I2V3D: Controllable image-to-video generation with 3D guidance

Zhiyuan Zhang[1], Dongdong Chen[2], Jing Liao[1‡]
[1]City University of Hong Kong [2]Microsoft GenAI
https://bestzzhang.github.io/I2V3D

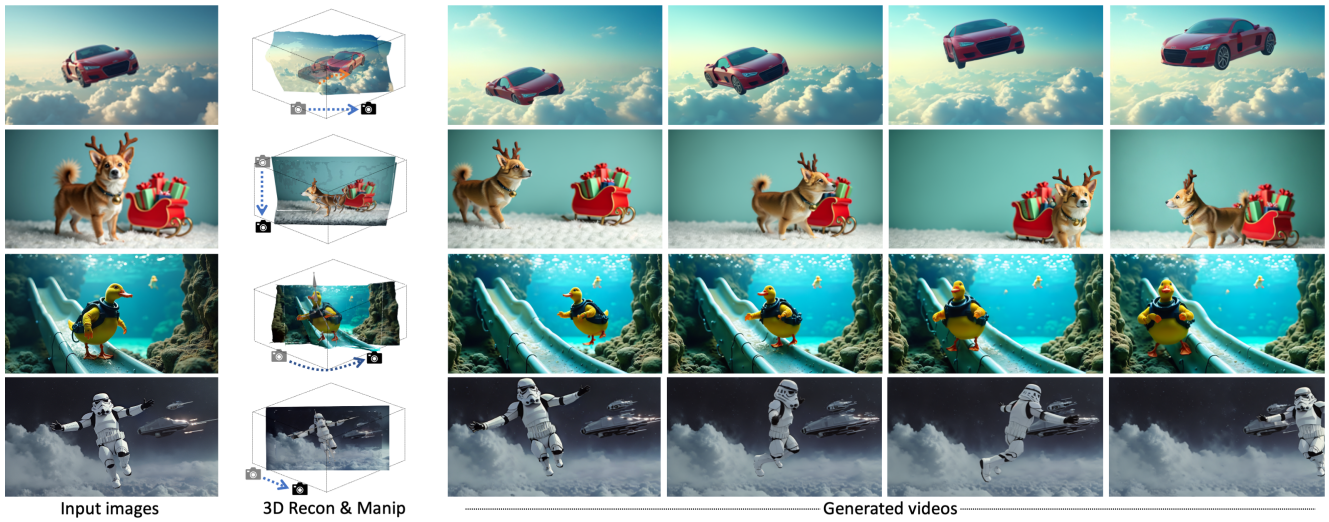

Figure 1. Starting with a single image, our method reconstructs the complete scene geometry and uses the CG pipeline to enable precise control of character animation (e.g., keyframe animation or skeleton control) and camera movement (e.g., the camera rotation in the 2nd and 3rd rows, and the camera panning and zooming in the 1st and 4th rows). We then apply geometric guidance, based on the coarse rendering results, to generate high-quality, controllable videos.

Abstract

We present I2V3D, a novel framework for animating static images into dynamic videos with precise 3D control, leveraging the strengths of both 3D geometry guidance and advanced generative models. Our approach combines the precision of a computer graphics pipeline, enabling accurate control over elements such as camera movement, object rotation, and character animation, with the visual fidelity of generative AI to produce high-quality videos from coarsely rendered inputs. To support animations with any initial start point and extended sequences, we adopt a two-stage generation process guided by 3D geometry: 1) 3D-Guided Keyframe Generation, where a customized image diffusion model refines rendered keyframes to ensure consistency and quality, and 2) 3D-Guided Video Interpolation, a training-free approach that generates smooth, high-quality video frames between keyframes using bidirectional guidance. Experimental results highlight the effectiveness of our framework in producing controllable, high-quality animations from single input images by harmonizing 3D geometry with generative models.

1. Introduction

Recent advancements in image-to-video generation models [3, 11, 31, 63] have significantly improved the ability to produce high-quality videos with dynamic motions. However, these methods still face substantial challenges in providing precise controllability during the generation process. For example, Stable Video Diffusion (SVD) [2] constrains the first frame of the generated video to be the given input image, offering no user control over subsequent frames. While subsequent text-based image-to-video methods [41, 59] allow rough motion control using textual prompts, they fall short of achieving precise control over object and camera movements as well. To address these limitations, bounding box-based [54, 60] and trajectory-

† Corresponding Author: jingliao@cityu.edu.hk

based [61] approaches enable users to control object movements via 2D signals such as bounding boxes or dragging gestures. However, these methods struggle with motion sequences requiring an understanding of 3D geometry and perspectives. Skeleton-based techniques [7, 20] provide an alternative by using skeletal structures as control signals which, while effective for human characters, are inherently limited to human-like subjects. These constraints underscore the inability of current image-to-video generation methods to achieve fine-grained 3D control, such as rotating an object by 180 degrees, as shown by the examples of the dog and stormtrooper in Fig. 1.

In contrast to existing generative models, traditional Computer Graphics (CG) pipelines offer a mature and versatile set of tools—such as rigging, keyframe animation, and inverse kinematics—that enable highly flexible and precise control over object motion and camera movement in 3D space. Tools like Blender, Maya, and Houdini provide robust frameworks for animation, rigging, and simulation, making them a preferred choice for achieving detailed control. However, generating photorealistic videos with these pipelines depends on the availability of high-quality 3D models for animation and rendering. Unfortunately, existing automatic 3D model reconstruction methods from a single image [55, 64] remain too coarse to meet the required standards for photorealism.

