CAMERA CONDITIONED VIDEO GENERATION WITH IMPROVED POSE FIDELITY

Anonymous authors
Paper under double-blind review

Figure 1: **FreeCam Results.** Given a source video and a target camera trajectory, our method generates a video that faithfully follows the specified camera path. With the reference coordinate system defined by the initial frame of the source video (highlighted in red), FreeCam enables novel-view video synthesis along arbitrary trajectories. The examples show generated videos following backward (first) and arc (second) camera trajectories.

ABSTRACT

Novel-view video generation from dynamic scenes has emerged as a compelling research direction with the advancement of video diffusion models. However, current approaches face key constraints that restrict flexibility. Specifically, methods exploiting Image-to-Video models as a baseline are constrained by the bias of the base model, limiting the target camera pose of the initial frame to remain close to the source. Limited diversity of camera trajectories in currently available datasets also confines trained models to generating output with limited camera trajectories. The generation results of projection-based methods that rely on depth estimation are affected by projection errors present in the depth-warped input video. In this paper, we present FreeCam, a camera trajectory conditioned Video-to-Video generation framework that enables depth-free novel-view video generation for a constraint-free camera trajectory. We introduce infinite homography warping that encodes 3D camera rotations directly in 2D latent space without depth, enabling high camera pose fidelity. Also, we augment existing multi-view datasets with identical initial frames into the dataset with arbitrary-trajectories and heterogeneous intrinsic parameters, enabling training on diverse camera motions and focal lengths. Our experimental evaluation demonstrates that FreeCam delivers enhanced trajectory precision over existing state-of-the-art approaches while preserving visual fidelity. Notably, despite training exclusively on synthetic data, FreeCam generalizes effectively to real-world videos. Through comprehensive ablation studies and comparative analyses, we confirm the complementary advantages of our proposed data processing pipeline and infinite homography warping technique, together establishing a novel framework for achieving precise and flexible camera motion control in video synthesis applications.

1 INTRODUCTION

Changing the camera viewpoint in post-production is a highly sought-after video editing technique, as it eliminates the need for costly reshoots and enables creative visual effects. The recent emergence of large-scale video diffusion models has triggered research on controlling camera trajectories for novel-view video generation Van Hoorick et al. (2024); YU et al. (2025); Bai et al. (2025a).

However, existing methods still suffer from several limitations. First, although models built upon Image-to-Video backbones have the benefit of preserving visual fidelity from the source video thanks to their pre-trained knowledge, this often imposes a restrictive constraint that the initial frame of the generated video must remain close to that of the source video. For example, GCD(Van Hoorick et al. (2024)) , which leverages a pretrained Image-to-Video diffusion model as its base model, generates a video with severe artifacts when given a distant camera pose for the first frame. Although the model is fine-tuned on a multi-view dynamic dataset, we conjecture that the base Image-to-Video model still tries to anchor the input image as the first frame even after fine-tuning, leading to artifacts.

Second, models are constrained by the narrow range of camera poses encountered during training. For instance, ReCamMaster Bai et al. (2025a) and SynCamMaster Bai et al. (2025b), both of which perform camera-controlled video generation built upon Text-to-Video diffusion model, generate first-frame preserving videos and videos with stationary cameras respectively, adhering to their respective training dataset configurations. Although the datasets used to train the model are both synthetically generated by the authors and could theoretically be freely manipulated, they still exhibit inherent constraints in camera pose diversity. We hypothesize that these dataset constraints arise from the fundamental requirement that camera viewpoints must maintain sufficient frustum overlap for the purpose of novel-view video generation, prohibiting the task from becoming pure generation. This geometric constraint explains why MultiCamVideo Dataset employs cameras that share the starting point of the arbitrary trajectories and why SynCamVideo Dataset utilizes stationary cameras randomly sampled on a hemispherical surface centered around the subject. While these configurations ensure sufficient visual overlap for stable training and inference, they inadvertently force models to internalize these geometric constraints, thereby limiting their ability to generalize to novel camera poses outside the training distribution.

Third, methods that explicitly incorporate depth projection face inherent performance degradation stemming from their dependency on depth estimation module. Although approaches such as YU et al. (2025) benefit from leveraging geometric information for novel view synthesis, the overall system performance becomes fundamentally constrained by the accuracy and reliability of the underlying depth predictor.

