CAVIA: CAMERA-CONTROLLABLE MULTI-VIEW VIDEO DIFFUSION WITH VIEW-INTEGRATED ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years there have been remarkable breakthroughs in image-to-video generation. However, the 3D consistency and camera controllability of generated frames have remained unsolved. Recent studies have attempted to incorporate camera control into the generation process, but their results are often limited to simple trajectories or lack the ability to generate consistent videos from multiple distinct camera paths for the same scene. To address these limitations, we introduce **Cavia**, a novel framework for camera-controllable, multi-view video generation, capable of converting an input image into multiple spatiotemporally consistent videos. Our framework extends the spatial and temporal attention modules into view-integrated attention modules, improving both viewpoint and temporal consistency. This flexible design allows for joint training with diverse curated data sources, including scene-level static videos, object-level synthetic multi-view dynamic videos, and real-world monocular dynamic videos. To our best knowledge, Cavia is the first of its kind that allows the user to precisely specify camera motion while obtaining object motion. To the best of our knowledge, Cavia is the first framework that enables users to generate multiple videos of the same scene with precise control over camera motion, while simultaneously preserving object motion. Extensive experiments demonstrate that Cavia surpasses state-of-the-art methods in terms of geometric consistency and perceptual quality.

1 INTRODUCTION

The rapid development of diffusion models has enabled significant advancements in video generative models. Early efforts have explored various approaches, either training a video model from scratch or by fine-tuning pre-trained image generation models with additional temporal layers (Stability, 2023; Wang et al., 2023a; Ho et al., 2022b; Singer et al., 2022; Ho et al., 2022a; Nan et al., 2024). The training data of these video models typically consist of a curated mixture of image (Schuhmann et al., 2022) and video datasets (Bain et al., 2021; Wang et al., 2023b;a; Nan et al., 2024). While substantial progress has been made in improving model architectures and refining training data, relatively little research has been conducted on the 3D consistency and camera controllability of generated videos.

To tackle this issue, several recent works (Wang et al., 2023c; He et al., 2024; Bahmani et al., 2024; Xu et al., 2024; Hou et al., 2024) have attempted to introduce camera controllability in video generation, aiming to ensure that generated frames adhere to viewpoint instructions, thereby improving 3D consistency. These works either enhance viewpoint control through better conditioning signals (Wang et al., 2023c; He et al., 2024; Bahmani et al., 2024) or by utilizing geometric priors, such as epipolar constraints (Xu et al., 2024) or explicit 3D representations (Hou et al., 2024). However, despite these efforts, the generated videos often lack precise 3D consistency or are restricted to static scenes with little to no object motion. Moreover, it remains challenging for monocular video generators to produce multi-view consistent videos of the same scene from different camera trajectories.

Since independently sampling multiple sequences often results in significantly inconsistent scenes, generating multiple video sequences simultaneously is desirable. However, this remains extremely challenging due to the scarcity of multi-view video data in the wild, leading to multi-view generations limited to inconsistent near-static scenes or synthetic objects. A concurrent work, CVD (Kuang et al., 2024), builds on multi-view static videos (Zhou et al., 2018) and warping-augmented monocular videos (Bain et al., 2021), but it can only generate videos with limited baselines, yielding inconsistent

results when object motion is present. Another concurrent work, Vivid-ZOO (Li et al., 2024a), leverages dynamic objects from Objaverse (Deitke et al., 2023b) dataset and renders multi-view videos to train a video generator. However, due to limited data sources, their results are primarily object-centric frames from fixed viewpoints, lacking realistic backgrounds.

To address these challenges, we propose **Cavia**, a novel framework that extends a monocular video generator (Stability, 2023) to generate multi-view consistent videos with precise camera control. We enhance the spatial and temporal attention modules to cross-view and cross-frame 3D attentions respectively, improving consistency across both viewpoints and frames. Our model architecture enables a novel joint training strategy that fully utilizes static, monocular, and multi-view dynamic videos. Static videos (Zhou et al., 2018; Yu et al., 2023; Xia et al., 2024; Reizenstein et al., 2021; Deitke et al., 2023b;a) are converted to multi-view formats to ensure the geometric consistency in the generated frames. We then incorporate rendered synthetic multi-view videos of dynamic 3D objects (Liang et al., 2024; Jiang et al., 2024; Li et al., 2024c) to teach the model to generate reasonable object motion. To prevent overfitting on synthetic data, we finetune the model on pose-annotated monocular videos (Wang et al., 2023b; Nan et al., 2024) to enhance performance on complex scenes. Our framework synthesizes cross-view and cross-frame consistent videos, and extensive evaluations on real and text-to-image generated images show its applicability across challenging indoor, outdoor, object-centric, and large-scene cases. We systematically measure the quality of the generated videos in terms of per-video and cross-view geometric consistency and perceptual quality. Our experiments demonstrate our superiority compared to previous works both qualitatively and quantitatively. Our experiments demonstrate superior performance compared to previous methods, both qualitatively and quantitatively. Additionally, we show that our method can extrapolate to generate four views during inference and enable 3D reconstruction of the generated frames.

Our main contributions can be summarized as follows,

- We propose a novel framework, **Cavia**, for generating multi-view videos with camera controllability. We introduce view-integrated attentions, namely cross-view and cross-frame 3D attentions, to enhance consistency across viewpoints and frames.
- We introduce an effective joint training strategy that leverages a curated mixture of static, monocular dynamic, and multi-view dynamic videos, ensuring geometric consistency, high-quality object motion, and background preservation in the generated results.
- Our experiments demonstrate superior geometric and perceptual quality in both monocular video generation and cross-video consistency compared to baseline methods. Additionally, our flexible framework can operate on four views at inference, offering improved view consistency and enabling 3D reconstruction of the generated frames.

2 Related Works

2.1 Camera Controllable Video Diffusion Models

Recent advancements in video diffusion models have significantly benefited from scaling model architectures and leveraging extensive datasets (Bain et al., 2021; Wang et al., 2023b;a), leading to impressive capabilities in generating high-quality videos (Stability, 2023; Ho et al., 2022b; Singer et al., 2022; Ho et al., 2022a; OpenAI). While large foundational video diffusion models exist, our work focuses on enhancing camera control over video diffusion processes, a rapidly growing area of research. AnimateDiff (Guo et al., 2023) and Stable Video Diffusion (SVD) (Stability, 2023) employ individual camera LoRA (Hu et al., 2021) models for specific camera motions. MotionCtrl (Wang et al., 2023c) improves flexibility by introducing camera matrices, while CameraCtrl (He et al., 2024), CamCo (Xu et al., 2024), and VD3D (Bahmani et al., 2024) enhance the camera control accuracy by introducing Plücker coordinates to the video models via controlnet (Zhang & Agrawala, 2023). To further improve the geometric consistency, CamCo (Xu et al., 2024) applies epipolar constraints and CamTrol (Hou et al., 2024) incorporates 3D Gaussians (Kerbl et al., 2023). However, these methods focus on monocular video generation, limiting their ability to sample multiple consistent video sequences of the same scene from distinct camera paths. CVD (Kuang et al., 2024) extends CameraCtrl (He et al., 2024) for multi-view video generation, but their results are constrained to simple camera and object motion. ViVid-Zoo (Li et al., 2024a) extends MVDream (Shi et al., 2023b) for multi-view purposes but is limited to object-centric results with fixed viewpoints. In contrast, our work explores

view-integrated attentions for more precise camera control over arbitrary viewpoints and introduces a joint training strategy leveraging data mixtures to improve novel-view performance in complex scenes.

2.2 Multi-view Image Generation

Early approaches such as MVDiffusion (Tang et al., 2023) focused on generating multiview images in parallel by employing correspondence-aware attention mechanisms, enabling effective cross-view information interaction, particularly for textured scene meshes. Recent approaches like Zero123++ (Shi et al., 2023a), Direct2.5 (Lu et al., 2024), Instant3D (Li et al., 2023), MVDream (Shi et al., 2023b), MVDiffusion++ (Tang et al., 2024), CAT3D (Gao et al., 2024), and Wonder3D (Long et al., 2024) have introduced single-pass frameworks for multiview generation, utilizing multiview self-attention to improve viewpoint consistency. Other works, such as SyncDreamer (Liu et al., 2023b), One-2-3-45 (Liu et al., 2024), Cascade-Zero123 (Chen et al., 2023) and ConsistNet (Yang et al., 2024a), incorporate multiview features into 3D volumes to facilitate 3D-aware diffusion models (Liu et al., 2023a; Watson et al., 2022). Meanwhile, techniques such as Pose-Guided Diffusion (Tseng et al., 2023), Era3D (Li et al., 2024b), Epidiff (Huang et al., 2024), and SPAD (Kant et al., 2024) have integrated epipolar-based features to facilitate enhanced viewpoint fusion within diffusion models. Finally, approaches like V3D (Chen et al., 2024b), IM-3D (Melas-Kyriazi et al., 2024), SV3D (Voleti et al., 2024) and Vivid-1-to-3 (Kwak et al., 2024) leverage priors from video diffusion models to achieve multiview generation with improved consistency. However, these methods focus on generating static 3D objects or scenes, while our work introduces vivid object motion into multiview dynamic video generation in complex scenes.

2.3 4D Generation

Recent efforts in 4D generation have explored various methods (Singer et al., 2023; Zhao et al., 2023; Bahmani et al., 2023; Zheng et al., 2023; Ling et al., 2023a) that use score distillation from video diffusion models to optimize dynamic NeRFs or 3D Gaussians for text- or image-conditioned scenes. Follow-up works (Jiang et al., 2023; Ren et al., 2023; Yin et al., 2023; Ren et al., 2024; Zeng et al., 2024; Pan et al., 2024) investigate video-to-4D generation, enabling controllable 4D scene generation from monocular videos. More recent methods (Liang et al., 2024; Xie et al., 2024; Zhang et al., 2024) utilize video diffusion models to address the spatial-temporal consistency required for efficient 4D generation. However, these approaches primarily focus on object-centric generation and face challenges in producing realistic results with complex backgrounds. In contrast, our work emphasizes generating multi-view, 3D-consistent videos for complex scenes.