In this paper, we propose a novel controllable image-to-video generation framework that combines the precise 3D control of traditional CG techniques with the photorealistic generation capabilities of advanced generative models. Given an input image, our approach consists of two main components: 1) 3D reconstruction to produce coarse 3D geometry guidance; and 2) 3D guided video generation. For the first component, our method reconstructs the entire scene in 3D from the given image. Foreground objects are extracted and converted into 3D meshes, enabling both simple and complex animations. For the background, we first inpaint occluded regions, then expand it using multi-view generation, and finally reconstruct it as a 3D mesh as well. By reconstructing the entire scene in 3D, our method facilitates camera movement and meaningful interactions between foreground objects and their surroundings. The resulting 3D animations are then rendered into coarse outputs, including RGB frames and depth maps, which serve as 3D guidance for generating high-quality videos.

Building upon the 3D guidance, the video generation pipeline employs a two-stage process: 1) 3D-guided keyframe generation and 2) 3D-guided video interpolation. In the first stage, we customize an image diffusion model to learn the appearance of the input image and utilize coarse renderings to guide the generation of keyframes. To enhance the generalization ability of the customized diffusion model, we augment the training process with novel-view images of the foreground object, ensuring its appearance is captured from various angles. Additionally, the temporal consistency between keyframes is further improved through extended attention mechanisms. In the second stage, we introduce a training-free, bidirectional 3D-guided interpolation method that generates high-quality, visually consistent videos between adjacent keyframes. This two-stage framework offers two key advantages. First, it allows users to define the starting point of the animation, without restricting them to the input image as the initial frame. Second, it alleviates the temporal error accumulation issue, enabling the generation of videos that extend beyond the typical window length of video diffusion models.

Extensive experiments demonstrate the superiority of our method in generating high-quality, temporally consistent videos with fine-grained controllability over camera and object motion. Our method enables users, especially professional users, to flexibly and precisely control the motions in generated videos by leveraging CG animation tools. However, compared with the traditional CG pipeline, our method has significantly lowered the professional threshold and reduced the time cost. It automates modelling and rendering by integrating AI-based 3D reconstruction and video generation with CG animation. Thus, users only need to specify the motion to customize the animation from a single image. Our key contributions are summarized as below:

- We propose a novel controllable image-to-video framework that combines the strengths of both worlds, i.e., the fine-grained controllability of 3D graphics pipelines and the photorealistic generation capabilities of generative models.
- We introduce a two-stage video generation pipeline with 3D guidance, comprising 3D-guided keyframe generation and video interpolation. This design allows for flexibility in specifying arbitrary starting points and generating extended video sequences.
- We incorporate multiple technical innovations to enhance quality, including multi-view augmentation for improved generalization, extended attention for better temporal consistency, and a bidirectional, training-free interpolation method for video synthesis.

2. Related Works

2.1. Single-View Reconstruction

3D reconstruction is a key challenge in computer vision and graphics. Recent advances in image-to-3D generation [18, 36, 51] have improved results, especially for objects with simple backgrounds. To further enhance reconstruction quality, recent studies [29, 50, 77] have explored leveraging sparse-view inputs. For example, InstantMesh [64], used in this paper, is a feed-forward sparse-view reconstruction model that utilizes novel views gen-

erated by Zero123++[46] to create 3D meshes. However, reconstructing entire scenes from a single image remains underdeveloped due to poor inpainting and warping quality[9, 14, 71, 73] or limited resolution [45, 49]. Recent advancements [58, 72] have demonstrated improvements by employing video diffusion models for multi-view generation. In this paper, we adopt ViewCrafter [72] for novel view generation and use a stereo model [55] to extract background meshes. Despite these advancements, reconstructed 3D meshes still remain coarse and lack fine details, motivating us to leverage generative models to produce photorealistic videos.

2.2. Image Editing with 3D Manipulation

Manipulating 3D objects in 2D images has evolved from early methods like Cuboid Proxies [78] and Stock 3D Models [24], which were limited to simple shapes or predefined assets. Recent approaches fall into two categories. One leverages image diffusion models to infer object poses after manipulation without explicit 3D reconstruction. For example, 3Dit [38] fine-tunes Zero-1-to-3 [33] for 3D-aware editing but struggles with diverse objects, while Neural Assets [62] enables multi-object pose control but is limited to rigid transformations. The other category uses 3D proxies for editing. Diffusion Handles [39] employs foreground depth but has constrained rotation, while 3DitScene [75] uses 3D Gaussian Splatting to refine 3D representations but still cannot handle non-rigid changes.

A recent work, ISculpting [69], adopts an approach similar to ours by lifting foreground objects into 3D, manipulating them, re-rendering them into 2D, and refining the results through a coarse-to-fine process. However, there are two key differences between their method and ours. First, their approach is designed specifically for image manipulation, and directly applying it to video generation results in significant flickering issues. Second, they only reconstruct the foreground and blend it with the background after manipulation. In contrast, our method reconstructs the entire scene geometry, including both foreground and background. This enables camera movement and supports meaningful interactions between foreground objects and their surroundings, providing greater flexibility and realism.

2.3. Controllable Video Generation

A major challenge in video generation is enabling precise control to align videos with user intentions. Recent work has introduced various control strategies. Some methods use dense spatial conditions, such as depth maps, sketches, or canny edges, to control video generation [8, 32, 56, 76]. Human poses are also commonly used as conditions to create dance videos [7, 20, 35]. However, finding a suitable video that matches the user's imagination and provides these control signals is often difficult.