To address these limitations, we present FreeCam, a novel depth-free framework for camera trajectory control in video generation. Unlike existing approaches that suffer from cascading errors due to their reliance on auxiliary depth estimation modules, our method bypasses depth information entirely while preserving essential geometric constraints. Specifically, we introduce an infinite homography warping module that effectively injects 3D rotational information into the 2D latent space, providing robust geometric conditioning for the generation process. Additionally, we introduce a data processing strategy that transforms datasets with constrained viewpoints into flexible trajectory formats, enabling our model to learn from diverse camera movements. Extensive experiments demonstrate that the synergistic combination of our dataset processing and the infinite homography warping module enables superior trajectory fidelity compared to state-of-the-art methods. We further validate the generalizability of our approach through successful application to challenging real-world videos.

Our contributions are summarized as follows:

- We propose a camera-controlled novel-view video generation framework, which is free from camera pose constraints.
- We introduce an infinite homography warping module, successfully incorporating camera trajectory in 2D latent space without using depth information.
- We present a data augmentation scheme that constructs paired videos with arbitrary trajectories and varying camera intrinsics.

- Experimental results demonstrate that the proposed warping module and data augmentation scheme not only synergistically improve novel-view video generation performance, but also generalize effectively to unconstrained, real-world video data.

2 Related Work

Camera-Controlled Text-to-Video Generation Camera-controlled text-to-video generation methods produce videos based on a camera pose and a text prompt that describes the scene the user intends to control. Several works introduce a plug-and-play module for camera trajectory conditioning He et al. (2025); Bai et al. (2025b), ControlNet-like encoder with spatiotemporal camera embeddings based on Plücker coordinates Bahmani et al. (2025b), and improved training schedules Bahmani et al. (2025a). Although the authors argue that their method can be extended to the camera control of real-world videos, the generated outputs do not incorporate dynamic camera motion, instead producing novel-view videos with static camera positions, since the model learns from a training dataset that primarily consists of static cameras.

Camera-Controlled Video-to-Video Generation Camera-controlled video-to-video generation methods produce videos conditioned on both an input video and a specified camera pose, preserving the temporal dynamics of the input video while reflecting the spatial movement dictated by the given camera trajectory. Van Hoorick et al. (2024) pioneered a controllable novel-view dynamic video synthesis method that generates videos from novel viewpoints by leveraging pretrained image-to-video (I2V) diffusion model priors. Although the model is trained on a multi-view dynamic dataset, direct camera displacement often leads to performance degradation. YU et al. (2025) project the source video onto the target camera with a desired trajectory, utilizing depth maps estimated from a monocular depth estimator. The projection results serve as input to a video inpainting model, which is fine-tuned on a dataset generated using a double reprojection scheme. As the approach relies on an external monocular depth estimation module, its performance is bound by the quality of the depth estimator. Bai et al. (2025a) perform camera-controlled video generation by conditioning Text-to-Video (T2V) model on camera trajectories. The model is trained using a synchronized multi-camera synthetic video dataset, MultiCamVideo. Although their approach does not employ image-to-video models that enforce initial frame preservation, the proposed method still preserves the first frame, since all rendered multi-view videos in the MultiCamVideo dataset are synchronized to share the same initial frame. The generation performance is either constrained by the accuracy of the depth estimator or remains limited by restrictions on the trajectory types.

3 Preliminary

3.1 Infinite Homography

The infinite homography $\mathbf{H}_\infty$ represents the homography induced by the plane at infinity. Given source and target camera intrinsic matrices $\mathbf{K}_{src}$ and $\mathbf{K}_{trg}$, rotation matrix $\mathbf{R}$, translation vector $\mathbf{t}$, and normal $\mathbf{n}$ of a plane, the infinite homography $\mathbf{H}_\infty$ can be derived from the plane-induced homography $\mathbf{H} = \mathbf{K}_{trg}(\mathbf{R} - \mathbf{t}\mathbf{n}^T/d)\mathbf{K}_{src}^{-1}$ by taking the limit as the distance d to the plane approaches infinity:

$$\mathbf{H}_\infty = \lim_{d\to\infty} \mathbf{H} = \mathbf{K}_{trg}\mathbf{R}\mathbf{K}_{src}^{-1}. \tag{1}$$

