mixed-source training approach with two-stage training stages.

In the first stage, training is conducted using pair-wise video frames, without incorporating the MAN module of the Reference UNet or the temporal module of the Denoising UNet. Different from existing methods [14, 37], we randomly mix stylized pair-wise frames by concatenating the initial frames along the spatial dimension and utilize the depth and edge-controlled anime SDXL model [26, 47], known as SDXL-Neta [18], to transfer the art style while preserving the image content. Additionally, to simulate the inference process in which the reference character art is irrelevant to the driving pose, we randomly select reference frames that are not involved in the target sequence.

In the second stage, both the MAN module and the temporal module are incorporated into our Mixed-Control Diffusion model, while the other parameters remain frozen during this phase. The training data in this stage consists of mixed MMD video clips and camera movement videos that do not include characters. Importantly, we randomly drop the pose and motion guidance during the two-stage training to enhance the robustness of our MikuDance.

4. Experiments

Datasets. To train our MikuDance, we collected an MMD video dataset comprising 3,600 animations created by artists, all rendered from 3D models. We split these videos into approximately 120,000 clips, which together include over 10.2 million frames. Additionally, we incorporated around 3,500 non-character camera movement videos in the second training stage. For quantitative evaluation, we used 100 MMD videos that were not included in the training set, with their first frames serving as reference images. We used Xpose [44] for character pose and DROID-SLAM [33] for camera pose extraction. For qualitative evaluation, all character art was randomly generated using SDXL-Neta [18] and the driving videos were unseen during training.

Implementation details. We implement MikuDance using the SD-1.5 framework [28] and PyTorch [25]. Experiments are conducted on 16 NVIDIA A800 GPUs. In the first training stage, the video frames are center-cropped and resized to a resolution of 768×768. Training is conducted for 120,000 steps with a batch size of 128. In the second training stage, we train the MAN module and the temporal module for 60,000 steps using 24-frame video sequences and a batch size of 16. Both learning rates are set to 1e-5, and the dropout ratio for the pose and scene motion guidance is set to 0.2. During inference, we use a DDIM sampler for 20 denoising steps. We adopt the temporal aggregation method described in [34] to generate long videos.

Evaluation metrics. Following DISCO [36], we evaluate the results from two aspects: image and video. To assess image quality, we report frame-wise FID [10], SSIM [38], LISPIS [48], PSNR [13], and L1. For video quality, we concatenate every consecutive 16 frames to form a sample, from which we report FID-VID [1] and FVD [35].

4.1. Qualitative Results

Comparison with the baselines. We compare our MikuDance with recent human video generation methods, including Animate Anyone (AniAny) [14], DISCO [36], MagicPose [3], and UniAnimate [37], all of which claim the capability to animate anime-style characters in their official reports. Additionally, we implemented AniAny* by fine-tuning the model on our MMD video dataset.

The results in Figure 5 show that AniAny, MagicPose, and UniAnimate fail to address the misalignment of character shape and scale, resulting in character distortion in their outputs. While DISCO uses independent ControlNets to process background and foreground features, its results suffer from scene collapse when animating character art. Although AniAny* is specifically fine-tuned on the anime-style dataset, its results show limited improvements and blurring in high-dynamic motion, as the pipeline fails to account for background scene motion. Notably, MikuDance