3 Method

3.1 Overview

Image-to-video generation takes a single image I_0 as input and outputs a video sequence $O_1, \cdots, O_n$. By introducing camera control, the model additionally takes in a sequence of camera information $C_1, \cdots, C_n$, which dictates the desired viewpoint changes for the output sequence. In the multi-view scenario, we extend each batch of the camera control signal and output video sequence to V sequences. In the following paragraphs, we present our proposed **Cavia** framework in detail. First, we outline the preliminaries of image-to-video diffusion and describe how camera controllability is introduced in monocular video generation. Then, we elaborate on the model design for multi-view consistent video generation. An overview of our framework is provided in Fig. 1.

3.2 Camera Controllable Video Diffusion Model

Preliminaries Our model builds on pre-trained Stable Video Diffusion (SVD) (Stability, 2023). SVD extends Stable Diffusion 2.1(Rombach et al., 2022) by adding temporal convolution and attention layers, following the VideoLDM architecture (Blattmann et al., 2023). SVD is trained with a continuous-time noise scheduler (Karras et al., 2022). In each iteration, the training data is perturbed by Gaussian noise $\mathbf{n}(t) \sim \mathcal{N}(0, \sigma^2(t)\mathbf{I})$ and the diffusion model is tasked with estimating the clean data $x_0 \sim p_0$. Let $p(\mathbf{x}; \sigma(t))$ denote the marginal probability of noisy data $\mathbf{x}_t = \mathbf{x}_0 + \mathbf{n}(t)$, the iterative refinement process of diffusion model corresponds to the probability flow ordinary

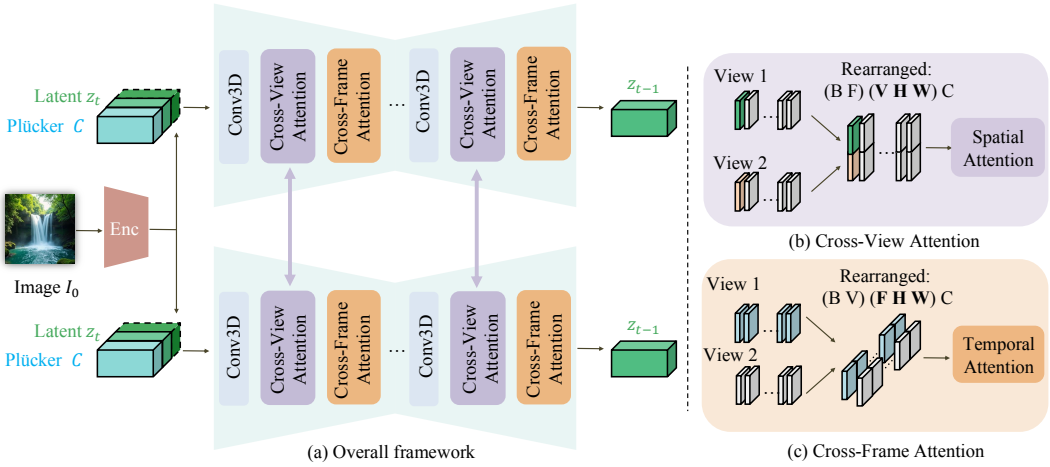


Figure 1: An overview of Cavia is shown in (a). We introduce view-integrated attention modules, namely cross-view attentions and cross-frame attentions, which enforce viewpoint and temporal consistency of the generated frames, respectively. As illustrated in (b) and (c), our view-integrated attention incorporates additional feature dimensions into the attention mechanism, enhancing consistency across views and frames.

differential equation (ODE):

$$d\mathbf{x} = -\dot{\sigma}(t)\sigma(t)\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t))dt. \tag{1}$$

$\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t))$ refers to the score function, which is parameterized by a denoiser $D_{\boldsymbol{\theta}}$ through $\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma) \approx (D_{\boldsymbol{\theta}}(\mathbf{x}; \sigma) - \mathbf{x}) / \sigma^2$. We follow the EDM-preconditioning framework (Karras et al., 2022; Stability, 2023) and parameterize $D_{\boldsymbol{\theta}}$ with a neural network $F_{\boldsymbol{\theta}}$ as follows,

$$D_{\boldsymbol{\theta}} = c_{\text{skip}}\mathbf{x} + c_{\text{out}}F_{\boldsymbol{\theta}}(c_{\text{in}}\mathbf{x}; c_{\text{noise}}). \tag{2}$$

During training, the network $F_{\boldsymbol{\theta}}$ is optimized using denoising score matching for $D_{\boldsymbol{\theta}}$:

$$\mathbb{E}\left[\|D_{\boldsymbol{\theta}}(\mathbf{x}_0 + \mathbf{n}; \sigma, \text{cond}) - \mathbf{x}_0\|_2^2\right]. \tag{3}$$

Camera Conditioning Although SVD is pre-trained on various high-quality video and image data, it does not natively support precise camera control instructions directly. To address this, we introduce camera conditioning to the model via Plücker coordinates (Jia, 2020), which is widely adopted as position embeddings in 360° unbounded light fields(Sitzmann et al., 2021). Plücker coordinates are defined as $P = (d', o \times d')$, where $\times$ is the cross product and d' refers to the normalized ray direction $d' = \frac{d}{||d||}$. Let camera extrinsic matrix be $E = [\mathbf{R}|\mathbf{T}]$ and intrinsic matrix be $\mathbf{K}$, the ray direction $d_{x,y}$ for 2D pixel located at (x, y) is formulated as $d = \mathbf{R}\mathbf{K}^{-1}\left(\begin{smallmatrix} x \\ y \\ 1 \end{smallmatrix}\right) + \mathbf{T}$. These spatial Plücker coordinates are concatenated channel-wise with the original latent inputs of SVD. We enlarge the convolution kernel of the first layer accordingly. The newly introduced matrices are zero-initialized to ensure training stability.

We utilize a relative camera coordinate system, where the first frame is positioned at the world origin with an identity matrix for rotation. The following frames are rotated accordingly. To stabilize training, we normalize the scale of the training sequences to a unit scale. This is implemented by resizing the maximum distance-to-origin in the multi-view camera sequence to 1.

Cross-frame Attention for Temporal Consistency Vanilla 1D temporal attention in the SVD backbone is insufficient for modeling large pixel displacements when the viewpoint changes (Shi et al., 2023b; Yang et al., 2024b). In vanilla 1D temporal attention, attention matrices are calculated over the frame number dimension, and latent features only interact with features from the same spatial location across frames. This limits information flow between different spatial-temporal locations. While this might not be a big issue for video generation with limited motion, viewpoint changes

typically cause significant pixel displacements, which calls for better architecture for more efficient information propagation.

To overcome this issue, we inflate the original 1D temporal attention modules in the SVD network into 3D cross-frame temporal attention modules, allowing for joint modeling of spatial-temporal feature coherence. The inflation operation can be achieved by rearranging the latent features before the attention matrix calculations. Consider the latent features of shape `(B V F C H W)` where `B`, `V`, `F`, `C` refer to the batch size, the number of views, the length of frames, and the feature dimension, respectively, instead of employing 1D attention mechanism on rearranged shape of `((B V H W) F C)`, our inflated attention operates on the rearranged shape `((B V) (F H W) C)`, integrating spatial features into the attention matrices. A visualization is provided in Fig. 1(c).

Since our rearrange operation only alters the sequence length of the attention inputs without modifying the feature dimensions, we can seamlessly inherit the pre-trained weights from the SVD backbone for our purpose. Thanks to this rearrange operation, our inflated temporal attention now calculates the similarity of spatial-temporal features simultaneously, accommodating larger pixel displacements while maintaining temporal consistency.

3.3 Consistent Multi-view Video Diffusion Model

Adding Plücker coordinates for camera control and introducing improved temporal attention allows the video diffusion model to generate reasonably consistent monocular videos. However, for multi-view generation, a monocular video diffusion model that generates samples independently cannot ensure view consistency across multiple sequences. To address this, we introduce novel design mechanisms and training strategies to extend the monocular video diffusion model to the multi-view generation task.

Cross-view Attention for Multi-view consistency To improve cross-view consistency in multi-view videos, we aim to encourage information exchange during the generation process. Since our temporal cross-frame attention modules already handle intra-view feature connections within each video sequence, we focus on exchanging inter-view signals through the spatial cross-view modules. Inspired by MVDream (Shi et al., 2023b), we introduce 3D cross-view attention modules, inflated from the spatial attention blocks of SVD (Stability, 2023). Specifically, we rearrange the `V` views such that frames at each corresponding timesteps are concatenated before being sent into the attention modules. In detail, we rearrange the latent features from shape `(B V F C H W)` to `(((B F) (V H W) C)` instead of `(((B V F) (H W) C)`. A visualization is provided in Fig. 1(b).