Alternative approaches allow users to input sparse conditions, such as bounding boxes [21, 34, 54] or trajectories [13, 30, 47, 70]. For example, Boximator [54] combines hard and soft bounding box constraints to control the position, shape, or motion path of an object. DragAnything [61] enables users to define motion trajectories for objects or backgrounds through simple interactions, such as selecting regions and drawing paths. Some methods also explore the integration of multiple conditions. For example, Direct-A-Video [66] supports controlling camera movements by specifying relative panning or zooming distances and defining object motions using bounding boxes. MotionCtrl [58] allows for fine-grained motion control by utilizing camera poses for camera motion and trajectories for object motion. MotionBooth [60] introduces a framework for customizing text-to-video models using only images of a specific object and employs training-free methods to control both learned subject and camera motions at the inference time. However, since the trajectories and bounding boxes used in these methods are defined in 2D space, they struggle to accurately represent 3D motions. In addition, crafting sophisticated movements with these controls is often challenging and unintuitive. A recent work, Generative Rendering [5], proposes 4D-guided video generation with styles specified by text prompts. However, their method is limited to human subjects with a stationary camera and does not support generating a specific object from an image. In contrast, our approach enables users to intuitively manipulate both cameras and specified objects within a 3D engine. Furthermore, our method supports fine-grained movements, such as 3D rotations, for general objects, offering greater precision and flexibility in video generation.

3. Method

Given an input image, our goal is to generate an animation video starting from any desired point, guided by precise 3D control signals specified by the user. As illustrated in Fig. 2, our proposed pipeline begins by reconstructing the entire 3D scene, creating an animation using a traditional computer graphics (CG) pipeline, and producing a high-quality video guided by 3D signals such as depth and coarsely rendered outputs. Specifically, we first extract 3D meshes for both the foreground objects and the background scene. Next, we utilize Blender, a widely adopted 3D engine, to compose the animation and render the video. Finally, we adopt a two-stage video generation process: 3D-guided keyframe generation followed by 3D-guided interpolation between keyframes. This approach provides two key benefits. First, it offers users the flexibility to define any starting frame, removing the constraint of being tied to the input image. Second, it prevents error accumulation artifacts, enabling the generation of videos that extend well beyond the temporal window typically supported by video

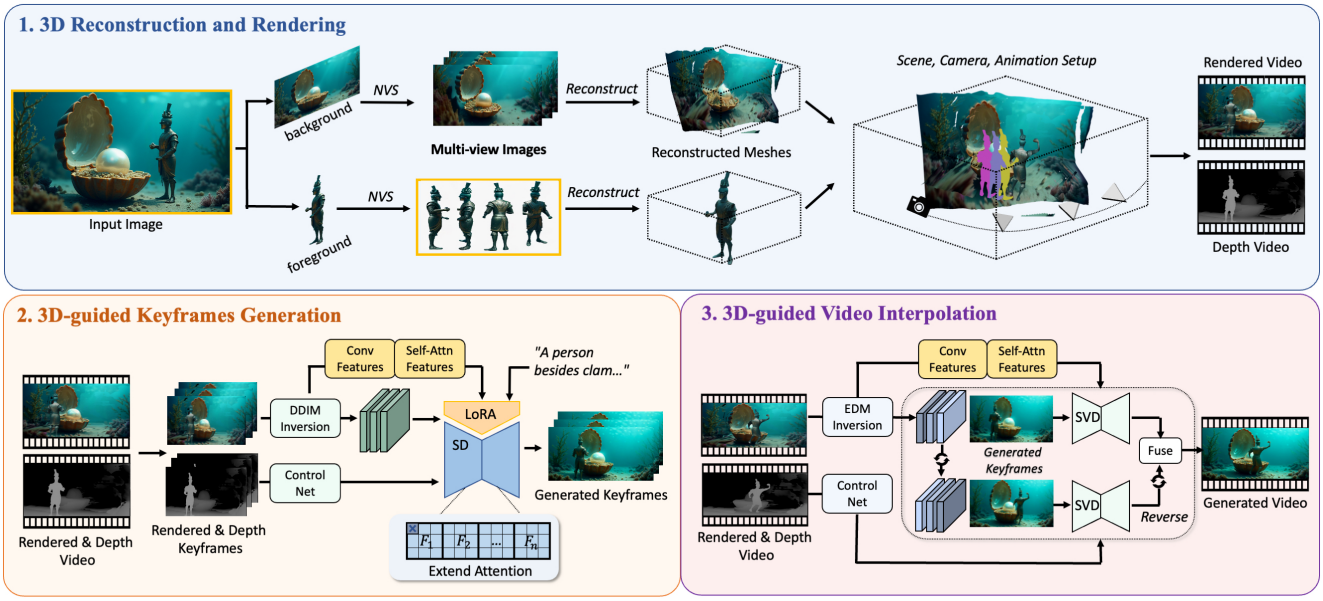

Figure 2. Our framework consists of three parts. First, we extract meshes from a single input image and use a 3D engine to create and preview a coarse animation. Next, we generate the keyframes using a 3D-guided process with an image diffusion model customized for the input image, incorporating multi-view augmentation and extended attention. Finally, we perform 3D-guided interpolation between generated keyframes to produce a high-quality, consistent video.

diffusion models, ensuring both high quality and temporal consistency.

3.1. 3D Reconstruction and Rendering

In this stage, we first lift the 2D image into a 3D scene and set up the camera and animation within a 3D engine. To create disentangled meshes, we isolate the foreground object using the object mask generated by SegAnything [25] and in-paint the object's region by using SDXL inpainting [40] to obtain a complete and clean background. For the foreground mesh, we use InstantMesh [64], a sparse-view feed-forward method that generates six novel-view images of the foreground, which are later reused in the customization stages. For the background, we begin by generating high-quality novel views and then apply Multi-View Stereo (MVS) reconstruction to obtain the 3D mesh. Specifically, we employ ViewCrafter [72], a state-of-the-art camera pose-conditioned video generative model, to produce novel views by setting predefined camera trajectories, such as rotating left/right or moving up/backward. A dense stereo model, Dust3r [55], is then used to reconstruct the background meshes from 1–4 of these novel-view images.