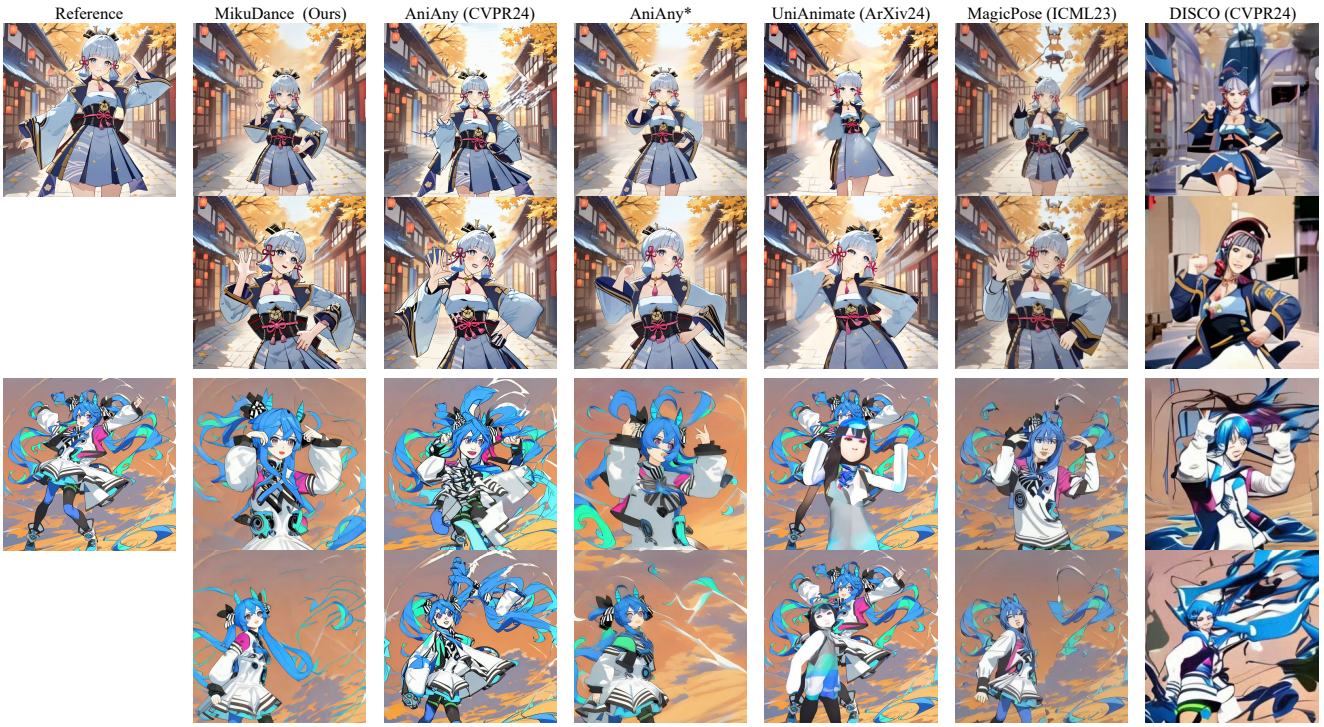

Figure 5. **Comparison with the baselines.** AniAny* is the fine-tuned version of the AniAny model, trained on our MMD video dataset.

Figure 6. **Experiments on high-dynamic motion guidance.** Large camera movements (left) and significant pose variations (right).

effectively handles complex reference and motion guidance, delivering high-quality and vivid animation results.

High-dynamic motion. A key highlight of MikuDance is its ability to handle high-dynamic motion guidance in animation, which goes beyond the simple actions used in existing methods. As shown in Figure 6, the character performs large dance movements with a fast-moving camera. Despite these challenging conditions, MikuDance, equipped with our Mixed Motion Modeling approach, demonstrates remarkable robustness, delivering high-fidelity animation results that effectively capture the dramatic visual impact.

Reference-guidance misalignments. Another key contribution of MikuDance is its implicit alignment of the reference character with motion guidance. As illustrated in Figure 7, two examples show significant spatial and scale misalignments between the guidance and the reference. In such cases, existing methods like AniAny struggle to animate the reference character effectively, whereas MikuDance successfully manages these complexities and generates coherent animations.

Various shapes and scales. MikuDance effectively handles variations in character shapes and scales. As shown in the left part of Figure 8, characters with distinct body shapes, various poses, and different clothing are precisely driven by the same motion guidance. In the right part of Figure 8, MikuDance demonstrates its ability to implicitly align characters of varying scales, preserving each character's unique features and producing reasonable animation results.

Generalizability on various art styles. As illustrated in Figure 9, MikuDance, leveraging our mixed-source training approach, can handle a wide range of art styles, including but not limited to celluloid, antiquity, and line sketch. This high level of generalizability opens up broad prospects for real-world applications.

Ablation study. We conduct ablation experiments to verify the key designs of our MikuDance, as shown in Figure 10, which include the mixed-control architecture (MIX), the MAN module, and the SMT strategy.