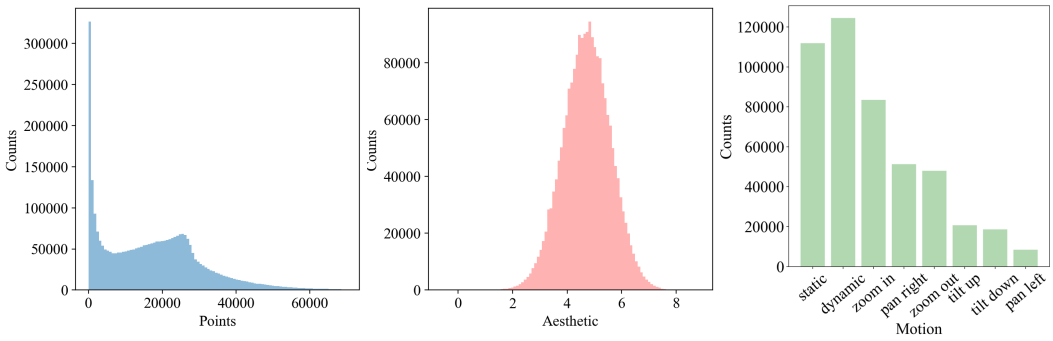
Since only the second-to-last dimension, representing token length, is extended while other dimensions remain unchanged, our inflated spatial attention can inherit the model weights from the monocular setting. This flexibility allows our model to leverage training data with varying numbers of views and facilitates extrapolation to additional views at inference. To handle multi-view generation, we introduce an additional view dimension to the input data. To maintain workflow simplicity, we absorb the view dimension into the batch dimension during processing of other blocks, ensuring flexibility in handling different numbers of views.

4 Joint Training Strategy on Curated Data Mixtures

Thanks to the view-integrated attention mechanism, which allows for inheriting the module weights with arbitrary viewpoint (V) numbers, our framework can leverage various data sources, including static, multi-view dynamic, and monocular videos. This is hard to achieve in previous methods. In this section, we first illustrate our joint training strategy, followed by details on the curated data mixtures that enable this strategy.

4.1 Joint Training Strategy for Videos with Varying Views

For videos capturing static scenes (Zhou et al., 2018; Yu et al., 2023; Xia et al., 2024; Reizenstein et al., 2021; Deitke et al., 2023b;a), we consider all frames to be temporally synchronized. An arbitrary subsequence of length $(F-1) \times V + 1$ from the original video can be reformatted into a V-view sequence with a shared starting frame and F total frames per view. Static scenes also allow



(a) Point Cloud Size (b) Aesthetic Score (c) Camera Motion Classification

Figure 2: Statistics of our curated monocular video dataset. We provide visualizations of (a) point cloud size, (b) aesthetic score, and (c) camera motion classification result.

frame order reversal, providing additional augmentation opportunities. To overcome the static issue where camera-controlled video models are only able to generate static scenes as in (Wang et al., 2023c; He et al., 2024), we further prepare multi-view dynamic videos by rendering animatable objects from Objaverse (Liang et al., 2024; Jiang et al., 2024). We design random smooth trajectories with diverse elevation and azimuth changes to avoid overfitting on simple camera movements. For each set of renderings, we assign a shared random forward-facing starting point for all V views.

To avoid the model overfitting on synthetic images with simple backgrounds, we include a portion of data from monocular in-the-wild videos (Wang et al., 2023b; Nan et al., 2024). Training multi-view camera control from monocular videos is extremely challenging. Although CamCo (Xu et al., 2024) and 4DiM (Watson et al., 2024) have explored joint training for monocular camera-controllable video generation, these approaches are unsuitable for multi-view scenarios. The concurrent work CVD (Kuang et al., 2024) explored homography warping to augment the monocular videos into pseudo-multi-view videos, but the limited realism of these augmentations restricts their ability to generate complex camera and object motion.

To overcome these issues, we choose to jointly train our model on monocular and multi-view videos to effectively utilize the abundant object motion information from all data sources. We annotate the monocular videos with camera poses using Particle-SfM (Zhao et al., 2022). Since in-the-wild monocular videos often contain noisy or unnatural content, we apply a rigorous filtering pipeline to remove unsuitable clips. These curated video clips, sourced from InternVid (Wang et al., 2023b) and OpenVid (Nan et al., 2024) datasets, provide rich object motion as well as complex backgrounds that mitigate the gap between scene-level static data and object-level dynamic data. We rearrange monocular videos as $V = 1$ samples so that all data items can be consumed by the model in the same way without bells and whistles. Thanks to our view-integrated attention modules, which accommodate varying token lengths introduced by the varying view numbers V, the training process remains unaffected when our data items contain both monocular and multi-view videos.

4.2 Data Curation Workflows

We begin by training our model extensively on static video data sourced from various publicly available datasets. Wild-RGBD (Xia et al., 2024) includes nearly 20,000 RGB-D videos across 46 common object categories. MVImgNet (Yu et al., 2023) comprises 219,188 videos featuring objects from 238 classes. DL3DV-10K (Ling et al., 2023b) provides 7,000 long-duration videos captured in both indoor and outdoor environments. CO3Dv2 (Reizenstein et al., 2021) contains 34,000 turntable-like videos of rigid objects, crowd-sourced by nonexperts using cellphone cameras. Objaverse (Deitke et al., 2023b) and Objaverse-XL (Deitke et al., 2023a) exhaustively crawl 10 million publicly available 3D assets. From these, we filtered out low-quality assets, such as those with incorrect textures or overly simplistic geometry, yielding a high-quality subset of 400,000 assets.

Similar to Diffusion4D (Liang et al., 2024) and Animate3D (Jiang et al., 2024), we filter the animatable objects from Objaverse's Sketchfab subset. We exclude objects with excessive motion, which might result in partial observations, as well as nearly static objects with minimal motion. This curation process helps us obtain 19,000 high-quality dynamic assets that can be rendered from

arbitrary viewpoints and timesteps, facilitating multi-view video generation. During each training iteration, we augment the frames with randomly selected background colors.

To improve the model's ability to generate object motion in the presence of complex backgrounds, we prepare monocular videos with camera pose annotations similar to CamCo (Xu et al., 2024). First, we use Particle-SfM (Zhao et al., 2022) to estimate the camera poses for randomly sampled frames from videos from InternVid (Wang et al., 2023b) and OpenVid (Nan et al., 2024). Inspired by CO3D (Reizenstein et al., 2021) and CamCo (Xu et al., 2024), we remove the videos where SfM fails to register all available frames or produces a point cloud with too few points or too many points. Fig. 2(a) shows the point count statistics. A point cloud with too few points indicates poor frame registration to a shared 3D representation, while too many points suggest a mostly static scene, which is undesirable as we focus on object motion. Additionally, non-registered frames may indicate potential scene changes. We then apply a rigorous filtering pipeline to ensure the quality of the video samples used for training. This includes filtering based on aesthetic scores, optical character recognition (OCR), and camera motion classification using optical flow. Videos containing detected character regions are aggressively removed. Fig. 2(b) and (c) present statistics on aesthetic score and camera motion classification results. Videos with low aesthetic scores or those classified as having static camera motion are excluded from the training set. Ultimately, we construct a dataset of 393,000 monocular videos annotated with camera poses. We provide a summary of the data sources used in Fig. 3. More details and analysis are provided in the appendix.

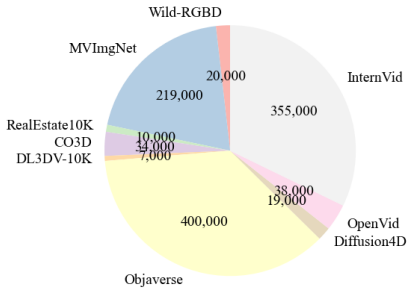


Figure 3: Sources of our training videos. We visualize the numbers of raw videos from each source. During training, we perform augmentations to the raw video sequences to avoid overfitting on certain camera motions.

5 Experiments

In this section, we present experimental results and analysis. Video comparisons are included in the supplementary material for optimal visual evaluation. It is important to note that for all qualitative and quantitative evaluations, neither the input images nor the camera trajectories were used during model training.

5.1 Quantitative Comparisons

3D Consistency of Frames We evaluate the 3D consistency of the generated videos using COLMAP (Schönberger & Frahm, 2016; Schönberger et al., 2016). COLMAP is widely adopted for 3D reconstruction methods where camera pose estimation is required for in-the-wild images. We configure the COLMAP following previous methods (Deng et al., 2022; Xu et al., 2024) for best few-view performance. A higher COLMAP error rate indicates poorer 3D consistency in the input images. Motivated by this, we report COLMAP errors as a measure of the 3D consistency of the frames. Each video is retried up to five times to reduce randomness. We randomly sample 1,000 video sequences from RealEstate10K (Zhou et al., 2018) test set for evaluation. Since we have ground truth 3D scenes, we use the ground truth camera pose sequences as the viewpoint instruction of the video model and compare the generated frames against the ground truth images. Similar to prior works (He et al., 2024; Xu et al., 2024), we extract the estimated camera poses and calculate the relative translation and rotation differences. Specifically, given two camera pose sequences, we convert them to relative poses and align the first frames to world origin. We then measure the angular errors in translation and rotation. Unlike previous works (He et al., 2024; Xu et al., 2024) that calculate the Euclidean distance of translation vectors, we use angular error measurements to ensure the camera pose scales are normalized, addressing scale ambiguity. As shown in Tab. 1, we calculate the area under the cumulative error curve (AUC) of frames whose rotation and translations are below certain thresholds (5°, 10°, 20°). Our method significantly outperforms existing baselines.

Figure 4: Per-video qualitative comparisons. The first frame in each reference set is the input image. Neither the image nor the camera trajectories were seen during model training. Video results are provided in supplementary for clearer qualitative comparisons.

Table 1: Quantitative comparison for monocular geometry consistency on RealEstate10K test set.