For a pixel $\mathbf{p}$ with known depth Z measured from the source camera, the projection $\mathbf{p}'$ to the target image is expressed as:

$$\mathbf{p}' = \mathbf{K}_{trg}\mathbf{R}\mathbf{K}_{src}^{-1}\mathbf{p} + \mathbf{K}_{trg}\mathbf{t}/Z = \mathbf{H}_\infty\mathbf{p} + \mathbf{K}_{trg}\mathbf{t}/Z \tag{2}$$

Notably, $\mathbf{H}_\infty$ does not depend on the translation and depth, enabling correspondence of image points at arbitrary depths when the camera undergoes pure rotation($\mathbf{t} = \mathbf{0}$). The term $\mathbf{K}_{trg}\mathbf{t}/Z$ represents the parallax induced by the plane at infinity. We condition our model on 2D latent features transformed by infinite homography, allowing the network to focus exclusively on learning parallax information from the training data. Further details about the infinite homography can be found in Hartley & Zisserman (2003).

3.2 MULTICAMVIDEO DATASET

The MultiCamVideo Dataset is a synthetic dataset introduced in Bai et al. (2025a), comprising synchronized multi-camera videos with their corresponding camera trajectories. The dataset contains 13,600 distinct dynamic scenes, each captured from 10 different camera viewpoints, yielding a total of 136,000 videos. The synthetic rendering employs four different focal lengths: 18mm, 24mm, 35mm, and 50mm, with the focal length remaining constant within each scene. Each rendered video consists of 81 frames with a resolution of 1280×1280 pixels. The dataset encompasses diverse trajectory types including pan, tilt, translation, arc, random, and static movements. Each trajectory type is parameterized with randomly sampled values for angle, distance, scene coverage, and speed. A notable feature of this data is that all 10 cameras within each scene share identical starting positions, ensuring consistent initialization across viewpoints. Building on these properties, we introduce a data augmentation approach utilizing the MultiCamVideo Dataset to eliminate potential bias from initial frame conditioning and camera intrinsics preservation. Additional details about MultiCamVideo Dataset are provided in Bai et al. (2025a)

3.3 WAN2.1 TEXT-TO-VIDEO MODEL

Wan2.1(Wan et al. (2025)) is an open-sourced Text-to-Video (T2V) diffusion model based on transformer architecture. During training, for a given video $\mathbf{V} \in \mathbb{R}^{B\times(1+F)\times H\times W\times 3}$, the Wan-VAE compresses its spatio-temporal dimensions to $[1 + F/4, H/8, W/8]$. Subsequent patchification further reduces the spatial resolution, yielding $\mathbf{z} \in \mathbb{R}^{B\times(f\times h\times w)\times d}$, where $f = 1 + F/4$, $h = H/16$, $w = W/16$. Given a video latent $\mathbf{z}_1$, a random noise $\mathbf{z}_0 \sim \mathcal{N}(0, I)$, and a sampled timestep $t \in [0, 1]$, an intermediate latent $\mathbf{z}_t$ is obtained as the training input. Following Rectified Flows (RFs) (Esser et al. (2024)), $\mathbf{z}_t$ is defined as a linear interpolation between $\mathbf{z}_0$ and $\mathbf{z}_1$, i.e., $\mathbf{z}_t = t\mathbf{z}_1 + (1 - t)\mathbf{z}_0$. The ground truth velocity $\mathbf{v}_t$ is $\mathbf{v}_t = \frac{d\mathbf{z}_t}{dt} = \mathbf{z}_1 - \mathbf{z}_0$. The model is trained to predict the velocity, thus, the loss function can be formulated as the mean squared error (MSE) between the model output and $\mathbf{v}_t$,

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_0,\mathbf{z}_1,c_{txt},t}\left[\|u(\mathbf{z}_t, c_{txt}, t; \theta) - \mathbf{v}_t\|^2\right], \tag{3}$$

where c_{txt} is the text embedding sequence, θ represents the model weights, and $u(\mathbf{z}_t, c_{txt}, t; \theta)$ denotes the output velocity predicted by the model. We use pretrained Wan2.1 as our base model, keeping the weights frozen while introducing additional trainable layers. We employ the same training objective as Wan2.1.

