DreamDance: Animating Character Art via Inpainting Stable Gaussian Worlds

Jiaxu Zhang[1,2,3] Xianfang Zeng[3†] Xin Chen[4]
Wei Zuo[3] Gang Yu[3‡] Guosheng Lin[2] Zhigang Tu[1‡]
[1]Wuhan University [2]Nanyang Technological University [3]StepFun [4]ByteDance
Project page: https://kebii.github.io/DreamDance

Abstract

This paper presents DreamDance, a novel character art animation framework capable of producing stable, consistent character and scene motion conditioned on precise camera trajectories. To achieve this, we re-formulate the animation task as two inpainting-based steps: Camera-aware Scene Inpainting and Pose-aware Video Inpainting. The first step leverages a pre-trained image inpainting model to generate multi-view scene images from the reference art and optimizes a stable large-scale Gaussian field, which enables coarse background video rendering with camera trajectories. However, the rendered video is rough and only conveys scene motion. To resolve this, the second step trains a pose-aware video inpainting model that injects the dynamic character into the scene video while enhancing background quality. Specifically, this model is a DiT-based video generation model with a gating strategy that adaptively integrates the character's appearance and pose information into the base background video. Through extensive experiments, we demonstrate the effectiveness and generalizability of DreamDance, producing high-quality and consistent character animations with remarkable camera dynamics.

1. Introduction

Animating character art is a fundamental challenge in the 2D animation industry, with a wide range of applications in film, game, and digital design. However, traditional 2D animation is a labor-intensive and time-consuming process that requires expertise in professional software such as MMD [10] and Live2D [23]. Recently, human video generation methods, particularly MikuDance [51], have revolutionized this challenging task, making it accessible to non-experts. Derived from previous methods [2, 13, 59], MikuDance performs two strategies to animate character art using driving videos, as illustrated in Figure 1.

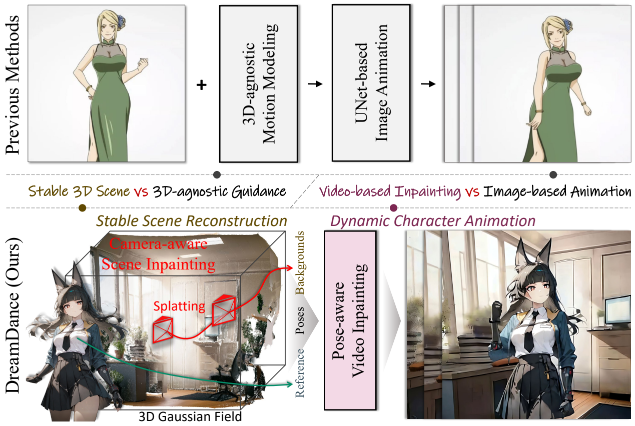
Previous Methods
+
3D-agnostic Motion Modeling
UNet-based Image Animation
Stable 3D Scene vs 3D-agnostic Guidance
Video-based Inpainting vs Image-based Animation
Stable Scene Reconstruction
Dynamic Character Animation
DreamDance (Ours)
Camera-aware Scene Inpainting
Splatting
3D Gaussian Field
Backgrounds
Poses
Reference
Pose-aware Video Inpainting

Figure 1. **We propose DreamDance**, a novel paradigm that re-formulates the character art animation task into two inpainting-based steps: *Camera-aware Scene Inpainting* for stable scene reconstruction and *Pose-aware Video Inpainting* for dynamic character animation.

The first strategy is motion modeling, which uses pose image sequences to drive characters and 2D scene flow to guide backgrounds. Similar to MikuDance, existing methods incorporate other motion guidance, such as optical flow [24, 42] and camera parameters [31, 44], to represent global camera movements. However, this motion guidance is entirely 3D-agnostic and struggles to provide consistent scene context, leading to scene distortion during large-scale camera movements. This inconsistency arises from the implicit inpainting process, where the scene dynamics exceed the area covered by the reference, requiring the model to hallucinate missing regions. Therefore, it is crucial to explore 3D-aware scene modeling for consistent camera control.

The second strategy is image animation, which utilizes the UNet-based Stable Diffusion model [29, 30] to animate the reference character art using mixed motion guidance. Additionally, some recent methods have developed incremental pose encoders [55, 59] and reference adapters [5, 40], using SVD [3] as a base model to achieve human image animation. However, due to the limitations of base model capacity and the ambiguity of mixed guidance, the animation results from these methods exhibit significant temporal jitters in both characters and backgrounds. There-

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

fore, it is crucial to introduce a more powerful video foundation model and redefine the animation process with explicit contextual scene guidance.

Unlike MikuDance and other relevant methods, we propose DreamDance, a new paradigm for animating in-the-wild character art. As illustrated in Figure 1, DreamDance reformulates the motion modeling and image animation processes into two inpainting-based steps: Camera-aware Scene Inpainting and Pose-aware Video Inpainting. These two components work synergistically to generate consistent, high-quality animation sequences from the reference character art and driving videos.

Camera-aware Scene Inpainting is presented for stable scene reconstruction. Inspired by existing 3D Gaussian methods [6, 39, 48], we leverage a pre-trained image inpainting model [53] to generate multi-view images and construct a large-scale 3D Gaussian field from the reference character art. This process utilizes both a pre-defined spiral camera trajectory and an extracted custom camera trajectory. A coarse background video is then rendered along the custom trajectory by splatting the stable Gaussian field. This background video contains consistent scene motion information and serves as a rough yet foundational video for the later character animation stage.

Pose-aware Video Inpainting is proposed for injecting dynamic character animation into the coarse scene video. Based on a video generation model, i.e., CogVideoX [47], we train a gated video inpainting model to refine the coarse backgrounds and inject the reference character according to the pose guidance. The gating strategy is designed to adaptively incorporate both the character's appearance and poses based on the denoising time step, ensuring character and background consistency throughout the animation. Additionally, we exploit a 3D Gaussian-free training approach to train the dynamic video inpainting model directly using background-degraded video datasets.

By leveraging this new paradigm, DreamDance animates diverse character art with stable scenes and precise camera movements, generating spatio-temporally consistent animations. We evaluate DreamDance using a wide range of reference character art and driving videos. Both qualitative and quantitative results demonstrate that DreamDance can generate high-quality animations, particularly with flexible and coherent scene dynamics.

