Reangle-A-Video: 4D Video Generation as Video-to-Video Translation

Hyeonho Jeong[1,2,*] Suhyeon Lee[1,*] Jong Chul Ye[1]
[1]KAIST [2]Adobe Research
{hyeonho.jeong, suhyeon.lee, jong.ye}@kaist.ac.kr

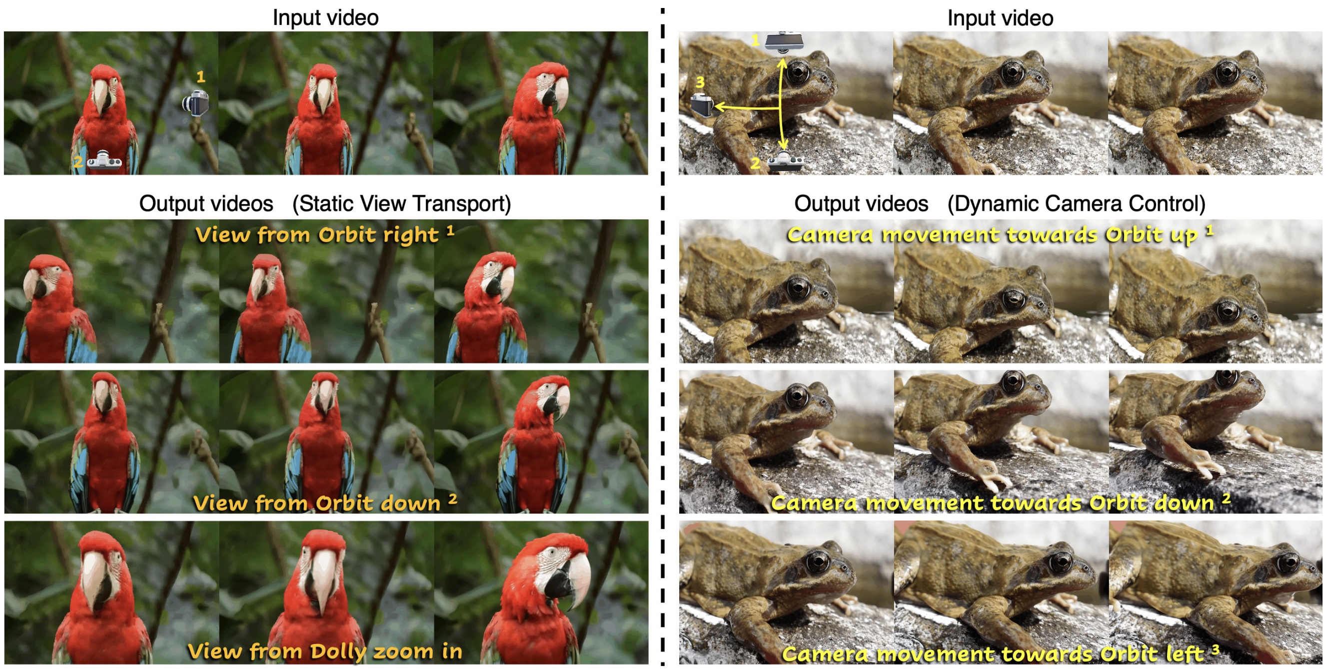

Figure 1. From a single monocular video of any scene, `Reangle-A-Video` generates synchronized videos from diverse camera viewpoints or movements without relying on any multi-view generative prior—*using only single fine-tuning of a video generator*. The first row shows the input video, while the rows below present videos generated by Reangle-A-Video. (Left): Static view transport results. (Right): Dynamic camera control results. Full video examples are available on our project page: *hyeonho99.github.io/reangle-a-video*

Abstract

We introduce Reangle-A-Video, a unified framework for generating synchronized multi-view videos from a single input video. Unlike mainstream approaches that train multi-view video diffusion models on large-scale 4D datasets, our method reframes the multi-view video generation task as video-to-videos translation, leveraging publicly available image and video diffusion priors. In essence, Reangle-A-Video operates in two stages. (1) Multi-View Motion Learning: An image-to-video diffusion transformer is synchronously fine-tuned in a self-supervised manner to distill view-invariant motion from a set of warped videos. (2) Multi-View Consistent Image-to-Images Translation: The first frame of the input video is warped and inpainted into various camera perspectives under an inference-time cross-view consistency guidance using DUSt3R, generating multi-view consistent starting images. Extensive experiments on static view transport and dynamic camera control show that Reangle-A-Video surpasses existing methods, establishing a new solution for multi-view video generation. Our codes and data are available at our project page .

1. Introduction

Diffusion-based video generators [5, 7, 9, 39, 44, 57, 73, 84, 91] are rapidly advancing, enabling the generation of visually rich and dynamic videos from text or visual inputs. Recent progress in video generative models highlights the

[*] indicates equal contribution

Figure 2. **Qualitative results on *static view transport (left)* & *dynamic camera control (right)***. Click with Acrobat Reader to play videos.

growing need for user control over object appearance [12, 41, 48, 76], object motion [10, 26, 38, 40, 55, 85, 93], and camera pose [30, 59, 74, 78, 81, 83, 87, 90]. However, a video itself inherently captures only a partial perspective of the world, which exists as a dynamic 4D environment.

To obtain dynamic 4D generative priors, previous works [50, 52, 79, 96] have extended 3D or video generators into 4D generative models using rendered synthetic assets [21, 22] that focus on animated objects or rigged characters. As a result, these methods are limited to object-level 4D synthesis and fail to generalize to real-world scenes. Recent studies have addressed this limitation by training foundational multi-view video diffusion models [4, 64, 67, 69, 71, 75, 77] on hybrid datasets combining indoor/outdoor static scenes [20, 53, 95], general single-view videos, and realistic scenes rendered by simulation engines [21, 22, 28, 62]. Although promising, these approaches often face challenges in real-world scenarios [69], are limited to specific domains (e.g., human-centric synthesis [64]), and most are not publicly available. More importantly, most existing methods generate multi-view videos from image or text inputs, rather than from user-input videos.

To address this, here we present Reangle-A-Video, an alternative solution for synchronized multi-view video generation that does not require specialized multi-view generative priors. Given any input video, we frame the task as *video-to-videos translation*, capitalizing on publicly available image and video diffusion models. Our unified framework supports both *static view transport*—resimulating a video from target viewpoints—and *dynamic camera control*, where the video gradually transitions to the target viewpoints. Both approaches offer six degrees of freedom (see Fig. 1 and 2).