4 METHOD

Our goal is to perform novel-view video synthesis using a given camera trajectory. Specifically, given a source video $\mathbf{V}^s \in \mathbb{R}^{F\times C\times H\times W}$, target camera trajectory $\mathbf{T} \in \mathbb{R}^{F\times 3\times 4}$ expressed relative to the source video's initial camera pose, and target camera intrinsic $\mathbf{K}^t \in \mathbb{R}^{3\times 3}$, our FreeCam generates novel view video $\mathbf{V}^t \in \mathbb{R}^{F\times C\times H\times W}$ that faithfully follows the target camera trajectory and target intrinsic configuration. Target camera trajectory is defined in a special Euclidean space $(\mathbf{R}, \mathbf{t}) \in SE(3)$, having rotation $\mathbf{R} \in \mathbb{R}^{3\times 3}$ and translation $\mathbf{t} \in \mathbb{R}^3$.

In Section 4.1, we present our model design tailored for the novel-view video synthesis task. In Section 4.2, we describe our data augmentation strategy that enhances existing synthetic data for novel-view video synthesis across unconstrained camera trajectories and varying intrinsics.

4.1 MODEL ARCHITECTURE

We adopt the Text-to-Video model Wan2.1(Wan et al. (2025)) as our base architecture. To incorporate camera controllability while preserving Wan2.1's robust video generation capabilities trained on extensive datasets, we freeze the pretrained weights of Wan2.1 and train only the newly introduced camera encoder and the Homography-Guided Attention Layers. The new attention layers are initialized using weights from the corresponding pretrained transformer blocks. We employ a camera encoder consisting of a linear layer with 16-dimensional input and d-dimensional output to encode camera poses. The input comprises a flattened 3×3 rotation matrix (9 parameters), three translation

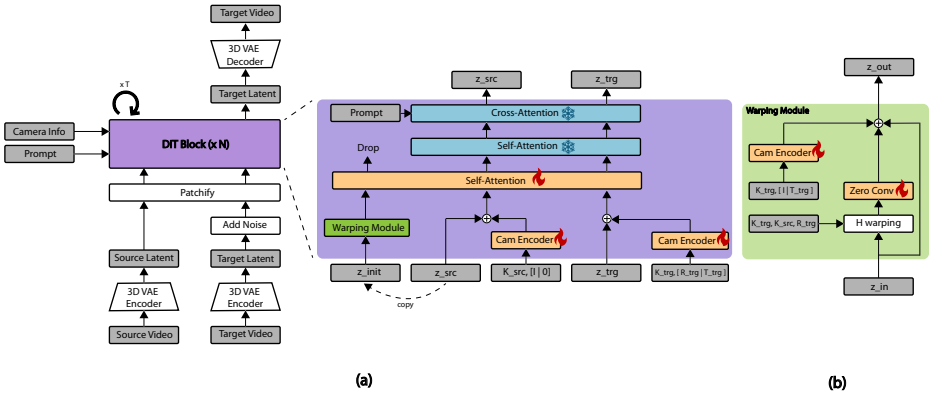

Figure 2: **Model Architecture Overview.** Our model builds upon Text-to-Video generation model, training only newly introduced parameters while freezing pretrained weights. **(a) Homography-Guided Attention Layer**: This layer performs per-frame attention ensuring temporal alignment by allowing target frames to reference corresponding source frames. The warped latents help the model understand rotation-induced view transformations. **(b) Warping Module**: This module warps latents using infinite homography, which approximates rotational transformations of 3D points in 2D space. This design simplifies 3D projection prediction problem to the simpler parallax prediction based on target translation.

parameters, and four intrinsic camera parameters (focal lengths f_x, f_y and principal point coordinates c_x, c_y). This camera encoder is shared across all transformer blocks and all types of camera encoding for consistent camera conditioning. The overall architecture of our model is illustrated in Figure 2.

Homography-Guided Attention Layer Our Homography-Guided Attention Layer performs per-frame attention by spatially concatenating information for target frame generation. The layer architecture is illustrated in Figure 2 (a). Specifically, given a source video latent $\mathbf{z}^s$ and a noisy target video latent $\mathbf{z}^t$, the concatenated information includes: $\mathbf{z}_i^t$, which is indexed from the target latent at the frame index i. $\mathbf{z}_i^s$, which is indexed from the source latent at the corresponding frame i, and $\mathbf{z}_i^w$, representing the warped version of the source latent at the initial frame ($\mathbf{z}_{init}^s$). The warping operation is performed using the camera intrinsics and i-th target camera pose.