Contributions of DreamDance are listed in three folds:

- Camera-aware Scene Inpainting and Pose-aware Video Inpainting are proposed to re-formulate the character art animation task, enabling explicit and consistent scene context modeling.
- A gating strategy is introduced into a fundamental video generation model to achieve adaptive video inpainting, enabling high-dynamic animation of character art within a stable Gaussian scene.
- Extensive experiments demonstrate the effectiveness and generalizability of DreamDance, achieving superior animation quality over state-of-the-art methods.

2. Related Work

2D character animation provides a vibrant platform for storytelling but has long been a challenge in the animation industry. Some previous methods construct animated 3D characters from reference images and re-project them into 2D videos [27, 33, 50, 58]. These methods require precise geometry, rigging, and motion editing, making them hard to automate, and often resulting in a loss of the original 2D style. Recent approaches like Textoon [8] and AniClipart [45] aim to generate animatable 2D characters using image and video generation models. However, they still require significant manual work, and the character's motion freedom is limited. In contrast, MikuDance [51] directly generates 2D animation through an image animation model, offering a promising solution, but it faces issues with scene distortion and artifacts due to its 3D-agnostic approach. Derived from MikuDance, we propose DreamDance, which introduces two inpainting steps for 3D context-aware and consistent character art animation in stable Gaussian scenes.

Human image animation has gained popularity in recent years [2, 20], with many methods building on pre-trained image and video generation models [3, 29]. For example, Animate Anyone [13] uses a reference-denoising UNet structure and a temporal module from AnimateDiff [7] to improve video consistency. DisCo [40] separates human subjects from backgrounds, allowing for more flexible combinations. MimicMotion [55] introduces a regional loss based on pose confidence to enhance human fidelity. Animate-X [35] extends the pipeline to anthropomorphic characters. However, these methods mainly focus on character actions and overlook scene dynamics.

Moving forward, Human4DiT [31] and HumanVid [44] address camera movements by incorporating a camera encoder. However, camera guidance alone cannot handle large-scale scene dynamics, as it is difficult for the model to consistently fill in missing areas. MIMO [22] and Animate Anyone 2 [14] capture environmental representations from the driving video and restore the backgrounds in the animation, but they focus on different goals than 2D character animation, where the scene comes from the reference character art. In this work, we reconstruct 3D Gaussian scenes by inpainting reference character art to support precise camera dynamics, and use an MM-DiT-based video foundation model to generate high-quality character animations.

Video inpainting typically focuses on two main tasks: object removal and text-guided inpainting. Traditional methods, like E^{2}FGVI [18] and FGT [52], use optical flow-guided feature propagation to reconstruct missing areas

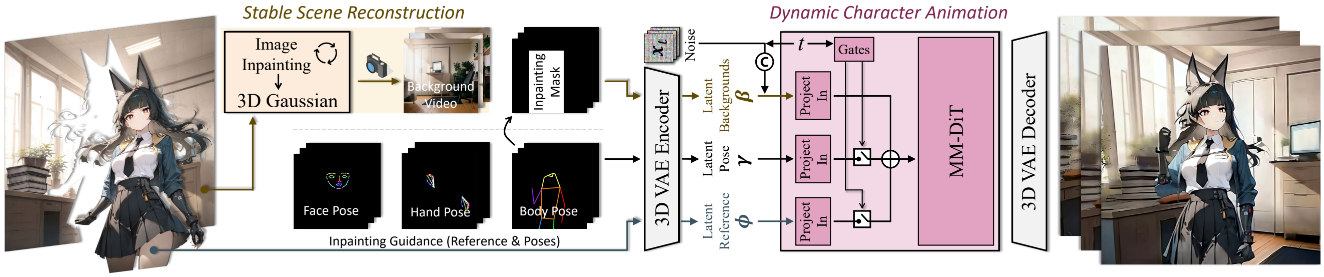

Figure 2. **Illustration of our DreamDance.** The reference character art is decomposed into foreground and background layers. The background image is used to reconstruct a stable 3D Gaussian scene through a wrap-and-inpaint scheme, enabling coarse background video rendering based on custom camera trajectories. The gated MM-DiT model then inpaints the background video based on the foreground character and the driving poses, generating dynamic character animations.

with coherent content. More recently, models like AVID [56] and CoCoCo [60] have integrated pre-trained generative inpainting models [29] with motion modules for text-guided video inpainting. Unlike these approaches, our video inpainting step focuses on filling in the coarse background with animated characters guided by pose videos. To achieve this, we propose a gating strategy within the DiT [26, 47] model to adaptively integrate the character's appearance and poses, enabling dynamic character animation.

3D Gaussian Splatting [15] utilizes the concept of Gaussian splats combined with spherical harmonics and opacity to represent 3D scenes. Later work incorporates image inpainting to generate multi-view images and reconstruct 3D Gaussian fields from a single image [6, 19, 39, 48, 61]. Inspired by these approaches, we reconstruct stable Gaussian scenes from the reference character art and render coarse background videos to improve scene consistency.

3. Method

As illustrated in Figure 2, given a character art $\mathcal{I}$ and a driving video $\mathcal{V}$, the goal of DreamDance is to animate the image $\mathcal{I}$ based on the human and camera motion in the video $\mathcal{V}$. Specifically, we utilize Xpose [46] to separately extract the pose sequences of the human body, face, and hand, and employ DPVO [36] 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 and the background are segmented from $\mathcal{I}$ using BiRefNet [57]. Next, we reconstruct a 3D Gaussian field from the reference background through multi-view image inpainting, using both a pre-defined spiral camera trajectory and the extracted camera trajectory. Then, a coarse background video is rendered by splitting the 3D Gaussian field according to the extracted camera poses, and an inpainting mask is generated from the driving pose sequence. Finally, all the references and guidance are processed using the pre-trained VAE and input into the gated video inpainting model for dynamic character animation.