In essence, our approach is based on the decomposition of a dynamic 4D scene into view-specific appearance (*the starting image*) and view-invariant motion (*image-to-video generation*). To capture view-invariant motion of the scene, we augment the training dataset with warped videos generated via repeated point-based warping of a single monocular video—these warped videos provide strong hints about the camera's perspectives. We then fine-tune a pre-trained image-to-video model [84] using a synchronized few-shot training strategy building on a masked diffusion loss [2, 17, 63, 90]. After training, dynamic camera control over video (Fig.1,2-*right*) is achieved by generating videos using the original first frame input with text that specifies the desired camera movement. On the other hand, static view transport of a video requires viewpoint-transported starting images. To achieve this, our method addresses several key technical challenges. First, we generate the starting images by inpainting warped images with image diffusion prior [49]. Second, inspired by test-time compute scaling [42, 70, 86], we enforce multi-view consistency in *inference-time*, using a stochastic control guidance with an off-the-shelf multi-view stereo reconstruction network [72]. We demonstrate the effectiveness of our approach on a variety of real-world scenes and metrics.

2. Related Work

Denoising-based Image and Video Generation. The introduction of Diffusion Transformers (DiT) [56] revolutionized image and video generation by replacing the traditional U-Net with a transformer-based backbone. Combined with larger-scale, high-quality, curated datasets, this shift enabled more efficient and scalable diffusion generators. In image generation, the PixArt family [13–15] extended DiT to text-to-image synthesis, showcasing its versatility. Stable Diffusion 3 [23] and Flux [49] further advanced the field with Multi-modal Diffusion Transformers (MM-DiT), enabling bidirectional text-image interactions. Following this trend, recent video diffusion models also integrate MM-DiT with spatio-temporal VAEs, achieving high-quality, long-form video generation. In our work, we adopt Flux [49] for image diffusion and CogVideoX [84] for video diffusion, both built on the MM-DiT architecture.

Multi-View Video Generation. Previous efforts on multi-view (4D) video generation [50, 52, 79, 96] have mainly focused on synthesizing 4D assets—animated objects and rigged characters [21, 22]. A common approach is to combine motion priors from video diffusion models with 3D priors from multi-view image models and train on curated, synthetic datasets of dynamic 3D objects [50, 79]. While

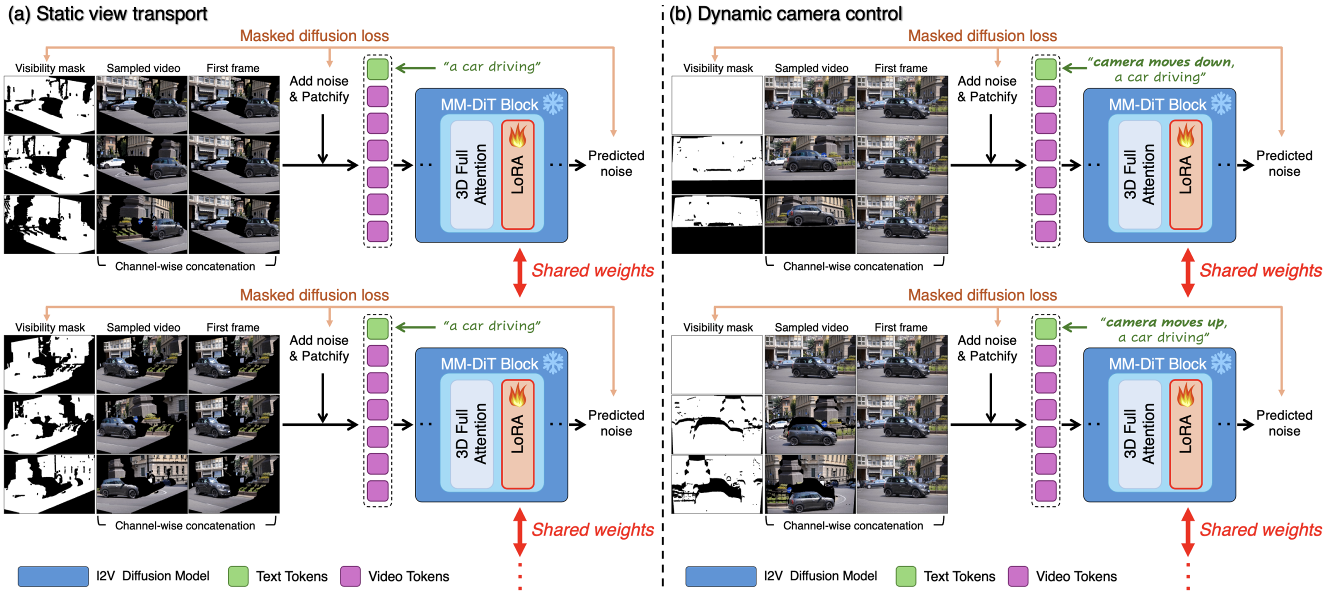

Figure 3. Multi-view motion learning pipelines for **(a) Static view transport** and **(b) Dynamic camera control.** For both tasks, we distill view-robust motion of the underlying scene to a pre-trained MM-DiT video model [84], using all visible pixels within the sampled videos. This few-shot, self-supervised training optimizes only the LoRA layers [35, 61], enabling lightweight training.

promising, these methods are limited to object-level 4D synthesis—rendered animations of single objects or animals—and produce multi-view videos with a static camera pose per video, failing to generalize to real-world videos and arbitrary viewpoints. More recently, to capture open-domain real-world scenes, several works [4, 45, 67, 69, 71, 77] have built foundational multi-view video generative priors by training on multi-view images/videos of real-world static scenes [53, 95], general single-view video datasets, and photorealistic scenes rendered by various engines [21, 28, 62]. These efforts aim to overcome the scarcity of dynamic, in-the-wild multi-view video data.

In contrast, our work reinterprets multi-view video generation as a *video-to-multi-view-videos* translation task. We demonstrate that synchronized few-shot fine-tuning of a video diffusion prior with visible pixels across arbitrary views is sufficient for multi-view video generation in real-world settings, without the need for expensive 4D diffusion models trained on large datasets.

Camera Controllable Video Generation. Our framework is also related to video generation methods that incorporate controllable camera trajectories [3, 16, 30, 34, 46, 51, 59, 69, 74, 78, 80, 81, 83, 87, 89, 90]. Most approaches fine-tune pre-trained text/image-to-video diffusion models on multi-view datasets of *static* scenes (e.g., RealEstate10K [95], ScanNet [20], DL3DV-10K [53]) to integrate additional camera control signals. They often rely on ControlNet-like hypernetworks [26, 29, 30, 74, 92], raw camera extrinsics [74], camera ray coordinates [30, 75], Plücker coordinate embeddings [3, 46, 81], or combinations of multiple conditions [59].