Methods	FID ↓	FVD ↓	COLMAP error↓	Rot. AUC ↑ (@5°/10°/20°)	Trans. AUC ↑ (@5°/10°/20°)
SVD	16.89	139.64	30.3%	14.4 / 22.8 / 35.3	0.2 / 1.0 / 3.2
MotionCtrl	21.09	119.06	55.0%	8.6 / 13.9 / 22.2	0.6 / 2.1 / 5.7
CameraCtrl	14.69	105.41	19.3%	21.4 / 32.9 / 48.4	0.3 / 1.3 / 4.4
Ours	**11.43**	**55.10**	**14.4%**	**22.9 / 34.5 / 50.1**	**5.1 / 12.7 / 24.6**

Multi-view Consistency Alongside evaluating the individual monocular frame pose accuracy using COLMAP-based metrics, we further assess the cross-video consistency of the corresponding frames from generated multi-view videos. We randomly sample 1,000 videos, each with 27 frames, from RealEstate10k (Zhou et al., 2018) test set and convert each video into a two-view sequence with 14 frames per view. The new camera pose sequences are generated by setting the 14th frame as the world origin and positioning the remaining frames relative to it. The scales of the scenes are normalized

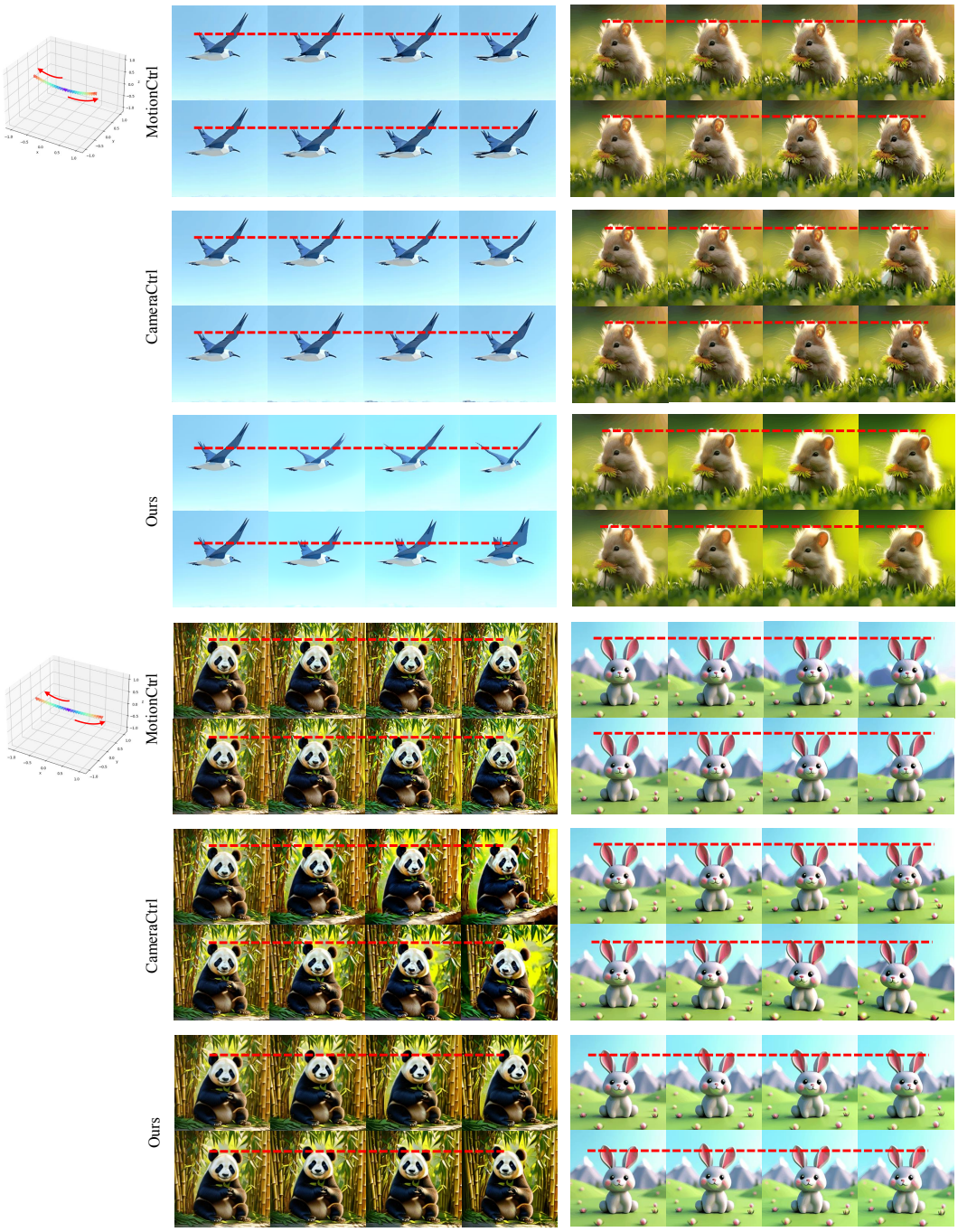


Figure 5: Qualitative comparisons for 2-view video generations. Each generation consists of two rows, where each row represents a sequence of generated frames, with columns showing frames at the same timestep. Neither the image nor the camera trajectories were used during model training. Red dotted lines are annotated to highlight object motion. Video results are included in the supplementary material for clearer comparisons.

so that the maximum distance from the origin is 1. Following CVD (Kuang et al., 2024), we adopt SuperGlue (Sarlin et al., 2020) to find correspondences and estimate the camera poses between each time-aligned set of frames. SuperGlue not only measures angular errors in the rotation and translation but also computes the epipolar error of the matched correspondences. We similarly collect the AUC for frame pairs with rotation and translation errors below specific thresholds (5°, 10°, 20°). The

Table 2: Quantitative comparison for 2-view video generation.

Scenes	Methods	FID ↓	FVD ↓	Rot. AUC ↑ (@5°/10°/20°)	Trans. AUC ↑ (@5°/10°/20°)	Prec. ↑	MS. ↑
Real10K	SVD	37.99	296.95	7.9 / 13.5 / 28.2	0.2 / 0.7 / 2.4	6.49	4.17
	MotionCtrl	29.23	277.05	8.1 / 16.5 / 29.4	1.5 / 5.3 / 16.1	11.45	5.90
	CameraCtrl	12.57	131.32	22.4 / **38.5** / **56.2**	0.6 / 2.5 / 8.2	19.49	11.25
	Ours	**8.82**	**94.86**	**23.9** / 37.4 / 52.9	**3.3** / **10.2** / **23.5**	**29.39**	**15.22**
General	MotionCtrl	47.31	313.92	4.9 / 11.3 / 21.9	0.7 / 2.4 / 8.2	8.12	3.93
	CameraCtrl	26.71	221.23	14.1 / 26.9 / 43.2	0.5 / 1.7 / 5.7	15.13	7.35
	Ours	**26.12**	**173.70**	**19.7** / **32.7** / **48.4**	**0.8** / **2.8** / **8.7**	**33.10**	**19.96**

epipolar errors for the estimated correspondences are summarized to the precision (P) and matching score (MS). As shown in Tab. 2, our method outperforms baselines greatly. The "Real10K" category means that the input images are taken from the corresponding RealEstate10K test sequence, while the "General" means that the input images are taken from 1,000 randomly sampled images in the test split of our monocular video dataset.

Visual Quality To assess the frame perceptual quality, we evaluate visual quality using FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018). FID and FVD measure the feature-space similarity of two sets of images and videos, respectively. In our case, they quantify the distribution distance between the generated frame sequences and the ground-truth frames. We provide monocular evaluations in Tab. 1 and multi-view evaluations in Tab. 2. As shown in these tables, our proposed framework enjoys the best visual quality. For both the "Real10K" and "General" categories, the ground-truth videos used to calculate these metrics are the video sequences corresponding to the input frames. These video sequences are from the test set split of the datasets and are not seen during training.

5.2 Qualtitative Comparison

We provide qualitative comparisons on RealEstate10k (Zhou et al., 2018) scenes in Fig. 4 and text-to-image generated images in Fig. 5. As shown in Fig. 4, our method produces videos with precise camera control, whereas MotionCtrl tends to generate overly smooth trajectories that simplify the viewpoint instructions, and CameraCtrl suffers from severe distortions at novel viewpoints. For example, in the first case, the camera instruction involves multiple panning operations, first panning left and then panning right. Still, MotionCtrl only pans left, ignoring the rest of the instructions. CameraCtrl's outputs, particularly in the first two cases, exhibit noticeable distortion, with the walls bending in the later frames. Additionally, in the third and fourth cases, where the camera trajectories cover a long distance, both MotionCtrl and CameraCtrl produce unrealistic hallucinations, introducing artifacts such as merging indoor and outdoor pixels or distorting input pixels to compensate for a lack of generation ability. In Fig. 5, we observe that MotionCtrl and CameraCtrl tend to generate static scenes without any object motion. Although their methods produce realistic novel views, the synthesized objects remain static. In contrast, our method generates vivid object motion while maintaining accurate camera control. We highlight the object motion in Fig. 5 using auxiliary red lines. We encourage readers to view the supplementary videos for optimal visual comparisons.

5.3 Ablation Studies and Applications

Due to the space limit, we refer readers to the Appendix for ablation studies and applications of our framework. We provide detailed ablation studies in Sec. D on our proposed framework. Additionally, we explore the 3D reconstruction of our generated frames and four-view generation capabilities in Sec. E. Videos are included in the supplementary material for optimal qualitative comparison.

6 Conclusion

In this paper, we propose Cavia, a novel framework for consistent multi-view camera-controllable video generation. Our framework incorporates cross-frame and cross-view attentions for effective camera controllability and view consistency. Our model benefits from joint training using static 3D scenes and objects, animatable objects, and in-the-wild monocular videos. Extensive experiments demonstrate the superiority of our method over previous works in terms of geometric consistency and perceptual quality.

REFERENCES

Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. *arXiv preprint arXiv:2311.17984*, 2023.

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1728–1738, 2021.

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*, pp. 22563–22575, 2023.

David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. *arXiv preprint arXiv:2312.12337*, 2023.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13320–13331, 2024a.

Yabo Chen, Jiemin Fang, Yuyang Huang, Taoran Yi, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Cascade-zero123: One image to highly consistent 3d with self-prompted nearby views. *arXiv preprint arXiv:2312.04424*, 2023.

Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024b.

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. *arXiv preprint arXiv:2307.05663*, 2023a.

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*, pp. 13142–13153, 2023b.

Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12882–12891, 2022.

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.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

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.

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.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022b.

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

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.

Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9784–9794, 2024.

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

Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. *arXiv preprint arXiv:2311.02848*, 2023.

Yanqin Jiang, Chaohui Yu, Chenjie Cao, Fan Wang, Weiming Hu, and Jin Gao. Animate3d: Animating any 3d model with multi-view video diffusion. *arXiv preprint arXiv:2407.11398*, 2024.

Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, and Igor Gilitschenski. Spad: Spatially aware multi-view diffusers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10026–10038, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023.

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.

Jeong-gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, and Kwang Moo Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6775–6785, 2024.

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*, 2024a.

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. *https://arxiv.org/abs/2311.06214*, 2023.

Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *arXiv preprint arXiv:2405.11616*, 2024b.

Ruining Li, Chuanxia Zheng, Christian Rupprecht, and Andrea Vedaldi. Puppet-master: Scaling interactive video generation as a motion prior for part-level dynamics. *arXiv preprint arXiv:2408.04631*, 2024c.

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.

Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. *arXiv preprint arXiv:2312.13763*, 2023a.

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. *arXiv preprint arXiv:2312.16256*, 2023b.

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024.

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. *arXiv preprint arXiv:2303.11328*, 2023a.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023b.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9970–9980, 2024.

Yuanxun Lu, Jingyang Zhang, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, Xun Cao, and Yao Yao. Direct2. 5: Diverse text-to-3d generation via multi-view 2.5 d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8744–8753, 2024.

Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, Natalia Neverova, Andrea Vedaldi, Oran Gafni, and Filippos Kokkinos. Im-3d: Iterative multiview diffusion and reconstruction for high-quality 3d generation. *International Conference on Machine Learning, 2024*, 2024.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

OpenAI. Sora — openai. `https://openai.com/index/sora/`. (Accessed on 09/22/2024).

Zijie Pan, Zeyu Yang, Xiatian Zhu, and Li Zhang. Fast dynamic 3d object generation from a single-view video. *arXiv preprint arXiv:2401.08742*, 2024.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10901–10911, 2021.

Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023.

Jiawei Ren, Kevin Xie, Ashkan Mirzaei, Hanxue Liang, Xiaohui Zeng, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, et al. L4gm: Large 4d gaussian reconstruction model. *arXiv preprint arXiv:2406.10324*, 2024.

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*, pp. 10684–10695, 2022.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4938–4947, 2020.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

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*, 2023a.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023b.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023.

Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 34:19313–19325, 2021.

Stability. Stable video diffusion: Scaling latent video diffusion models to large datasets. `https://stability.ai/research/stable-video-diffusion-scaling-latent-video-diffusion-models-to-large-datasets`, 2023.

Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv*, 2023.

Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. *arXiv preprint arXiv:2402.12712*, 2024.

Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *CVPR*, 2023.

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.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024.

Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. 2023a.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023b.

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*, 2023c.

Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022.

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.

Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. *arXiv preprint arXiv:2401.12592*, 2024.

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.

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.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7079–7088, 2024a.

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*, 2024b.

Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225*, 2023.

Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9150–9161, 2023.

Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. Stag4d: Spatial-temporal anchored generative 4d gaussians. *arXiv preprint arXiv:2403.14939*, 2024.

Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. 4diffusion: Multi-view video diffusion model for 4d generation. *arXiv preprint arXiv:2405.20674*, 2024.

Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In *European Conference on Computer Vision*, pp. 523–542. Springer, 2022.

Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023.

Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. A unified approach for text-and image-guided 4d scene generation. *arXiv preprint arXiv:2311.16854*, 2023.

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.

A Additional Implementation Details

Our training is divided into static stage and dynamic stage. Our static stage is trained for around 500k iterations and our dynamic stage is trained for roughly 300k iterations. The effective batch size is 128 and the learning rate is 1e-4. Our video length is 14 frames for each view with the first frame shared across views. Our model is fine-tuned at 256×256 spatial resolution from the SVD 1.0 checkpoint. The training data are prepared by first center-cropping the original videos and then resizing each frame to the shape of 256×256. In the dynamic stage, 30% of iterations are used to train on monocular videos. During static training, the strides of frames are randomly sampled in the range of `[1, 8]`. For monocular videos, the strides are sampled in the range of `[1, 2]`. For dynamic multi-view object renderings, the strides are fixed to 1 to use all rendered frames since we already introduced randomness in the frame rate during rendering. At inference time, the decoding chunk is set to 14 so all frames are decoded altogether. We sample 25 steps to obtain all our results.

B Additional Data Curation Details

In this section, we provide additional details on our data processing and curation pipelines.

Static 3D Objects Our static objects data comprises multi-view images rendered from the Objaverse (Deitke et al., 2023b) and Objaverse-XL(Deitke et al., 2023a) dataset. Similar to InstantMesh, we use a filtered high-quality subset of the original dataset to train our model. The filtering goal is to remove objects that satisfy any of the following criteria: (i) objects without texture maps, (ii) objects with rendered images occupying less than 10% of the view from any angle, (iii) including multiple separate objects, (iv) objects with no caption information provided by the Cap3D dataset, and (v) low-quality objects. The classification of "low-quality" objects is determined based on the presence of tags such as "lowpoly" and its variants (e.g., "low poly") in the metadata. By applying our filtering criteria, we curated approximately 400k high-quality instances from the initial pool of 800k objects in the Objaverse dataset.

For each 3D object, we use Blender's EEVEE renderer to render an 84-frame RGBA orbit at 512×512 resolution. we adaptively position the camera to a distance sufficient to ensure that the rendered object content makes good and consistent use of the image extents without being clipped in any view. For each frame, the azimuths can be irregularly spaced, and the elevation can vary per view. Specifically, the sequence of camera elevations for each orbit is obtained from a random weighted combination of sinusoids with different frequencies. The azimuth angles are sampled regularly, and then a small amount of noise is added to make them irregular. The elevation values are smoothed using a simple convolution kernel and then clamped to a maximum elevation of 89 degrees.

Static 3D Scenes Our static scenes data are sourced from RealEstate10k (Zhou et al., 2018), WildRGBD (Xia et al., 2024), MVImgNet (Yu et al., 2023), CO3Dv2 (Reizenstein et al., 2021), and DL3DV-10K (Ling et al., 2023b). For RealEstate10k, we use the train/test split released by PixelSplat (Charatan et al., 2023). During training, we sample every 8 original frames to construct the training sequences. For DL3DV-10K, we construct training sequences from the publicly available 7k subset. Since each video is very long for the DL3DV-10k dataset, we offline randomly sample multiple sequences from a single ground truth video to obtain multiple training data items. For CO3Dv2, we remove the video sequences that contain whole-black images to avoid temporally inconsistent frames. For WildRGBD and MVImgNet we use all classes available and removed sequences whose lengths are not enough for two-view training (shorter than 27 frames).

Dynamic 3D Objects Our dynamic 3D objects are similarly rendered as the static 3D objects. The filtering pipelines remain mostly the same as the static objects, except that we introduce additional workflows to consider object motion. Inspired by previous works (Liang et al., 2024; Jiang et al., 2024; Li et al., 2024a) that employ animatable objects from Objaverse. We render multiple fixed-view videos to examine the motion quality of the objects. We utilize lpips (Zhang et al., 2018) to measure the similarity of nearby frames and consider an object to be static if lpips similarity is above a certain threshold. Additionally, we render the alpha masks of the object and use this as an indicator of whether the object has moved out of the visible regions. Consequently, we remove objects with too large or sudden movements as well as objects with little-to-no motion. These filterings result in

Table 3: Ablation Studies on each of our introduced modules. "w/o Plücker" refers to replacing the Plücker coordinate conditioning with one-dimensional conditioning as in MotionCtrl. "w/o Cross-frame" refers to replacing the Cross-frame attention with vanilla 1D temporal attention. "w/o Cross-view" refers to replacing the Cross-view attention with vanilla spatial attention. "Ours (Static)" means the model is only trained on static video datasets. "Ours (w/o Mono)" means that the model is fine-tuned on synthetic multi-view datasets, but is not trained with monocular video datasets. "Ours (Full)" means that the model is trained on all available data sources.

Scenes	Methods	FID↓	FVD↓	Rot. AUC ↑ (@5°/10°/20°)	Trans. AUC ↑ (@5°/10°/20°)	Prec. ↑	MS. ↑
Real10K	w/o Plücker	12.75	195.84	12.1 / 21.9 / 35.5	1.6 / 5.8 / 16.4	14.74	10.02
	w/o Cross-frame	17.04	154.54	21.4 / 34.8 / 50.1	3.8 / 11.1 / 24.2	25.67	12.70
	w/o Cross-view	9.45	106.82	22.8 / 36.7 / 52.4	2.7 / 8.7 / 22.1	27.57	14.65
	Ours	**8.82**	**94.86**	**23.9 / 37.4 / 52.9**	**3.3 / 10.2 / 23.5**	**29.39**	**15.22**
General	w/o Cross-frame	71.39	249.02	9.8 / 19.1 / 32.7	0.5 / 1.9 / 6.6	13.20	8.97
	w/o Cross-view	30.89	246.68	14.9 / 27.4 / 42.9	1.2 / 4.3 / 12.2	17.58	9.59
	Ours (Static)	27.20	185.58	15.9 / 28.7 / 44.1	1.4 / 4.6 / 12.9	21.75	12.04
	Ours (w/o Mono)	35.79	243.05	15.0 / 27.1 / 42.6	0.3 / 1.3 / 4.2	18.55	10.78
	Ours (Full)	**26.12**	**173.70**	**19.7 / 32.7 / 48.4**	**0.8 / 2.8 / 8.7**	**33.10**	**19.96**

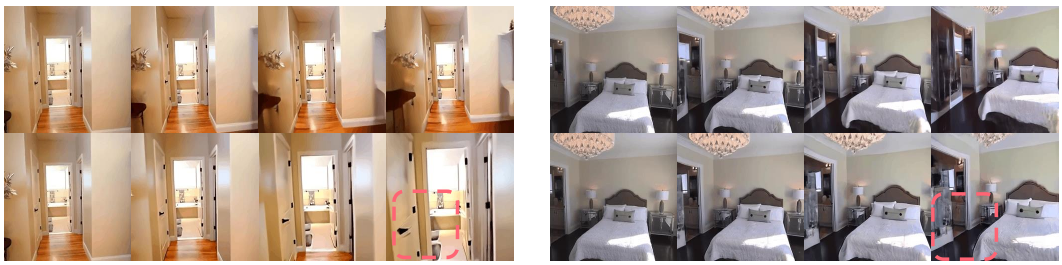
19,000 objects. Our rendering strategy is also very similar to that of static 3D objects, introducing random elevation and azimuth changes to complicate the trajectories, except that we additionally introduce a random frame stride at rendering to augment the object motion. The stride is sampled individually for each object from the range `[1, 3]`. A larger the stride leads to renderings with faster object motion.