Once the meshes are generated, they are imported into a 3D engine to create animations. Users can leverage familiar 3D content creation workflows, including rigging objects (designing skeletal structures for movement), defining camera paths, and animating characters. For rigging, users can either manually design skeletal frameworks or utilize automated rigging and animation tools like Mixamo. Additionally, we provide a custom Blender script to automate the generation of camera trajectories. This script navigates through the views selected during the Multi-View Stereo (MVS) reconstruction, simplifying the process for novice users to create animations aligned with their reconstructed 3D scenes. Finally, the scene is rendered with baked textures and a depth video, providing 3D geometric signals to guide the subsequent video generation.

3.2. 3D-guided Keyframes Generation

After acquiring the 3D geometric signals, our objective is to generate consistent keyframes that preserve the 3D geometry of the rendered frames while maintaining the visual fidelity of the input image. To retain the input image's appearance, we begin by customizing a StableDiffusion (SD) model [43] using LoRA [19]. However, fine-tuning SD with LoRA on a single image tends to overfit to the given view, limiting its ability to generalize across different views. To mitigate this issue, we customize the LoRA using multi-view images to improve generalization. To ensure that the geometry of the generated images aligns with the rendered frames, we use depth maps and coarsely rendered frames as control conditions. Furthermore, to enhance temporal consistency across the generated keyframes, we incorporate Extended Attention, introducing additional cross-frame interactions for improved coherence.

3.2.1. LoRA Customization with Multi-View Images

LoRA is an efficient model adaptation technique that freezes the pre-trained weight matrices W_0 and integrates additional trainable matrices ΔW, which are decomposed into two low-rank matrices A and B, significantly reducing the number of trainable parameters. In our implementation, we apply LoRA for the attention layers and feedforward layers of the self-attention and cross-attention block in SD.

During each training step, we use the input image along with a randomly selected augmented image of the foreground object from the multiple novel-view images generated by InstantMesh [64]. To facilitate learning, we create simple prompts, such as "a {object} in {environment}" to describe the input image, and "a {object}" to describe the foreground object in all augmented images. The training loss is defined as follows:

$$\mathbb{E}_{z,z',\epsilon,t,c}\Big[\|(\epsilon-\epsilon_\theta(z_t,c_i,t))\|_2^2 + \lambda\|(\epsilon-\epsilon_\theta(z'_t,c_{fg},t))\cdot M\|_2^2\Big] \tag{1}$$

where z_t and z'_t are the noisy latents at time step t for the input image and the augmented image, respectively. Similarly, c_i and c_{fg} are the text prompts for the input image and the foreground object, respectively. M represents the mask for the foreground object in the augmented image, ϵ denotes the added noise, and ϵ_θ refers to the denoising network. λ is a hyperparameter that controls the weights of the augmentation loss.

3.2.2. Geometric Guidance for Generation Control

Following previous research [69], we use both depth control and rendered feature control to preserve the geometry of the rendered keyframes. First, we invert the rendered keyframes using DDIM inversion [48] to obtain the inverted latents z_{inv}. These z_{inv} latents are then used as the initial noise for keyframe generation. During each denoising step, we apply depth control through ControlNet [74] and inject the self-attention features sa (consists of q and k) and the convolutional features $conv$ from the residual block, which are extracted during the inversion process. The depth control provides coarse 3D geometric information, while the rendered features (sa and $conv$) further complement the layout and fine-grained geometric structure of nearby planes, as shown in previous work [26]. Additionally, we mask the injection of rendered features in the invisible regions to account for imperfections in the reconstructed background mesh.

3.2.3. Extended Attention for Consistency Enhancement

To enhance temporal consistency between keyframes, we further leverage the extended attention mechanism proposed in prior style-aligned generation [17] and T2V editing works [6, 15] . This mechanism modifies the self-attention layers such that each frame not only attends to its own features but also features from other keyframes. Specifically, the Extended Attention is calculated as follows:

$$\text{ExtAttn}\left(q_i;k_{1\ldots n},v_{1\ldots n}\right)=\mathcal{S}\left(\frac{q_i k_{1\ldots n}^T}{\sqrt{d}}\right)\cdot v_{1\ldots n} \tag{2}$$

Here, q_i represents the query features, while $k_{1\ldots n}$ and $v_{1\ldots n}$ denote the concatenated key and value features of all frames, respectively. The function $\mathcal{S}(\cdot)$ represents the *Softmax* operation. By capturing both spatial and temporal context through this approach, the method facilitates better interaction among keyframes, encourages them to share a consistent global appearance.

3.3. 3D-guided Video Interpolation

To generate a consistent and high-quality video from keyframes, we utilize an image-to-video interpolation approach that ensures temporal coherence. Additionally, geometric guidance is applied to maintain motion alignment with the rendered video.

Inspired by [12], we adopt a dual-trajectory denoising approach to leverage bidirectional contextual information. For each denoising step, we pass the latents through two parallel denoising trajectories. One is a forward denoising trajectory, conditioned on the first frame, and the other is a time-reversed denoising trajectory, where the latents are reversed along the time dimension and conditioned on the last frame. After denoising, the outputs from the forward and time-reversed trajectories are fused using a weighted average, proportional to the distance between the first frame and the last frame.

To improve alignment with the geometry of the rendered videos, we apply similar geometric guidance during interpolation. Specifically, we use a depth ControlNet released by the community [1], complemented with rendered feature control. First, we invert the rendered video using EDM [23] to obtain the initial latents. During denoising, we use rendered depth control and override convolutional and self-attention features with those obtained during inversion. These steps ensure that the interpolated video aligns better with the geometry of the rendered video while maintaining temporal consistency.