3.1. Preliminaries

Diffusion Denoising Probabilistic Models. Diffusion-based generative models [11, 34] represent the data distribution by constructing a Markov chain. 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 $\{\boldsymbol{x}_t\}_{t=0}^T$:

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

where $\alpha_t \in (0,1)$ are constant hyper-parameters. When α_t is small enough, $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The reverse process takes a noisier data distribution $\boldsymbol{x}_t$ and generates a less noisy distribution $\boldsymbol{x}_{t-1}$ using a noise predictor, 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, we utilize the pre-trained CogVideoX, an MM-DiT-based video diffusion model, as the base model to achieve pose-aware video inpainting in DreamDance.

3D Gaussain Splatting (3DGS). Prior works [15, 62] propose to represent a 3D scene as a set of scaled 3D Gaussian primitives $\{\mathcal{G}_k|k=1,...,K\}$ and render scene images using volume splitting. The geometry of each scaled 3D Gaussian $\mathcal{G}_k$ is parameterized by an opacity $\alpha_k \in [0,1]$, center $\boldsymbol{o}_k \in \mathbb{R}^{3\times 1}$, and covariance matrix $\mathbf{\Sigma}_k \in \mathbb{R}^{3\times 3}$.

To render an image for a given camera defined by rotation $\mathbf{R} \in \mathbb{R}^{3\times 3}$ and translation $\mathbf{t} \in \mathbb{R}^3$, the 3D Gaussians are first transformed into camera coordinates:

$$\boldsymbol{o}_k^{'} = \mathbf{R}\boldsymbol{p}_k + \mathbf{t}, \quad \mathbf{\Sigma}_k^{'} = \mathbf{R}\mathbf{\Sigma}_k\mathbf{R}^T. \tag{3}$$

Then, they are projected to ray space via a local affine transformation. Finally, 3DGS utilizes spherical harmonics to model view-dependent color $\boldsymbol{c}_k$ and renders images via alpha blending according to the primitive's depth order:

$$c(x) = \sum_{k=1}^{K} \boldsymbol{c_k}\alpha_k\mathcal{G}_k^{2D}(x)\prod_{j=1}^{k-1}\left(1-\alpha_j\mathcal{G}_j^{2D}(x)\right), \tag{4}$$

where $\mathcal{G}^{2D}$ is the scaled 2D Gaussian, obtained by removing the third row and column of the ray space covariance matrix. In this work, we reconstruct a 3D Gaussian field by inpainting the reference image and then rendering the coarse background video using volume splitting.

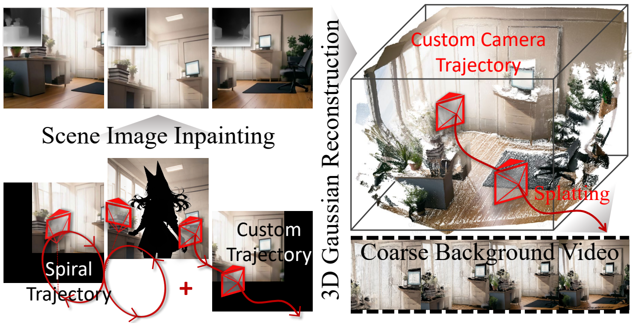

Figure 3. **Camera-aware Scene Inpainting** for stable scene reconstruction. We use both the pre-defined spiral camera trajectory and the custom camera trajectory to reconstruct a 3D Gaussian field via the warp-and-inpaint scheme.

3.2. Stable Scene Reconstruction

Existing 3D-agnostic motion guidance makes consistent background generation during large-scale camera movements an ill-posed problem [10]. Therefore, we reconstruct stable Gaussian scenes to facilitate character art animation.

Camera-aware Scene Inpainting. As illustrated in Figure 3, inspired by the warp-and-inpaint scheme [32, 39], we use a pre-defined spiral camera trajectory to reconstruct the 3D scene. Firstly, at the starting point of the camera trajectory, we use LLaVA [21] to generate detailed descriptions of the reference background, and use Fooocus [53] to inpaint the empty regions left by the removed character. Afterward, we estimate the global depth map on this complete background using DepthPro [4]. Next, as the camera moves along the spiral trajectory, we warp the background image to each new viewpoint using its depth map, and then fill the empty regions through image inpainting. After this warp-and-inpaint process, we obtain a set of RGBD images, which are then used to train a stable 3D Gaussian field. The spiral trajectory can be formulated as:

$$\mathbf{P}(t) = \begin{bmatrix} r \cdot sin(2\pi t) \cdot cos(2\pi t) \\ r \cdot sin(2\pi t) \cdot sin(2\pi t) \\ -sin(2\pi t) \end{bmatrix}, \\ \mathbf{R}(t) = \begin{bmatrix} norm(\boldsymbol{o} - \boldsymbol{p}_t) \times \mathcal{U} \\ \mathcal{U} \\ norm(\boldsymbol{o} - \boldsymbol{p}_t) \end{bmatrix}^{-1}, \tag{5}$$

where $\mathbf{P}$ is the position and $\mathbf{R}$ is the rotation of the camera, $t \in [0,1]$ is the camera time step, and r is the radius of the field. $\mathcal{U}$ is the up vector $[0,1,0]$ and $\boldsymbol{o}$ is the camera looking point. This well-defined spiral trajectory effectively covers most of the missing regions and generates comprehensive multi-view images. However, the custom camera movements, provided by the user or extracted from the driving video, may differ significantly from the spiral trajectory. Therefore, we expand the 3D Gaussian field according to the custom trajectory using the warp-and-inpaint strategy again. Before this process, the camera rotations are standardized based on the first camera frame to ensure consistency with the spiral trajectory. Finally, based on the custom camera trajectory, a background video is rendered through volume splatting at each camera step.

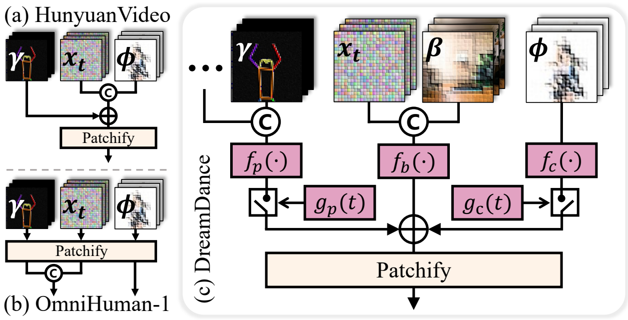

Figure 4. **The gating strategy** in our MM-DiT model and its comparison with the mainstream condition incorporation methods.