Monocular Videos Our monocular video filtering pipeline involves filtering according to Particle-SfM output, OCR, aesthetic score, and camera motion. As mentioned in Sec. 4.2, we first attempt to annotate the camera poses for the video frames using Particle-SfM (Zhao et al., 2022). Take InternVid (Wang et al., 2023b) as an example, roughly 10 million video clips are processed and around 3 million samples are successfully processed by Particle-SfM. For each video, we start from the first frame and randomly select a frame stride of 1 or 2. The total number of images sent to Particle-SfM is 32 images. Our point count filtering is empirically implemented as a cut-off at 1,000 points and 40,000 points. Point clouds with too few points are removed due to the concern that the frames are poorly registered. Point clouds with too many points are avoided because their limited object motion. This aggressive filtering results in around 2 million samples for further processing. We then evaluate all the video clips using OCR detection algorithms and remove the samples whose detected text regions are larger then 10^{-4} of the image resolution (*i.e.* 6 pixels). This process results in 604,000 samples. The next step is filtering with aesthetic scores and videos with aesthetic score annotations smaller than 4 are removed. 467,000 videos are left after these filtering process. Finally, we employ a camera motion classifier extended from the Open-Sora pipeline[1]. The main motivation is that optical-flow on consecutive frames can be summarized to a global motion vector, assuming the most parts of the scene is moving in a uniform direction. Optical flow is first obtained using `cv2.calcOpticalFlowFarneback` for each consecutive frame pairs. Then, the magnitudes and directions are calculated via `cv2.cartToPolar`. These magnitudes and directions are classified into 8 categories: static, zoom out, zoom in, pan left, tilt up, pan right, tilt down, and unknown. The results of the frame pairs are summarized to obtain the final result of each video clip. When a certain type appears more than 50%, the type for the whole video clip is determined directly. We aggressively classify a video clip as static if any of its frame pairs is categorized into static or unknown. Finally, we obtain 355,000 clips that satisfy our needs. The process is similarly applied to OpenVid (Nan et al., 2024)'s Panda-70M subset Chen et al. (2024a) and we obtained 38,000 clips. In summary, our monocular video dataset consists of 393,000 clips.

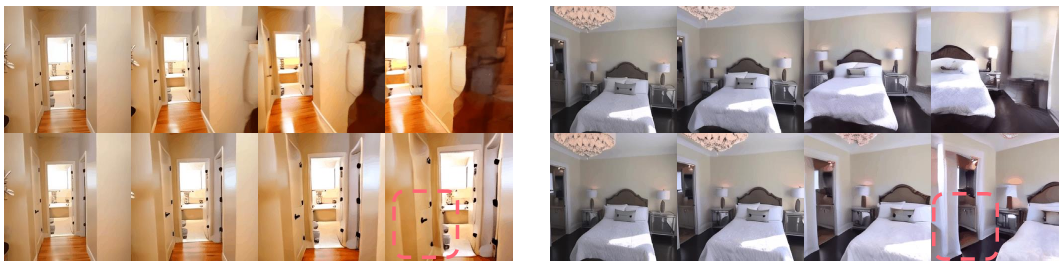
[1] `https://github.com/hpcaitech/Open-Sora/tree/main/tools/caption/camera_motion`

Model	Plücker Coordinates	Cross-frame Attention	Cross-view Attention
w/o Plücker	×	×	×
w/o Cross-frame	✓	×	×
w/o Cross-view	✓	✓	×
Ours	✓	✓	✓

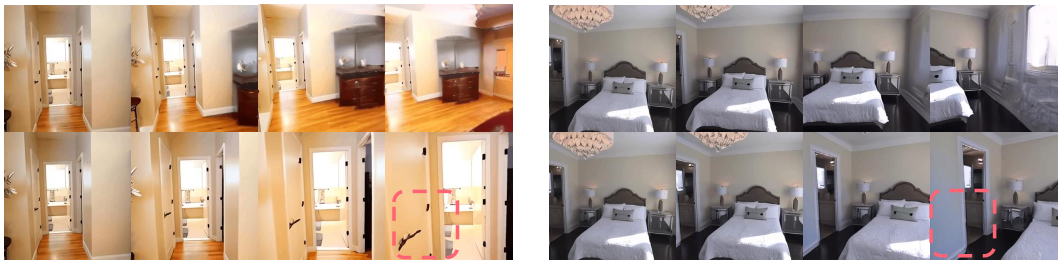
Table 4: Illustration of the model variants in the ablation studies.



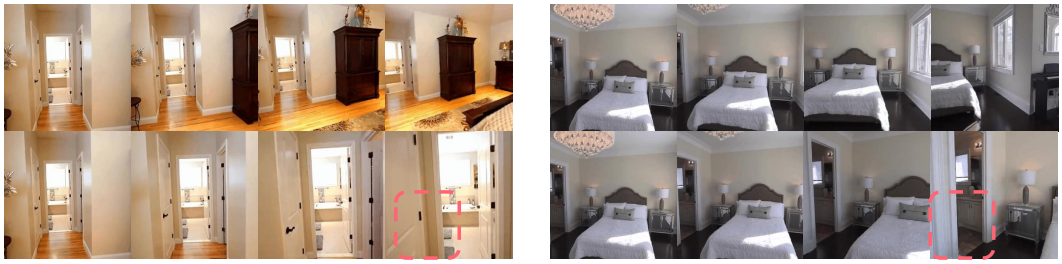
(a) Without Plücker



(b) Without Cross-frame Attention



(c) Full Model

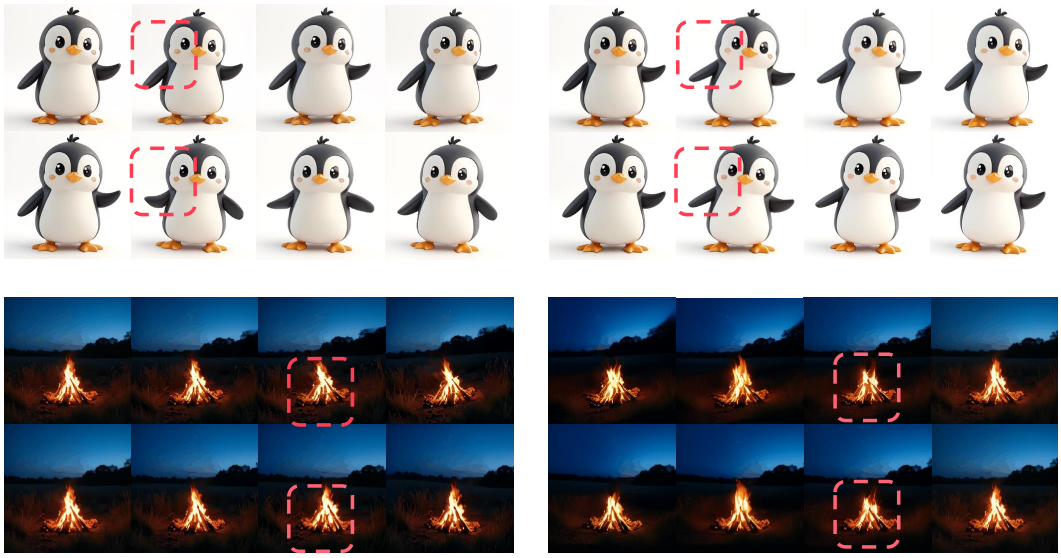


(d) Reference

Figure 6: Ablation studies on Plücker coordinates and Cross-frame Attention. Video results are provided in supplementary for clearer qualitative comparisons.

C Evaluation Details

For MotionCtrl and CameraCtrl, we use the open-source checkpoints trained from SVD released by the authors. These checkpoints are designed for image-to-video tasks so we can have fair compar-



(a) Without Cross-view Attention (b) With Cross-view Attention

Figure 7: Ablation studies on Cross-view Attention. Video results are provided in supplementary for clearer qualitative comparisons.

(a) Without Monocular Joint Training (b) With Monocular Joint Training

Figure 8: Ablation studies on the joint training strategy on monocular videos. Video results are provided in supplementary for clearer qualitative comparisons.

isons. We use "clean-fid"[2] and "common-metrics-on-video-quality"[3] for obtaining FID and FVD, respectively. Our FVD results are reported in VideoGPT (Yan et al., 2021) format. Our COLMAP is configured following DSNeRF (Deng et al., 2022) and (Xu et al., 2024) to improve the few-view reconstruction performance. Concretely speaking, we enable `--SiftMatching.max_num_matches 65536` to support robust feature matching. To ensure that the SfM results best align with our videos, we set `--ImageReader.single_camera 1` since most videos in our datasets consist of frames captured from a single camera.