Since these methods derive multi-view priors primarily from *static* scenes, they typically support camera-controlled generation from *text* or *image* inputs and are not designed to produce multiple videos that are mutually consistent. In contrast, our work focuses on (i) enabling camera control over an input *video* and, more importantly, (ii) generating multiple videos that remain consistent with each other. This is more challenging, as our framework must replicate the inherent dynamics of the input video, such as object motion, while ensuring consistency across the generated videos.

3. Reangle-A-Video

In this section, we provide a comprehensive discussion of our framework. Given an input video describing an arbitrary 4D scene, we aim to generate synchronized multi-view videos of the same scene without relying on multi-view generative priors. Leveraging pre-trained latent image and video diffusion models (Sec. 3.1), our approach decomposes the dynamic 4D scene into view-specific appearance *(starting image)* and view-invariant motion *(image-to-video)*, addressing each component separately. We first embed the scene's view-invariant motion into a pre-trained video diffusion model using our novel *self-supervised training with data augmentation* strategy. Initially, to capture diverse perspectives from a single monocular video, we repeatedly perform point-based warping to generate a set of warped videos (Sec. 3.2). These videos, together with the original video, form the training dataset for fine-tuning a pre-trained image-to-video diffusion model with a masked diffusion objective (Sec. 3.3). To achieve *(b) dynamic camera control*, we sample videos using the fine-tuned model with the original first frame as input. In contrast, for *(a) static view transport*, we generate view-transported starting images by inpainting the warped first frames under an inference-time view consistency guidance using an off-the-shelf multi-view stereo reconstruction network (Sec. 3.4).

3.1. Preliminary: Latent Diffusion Models

Diffusion models [33, 65, 66] generate clean images or videos from Gaussian noise via an iterative denoising process. This process reverses a fixed, time-dependent forward diffusion process, which gradually corrupts the data by adding Gaussian noise. In latent diffusion models [8, 60], this process operates in the lower-dimensional latent space of a pre-trained VAE [43], comprised of an encoder $\mathcal{E}(\cdot)$ and a decoder $\mathcal{D}(\cdot)$. Given a clean sample, $\boldsymbol{x}_0 \sim p_{\text{data}}(\boldsymbol{x})$, it is first compressed to a latent representation $\boldsymbol{z}_0 = \mathcal{E}(\boldsymbol{x}_0)$. Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ is then added to produce intermediate noisy latents via the forward process $\boldsymbol{z}_t = \alpha_t \boldsymbol{z}_0 + \sigma_t \boldsymbol{\epsilon}$, where t denotes the diffusion timestep, and α_t, σ_t are noise scheduler parameters. The training objective is to learn a denoiser network $\boldsymbol{\epsilon}_\theta$ that minimizes:

$$\mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(0,I), \boldsymbol{z}_t \sim p, t, c}\big[\, \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, c)\|_2^2 \,\big], \tag{1}$$

with condition c provided by a text prompt, image, or both. Once trained, the diffusion model generates clean samples by iteratively denoising a pure Gaussian noise.

3.2. Stage I: Point-based Warping for Training Data Augmentation

Given an input video $\boldsymbol{x}^{1:N}$ of N frames, our goal is to construct a training dataset for the subsequent fine-tuning stage, by lifting pixels into time-aware 3D point clouds and reprojecting them onto the image plane with target perspectives.

For each frame $\boldsymbol{x}^i, i \in \{1, ..., N\}$, we first estimate its depth map $\mathbf{D}^i$ using a monocular depth estimator [82]. The corresponding point cloud $\mathcal{P}^i$ is then generated from the RGBD image $[\boldsymbol{x}^i, \mathbf{D}^i]$ as follows:

$$\mathcal{P}^i = \phi_{2\to 3}([\boldsymbol{x}^i, \mathbf{D}^i], \mathbf{K}, \mathbf{P}^i_{\text{src}}), \tag{2}$$

where $\mathbf{K}$ is the camera intrinsic matrix, $\mathbf{P}^i_{\text{src}}$ is the extrinsic matrix for frame $\boldsymbol{x}^i$, and $\phi_{2\to 3}$ is the function that lifts the RGBD image into a point cloud. Note that $\mathbf{K}$ and $\mathbf{P}^i_{\text{src}}$ are set by convention as in [18], since they are intractable in open-domain videos.

Next, we define M target extrinsic matrix trajectories $\Phi_{1:M}$, where each trajectory is given by

$$\Phi_j = \{\mathbf{P}^1_j, ..., \mathbf{P}^N_j\},\ j \in \{1, ..., M\}. \tag{3}$$

Each extrinsic matrix $\mathbf{P}^i_j$ comprises a rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times 3}$ and a translation vector $\mathbf{t} \in \mathbb{R}^{3\times 1}$, which together transform the point cloud $\mathcal{P}^i$ into the target camera coordinate system. In the *static view transport* setting, each target trajectory Φ_j is constant across all frames, i.e., $\mathbf{P}^1_j = \mathbf{P}^2_j = ... = \mathbf{P}^N_j$; for *dynamic camera control*, each frame's camera pose $\mathbf{P}^i_j$ is determined by incrementally moving and rotating from the previous pose $\mathbf{P}^{i-1}_j$, with the first target pose set to the pose of the first input frame ($\mathbf{P}^1_j = \mathbf{P}^1_{\text{src}}$).

Finally, we reproject each point cloud $\mathcal{P}^i$ to the image plane under the target perspective using $\mathbf{K}$ and $\mathbf{P}^i_j$ via the function $\phi_{3\to 2}$:

$$(\hat{\boldsymbol{x}}^i_j, \boldsymbol{m}^i_j) = \phi_{3\to 2}(\mathcal{P}^i, \mathbf{K}, \mathbf{P}^i_j), \tag{4}$$

where $\hat{\boldsymbol{x}}^i_j$ is the rendered warped image and $\boldsymbol{m}^i_j$ is the corresponding visibility mask (1 for visible surfaces, 0 for invisible regions). Repeating this process constructs our training dataset Ω, which consists of M pairs of warped videos and corresponding visibility masks: $\Omega = \{(\hat{\boldsymbol{x}}^{1:N}_j, \boldsymbol{m}^{1:N}_j) \mid j = 1, ..., M\}$. We also add the input video $\boldsymbol{x}^{1:N}$ with uniform visibility masks (all pixels set to 1), resulting in a total of $M{+}1$ video-mask pairs.