4. Experiments

4.1. Implementation Details

To customize the input image, we fine-tune the SDXL model [40] with LoRA rank 32 for 500 steps using a learning rate of 1e-4 and the augmentation weight λ set to 1.0. For rendered feature control in both 3D-guided keyframe generation and video interpolation, we use self-attention features and convolutional features. For keyframe generation, which uses our customized model, we inject these

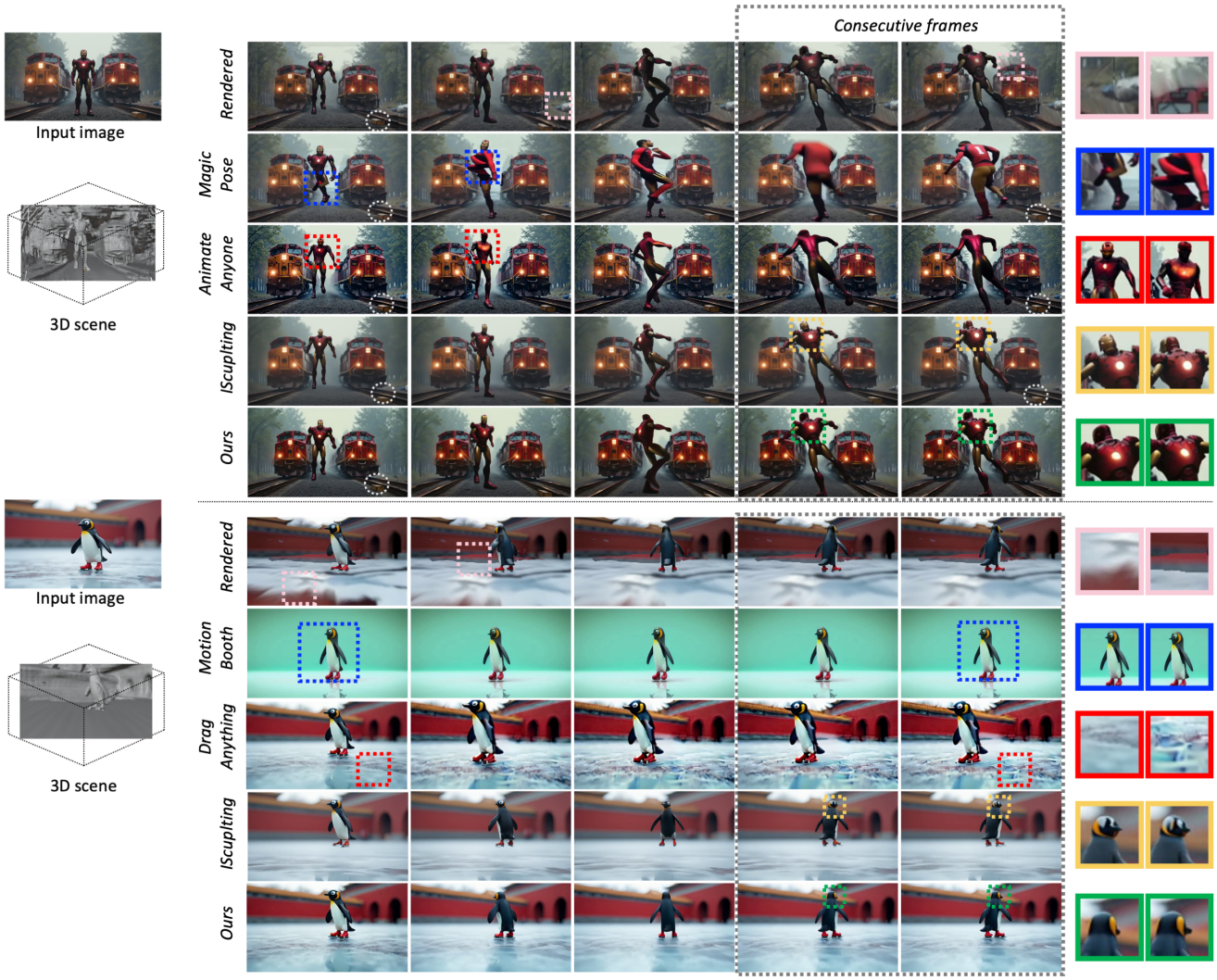

Figure 3. Qualitative comparison with baselines: (1st) human-like characters, (2nd panel) non-human objects. For human-like characters, MagicPose [7] struggles with pose control (blue), and AnimateAnyone [20] fails to preserve appearance (red). For non-human objects, MotionBooth [60] shows overfitting (blue), and DragAnything [61] shows error accumulation (red). ISculpting [69] exhibits frame inconsistency (yellow) for both categories. Our method outperforms them by following the geometry guidance of coarse renderings but resolves their artifacts (pink).

features from all the upsampling blocks and set $\tau_{\text{conv}} = 0.4$ and $\tau_{\text{sa}} = 0.0$. For video interpolation, which uses the pre-trained SVD [2], we use sa features from layers 4–11 and $conv$ features from layer 4, following previous training-free methods [26, 52]. Here, we set τ_{conv} and τ_{sa} to 0.0. The image and video are both of the resolution 1024×576. It takes approximately 20 seconds to generate a keyframe and about 310 seconds to generate 16 video frames on a 3090Ti GPU.

4.2. Experiment Setup

4.2.1. Baselines and Dataset

We compare our method with ISculpting [69] for both human-like characters and nonhuman-like objects. Additionally, we include comparisons with two other methods in each category: 1) For human-like characters, we evaluate our method against two skeleton-driven approaches: AnimateAnyone [20] and MagicPose [7]. To use these methods, we extract skeletons from the rendered video using DwPose [67]. These skeletons are then input to AnimateAnyone and MagicPose, which generate animations for the input images based on the skeleton data. 2) For nonhuman-like objects, we compare our method with two approaches that support both camera and object movement control. In DragAnything [61], pixel trajectories are detected using Co-Track [22]. Human interaction is simulated by randomly selecting one foreground trajectory to control the object's movement and four background trajectories to control the camera movement. In Motionbooth [60], bounding boxes are extracted from the foreground object mask, while the camera's movements along x and y directions