[2] https://github.com/GaParmar/clean-fid
[3] https://github.com/JunyaoHu/common_metrics_on_video_quality

D ABLATION STUDIES

In this section, we conduct extensive evaluations for ablation studies. We provide video comparisons in the supplementary. We provide thorough quantitative comparisons in Tab. 3 to illustrate the importance of our proposed components. In Tab. 4, we illustrate the differences between model variants used for ablation studies. The models are evaluated using RealEstate10K camera trajectories. For the "Real10K" and "General" categories, the testing images are from our test set split of RealEstate10K and InternVid, respectively. Our full model enjoys the best perceptual quality and geometric consistency.

We first examine the importance of Plücker coordinates conditioning and the cross-frame attention modules. As shown in Fig. 6, model variants without cross-frame attention contains severe distortion artifacts, such as the bent walls. The model variant without Plücker coordinates results in simplified camera motion that ignores the complex camera viewpoint instructions.

We then evaluate the model variant without cross-view attention. As shown in Fig. 7, we observe that removing the cross-view attention module results in multiple individual video samples that contain different object motions. For example, the penguin moves differently in the first case, and the wood sticks in the fire appear differently in the second case. This behavior is not desirable because our goal is to obtain multiple videos from different camera paths of the same scene.

Finally, we examine the importance of our monocular video joint training strategy. As shown in Fig. 8, we observe that when overfitting on dynamic objects from Objaverse, the generated results tend to contain frames with simplified backgrounds. This is mainly because, during the training, all data samples from Objaverse are implemented with single random color backgrounds. Our model benefits from joint training on monocular videos and preserves the ability to generate complex backgrounds when object motion is present.

E APPLICATIONS

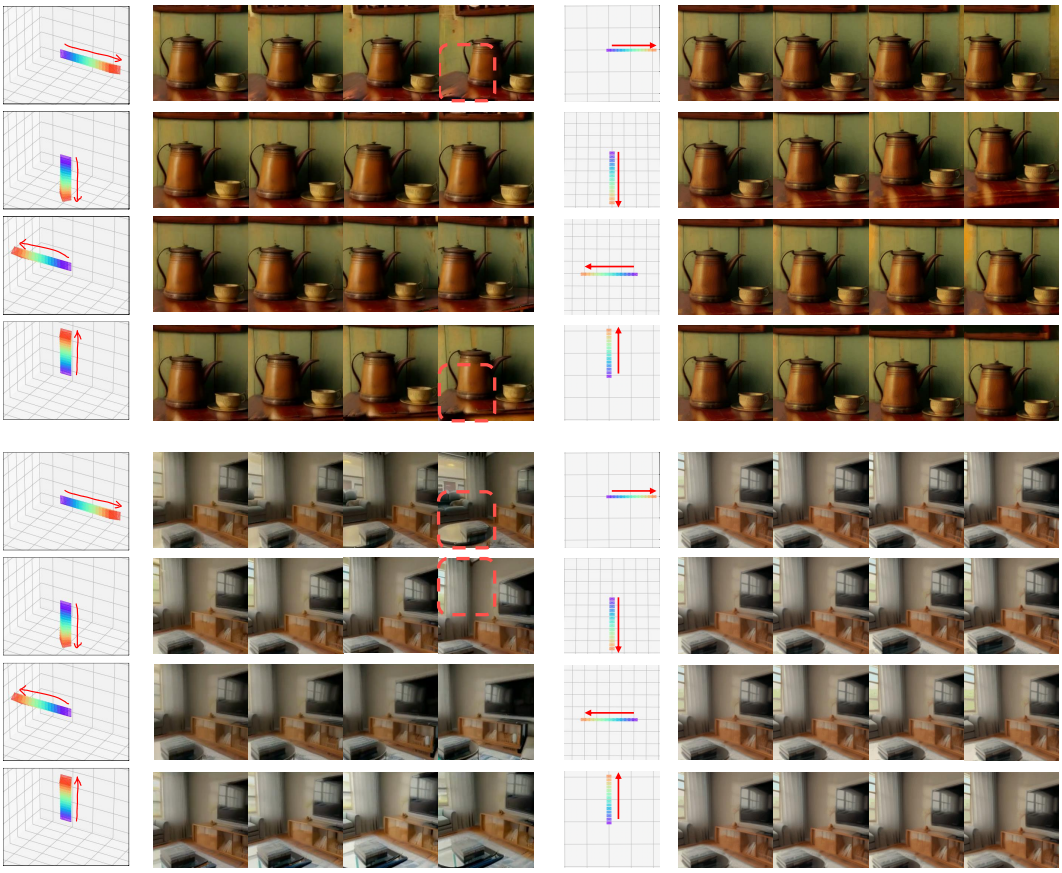
In this section, we provide additional results on four-view inference and 3D reconstruction of our generated frames.

E.1 ADVANCING TO FOUR VIEWS AT INFERENCE

Our cross-view attention design enables us to extrapolate to more views straightforwardly at inference time. This design is more efficient compared with the concurrent work CVD (Kuang et al., 2024) which requires enumeration of viewpoint pairs at inference time. We conduct a side-by-side comparison for 4-view generation in Fig. 9. Our method enjoys better consistency and shows more realistic results than CVD (Kuang et al., 2024). In comparison, CVD tends to produce artifacts at border regions. For example, the structure of the wall (first case) and the window (second case) change when the viewpoint changes. The results from CVD are taken from their author's website. We provide video comparisons in the supplementary. We also provide more 4-view generation results from Cavia in our supplementary.

E.2 3D RECONSTRUCTION OF GENERATED FRAMES

We further perform 3D reconstruction on our generated frames. We render our reconstructed 3D Gaussians from an elliptical trajectory consisting of 16 novel views. We provide a side-by-side comparison with the concurrent work CVD (Kuang et al., 2024) in Fig. 10. Compared with the results of CVD, our frames are more geometrically consistent and result in clearer 3D reconstruction and fewer floaters. For example, the results from CVD produce floaters on the cupboard regions and generate blurry artifacts for the wall and the TV due to inconsistencies. We provide video comparisons in the supplementary for clearer comparisons. We also provide additional 3D reconstruction results of Cavia's generated frames in the supplementary.



(a) CVD (b) Ours

Figure 9: Four-view video comparison. The result of CVD is taken from their website. CVD tends to generate black border pixels, potentially due to its homography warping augmentations during training. In comparison, our method produces frames with better geometric consistency and perceptual quality.

F LIMITATIONS

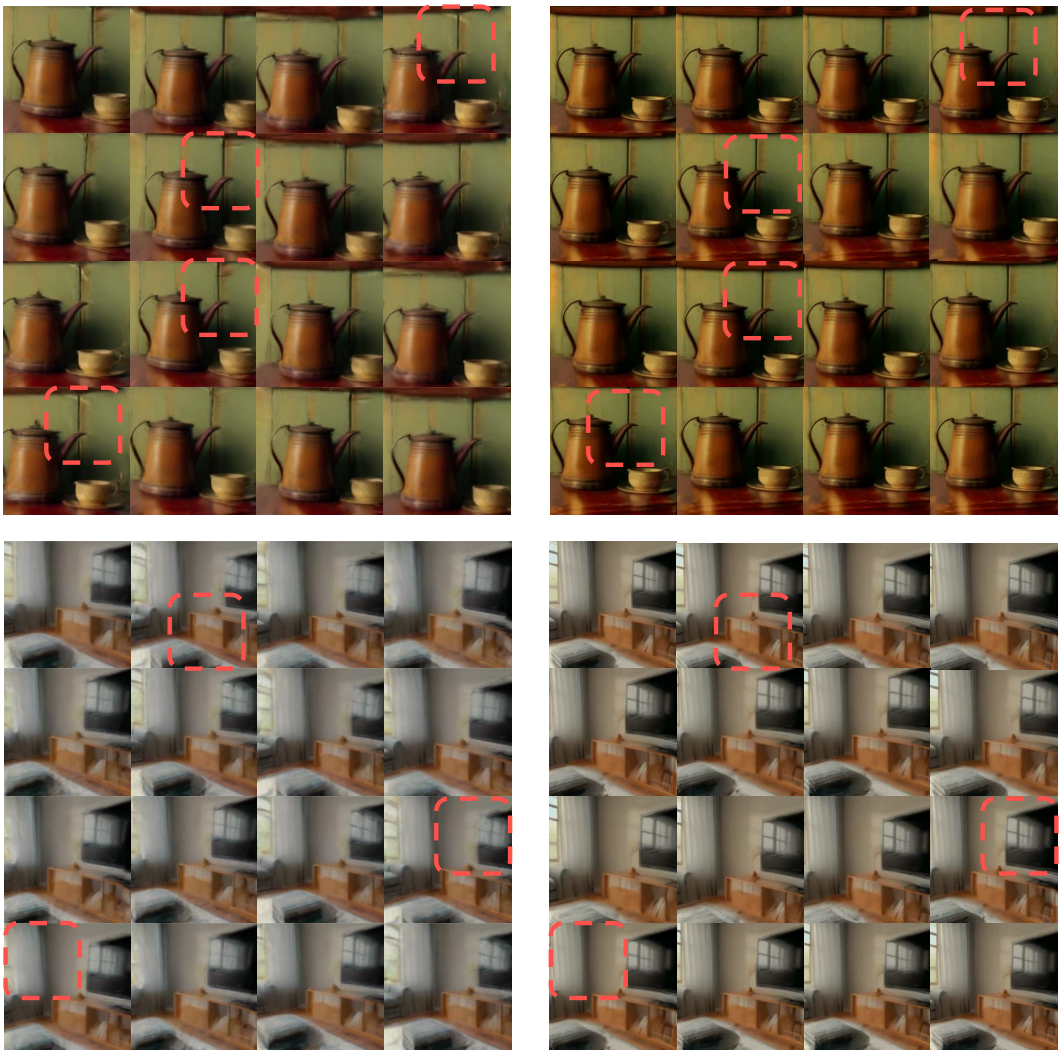
Our framework has limited ability to generate large object motion, mainly due to the limitation of the base video generator SVD (Stability, 2023). We will explore better base models in future works. Moreover, our data curation pipelines assume a simple camera model using shared camera intrinsic across frames. While enabling easier data preparation, this limits our model from generalizing to complex camera intrinsic changes at inference time, which is widely adopted in cinematography. Additionally, for simplicity, our framework is trained with normalized scales of scenes, which can be further improved if potentially calibrated with metric scale. We will explore calibration techniques for better quality if a well-generalizable metric depth estimator becomes publicly available.