3.3. Stage II: Multi-View Motion Learning

In this stage, we fine-tune a pre-trained image-to-video diffusion transformer to learn view-invariant motion from the dataset Ω that captures diverse scene perspectives and corresponding pixels. Here, "motion" refers to the nuanced dynamics of each object—including its type, direction, and speed—as well as any inherent camera movement when the original video's camera pose is not fixed.

To achieve lightweight fine-tuning while preserving the original model's prior, we employ Low-Rank Adaptation (LoRA) [35, 61]. LoRA augments an attention layer by adding a residual path comprised of two low-rank matrices, $\theta_B \in \mathbb{R}^{d\times r}$, $\theta_A \in \mathbb{R}^{r\times k}$, to the original weight $\theta_0 \in \mathbb{R}^{d\times k}, r \ll \min(d, k)$. The modified forward pass is defined as:

$$\theta = \theta_0 + \alpha\Delta\theta = \theta_0 + \alpha\theta_B\theta_A, \tag{5}$$

where α controls the strength of the LoRA adjustment. We implement this approach on a MM-DiT-based image-to-video diffusion model [84]—which incorporates 3D full-attention in its MM-DiT blocks—by attaching LoRA layers to these 3D full-attention layers.

We then randomly sample a video $\boldsymbol{x}^{1:N}$ and its corresponding visibility mask $\boldsymbol{m}^{1:N}$ from the dataset Ω prepared in Sec. 3.2. Note that if $\boldsymbol{x}^{1:N}$ is the original input video, its mask $\boldsymbol{m}^{1:N}$ is uniformly filled with 1. The video is compressed into latent space $\boldsymbol{z}^{1:N}_0 = \mathcal{E}(\boldsymbol{x}^{1:N})$, and the mask is downsampled to $\boldsymbol{m}^{1:N}_{\text{down}}$ using spatio-temporal downsampling (see Sup. Sec.A.4). The video latents are then noised, patchified, and unfolded into a long sequence to form video tokens. These, along with the text tokens, are fed into the video denoiser $\boldsymbol{\epsilon}_\theta$. Since most sampled videos are warped and contain significant black regions, naive LoRA fine-tuning with the standard diffusion loss

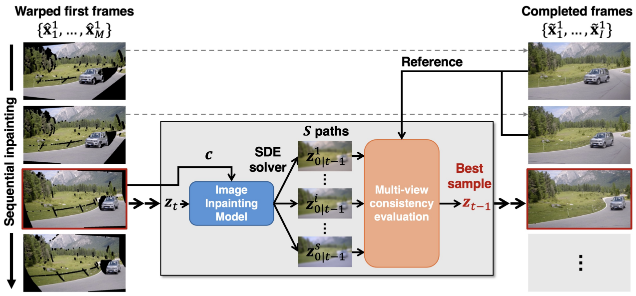

Figure 4. **Multi-view consistent image inpainting using stochastic control guidance.** In experiments, we set $S = 25$.

(Eq. (1)) inevitably degrades the original model prior and makes the model generate warped videos (see Supp. Sec. B.2). Thus, leveraging the compositionality of diffusion objectives [2, 17, 25, 63, 90, 94], we fine-tune the LoRA layers $\Delta\theta$ using a masked diffusion loss that excludes invisible regions in it's loss computation:

$$\mathbb{E}_{\boldsymbol{\epsilon}, \boldsymbol{z}_t^{1:N}, \boldsymbol{m}_{\text{down}}^{1:N}, t, c}\Big[\big\|\boldsymbol{\epsilon} \odot \boldsymbol{m}_{\text{down}}^{1:N} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t^{1:N}, t, c) \odot \boldsymbol{m}_{\text{down}}^{1:N}\big\|_2^2\Big], \tag{6}$$

where condition c includes both the input text and image, with the latter being the first frame of the sampled video. [1]

For text input, *(b) dynamic camera control* necessitates explicit specification of the camera movement type. As shown in Fig. 3-*(b)*, during the dynamic camera control training, the image-to-video model always starts with the same first frame—the original video's first frame—which is uniformly used across warped videos (see Sec. 3.2). As a result, the model cannot infer the desired camera movement from the uniform starting frame alone. To resolve this, we explicitly include the camera movement type in the text input during both training and inference. Our LoRA layers, attached to the 3D full-attention modules, enable interactions between text and video tokens, learning to distinguish between different camera movements and map the corresponding text tokens to the appropriate video tokens.

The proposed training pipelines are outlined in Fig. 3. This synchronized few-shot training strategy prevents the video model from overfitting to a specific view. Instead, it learns general scene motion, enabling video generation even from viewpoints that were unseen during the training (see Sec. 4.3), while ensuring that all output videos are synchronized (e.g., with consistent object motion speeds).

Once training is complete, for the *(b) dynamic camera control* setup, we sample videos directly using the original video's first frame as the input image and text specifying the desired camera movement. However, to achieve *(a) static view transport*, we require a starting input image that captures the scene from the desired viewpoint—a process addressed in the next stage.

[1]Following [84], the image input is encoded into latent space, then concatenated along the channel dimension with the noisy video latents.

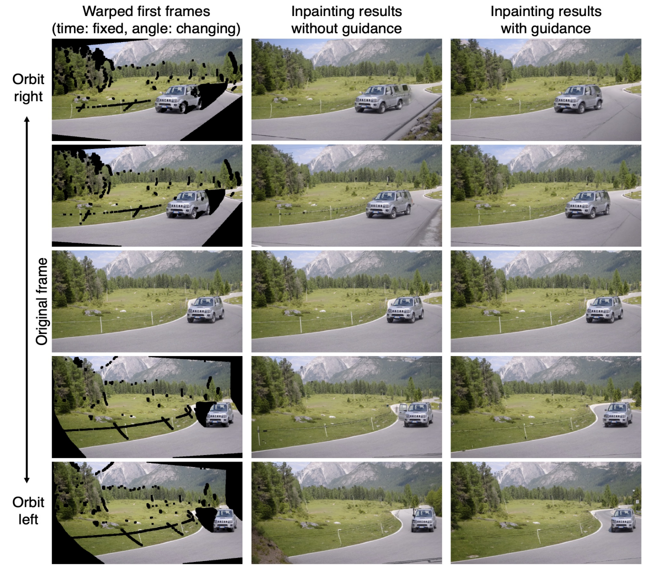

Figure 5. **Qualitative inpainting comparisons.** We compare naive inpainting to inpainting with stochastic control guidance.