Methods	Consistency ↑	CLIP-I ↑	SSIM ↑	D-RMSE ↓
		Nonhuman-like objects		
DragAnything	0.991	0.839	0.502	0.167
MotionBooth	0.974	0.755	0.519	0.256
ISculpting	0.986	0.906	0.570	0.161
Ours	**0.994**	**0.906**	**0.757**	**0.102**
		Human-like characters		
AnimateAnyone	0.985	0.874	0.407	0.178
MagicPose	0.952	0.85	0.425	0.210
ISculpting	0.982	**0.916**	0.507	0.155
Ours	**0.991**	0.901	**0.735**	**0.117**

Table 1. Quantitative comparison with DragAnything [61], MotionBooth [60], ISculpting [69] on Nonhuman-like objects and AnimateAnyone [20], MagicPose [7] and ISculpting [69] on Human-like characters.

are calculated by averaging the motion vectors of the background. To evaluate our methods against baselines, we created 30 animations, including 15 human-like characters and 15 nonhuman-like objects. Each animation consists of at least 61 frames (4 batches for a 16-frame diffusion model), which is the video length used in our evaluation.

4.2.2. Metrics

We evaluate the results from three aspects. To measure temporal consistency, we calculate the CLIP similarity [42] between consecutively generated frames, following previous methods [10]. For visual similarity, we use the CLIP-I score proposed in [44], which calculates the cosine similarity of the CLIP image embeddings between the generated frames and the input image. Finally, to assess the alignment between the generated and rendered videos, we employ the SSIM [57] and D-RMSE [69] metrics. D-RMSE measures the RMSE between the estimated depth of the rendered frames and the frames generated by each method, and we use DepthAnything [65] as the depth estimator.

4.3. Qualitative and Quantitative Comparison

Fig. 3 compares our method with several baselines. For human-like videos, MagicPose [7] and AnimateAnyone [20] struggle to accurately follow reference poses or preserve appearances, as their training datasets primarily feature faces and upper bodies, while ISculpting [69] produces flickering frames. Additionally, none of these methods can handle camera movement (e.g., the box highlighted with a white circle remains static despite the camera moving). For nonhuman-like objects, MotionBooth [60] overfits, resulting in nearly static videos, while DragAnything [61] is limited to 2D movements and accumulates color errors. ISculpting [69] again exhibits inconsistencies in appearance. In contrast, our method successfully animates both human-like characters and nonhuman-like objects according to the reference motion, preserves the original appearance, and achieves superior temporal consistency.

Beyond qualitative improvements, our method also excels in quantitative evaluations. As shown in Tab. 1, it outperforms baselines in temporal consistency (0.994 and 0.991) and structural similarity (0.102 and 0.117 for D-RMSE; 0.757 and 0.735 for SSIM). Additionally, it achieves comparable visual similarity (around 0.90) to the input image, similar to ISculpting [69], which directly blends the input image background during denoising. These results demonstrate our method's ability to generate temporally consistent animations that closely align with both the input image and the rendered video. Apart from the quantitative evaluations, we also conduct a user study to evaluate the effectiveness of our approach, which is included in our supplementary A.

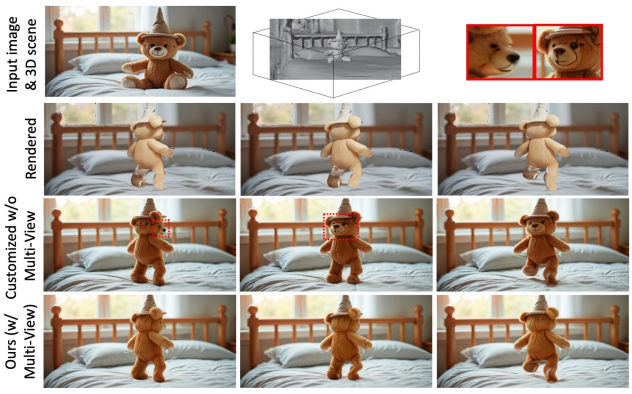

Figure 4. Ablation on LoRA customization with multi-view image augmentation. The red boxes highlight overfitting to the frontal view.

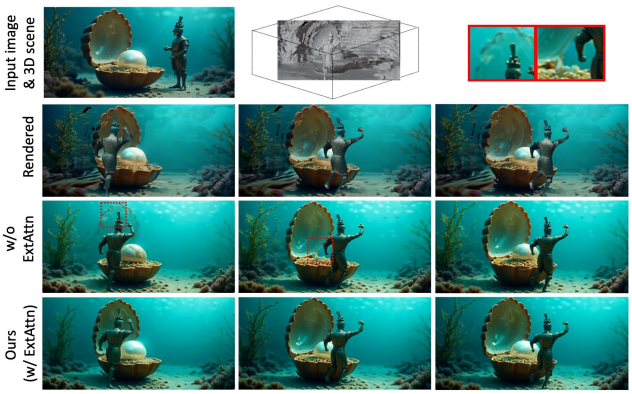

Figure 5. Ablation on extended attention for consistency enhancement. The red boxes highlight inconsistencies between individual generated frames.