The reason we do not directly use the custom camera trajectory to reconstruct the 3D scene is that it may be excessively dynamic, potentially resulting in a discontinuous and unstable 3D scene. Additionally, since the reconstructed 3D scene often suffers from fidelity issues, the rendered background video may contain blurring, distortions, and black voids. To address these challenges, we introduce a pose-aware video inpainting strategy in the next step, which not only integrates the animated character but also refines the coarse background video for improved visual quality.

3.3. Dynamic Character Animation

Based on the coarse background video generated in the first stage, we implement pose-aware video inpainting to achieve dynamic character animation. Previous UNet-based reference-denoising architectures lack the ability to model video coherence effectively [10]. Therefore, we introduce an MM-DiT-based video foundation model along with a gating strategy to enable pose-guided character integration and ensure temporal consistency.

Pose-aware Video Inpainting. As illustrated in Figure 2, we divide the input references and guidance into three sets. The first is the background set, which includes the coarse background video, and an inpainting mask video generated based on the region of the driving pose. The second is the pose set, consisting of the driving face, hand, and body pose videos. The final set is the reference character. All elements in these three sets are encoded by the pre-trained 3D VAE, and then stacked along the channel dimension to obtain the latent background $\boldsymbol{\beta}$, pose $\boldsymbol{\gamma}$, and character $\boldsymbol{\phi}$. Next, the latent background is concatenated with the base latent noise $\boldsymbol{x}_t$ along the channel dimension. Then, three convolutional layers are applied to project each of the three latent features to the same channels, respectively.

Existing DiT-based models for character image animation typically use simple feature concatenation or addition to inject the latent reference and guidance [16, 20]. However, unlike these methods, the goal of our model is to inpaint the base background video using the posed character. Obviously, character appearance and pose information should be prioritized during the initial denoising steps, while at later steps, the model should focus more on overall video refinement. To achieve this, we propose two denoising step-based gates that adaptively inject the latent character and pose into the base latent noise according to the denoising step t. Each gate consists of a Linear layer followed by a $tanh$ activation function. This gating strategy can be formulated as:

$$\begin{aligned} \boldsymbol{x}_t^{'} =& f_b\left([\boldsymbol{\beta}, \boldsymbol{x}_t]\right) \\ &+ tanh\left(g_p(t)\right) \cdot f_p(\boldsymbol{\gamma}) + tanh\left(g_c(t)\right) \cdot f_c(\boldsymbol{\phi}), \end{aligned} \tag{6}$$

where $f(\cdot)$ denotes the convolutional Project-In layers, and $g(\cdot)$ represents the Linear layers. A detailed comparison of the structural differences between existing methods and our gating strategy is shown in Figure 4. Finally, we use MM-DiT from CogVideoX [47] to perform the diffusion denoising steps. Additionally, the reference character is embedded using the CLIP image encoder [28] and serves as the text hidden states in the cross-attention operations of MM-DiT. This process is commonly used in existing work and is therefore omitted from Figure 2. The resulting latent output is decoded through the 3D VAE Decoder to generate the character art animation.

3D Gaussian-free Training Approach. We perform supervised fine-tuning to train the gated video inpainting model in DreamDance, starting from the image-to-video model CogVideoX-5B. Given that constructing 3D Gaussian fields is time-consuming, we inpaint the character region and apply random down-sampling to the original video background to simulate the coarse video rendered from the 3D Gaussian scene. The down-sampling approach includes adding black blocks, introducing noise, blurring, and applying random perspective transformations. Additionally, following [10], we randomly generate stylized pair-wise images by concatenating the initial frames along the spatial dimension and use the depth and edge-controlled SDXL-Neta model [17] to transfer the art style. Then, the stylized frames are repeated along the temporal dimension to construct a fake video for training. 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 video clips.

During the training, we found that the supervision of the background region in the video was too strong, causing the bodies of the inpainted characters to be incomplete. To address this issue, we use inpainting masks to reweight the loss and fill the character bounding boxes of the background videos with black in the early training stages, thereby enhancing dynamic character learning and eliminating the model from overfitting to the backgrounds.

4. Experiments

Datasets. To train DreamDance, we collected an MMD video dataset comprising 4,800 animations created by artists, which is comparable to that of MikuDance. We split these videos into approximately 150,000 clips, which together include over 14.8 million frames. For the quantitative evaluation, we used 100 MMD videos that were not included in the training set, with their first frames serving as reference images. For the qualitative evaluation, all character art was randomly generated using SDXL-Neta [17], and the driving videos were not seen during training.

Implementation details. We implement DreamDance using the code base of VistaDream [39] and Finetrainers [25]. Experiments are conducted on 32 NVIDIA A800 GPUs. During training, the videos are center-cropped and resized to a resolution of 768 × 768, and the length is sampled to 48 frames. Training is conducted for 60,000 steps with a batch size of 32. The learning rates are set to 1e-4, and the dropout ratio for the character and pose guidance is set to 0.1. During inference, we use a DDIM sampler for 50 denoising steps. We adopt the temporal aggregation method described in [37] to generate long videos. The code will be released in the final version.

Evaluation metrics. Following MikuDance [10], we evaluate the results from two aspects: image and video. To assess image quality, we report frame-wise FID [9], SSIM [43], LISPIS [54], PSNR [12], and and L1. For video quality, we concatenate every consecutive 16 frames to form a sample, from which we report FID-VID [1] and FVD [38].

4.1. Qualitative Results

Comparison with image animation baselines, including MikuDance [10], as well as recent human animation methods, such as Animate Anyone (AniAny) [13], UniAnimate [41], MimicMotion [55], and DisCo [40].

The results in Figure 5 demonstrate that AniAny, UniAnimate, MimicMotion, and DisCo struggle with a strong shape prior on the human body, which leads to substantial character distortion and fidelity issues in their outputs. Moreover, the backgrounds in their generated videos remain nearly static or excessively blurry, resulting in monotonous and visually flat effects. While DisCo employs an independent ControlNet to process the backgrounds, it suffers from scene collapse when animating character art. MikuDance shows significant improvements in animating character art, but the backgrounds still exhibit inconsistencies when confronted with large-scale camera movements. Notably, thanks to the explicit reconstruction of the 3D scene and the MM-DiT-based inpainting paradigm, our DreamDance achieves precise camera control, coherent scene dynamics, and consistent animation generation, producing high-quality and vivid 2D animation results.