3.4. Stage III: Multi-View Image Inpainting

We achieve multi-view consistent starting image generation using a warp-and-inpaint approach. In Stage I, given an original image $\boldsymbol{x}^1$ (the first frame of the input video), we render a set of warped images $\hat{\boldsymbol{x}}^1_{1:M}$ from M target viewpoints, along with corresponding binary masks $\boldsymbol{m}^1_{1:M}$ that indicate invisible surfaces. These missing regions are then filled via inpainting using image diffusion prior. While state-of-the-art image generators [15, 23, 49] yield plausible inpainting results on single warped images, independently inpainting each warped view fails to ensure cross-view consistency (see Fig. 5, 2nd column). Motivated by inference-time scaling practices [36, 42, 70, 86], we address the emerging inconsistencies by introducing a stochastic control guidance that enforces multi-view consistency during the denoising process using the DUSt3R multi-view stereo reconstruction model [72]. Specifically, we sequentially inpaint warped images by leveraging previously inpainted views for consistency, guided by stochastic control guidance. At each denoising step, we generate S clean estimates from S stochastic paths and compute a multi-view consistency score for each estimate with respect to the previously inpainted images $\{\tilde{\boldsymbol{x}}^1_1, ..., \tilde{\boldsymbol{x}}^1_I\}$. Following [1], the score is calculated by projecting the images into a shared view using DUSt3R [72] and then computing the DINO feature similarity [11] on the projected features. We select the path with the highest score, add new noise to generate S new paths, and iterate until the inpainting is complete [2]. We present detailed algorithms in Supp. Alg. 1 and 2.

[2]The process is repeated sequentially for all views, with the first two inpainted concurrently. To ensure initial consistency, we generate S candidate versions per view at every sampling step, evaluate multi-view consistency for all S^2 pairs, and select the optimal pair for the next timestep.

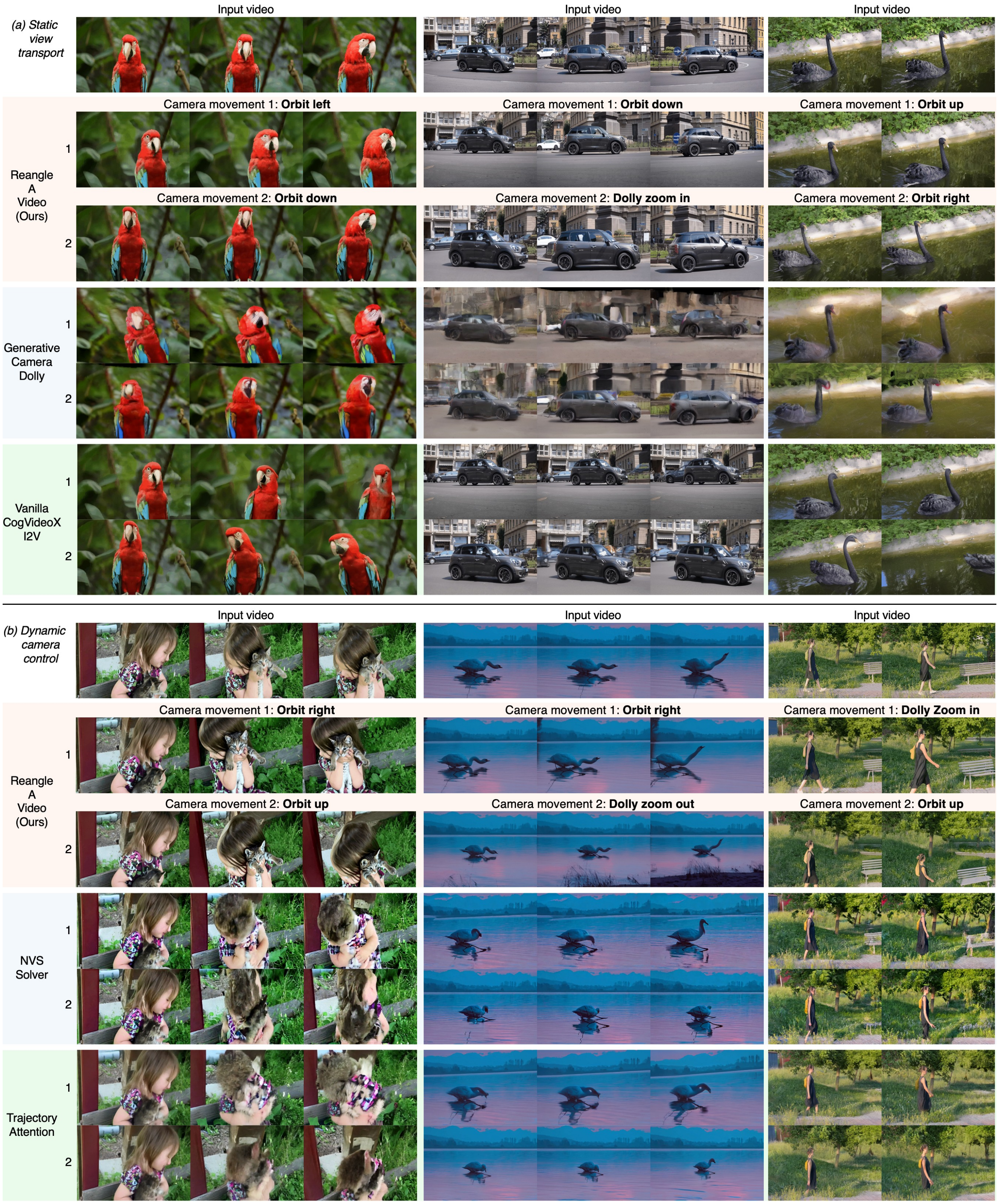

Figure 6. **Qualitative comparisons.** Top half shows *(a) Static view transport* and bottom half presents *(b) Dynamic camera control* results. The first row in each half displays the input videos, and for each input video, two generated videos corresponding to target cameras (1 and 2) are shown for each method. Across baseline, same camera parameters were used for each 1,2. Visit our page for full-video results.

Table 1. **Quantitative comparison results**. Reangle-A-Video is evaluated in two modes—**(a) Static View Transport** and **(b) Dynamic Camera Control**—against the baselines. VBench metrics [37] are presented on the left, other metrics [1, 31, 68] appear on the right.