4.4. Ablation study

Ablation on Multi-View Customization. Before keyframe generation, our method customizes the image diffusion model using multi-view images of the foreground object for training augmentation. As shown in the third row of Fig. 4, customizing LoRA on a single input image overfits by generating only the frontal view when the object rotates

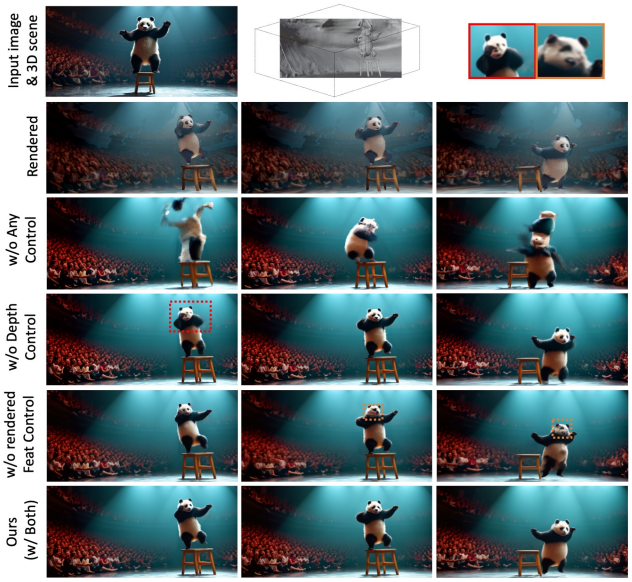

Figure 6. Ablation on 3D-guided video interpolation.

to its back, losing the ability to generalize to diverse unseen views. In contrast, our multi-view customization effectively captures the object's appearance from different angles and generalizes much better.

Ablation on Extended Attention. As discussed in Sec.3.2.3, we utilize extended attention to share global appearance information between keyframes. Generating frames individually often results in inconsistencies, such as the incomplete shell and missing pearl in the third row of Fig.5. In contrast, our method enhances consistency by generating the keyframes jointly with Extended Attention.

Ablation on 3D-Guided Interpolation. As discussed in Sec.3.2.2, ControlNet [74] directly captures 3D geometric information, while the rendered features enhance layout and fine-grained structure. As shown in third row of Fig. 6, interpolation without any guidance fails to generate meaningful content. When only rendered feature control is applied without depth control, the model lacks direct geometric control, as illustrated by the panda's hand highlighted in the red box. Conversely, without rendered feature control, the objects fail to accurately capture detailed features within the same plane, such as the panda's eyes highlighted by the orange box. In contrast, our method combines both depth control and rendered feature control, achieving better alignment with the rendered video. Apart from these ablations, more ablations are included in our supplementary B.

5. Dicussion

5.1. Balancing the 3D Control and Video Prior

Designed mainly for professional CG users, our framework aims to offer fine-grained animation control while streamlining the production process by automating modeling and rendering. Although our showcased animations primarily focus on object animations, users have the flexibility to incorporate additional dynamic effects like fluid simulations and cloth animations into the 3D scene, which are compatible with our pipeline. Alternatively, users can adjust the intensity of 3D guidance (by modifying feature injection and ControlNet scale) to leverage more motion priors of the video generation model to increase dynamism. As demonstrated in our supplementary videos, gradually reducing the 3D guidance from 1.0 to 0.5 and then to 0.2 increases the extent of dynamic motions in the background, such as water waves and flag flutters.

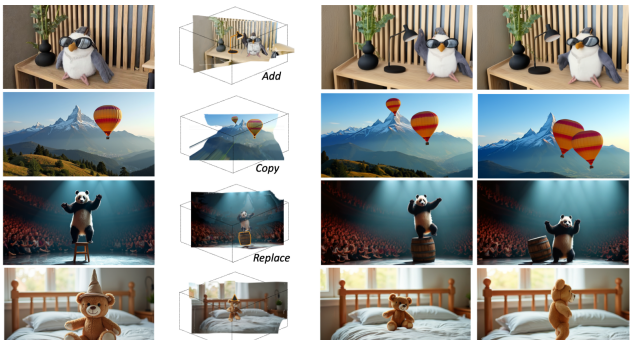

Figure 7. More applications of our method include adding, copying, replacing, and editing objects in 3D scenes to guide video generation.

5.2. Video Generation in New Composed Scene

By introducing 3D geometry control for image-to-video generation, it provides users with greater freedom to specify video content, in addition to object animation and camera movement. As demonstrated in Fig. 7, users can further add, copy, replace, and edit 3D objects to compose a new 3D scene that guides video generation.

6. Conclusion

In this paper, we present a novel solution to the longstanding challenge of precise and controllable video generation from a static image. By integrating 3D modeling tools and advanced generative models, our framework bridges the connection between traditional computer graphics and modern diffusion-based synthesis, enabling precise 3D control in the image-to-video generation process. Our proposed framework can achieve high-quality, temporally consistent animations while maintaining control over object and camera movements in 3D space. Experimental results validate the effectiveness of our approach, highlighting its general applicability to a wide range of scenarios, from moving and rotating objects, animating characters, to editing or composing new scenes in video generation.

Acknowledgement

The work described in this paper was fully supported by a GRF grant from the Research Grants Council (RGC) of the Hong Kong Special Administrative Region, China [Project No. CityU 11208123].

References

[1] temporal-controlnet-depth-svd-v1, 2024. 5

[2] 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, 6, 3

[3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1

[4] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, and Ning Yu. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *CVPR*, 2025. Licensed under Modified Apache 2.0 with special crediting requirement. 3

[5] Shengqu Cai, Duygu Ceylan, Matheus Gadelha, Chun-Hao Paul Huang, Tuanfeng Yang Wang, and Gordon Wetzstein. Generative rendering: Controllable 4d-guided video generation with 2d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7611–7620, 2024. 3

[6] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 5

[7] 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, 6, 7, 1

[8] Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023. 3

[9] 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. 3

[10] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 7

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1, 2

[12] Haiwen Feng, Zheng Ding, Zhihao Xia, Simon Niklaus, Victoria Abrevaya, Michael J Black, and Xuaner Zhang. Explorative inbetweening of time and space. In *European Conference on Computer Vision*, pages 378–395. Springer, 2025. 5, 1

[13] Wanquan Feng, Tianhao Qi, Jiawei Liu, Mingzhen Sun, Pengqi Tu, Tianxiang Ma, Fei Dai, Songtao Zhao, Siyu Zhou, and Qian He. I2vcontrol: Disentangled and unified video motion synthesis control. *arXiv preprint arXiv:2411.17765*, 2024. 3