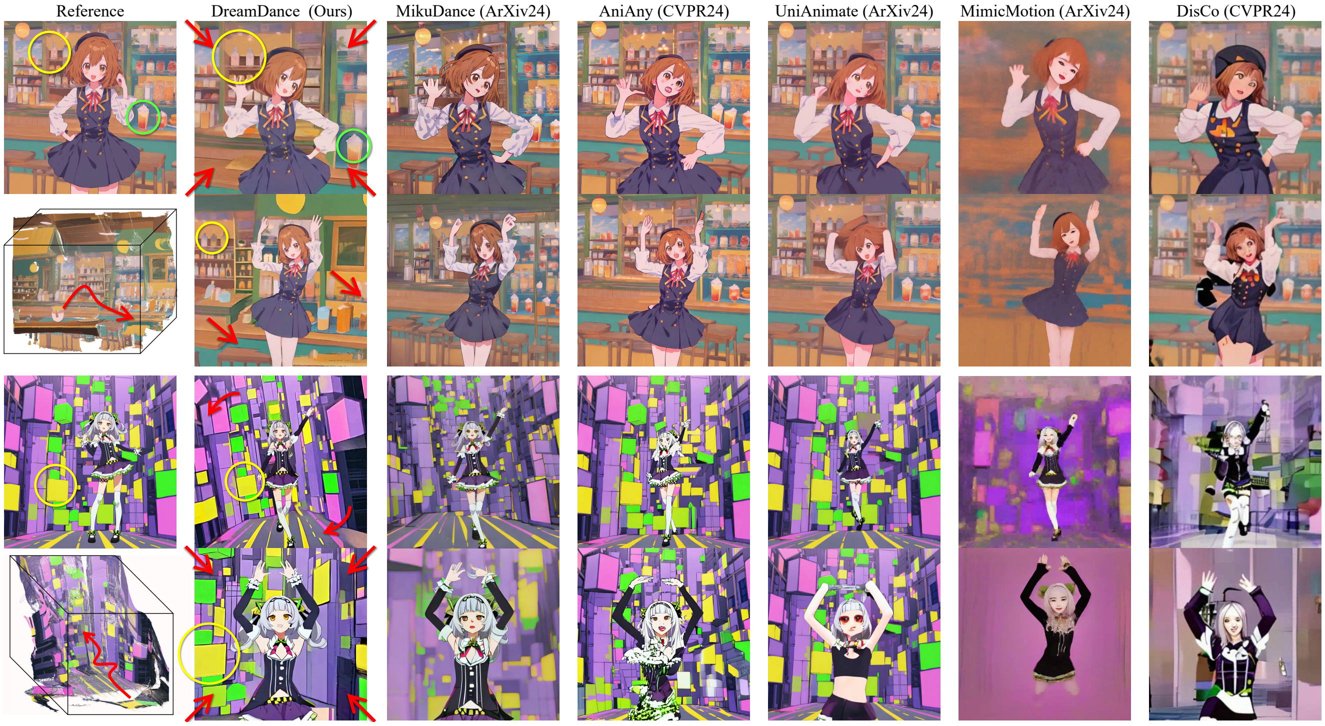

Figure 5. **Comparison with image animation baselines.** The red arrows represent the approximate direction of camera movement, while the circles highlight significant correspondences. The black boxes contain the reconstructed 3D Gaussian scenes of our DreamDance.

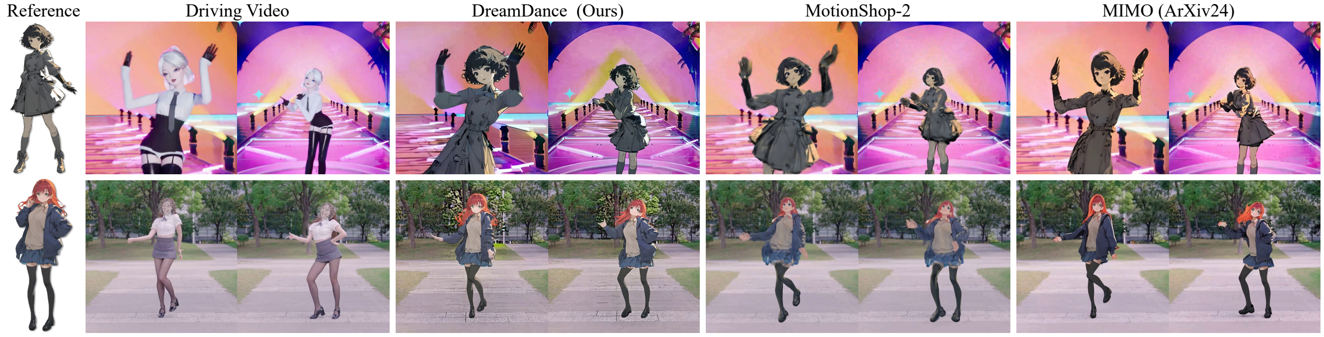

Figure 6. **Comparison with character replacement baselines.** MIMO and MotionShop-2 only support full-body reference images.

Comparison with character replacement baselines. One valuable application of DreamDance is its ability to directly replace humans in driving videos with reference characters. We compare it with MotionShop-2 [49] and MIMO [22], which are 3D and 2D-based methods, respectively. As shown in Figure 6, MotionShop-2 exhibits noticeable character distortion due to the unresolved challenges of 3D character reconstruction, while MIMO fails to effectively preserve the attributes of the reference characters. Additionally, both MotionShop-2 and MIMO only support full-body character images. In contrast, our DreamDance seamlessly integrates the 2D character into the driving video without disrupting the harmony of the scene. This application opens up broad prospects for DreamDance in creating flexible video content.

High-dynamic and precise camera control. A key highlight of DreamDance is its ability to animate characters with high-dynamic camera movements while maintaining scene coherence through precise camera control. Distinct from MikuDance, which relies on 2D flow for scene motion guidance, and AniAny, which always outputs static backgrounds, DreamDance explicitly reconstructs stable 3D scenes. This approach avoids the context ambiguity that typically arises in the scene inpainting process, ensuring more consistent and immersive character animation with high-dynamic motion, as demonstrated in Figure 7.