Camera embeddings are added to their corresponding latents prior to spatial concatenation. The target camera embedding is obtained by passing the user-specified target intrinsics $\mathbf{K}^t$, rotation $\mathbf{R}^t$, and translation $\mathbf{t}^t$ through the camera encoder. For the source camera embedding, we concatenate the flattened source intrinsics with the identity pose $[\mathbf{I}|\mathbf{0}]$, replicate this across frames, and encode it using the camera encoder.

The resulting concatenated latents have shape $\mathbf{z}^c \in \mathbb{R}^{bf\times 3hw\times d}$, where frames are processed as individual batch items within the attention mechanism. This structure ensures temporal alignment by allowing target frames to reference corresponding source frames at matching timestamps. Moreover, the warped latent features facilitate the model's understanding of rotation-induced view transformations. After passing through the Homography-Guided Attention Layer, the concatenated features $\mathbf{z}^c$ are split and reshaped to $\mathbb{R}^{b\times fhw\times d}$ format to serve as input to Wan2.1's Self-Attention Layer. During processing through the pretrained Wan2.1 layers, the paired source and target latents are treated as a unified batch, and do not use warped latents. More detailed model architecture can be found in Appendix A.

Warping Module Motivated by Equation (2), we condition the attention layer on the warped latent using infinite homography, which approximates the rotational transformation of 3D points in 2D space. Our warping module is illustrated in Figure 2 (b). Since target camera poses are expressed relative to the source video's first frame, the Warping Module warps the source latent of the initial frame using the target camera poses and intrinsics. The module is designed to reflect the components in Equation (2). Specifically, $\mathbf{z}_{init}$ is first warped using infinite homography derived from camera intrinsics and target poses. The warped result is then added to the original $\mathbf{z}_{init}$ through a zero convolution layer as a residual connection. Subsequently, camera embeddings encoding tar-

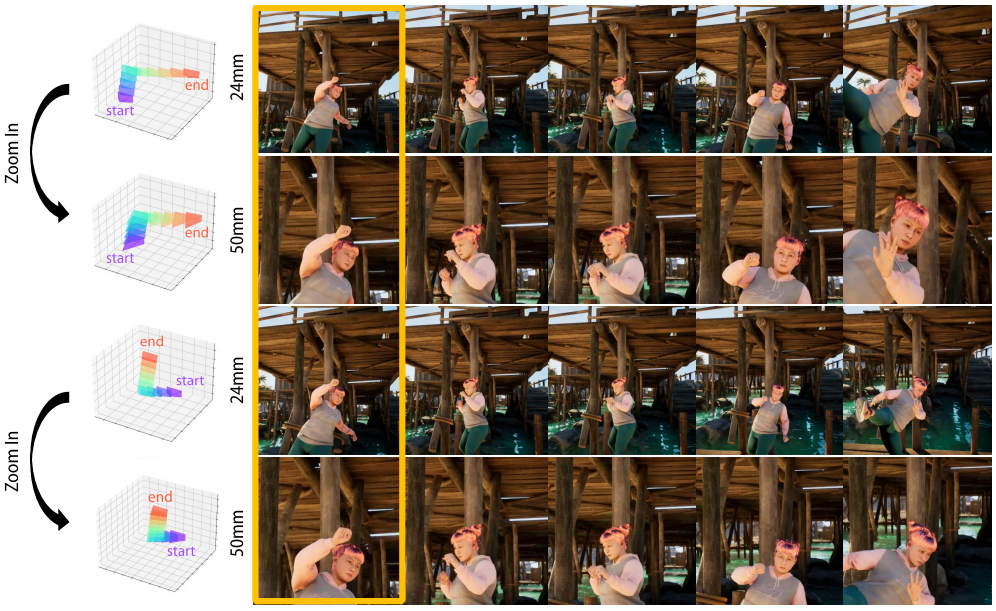

Figure 3: **Augmented Dataset Examples**. **Rows 1&3**: Examples of our trajectory augmentation. This ensures temporal alignment while introducing variability in initial frame selection (highlighted in yellow). **Rows 2&4**: Examples of our focal length augmentation. This helps the model learn the underlying relationship between focal length variations and their corresponding effects on video generation.

get translation and intrinsics are added, representing the second term in Equation (2). This design simplifies projection estimation under target camera poses to parallax estimation induced by target translation. The effectiveness of this module is validated through ablation studies in Table 3.