[14] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[15] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 5, 1

[16] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, Wenping Wang, and Yuan Liu. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. *arXiv preprint arXiv:2501.03847*, 2025. 3

[17] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024. 5

[18] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4

[20] 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. 2, 3, 6, 7, 1

[21] Hsin-Ping Huang, Yu-Chuan Su, Deqing Sun, Lu Jiang, Xuhui Jia, Yukun Zhu, and Ming-Hsuan Yang. Fine-grained controllable video generation via object appearance and context. *arXiv preprint arXiv:2312.02919*, 2023. 3

[22] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proc. arXiv:2410.11831*, 2024. 6

[23] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022. 5

[24] Natasha Kholgade, Tomas Simon, Alexei Efros, and Yaser Sheikh. 3d object manipulation in a single photograph using stock 3d models. *ACM Transactions on graphics (TOG)*, 33 (4):1–12, 2014. 3

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 4

[26] Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhu Chen. Anyv2v: A plug-and-play framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024. 5, 6

[27] Black Forest Labs. Flux, 2024. 2

[28] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 3

[29] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 2

[30] Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Yuexian Zou, and Ying Shan. Image conductor: Precision control for interactive video synthesis. *arXiv preprint arXiv:2406.15339*, 2024. 3

[31] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024. 1

[32] Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. *arXiv preprint arXiv:2404.09967*, 2024. 3

[33] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 3

[34] Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. *arXiv preprint arXiv:2401.00896*, 2023. 3

[35] 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. 3

[36] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8446–8455, 2023. 2

[37] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 1

[38] Oscar Michel, Anand Bhattad, Eli VanderBilt, Ranjay Krishna, Aniruddha Kembhavi, and Tanmay Gupta. Object 3dit: Language-guided 3d-aware image editing. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[39] Karran Pandey, Paul Guerrero, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, and Niloy J Mitra. Diffusion handles enabling 3d edits for diffusion models by lifting activations to 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7695–7704, 2024. 3

[40] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 4, 5, 3

[41] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. *arXiv preprint arXiv:2410.13720*, 2024. 1

[42] 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. 7

[43] 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. 4

[44] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 7

[45] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*, 2023. 3

[46] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3

[47] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[48] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[49] Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, João F Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. *arXiv preprint arXiv:2406.04343*, 2024. 3

[50] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European*

Conference on Computer Vision, pages 1–18. Springer, 2025. 2

[51] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 2

[52] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 6

[53] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3

[54] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*, 2024. 1, 3

[55] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2, 3, 4

[56] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[57] 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. 7

[58] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. *arXiv preprint arXiv:2312.03641*, 2023. 3

[59] Zijian Zhang Rox Min Zuozhuo Dai Jin Zhou Jiangfeng Xiong Xin Li Bo Wu Jianwei Zhang Kathrina Wu Qin Lin Aladdin Wang Andong Wang Changlin Li Duojun Huang Fang Yang Hao Tan Hongmei Wang Jacob Song Jiawang Bai Jianbing Wu Jinbao Xue Joey Wang Junkun Yuan Kai Wang Mengyang Liu Pengyu Li Shuai Li Weiyan Wang Wenqing Yu Xinchi Deng Yang Li Yanxin Long Yi Chen Yutao Cui Yuanbo Peng Zhentao Yu Zhiyu He Zhiyong Xu Zixiang Zhou Zunnan Xu Yangyu Tao Qinglin Lu Songtao Liu Dax Zhou Hongfa Wang Yong Yang Di Wang Yuhong Liu Weijie Kong, Qi Tian and along with Caesar Zhong Jie Jiang. Hunyuanvideo: A systematic framework for large video generative models, 2024. 1

[60] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *arXiv preprint arXiv:2406.17758*, 2024. 1, 3, 6, 7

[61] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pages 331–348. Springer, 2025. 2, 3, 6, 7, 1

[62] Ziyi Wu, Yulia Rubanova, Rishabh Kabra, Drew A Hudson, Igor Gilitschenski, Yusuf Aytar, Sjoerd van Steenkiste, Kelsey R Allen, and Thomas Kipf. Neural assets: 3d-aware multi-object scene synthesis with image diffusion models. *arXiv preprint arXiv:2406.09292*, 2024. 3

[63] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 1

[64] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 2, 4, 5

[65] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 7

[66] 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. *arXiv preprint arXiv:2402.03162*, 2024. 3

[67] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 6

[68] 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

[69] Jiraphon Yenphraphai, Xichen Pan, Sainan Liu, Daniele Panozzo, and Saining Xie. Image sculpting: Precise object editing with 3d geometry control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4241–4251, 2024. 3, 5, 6, 7, 1

[70] 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. 3

[71] 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. 3

[72] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 3, 4

[73] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Text2nerf: Text-driven 3d scene generation with neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 3

[74] 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. 5, 8

[75] Qihang Zhang, Yinghao Xu, Chaoyang Wang, Hsin-Ying Lee, Gordon Wetzstein, Bolei Zhou, and Ceyuan Yang. 3ditscene: Editing any scene via language-guided disentangled gaussian splatting. *arXiv preprint arXiv:2405.18424*, 2024. 3

[76] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 3

[77] Xin-Yang Zheng, Hao Pan, Yu-Xiao Guo, Xin Tong, and Yang Liu. Mvd^2: Efficient multiview 3d reconstruction for multiview diffusion. *arXiv preprint arXiv:2402.14253*, 2024. 2

[78] Youyi Zheng, Xiang Chen, Ming-Ming Cheng, Kun Zhou, Shi-Min Hu, and Niloy J Mitra. Interactive images: Cuboid proxies for smart image manipulation. *ACM Trans. Graph.*, 31(4):99–1, 2012. 3