To evaluate the mixed-control design, we implemented a pipeline (w/o MIX) inspired by AniAny, utilizing an in-

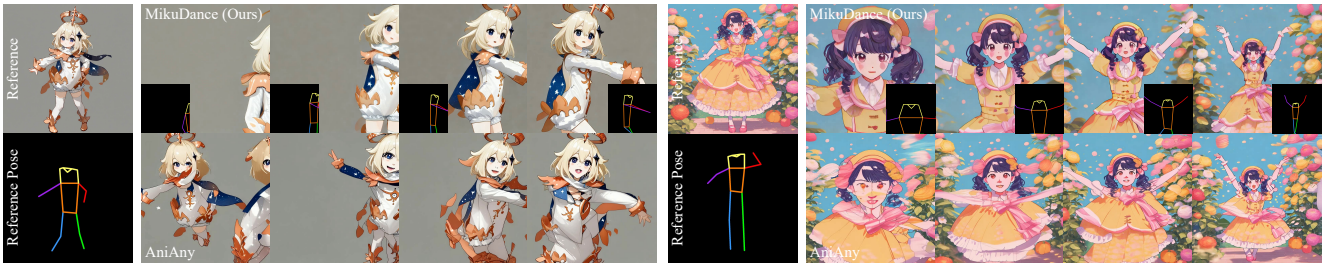

Figure 7. **Experiments on misalignments between reference images and driving poses.** Spatial (left) and scale (right) misalignments.

Figure 8. **Experiments on various shapes (the left part) and scales (the right part) of the reference character art.**

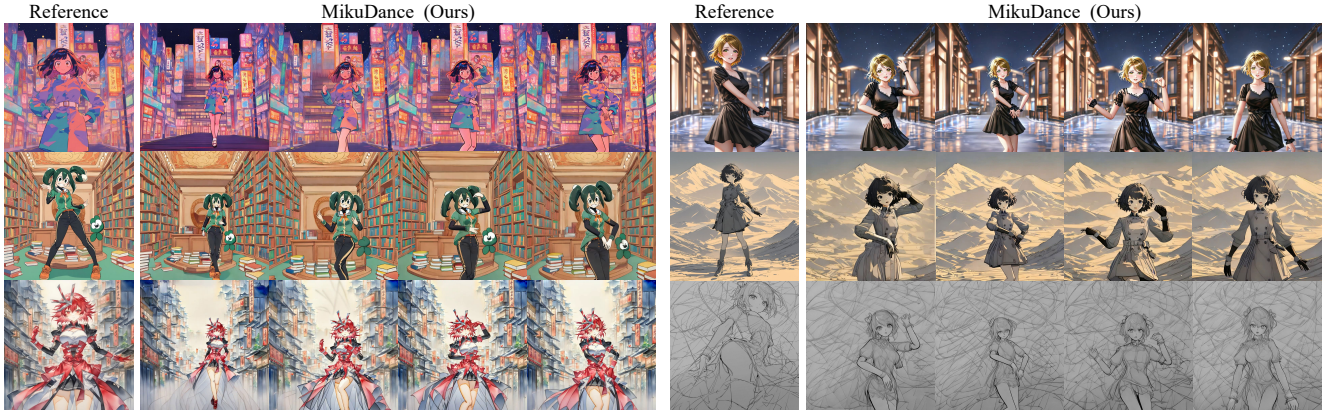

Figure 9. **Experiments on various styles of the reference character art.** Please see Appendix for more results.

dependent Reference UNet to process the reference image and two ControlNets to separately adapt the character and scene motion guidance. The results indicate that this conventional pipeline fails to account for scale differences between the character art and the driving guidance, leading to a mismatched appearance of the character's face and pose.

To evaluate the effectiveness of the MAN module, we implemented a pipeline without MAN (w/o MAN) that simply concatenates scene motion with character motion and inputs them together into the Reference UNet. While this approach yields better results than a pipeline without scene motion guidance (w/o SMT), it remains inferior to the results achieved by MikuDance. This is because the MAN module injects global motion through spatial-aware normalization, effectively complementing the local motion.