4.2 Data Preparation

In this section, we describe our data augmentation strategy that enhances existing synthetic data for novel-view video synthesis with unconstrained camera trajectories and varying intrinsics. We utilize the MultiCamVideo Dataset Bai et al. (2025a) for augmentation and will refer to the augmented version as AugMCV (Augmented MultiCamVideo) Dataset for brevity.

4.2.1 Trajectory Augmentation

As discussed in Section 3.2, the MultiCamVideo dataset provides 10 arbitrary trajectories of a single scene, each capturing different viewpoints. However, all trajectories originate from an identical initial viewpoint. Consequently, when randomly sampling two videos from the same scene, their first frames are always identical. We empirically observe that models trained on this dataset exhibit a bias toward reproducing the source video's first frame, even when conditioned to start from a different viewpoint.

To mitigate the bias, we propose a novel augmentation strategy to enhance the MultiCamVideo dataset. Our key observation is that while all trajectories share the same starting frame, their terminal frames diverge significantly across different videos. Leveraging this property, we randomly sample two distinct videos from the same scene and construct an augmented sequence by reversing the first video and concatenating it with the second. Since the final frame of the reversed video coincides with the initial frame of the second video, we remove the redundant first frame of the latter, resulting in an augmented video of 161 frames (81 + 80). Figure 3 (rows 1 and 3) shows examples of our trajectory augmentation approach. This approach ensures temporal alignment while introducing variability in initial frame selection, thereby enabling models to learn more generalizable trajectory representations without dataset-specific biases.

Table 1: Experimental results on the test set of AugMCV dataset. The best and second-best results are **bold** and underlined, respectively.

Method	Shared Intrinsics			Different Intrinsics			Mixed		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
ReCamMaster	21.130	0.617	0.420	19.665	0.558	0.510	20.442	0.589	0.462
TrajectoryCrafter	21.228	0.660	0.296	19.557	0.586	0.390	20.489	0.627	0.337
Ours	**22.677**	**0.718**	**0.246**	**22.261**	**0.699**	**0.270**	**22.494**	**0.710**	**0.257**

4.2.2 Intrinsic Augmentation

As discussed in Section 3.2, the MultiCamVideo Dataset incorporates four distinct focal lengths, with 3,400 scenes allocated to each focal length configuration. While the dataset inherently encompasses multiple intrinsic parameters, our empirical analysis reveals that models trained on this dataset demonstrate a systematic bias toward generating videos that preserve the focal length characteristics of the source video. This limitation arises from the training paradigm wherein both source and target videos are sampled from identical scenes, consequently maintaining consistent focal length parameters across video pairs. Such consistency prevents the model from learning the underlying relationship between focal length variations and their corresponding effects on video generation. To alleviate this bias, we introduce intrinsic augmentation.

Specifically, given the trajectory-augmented scene with focal length f_{scene}, we random sample the new focal length $f_{\text{new}} \in \{x \in \{18\text{mm}, 24\text{mm}, 35\text{mm}, 50\text{mm}\} | x > f_{scene}\}$, and apply intrinsic augmentation process to the input video. Figure 3 (rows 2 and 4) illustrates the results of applying focal length augmentation to video sequences obtained from Section 4.2.1. Intrinsic augmentation can be achieved by simple resize and crop. Detailed process is described in the pseudo code in the Appendix B.2.

4.2.3 Video Pair Selection

To train Video-to-Video model for novel-view video generation, paired video data is required. The source video represents the user input, while the target video serves as the ground truth for the generated output following a specified trajectory. Video pairs are constructed by sampling two videos from identical scenes. To align with the pre-training strategy of Wan2.1(Wan et al. (2025)) base model, we sample 81 frames from temporally synchronized video pairs with identical starting timestamps. For each video, we apply focal length augmentation with a probability of 0.5. Although we augment focal lengths only in the ascending direction, using augmented videos as either source or target ensures that the selected video pairs encompass both focal length increase and decrease scenarios. Figure 3 illustrates representative video pairs (rows 1 and 3) alongside their corresponding focal length augmented versions (rows 2 and 4). Source and target videos are sampled from these four candidates. This sampling strategy enables coverage of zoom-in, zoom-out, and arbitrary trajectory patterns.

5 Experiments

5.1 Evaluation Set

We evaluate the accuracy of camera trajectories and the quality of generated videos on two distinct datasets.