G ADDITIONAL DYNAMIC RESULTS

We provide additional dynamic results as suggested by reviewers in Fig. 11. The gif version are submitted with the original supplementary webpage.

H DATA SAMPLES

We provide visualizations for the annotations from Particle-SfM in Fig. 12. The good one is an example of a high-quality pose estimation result, while the bad one is a failure example.



(a) CVD (b) Ours

Figure 10: 3D Reconstruction comparison. We render the reconstructed 3D Gaussians from an elliptical trajectory consisting of 16 novel views. The result of CVD is taken from their website. CVD's reconstruction results suffer from floaters and blurry artifacts due to the inconsistency in their generated frames. In comparison, our method produces sharper results with clearer visual quality.

We also provide examples of each data source in Fig. 13. Thanks to our joint training strategy, our model supports data samples with arbitrary viewpoint numbers. As a result, our model is jointly trained on static scenes, dynamic objects, and monocular videos.

I MOTION COMPARISON WITH SVD

As suggested by the reviewers, we provide side-by-side comparisons with our base model (SVD). As shown in Fig. 14, the frames generated by our method exhibit rich object motion that is comparable to, if not better than, that of SVD. For example, our penguin, giraffe, and bird examples are behaving realistically, while SVD presents limited object motion.

Figure 11: We provide additional dynamic results as suggested by reviewers. The gif version are submitted with the original supplementary webpage.

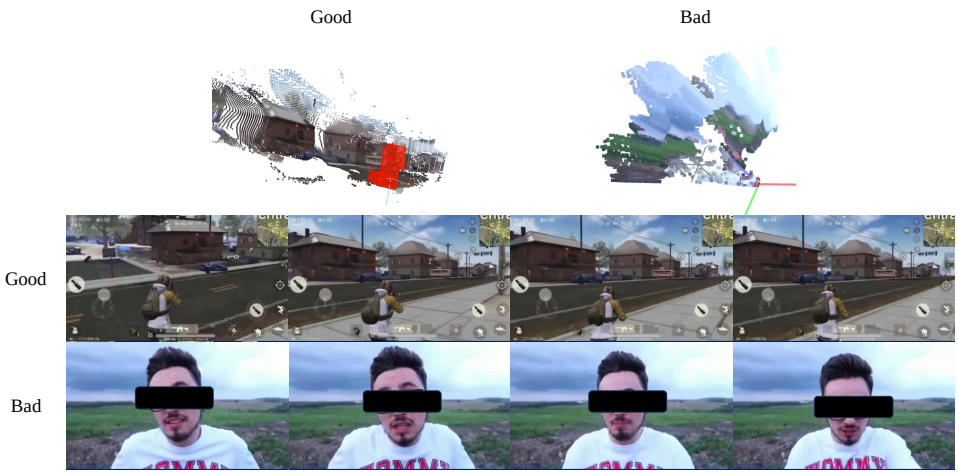


Figure 12: Examples of the annotations produced by Particle-SfM. We add black boxes to the original frames to avoid revealing the identity.

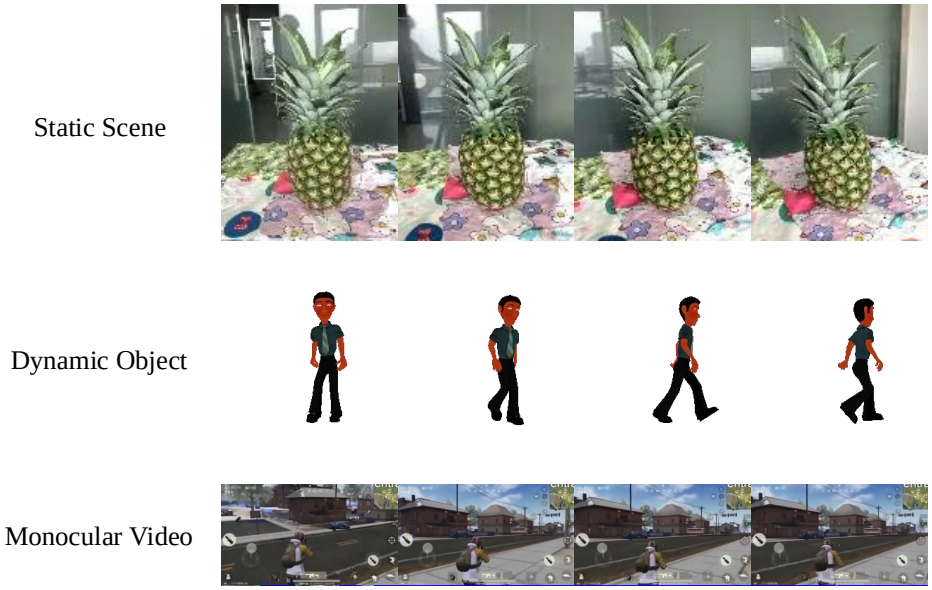


Figure 13: Examples of our data samples from various sources.

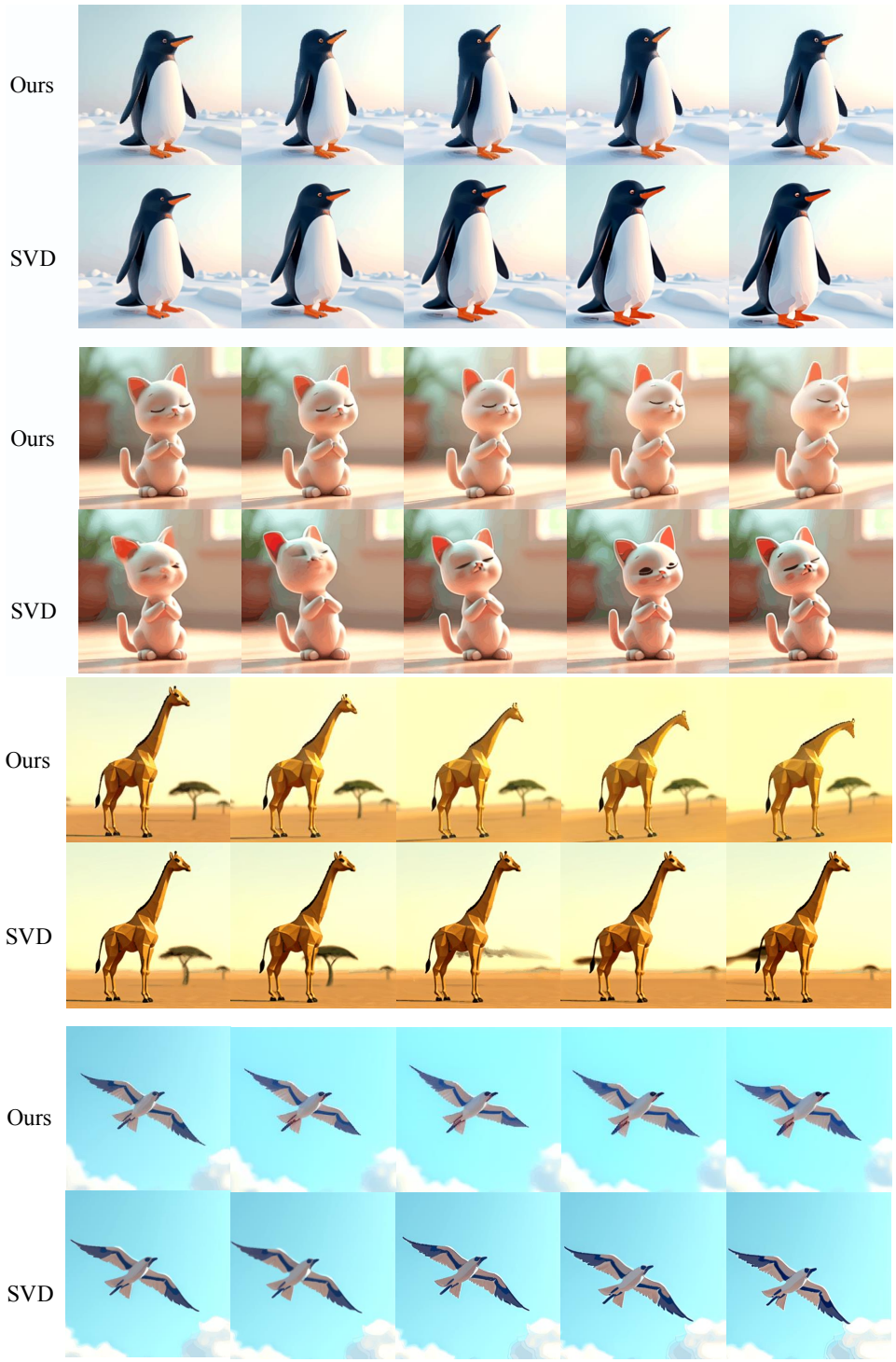


Figure 14: We compare the object motion against our base model (SVD).