(a) Static View Transport	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	MEt3R↓	FID↓	FVD↓
Generative Camera Dolly [69]	0.8849	0.9107	0.8731	0.9119	0.7608	0.4052	0.5524	0.1237	155.15	5264.7
Vanilla CogVideoX [84]	0.9448	**0.9398**	0.9738	0.9859	0.7286	0.4997	**0.6475**	0.0539	79.621	3664.2
Reangle-A-Video (Ours)	**0.9516**	0.9327	**0.9764**	**0.9907**	**0.7657**	**0.5160**	0.6414	**0.0412**	**53.448**	**2690.9**
(b) Dynamic Camera Control										
NVS-Solver [87]	0.9037	0.9325	0.9049	0.9521	0.8809	0.5023	**0.6411**	0.1090	95.815	3516.5
Trajectory Attention [78]	0.8984	0.9288	0.9342	0.9658	**0.8889**	0.4854	0.5990	0.0965	109.20	3624.9
Reangle-A-Video (Ours)	**0.9140**	**0.9364**	**0.9386**	**0.9794**	0.8884	**0.5238**	0.6271	**0.0648**	**74.194**	**3019.7**

4. Experiments

4.1. Implementation Details

We experiment with 28 publicly sourced videos [27, 58], covering diverse scenes with varying objects, scene motions, and physical environments. Each video consists of 49 frames at a resolution of 480×720. On average, we generate videos from 3.5 different camera viewpoints/movements per input, resulting in a total of 98 generated videos.

For **Stage I**, Depth Anything V2 model [82] is used to estimate the depth maps of the input videos. For **Stage II** and **III**, we use the CogVideoX-5b image-to-video diffusion model [84], and FLUX text-to-image model [49] with a pre-trained inpainting ControlNet [19] attached to it.

Upon video model fine-tuning, we set the LoRA rank to 128 and optimize the LoRA layers over 400 steps, where the fine-tuned parameters account for only about 2% of the original Video DiT parameters [3]. We use AdamW [54] optimizer with a learning rate of $1e{-}4$ and a weight decay of $1e{-}3$. The fine-tuning takes about an hour. For the *static view transport* mode, we set $M = 12$ warped videos; for *dynamic camera control*, we set $M = 6$ (see Supp. Sec. A.2). For video sampling, we apply 40 sampling steps with a CFG [32] scale of 6.0. During multi-view image inpainting, we first resize the input to 1024×1024 and then restore it to its original resolution. We use MEt3R [1] as the reward function for stochastic control with $S = 25$. To introduce stochasticity for the path control, we employ an SDE-based sampler with 50 steps, ensuring the same marginal distribution as the original ODE of the flow trajectory (see Supp. Sec. A.6). All experiments, including inference and fine-tuning, are performed using 40GB A100 GPUs.

4.2. Comparisons

4.2.1. Baselines

Multi-view/camera synchronized video generation from an input video remains largely underexplored. For *(a) Static view transport*, the closest works include **Generative Camera Dolly** [69] and the closed-source multi-view video foundation model CAT4D [77]. Although GS-DiT [6] aligns with our framework, its code was unavailable during our research. In addition to GCD, we evaluate a baseline using naive CogVideoX-I2V inference (**Vanilla CogVideoX**)[84], which employs the same input frame as our approach. For *(b) Dynamic camera control* over videos, we compare against two state-of-the-art methods, **NVS-Solver** [87] and **Trajectory Attention** [78]. While Recapture [90] is closely related, its code is not available, and it employs the proprietary multi-view image foundation model CAT3D [24].

4.2.2. Qualitative Results

Comprehensive qualitative results are shown in Fig. 6. For *(a) Static View Transport*, Generative Camera Dolly [69] struggles with real-world videos (trained on synthetic data [28]), while Vanilla CogVideoX [84] fails to capture the input video's motion. In contrast, Reangle-A-Video accurately reproduces the input motion from the target viewpoint. For *(b) Dynamic Camera Control*, NVS-Solver [87] and Trajectory Attention [78] either confuse foreground objects (e.g., a girl's and a cat's head in the first video), fail to capture precise motion (e.g., neck movement in the second video), or miss background elements (e.g., a bench in the third video). Our method faithfully regenerates the input video's motion, preserves object appearance, and accurately follows the target camera movement.

4.2.3. Quantitative Results

Automatic metrics. For automated evaluation, we first use VBench [37], which assesses generated videos across various disentangled dimensions. As shown in Tab. 1-left, our method outperforms baselines in most metrics for both *static view transport* and *dynamic camera control* modes. Notably, compared to vanilla CogVideoX I2V—which uses the same input image—our approach maintain robust performance even when fine-tuned with warped videos. Additionally, Tab. 1-right presents evaluation results using FID [31], FVD [68], and the recently proposed multi-view consistency metric MEt3R [1].

Human evaluation. We further assess Reangle-A-Video against baselines via a user study involving 36 participants who compare our results with randomly selected baselines. For the *static view transport* evaluation, participants rate: (i) accuracy of the transported viewpoint, (ii) preservation of the input video's motion. For the *dynamic camera control*

[3] To enable video model fine-tuning within a 40GB VRAM constraint, we employ gradient checkpointing, which reduces memory usage at the expense of slower gradient back-propagation.

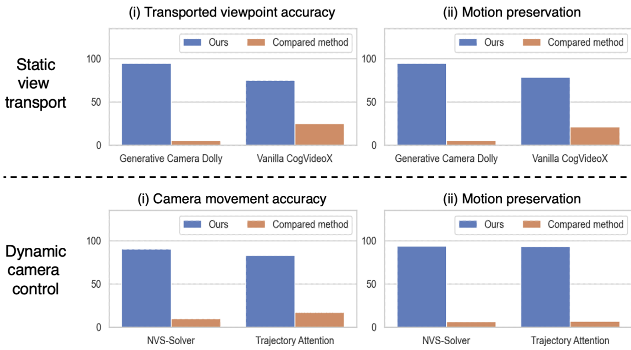

Figure 7. **User study results.** Top: Static view transport results. Bottom: Dynamic camera control results.

Table 2. Quantitative evaluation of multi-view consistency in image inpainting with and without stochastic control guidance.

Warped image inpainting	MEt3R ↓	SED ↓	TSED ↑
w/o stochastic control guidance	0.1431	1.1966	0.5241
w/ stochastic control guidance	**0.1184**	**1.1844**	**0.5588**

evaluation, participants assess: (i) the accuracy of the target camera movement, (ii) the preservation of the input video's motion in the output. As reported in Fig. 7, our method outperforms the baselines in all aspects.