Ablation study. In Figure 8 and Figure 10, we conduct ablation experiments to verify the key designs of our DreamDance, which include the camera-aware scene inpainting method, the MM-DiT-based model architecture, and the gating strategy for video inpainting.

To demonstrate the necessity of the camera-aware scene inpainting in DreamDance, we implemented a baseline (w/o 3DGS) that animates the character art directly using the

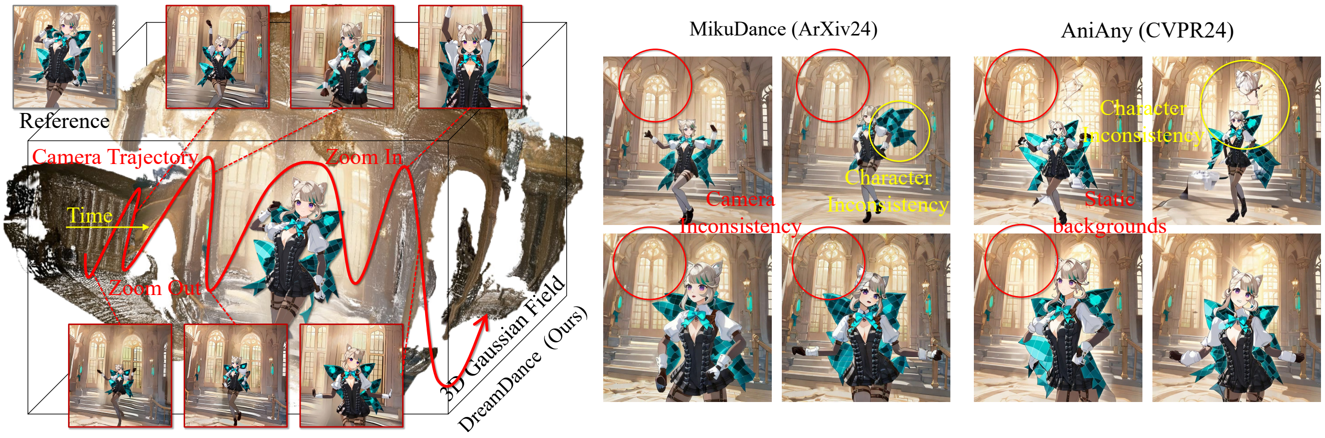

Figure 7. **High-dynamic and precise camera control** of our DreamDance. MikuDance exhibits inconsistencies due to its 3D-agnostic motion guidance, while AniAny produces static backgrounds. In contrast, DreamDance generates coherent and vivid animations.

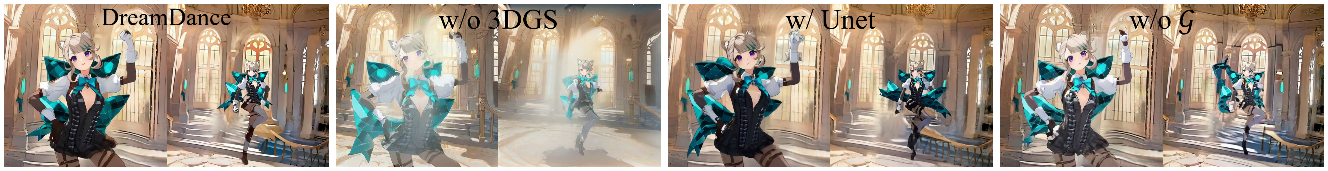

Figure 8. **Ablation experiments.** "w/o 3DGS", "w/ UNet", and "w/o $\mathcal{G}$" are defined in Section 4.1.

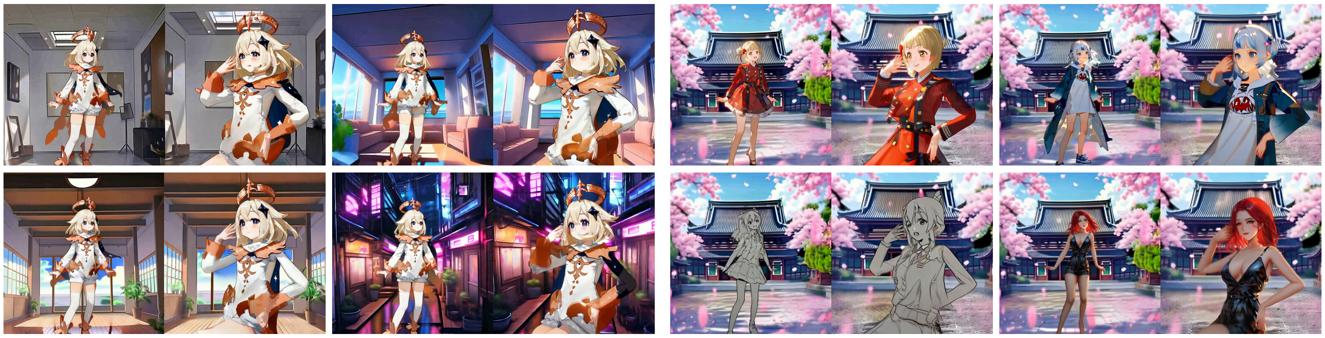

Figure 9. **Generalizability on various scenes (left) and characters (right).** Please see our demo video for a clearer understanding.

MM-DiT model, bypassing the 3D Gaussian reconstruction process. The results in Figure 8 indicate that this model fails to generate consistent backgrounds due to the implicit inpainting of unknown regions, leading to flickering and blurring in the generated video. Moreover, as shown in Figure 10, reconstructing the 3D scene without the spiral camera trajectory may result in discontinuous Gaussian fields.

To evaluate the MM-DiT-based model architecture, in Figure 8, we implemented a UNet-based video inpainting model. Since this model has limited capabilities in spatio-temporal modeling and a smaller pre-training scale compared to MM-DiT, its results are inferior to those of DreamDance. To assess the effectiveness of the gating strategy in our pose-aware video inpainting step, we conducted an experiment in which the latents were directly added to the noise without re-weighting them using the gates (w/o $\mathcal{G}$). This ablation model was trained on the same dataset with the same training settings as DreamDance. However, its results exhibited an underfitting phenomenon, with the videos appearing blurry during high-dynamic motion. Moreover, as shown in Figure 10, our gated video inpainting model also enhances the quality of the backgrounds provided by the rendered 3D scene for character art animation.