To evaluate the SMT strategy, we conducted three experiments: one pipeline without incorporating scene motion (w/o SMT), and two pipelines that replaced the scene motion with Plücker embedding (w/ Plücker) and optical flow (w/ Flow), respectively. However, the results from these alternative approaches were inferior to our SMT strategy, showing noticeable artifacts and inconsistencies in the dynamic backgrounds. The pixel-wise scene motion extracted by SMT proved to be a more effective representation for guiding background motion due to its domain consistency with the character motion guidance.

Compared to the ablative studies introduced above, our MikuDance effectively addresses the misalignment and high-dynamic challenges in animating character art.

4.2. Quantitative Results

Table 1 presents quantitative comparisons between MikuDance and the baseline methods. It is important to note that the metrics reported in our paper are lower than those in

Figure 10. **Ablation experiments on the key designs of MikuDance.** MIX, SMT, MAN, Plücker, and Flow are defined in Section 4.1.

Table 1. **Quantitative comparisons with baselines and ablative experiments.** AniAny* is the fine-tuned version of the AniAny model. MIX, SMT, MAN, Plücker, and Flow are defined in Section 4.1. The best results are highlighted in bold, and the second-best are underlined. MikuDance achieves superior results across all metrics.

Methods	Image					Video	
	FID↓	SSIM↑	PSNR↑	LISPIS↓	L1↓	FID-VID↓	FVD↓
AniAny [14]	43.945	0.488	12.530	0.548	7.307E-05	38.179	846.414
AniAny*	28.833	0.526	13.610	0.517	6.229E-05	26.764	575.304
DISCO [36]	59.221	0.313	10.732	0.615	9.248E-05	46.852	923.921
MagicPose [3]	44.258	0.424	12.357	0.554	7.767E-05	41.347	886.691
UniAnimate [37]	47.328	0.417	12.074	0.571	7.930E-05	40.924	882.245
MikuDance w/o MIX	27.315	0.523	14.004	0.528	5.860E-05	24.124	541.453
MikuDance w/o MAN	24.985	0.542	14.501	0.505	5.753E-05	23.366	509.342
MikuDance w/o SMT	25.472	0.534	14.312	0.512	5.911E-05	23.362	517.673
MikuDance w/ Plüc.	25.918	0.538	14.261	0.510	6.011E-05	23.471	521.853
MikuDance w/ Flow	26.141	0.516	14.088	0.523	5.925E-05	23.079	505.533
MikuDance (Ours)	**24.597**	**0.576**	**14.592**	**0.493**	**5.726E-05**	**22.868**	**502.380**

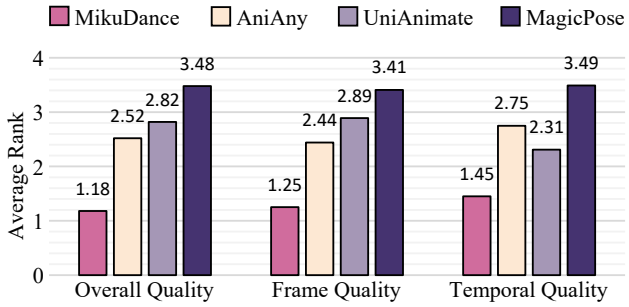

Figure 11. **User Study.** The smaller value means the better quality.

previous studies, as the entire scene in our testing videos is highly dynamic, unlike the static backgrounds used in earlier datasets. Nevertheless, the results demonstrate that MikuDance achieves state-of-the-art performance across all image and video metrics. Additionally, the ablation results confirm the effectiveness of the key design elements in MikuDance. In summary, by incorporating the proposed mixed motion dynamics techniques, MikuDance can animate a wide range of characters and generate high-quality image and video results.

User study. We invited 50 volunteers and gave them 20 videos to evaluate the performance of our MikuDance against the baseline methods. Each video includes one motion guidance and four anonymous animation results. We ask users to rank the four results in overall quality, frame quality, and temporal quality. After excluding abnormal questionnaires, the average rank of the methods is summarized in Figure 11. Our MikuDance outperforms the baseline methods by a large margin, and more than 97% of users prefer the animation generated by our MikuDance.