4.3. Ablation Studies

First, we ablate the stochastic control guidance by comparing it with naive inpainting (where each image is inpainted independently). Tab. 2 evaluates multi-view accuracy using the MEt3R [1], SED, and TSED [88] metrics (with $T_e = 1.25$ and $T_m = 10$ for SED). Fig. 5 shows that our method fills target (invisible) regions consistently across multiple views. Next, we ablate the necessity of our data augmentation strategy (i.e., the use of warped videos) in learning view-invariant motion. Fig. 8 compares fine-tuning with only the original input video versus using both the original and warped videos. The results indicate that relying solely on the original video fails to accurately capture motion—for example, the rhino moving in front of the tree. For quantitative assessment on input video's motion preservation, we employ user study (see Supp. Sec. B.1.) In Fig. 9-top, we demonstrate unseen view video generation. We exclude specific warped view videos (vertical up/down orbits) from the training dataset and fine-tune the video model without them. Then, using an inpainted first frame as input, we generate a video from that omitted view. In Fig. 9-bottom, we showcase novel view video generation guided by an edited first frame. Starting with an inpainted image representing the scene from a target viewpoint, we apply FlowEdit [47] to modify the image, then generate the novel-view video with our fine-tuned model. In both cases, the generated videos faithfully follow the input video's motion, demonstrating that our few-shot training strategy is robust in both view and appearance.

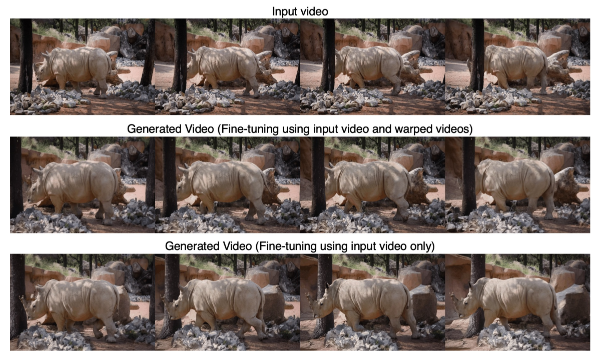

Figure 8. Novel view video generation with and without using warped video for training (target viewpoint: dolly zoom in).

Figure 9. Top: Unseen view video generation. Bottom: Novel view video generation using an appearance-edited first frame.

5. Conclusion

We have presented an approximate yet effective solution for generating synchronized multi-view videos from a single monocular real-world video—without relying on any multi-view generative priors. Our approach enables an image-to-video diffusion model to learn view-invariant scene motion through self-supervised fine-tuning with video augmentation. Then, we sample videos from the fine-tuned model using a set of multi-view consistent first-frame images, where the images are generated via a warp-and-inpaint process that enforces multi-view consistency during inference-time.

Limitations and future work. Input image quality is crucial in image-to-video generation, as the output cannot exceed the first frame's quality. While our approach ensures *multi-view consistency* in the warp-and-inpaint scheme, the warping stage is inherently prone to artifacts caused by camera and depth errors (Supp. Sec. C). We believe that improved depth models could enhance both the warped and inpainted images. Another limitation of our work is the need for scene-specific model tuning to achieve 4D video synthesis. A promising future direction is to extend existing video datasets with warped videos to train 4D foundation models.

Acknowledgments

We would like to thank Geon Yeong Park and Mingi Kwon for their valuable discussions at the early stage of the project. This work was supported by the National Research Foundation of Korea under Grant RS-2024-00336454, and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2024-00457882, AI Research Hub Project; No. RS-2025-02304967, AI Star Fellowship (KAIST)).

References

[1] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. *arXiv preprint arXiv:2501.06336*, 2025. 5, 7, 8

[2] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. 2, 5

[3] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. *arXiv preprint arXiv:2411.18673*, 2024. 3

[4] Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. *arXiv preprint arXiv:2412.07760*, 2024. 2, 3

[5] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1

[6] Weikang Bian, Zhaoyang Huang, Xiaoyu Shi, Yijin Li, Fu-Yun Wang, and Hongsheng Li. Gs-dit: Advancing video generation with pseudo 4d gaussian fields through efficient dense 3d point tracking. *arXiv preprint arXiv:2501.02690*, 2025. 7

[7] 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

[8] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 4

[9] 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

[10] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. *arXiv preprint arXiv:2501.08331*, 2025. 2

[11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 5

[12] Hila Chefer, Shiran Zada, Roni Paiss, Ariel Ephrat, Omer Tov, Michael Rubinstein, Lior Wolf, Tali Dekel, Tomer Michaeli, and Inbar Mosseri. Still-moving: Customized video generation without customized video data. *ACM Transactions on Graphics (TOG)*, 43(6):1–11, 2024. 2

[13] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 2

[14] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[15] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart-δ: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024. 2, 5

[16] Soon Yau Cheong, Duygu Ceylan, Armin Mustafa, Andrew Gilbert, and Chun-Hao Paul Huang. Boosting camera motion control for video diffusion transformers. *arXiv preprint arXiv:2410.10802*, 2024. 3

[17] Gene Chou, Kai Zhang, Sai Bi, Hao Tan, Zexiang Xu, Fujun Luan, Bharath Hariharan, and Noah Snavely. Generating 3d-consistent videos from unposed internet photos. *arXiv preprint arXiv:2411.13549*, 2024. 2, 5

[18] 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. 4

[19] Alimama Smart Creative and AI Application Team. Flux inpainting controlnet, 2025. URL https://huggingface.co/alimama-creative/FLUX.1-dev-Controlnet-Inpainting-Beta. 7

[20] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 3

[21] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 3

[22] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[23] 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. 2, 5

[24] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 7

[25] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Mdtv2: Masked diffusion transformer is a strong image synthesizer. *arXiv preprint arXiv:2303.14389*, 2023. 5

[26] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, et al. Motion prompting: Controlling video generation with motion trajectories. *arXiv preprint arXiv:2412.02700*, 2024. 2, 3

[27] Canva Germany GmbH. pexels, 2024. URL https://www.pexels.com/videos/. 7

[28] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3749–3761, 2022. 2, 3, 7

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

[30] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 2, 3

[31] 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. 7

[32] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 7

[33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4

[34] Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024. 3

[35] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 3, 4

[36] Yujia Huang, Adishree Ghatare, Yuanzhe Liu, Ziniu Hu, Qinsheng Zhang, Chandramouli S Sastry, Siddharth Gururani, Sageev Oore, and Yisong Yue. Symbolic music generation with non-differentiable rule guided diffusion. *arXiv preprint arXiv:2402.14285*, 2024. 5