We visualize the values of the adaptive gates across the denoising time steps in the right part of Figure 10. As the denoising steps progress from 0 to 50, both the character gate and the pose gate decrease. This supports our conjecture that the model requires more information about the character's appearance and pose during the early denoising steps, whereas in the later stages, it prioritizes optimizing the quality of the existing latent videos.

Generalizability on various scenes and characters. Beyond reconstructing the scene from the reference character art, DreamDance supports scene reconstruction from custom images and the animation of the reference character across various scenes. As shown in the left part of Figure 9, DreamDance effectively integrates animated characters into diverse 3D Gaussian scenes. On the other hand, DreamDance is also capable of handling multiple characters in a wide range of art styles, including but not limited to cellu-

Table 1. **Quantitative comparisons with baselines and ablative experiments.** 'UNet' and 'w/o $\mathcal{G}$' are defined in Section 4.1. 'Foreground-only' refers to replacing the reference backgrounds with white images and evaluating only the character animations. The best results are highlighted in bold, and the second-best are underlined. DreamDance achieves superior results across most metrics.

									Foreground-only	
	Methods	**SSIM**$_{\uparrow}$	**PSNR**$_{\uparrow}$	**LISPIS**$_{\downarrow}$	**L1**$^{E-05}_{\downarrow}$	**FID**$_{\downarrow}$	**FID-VID**$_{\downarrow}$	**FVD**$_{\downarrow}$	**FID**$_{\downarrow}$	**FVD**$_{\downarrow}$
SVD	UniAnimate [41]	0.417	12.074	0.571	7.930	47.328	40.924	882.245	29.818	381.485
	MimicMotion [55]	0.325	12.264	0.600	9.313	60.210	44.517	903.674	30.125	407.856
SD	DisCo [40]	0.313	10.732	0.615	9.248	59.221	46.852	923.921	31.221	564.892
	AniAny [13]	0.488	12.530	0.548	7.307	43.945	38.179	846.414	27.927	326.842
	MikuDance [51]	0.576	14.592	0.493	5.726	**24.597**	22.868	502.380	**14.835**	194.124
	DreamDance UNet	0.612	16.721	0.383	4.622	32.923	19.387	477.235	15.227	221.126
DiT	DreamDance w/o $\mathcal{G}$	0.626	17.135	0.378	4.601	30.794	17.198	441.057	16.831	217.946
	DreamDance (Ours)	**0.699**	**17.964**	**0.355**	**4.109**	29.659	**16.411**	**430.136**	16.102	**188.852**

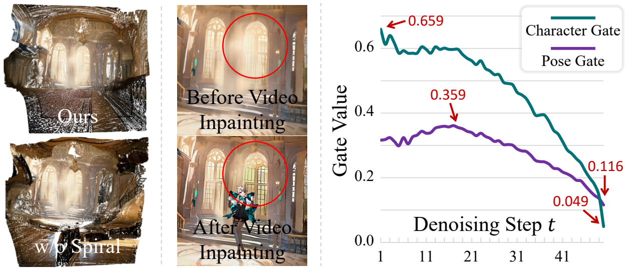

Figure 10. **Ablations on spiral trajectory (left), Scene enhancement (middle), and visualization of the gate values (right).**

loid, antiquity, and line sketch, as demonstrated in the right part of Figure 9. This high level of flexibility opens up vast possibilities for 2D animation applications.

4.2. Quantitative Results

Table 1 presents quantitative comparisons and the results demonstrate that DreamDance achieves state-of-the-art performance across most image and video metrics. Additionally, the ablation results confirm the effectiveness of the key design elements in the dynamic character animation stage of DreamDance. To isolate character quality from background effects, we conducted evaluations on foreground-only results, as shown in the right part of Table 1. In this setup, we replaced the backgrounds of reference character art with white images for baseline evaluation. For DreamDance, we provided white background videos for character animation inpainting. Under these conditions, DreamDance consistently achieved the best video temporal quality.

User study. We invited 50 volunteers to evaluate DreamDance against baseline methods on two tasks: image animation and character replacement. Each participant reviewed 15 videos, each containing one pose guidance and three or four anonymous animation results. They ranked the results based on character quality, background quality, and temporal consistency. After filtering out abnormal responses, the average rankings are summarized in Figure 11. For image animation, DreamDance significantly outperforms baseline methods in background and temporal quality, with over 77.27% of users preferring its animations. For character replacement, DreamDance achieves the highest character and temporal quality, favored by more than 59.09% of users.

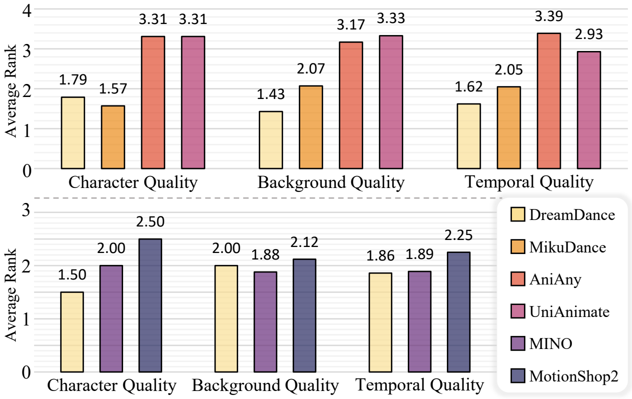

Figure 11. **User Study** for image animation (top) and character replacement (bottom). The smaller value means the better quality.

5. Conclusions

In this work, we propose DreamDance, a new inpainting-based pipeline for animating character art. DreamDance integrates two key techniques: Camera-aware Scene Inpainting and Pose-aware Video Inpainting. Camera-aware Scene Inpainting reconstructs stable Gaussian scenes, allowing for the rendering of coarse yet context-coherent background videos. Pose-aware Video Inpainting then adaptively incorporates pose-guided characters into the background, refining the video quality and ensuring consistent animation for stylized character art. Extensive experiments demonstrate that DreamDance outperforms baseline methods, achieving state-of-the-art results in character art animation.