5. Conclusions

In this work, we propose MikuDance, a new animation pipeline designed to generate high-dynamic animations for in-the-wild character art. MikuDance incorporates two key techniques: Mixed Motion Modeling and Mixed-Control Diffusion. Mixed Motion Modeling enables the representation of large-scale character and scene motions within a unified reference space, while Mixed-Control Diffusion addresses misalignment between characters and motion guidance. To support diverse art styles, we also employ a mixed-source training approach to enhance generalizability. Extensive experiments demonstrate that MikuDance achieves state-of-the-art performance compared to baseline methods.
Limitations. We acknowledge that some generated animations exhibit background distortions and artifacts. This issue stems from the 3D-agnostic challenge in image animation, making scene reconstruction in dynamic cameras an ill-posed problem that requires further investigation.

Acknowledgements. This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China No. 624B2110, the National Key Research and Development Program of China No. 2024YFC3015600, and the Fundamental Research Funds for Central Universities No.2042023KF0180 & No.2042025KF0053. The numerical calculations are supported by the supercomputing system at the Supercomputing Center of Wuhan University and StepFun.

References

[1] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, page 2, 2019. 5

[2] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5968–5976, 2023. 2

[3] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *Forty-first International Conference on Machine Learning*, 2023. 2, 3, 5, 8

[4] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2

[5] Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Animateanything: Fine-grained open domain image animation with motion guidance. *arXiv e-prints*, pages arXiv–2311, 2023. 2

[6] Daniel Geng and Andrew Owens. Motion guidance: Diffusion-based image editing with differentiable motion estimators. In *The Twelfth International Conference on Learning Representations*, 2024. 4

[7] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 1

[8] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[9] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 2

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[11] Yu Higuchi. Mikumikudance. https://sites.google.com/view/evpvp. 1

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[13] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 5

[14] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 1, 2, 3, 4, 5, 8

[15] Yan-Bin Jia. Plücker coordinates for lines in the space. *Problem Solver Techniques for Applied Computer Science, Com-S-477/577 Course Handout*, 3, 2020. 2

[16] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22623–22633. IEEE, 2023. 2

[17] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 3

[18] Neta.art Lab. neta-art-xl-1.0. https://huggingface.co/neta-art/neta-art-xl-1.0. 5

[19] Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24142–24153, 2024. 1

[20] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5904–5913, 2019. 2

[21] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 1

[22] Tetsuya Nakajo. Live2d. https://www.live2d.com/. 1

[23] Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. *arXiv preprint arXiv:2405.20222*, 2024. 4

[24] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 4

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming

Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[26] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 5

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4

[28] 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. 3, 5

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2, 3

[30] Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. Human4dit: Free-view human video generation with 4d diffusion transformer. *arXiv preprint arXiv:2405.17405*, 2024. 1, 2

[31] 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

[32] Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. Animate-x: Universal character image animation with enhanced motion representation. *arXiv preprint arXiv:2410.10306*, 2024. 2

[33] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 3, 5

[34] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 5

[35] 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. 5

[36] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9326–9336, 2024. 1, 2, 3, 5, 8

[37] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *arXiv preprint arXiv:2406.01188*, 2024. 1, 2, 5, 8

[38] 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. 5

[39] Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, et al. Humanvid: Demystifying training data for camera-controllable human image animation. *arXiv preprint arXiv:2407.17438*, 2024. 2

[40] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 1

[41] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023. 2

[42] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 2

[43] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 2

[44] Jie Yang, Ailing Zeng, Ruimao Zhang, and Lei Zhang. Xpose: Detecting any keypoints. *arXiv preprint arXiv:2310.08530*, 2023. 3, 5

[45] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 2

[46] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 2

[47] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 4, 5

[48] 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. 5

[49] Jingkai Zhou, Benzhi Wang, Weihua Chen, Jingqi Bai, Dongyang Li, Aixi Zhang, Hao Xu, Mingyang Yang, and

Fan Wang. Realisdance: Equip controllable character animation with realistic hands. *arXiv preprint arXiv:2409.06202*, 2024. 2

[50] Bingwen Zhu, Fanyi Wang, Tianyi Lu, Peng Liu, Jingwen Su, Jinxiu Liu, Yanhao Zhang, Zuxuan Wu, Yu-Gang Jiang, and Guo-Jun Qi. Poseanimate: Zero-shot high fidelity pose controllable character animation. *arXiv preprint arXiv:2404.13680*, 2024. 1

[51] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. *arXiv preprint arXiv:2403.14781*, 2024. 1, 2