[37] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024. 7

[38] Hyeonho Jeong, Jinho Chang, Geon Yeong Park, and Jong Chul Ye. Dreammotion: Space-time self-similar score distillation for zero-shot video editing. *arXiv preprint arXiv:2403.12002*, 2024. 2

[39] Hyeonho Jeong, Chun-Hao Paul Huang, Jong Chul Ye, Niloy Mitra, and Duygu Ceylan. Track4gen: Teaching video diffusion models to track points improves video generation. *arXiv preprint arXiv:2412.06016*, 2024. 1

[40] Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9212–9221, 2024. 2

[41] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6689–6700, 2024. 2

[42] Jaemin Kim, Bryan S Kim, and Jong Chul Ye. Free˜2 guide: Gradient-free path integral control for enhancing text-to-video generation with large vision-language models. *arXiv preprint arXiv:2411.17041*, 2024. 2, 5

[43] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[44] 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. 1

[45] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. *arXiv preprint arXiv:2405.17414*, 2024. 3

[46] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. *Advances in Neural Information Processing Systems*, 37:16240–16271, 2025. 3

[47] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. *arXiv preprint arXiv:2412.08629*, 2024. 8

[48] Gihyun Kwon and Jong Chul Ye. Tweediemix: Improving multi-concept fusion for diffusion-based image/video generation. *arXiv preprint arXiv:2410.05591*, 2024. 2

[49] Black Forest Labs. Flux, 2024. URL https://blackforestlabs.ai/. 2, 5, 7

[50] Bing Li, Cheng Zheng, Wenxuan Zhu, Jinjie Mai, Biao Zhang, Peter Wonka, and Bernard Ghanem. Vivid-zoo: Multi-view video generation with diffusion model. *arXiv preprint arXiv:2406.08659*, 2024. 2

[51] Teng Li, Guangcong Zheng, Rui Jiang, Tao Wu, Yehao Lu, Yining Lin, Xi Li, et al. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. *arXiv preprint arXiv:2502.10059*, 2025. 3

[52] Hanwen Liang, Yuyang Yin, Dejia Xu, Hanxue Liang, Zhangyang Wang, Konstantinos N Plataniotis, Yao Zhao, and Yunchao Wei. Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models. *arXiv preprint arXiv:2405.16645*, 2024. 2

[53] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 2, 3

[54] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[55] Geon Yeong Park, Hyeonho Jeong, Sang Wan Lee, and Jong Chul Ye. Spectral motion alignment for video motion transfer using diffusion models. *arXiv preprint arXiv:2403.15249*, 2024. 2

[56] 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. 2

[57] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models, 2024. 1

[58] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 7

[59] Stefan Popov, Amit Raj, Michael Krainin, Yuanzhen Li, William T Freeman, and Michael Rubinstein. Camctrl3d: Single-image scene exploration with precise 3d camera control. *arXiv preprint arXiv:2501.06006*, 2025. 2, 3

[60] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2021. 4

[61] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning. *Low-rank adaptation for fast text-to-image diffusion fine-tuning*, 3, 2023. 3, 4

[62] Andrew Sanders. *An introduction to Unreal engine 4*. AK Peters/CRC Press, 2016. 2, 3

[63] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *Advances in Neural Information Processing Systems*, 36:39443–39469, 2023. 2, 5

[64] Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. 360-degree human video generation with 4d diffusion transformer. *ACM Transactions on Graphics (TOG)*, 43(6):1–13, 2024. 2

[65] 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. 4

[66] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 4

[67] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024. 2, 3

[68] 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. 7

[69] Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *European Conference on Computer Vision*, pages 313–331. Springer, 2024. 2, 3, 7

[70] Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7280–7290, 2023. 2, 5

[71] Chaoyang Wang, Peiye Zhuang, Tuan Duc Ngo, Willi Menapace, Aliaksandr Siarohin, Michael Vasilkovsky, Ivan Skorokhodov, Sergey Tulyakov, Peter Wonka, and Hsin-Ying

Lee. 4real-video: Learning generalizable photo-realistic 4d video diffusion. *arXiv preprint arXiv:2412.04462*, 2024. 2, 3

[72] 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, 5

[73] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, pages 1–20, 2024. 1

[74] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 3

[75] Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David J Fleet. Controlling space and time with diffusion models. *arXiv preprint arXiv:2407.07860*, 2024. 2, 3

[76] Yujie Wei, Shiwei Zhang, Hangjie Yuan, Xiang Wang, Haonan Qiu, Rui Zhao, Yutong Feng, Feng Liu, Zhizhong Huang, Jiaxin Ye, et al. Dreamvideo-2: Zero-shot subject-driven video customization with precise motion control. *arXiv preprint arXiv:2410.13830*, 2024. 2

[77] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. *arXiv preprint arXiv:2411.18613*, 2024. 2, 3, 7

[78] Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. Trajectory attention for fine-grained video motion control. *arXiv preprint arXiv:2411.19324*, 2024. 2, 3, 7

[79] Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470*, 2024. 2

[80] Dejia Xu, Yifan Jiang, Chen Huang, Liangchen Song, Thorsten Gernoth, Liangliang Cao, Zhangyang Wang, and Hao Tang. Cavia: Camera-controllable multi-view video diffusion with view-integrated attention. *arXiv preprint arXiv:2410.10774*, 2024. 3

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

[82] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 4, 7

[83] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 2, 3

[84] 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. 1, 2, 3, 4, 5, 7

[85] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8466–8476, 2024. 2

[86] Po-Hung Yeh, Kuang-Huei Lee, and Jun-Cheng Chen. Training-free diffusion model alignment with sampling demons. *arXiv preprint arXiv:2410.05760*, 2024. 2, 5

[87] Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. *arXiv preprint arXiv:2405.15364*, 2024. 2, 3, 7

[88] Jason J Yu, Fereshteh Forghani, Konstantinos G Derpanis, and Marcus A Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7094–7104, 2023. 8

[89] 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

[90] David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. *arXiv preprint arXiv:2411.05003*, 2024. 2, 3, 5, 7

[91] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, pages 1–15, 2024. 1

[92] 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. 3

[93] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 273–290. Springer, 2024. 2

[94] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023. 5

[95] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 2, 3

[96] Qi Zuo, Xiaodong Gu, Lingteng Qiu, Yuan Dong, Zhengyi Zhao, Weihao Yuan, Rui Peng, Siyu Zhu, Zilong Dong, Liefeng Bo, et al. Videomv: Consistent multi-view generation based on large video generative model. *arXiv preprint arXiv:2403.12010*, 2024. 2