Limitations. We acknowledge that some generated animations exhibit artifacts, particularly in character details such as the hands and clothing. This issue arises from the limitations of the datasets and the base model. Additionally, extracting precise camera parameters from real-world videos remains a challenge, often requiring manual adjustments.

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. 1, 2

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1, 2

[4] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 4

[5] Di Chang, Hongyi Xu, You Xie, Yipeng Gao, Zhengfei Kuang, Shengqu Cai, Chenxu Zhang, Guoxian Song, Chao Wang, Yichun Shi, et al. X-dyna: Expressive dynamic human image animation. *arXiv preprint arXiv:2501.10021*, 2025. 1

[6] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023. 2, 3

[7] 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

[8] Chao He, Jianqiang Ren, and Liefeng Bo. Textoon: Generating vivid 2d cartoon characters from text descriptions. *arXiv preprint arXiv:2501.10020*, 2025. 2

[9] 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

[10] Yu Higuchi. Mikumikudance. https://sites.google.com/view/evpvp. 1, 4, 5

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 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. 5

[13] 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, 5, 8

[14] Li Hu, Guangyuan Wang, Zhen Shen, Xin Gao, Dechao Meng, Lian Zhuo, Peng Zhang, Bang Zhang, and Liefeng Bo. Animate anyone 2: High-fidelity character image animation with environment affordance, 2025. 2

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 3

[16] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 4

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

[18] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17562–17571, 2022. 2

[19] Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N Plataniotis, Sergey Tulyakov, and Jian Ren. Wonderland: Navigating 3d scenes from a single image. *arXiv preprint arXiv:2412.12091*, 2024. 3

[20] Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, and Chao Liang. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. *arXiv preprint arXiv:2502.01061*, 2025. 2, 4

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 4

[22] Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Mimo: Controllable character video synthesis with spatial decomposed modeling. *arXiv preprint arXiv:2409.16160*, 2024. 2, 6

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

[24] 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. In *European Conference on Computer Vision*, pages 111–128. Springer, 2024. 1

[25] Sayak Paul. Finetrainers. https://github.com/a-r-r-o-w/finetrainers. 5

[26] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[27] Lingteng Qiu, Shenhao Zhu, Qi Zuo, Xiaodong Gu, Yuan Dong, Junfei Zhang, Chao Xu, Zhe Li, Weihao Yuan, Liefeng Bo, et al. Anigs: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. *arXiv preprint arXiv:2412.02684*, 2024. 2

[28] 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. 5

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3

[30] 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. 1

[31] 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

[32] Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199*, 2024. 4

[33] Harrison Jesse Smith, Qingyuan Zheng, Yifei Li, Somya Jain, and Jessica K Hodgins. A method for animating children's drawings of the human figure. *ACM Transactions on Graphics*, 42(3):1–15, 2023. 2

[34] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[35] 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

[36] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[37] 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

[38] 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

[39] Haiping Wang, Yuan Liu, Ziwei Liu, Wenping Wang, Zhen Dong, and Bisheng Yang. Vistadream: Sampling multiview consistent images for single-view scene reconstruction. *arXiv preprint arXiv:2410.16892*, 2024. 2, 3, 4, 5

[40] 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, 5, 8

[41] 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. 5, 8

[42] Yaohui Wang, Xin Ma, Xinyuan Chen, Cunjian Chen, Antitza Dantcheva, Bo Dai, and Yu Qiao. Leo: Generative latent image animator for human video synthesis. *International Journal of Computer Vision*, pages 1–13, 2024. 1

[43] 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

[44] 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. 1, 2

[45] Ronghuan Wu, Wanchao Su, Kede Ma, and Jing Liao. Aniclipart: Clipart animation with text-to-video priors. *International Journal of Computer Vision*, pages 1–17, 2024. 2

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

[47] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 3, 5

[48] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. *arXiv preprint arXiv:2406.09394*, 2024. 2, 3

[49] Junfei Zhang, Xiaodan Ye, Chao Xu, Feng Wang, Qing Ran, Kejie Qiu, Guangyuan Wang, Jianfeng Luo, Junyao Wu, Gang Cheng, Zilong Dong, and Liefeng Bo. Motionshop-2. https://aigc3d.github.io/motionshop/. 6

[50] Jiaxu Zhang, Shaoli Huang, Zhigang Tu, Xin Chen, Xiaohang Zhan, YU Gang, and Ying Shan. Tapmo: Shape-aware motion generation of skeleton-free characters. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[51] Jiaxu Zhang, Xianfang Zeng, Xin Chen, Wei Zuo, Gang Yu, and Zhigang Tu. Mikudance: Animating character art with mixed motion dynamics. *arXiv preprint arXiv:2411.08656*, 2024. 1, 2, 8

[52] Kaidong Zhang, Jingjing Fu, and Dong Liu. Flow-guided transformer for video inpainting. In *European Conference on Computer Vision*, pages 74–90. Springer, 2022. 2

[53] Lvming Zhang. Fooocus. https://github.com/lllyasviel/Fooocus. 2, 4

[54] 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

[55] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance, 2024. 1, 2, 5, 8

[56] Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang, Yinan Zhao, Peter Vajda, Dimitris Metaxas, and Licheng Yu. Avid: Any-length video inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7162–7172, 2024. 3

[57] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *arXiv preprint arXiv:2401.03407*, 2024. 3

[58] Jie Zhou, Chufeng Xiao, Miu-Ling Lam, and Hongbo Fu. Drawingspinup: 3d animation from single character drawings. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–10, 2024. 2

[59] 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

[60] Bojia Zi, Shihao Zhao, Xianbiao Qi, Jianan Wang, Yukai Shi, Qianyu Chen, Bin Liang, Kam-Fai Wong, and Lei Zhang. Cococo: Improving text-guided video inpainting for better consistency, controllability and compatibility. *arXiv preprint arXiv:2403.12035*, 2024. 3

[61] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10324–10335, 2024. 3

[62] Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Ewa volume splatting. In *Proceedings Visualization, 2001. VIS'01.*, pages 29–538. IEEE, 2001. 3