AugMCV Dataset Experiments are conducted on 168 scenes from the AugMCV test split. Each scene includes 10 target camera trajectories, along with the corresponding ground-truth video for each trajectory. By using a static-camera clip as input, we generate one video for each target trajectory, resulting in a total of 1,680 generated videos. Among the 168 test scenes, 72 scenes use source and target videos with different camera intrinsics, while the remaining 96 scenes have identical intrinsics between source and target videos. These videos are then compared to their respective ground truth videos using performance metrics such as PSNR, SSIM, and LPIPS Zhang et al. (2018).

Figure 4: Qualitative comparison on the test set of the AugMCV dataset. ReCamMaster(RCM) fails in viewpoint transformation, keeping the initial frame of the source video, while TrajectoryCrafter exhibits inaccurate projection performance. In contrast, our methodology achieves high visual fidelity to the target video. Best viewed in zoom

WebVid Dataset A random sample of 100 source videos is selected from the WebVid Bain et al. (2021) dataset to evaluate performance in real-world scenarios. For each source video, synthetic videos are generated using 20 different camera trajectories. Of these, ten camera trajectories maintain an initial camera pose identical to the source video's first frame, while the remaining ten camera trajectories employ an initial camera pose intentionally deviated from the source video's first frame. This procedure yields a total of 2,000 generated synthetic videos. Video and frame-level fidelity are assessed using FID Heusel et al. (2017) and FVD Unterthiner et al. (2019), and rotation and translation errors He et al. (2024) are reported for the target trajectories and the generated videos.

5.2 Experiments on AugMCV Dataset

Qualitative Results Fig. 4 presents a qualitative comparison of our methodology against baseline models for AugMCV test set. It shows the generated videos (rows 2-4) when the source video (row 1) is transformed according to the camera trajectory, compared with the ground-truth target video (row 5) corresponding to the input camera trajectory. ReCamMaster preserves the initial frame of the source video unchanged due to its limitations of training data that starts from the same initial frame. TrajectoryCrafter, which employs projection-based methods, successfully performs viewpoint transformation of the initial frame but fails to reflect the appearance of the source video across all frames due to inaccurate projections from the depth estimator. In contrast, our methodology, benefiting from the warping module and trajectory-intrinsic augmentation, successfully transforms the initial frame to align with the target camera trajectory and maintains consistency with the target video's viewpoint throughout the remaining trajectory.

Quantitative Results Table 1 presents quantitative evaluation results for three scenarios: (1) source and target videos with identical camera intrinsics, (2) source and target videos with different camera intrinsics, and (3) a mixed setting with both types. Across all scenarios, our method consistently outperforms the baseline approaches in terms of PSNR, SSIM, and LPIPS metrics.

5.3 Experiments on WebVid Dataset

Qualitative Results Figure 5 shows qualitative results on First-Frame Asynchronized (FF-Async) scenarios using videos from the WebVid dataset as source video. As shown in row 2 of the Figure 5, ReCamMaster (RCM) preserves the first frame of the source video unchanged, similar to the experimental results conducted in Section 5.2. This is because RCM is trained with bias inherent in the

Figure 5: **Qualitative Comparison.** ReCamMaster (RCM) fails to reflect changes in the initial frame pose, regardless of whether trajectory interpolation is applied. Trajectory Crafter exhibits noticeable distortions in facial regions due to projection errors. In contrast, our method achieves natural pose transitions while maintaining high visual quality throughout the sequence. Best viewed in zoom

Table 2: Quantitative comparison on WebVid dataset. We compare methods on First-Frame Synchronized (FF-Sync) and First-Frame Asynchronized (FF-Async) settings. Best scores per metric are in **bold**.

Method	FF-Sync				FF-Async			
	Rot.↓	Trans.↓	FID↓	FVD↓	Rot.↓	Trans.↓	FID↓	FVD↓
ReCamMaster	9.673	1.466	40.612	308.697	4.843	1.599	30.828	297.328
ReCamMaster w/ interp.	-	-	-	-	7.076	0.589	39.248	295.353
TrajectoryCrafter	5.595	**0.502**	**32.220**	287.805	3.437	1.467	29.534	291.954
Ours	**3.605**	0.510	32.906	**282.703**	**2.718**	**0.365**	**26.497**	**291.202**

MultiCamVideo Dataset, which favors generating videos that start with the same initial frame as the source video. To mitigate this limitation, we additionally employed frame interpolation techniques. Specifically, we interpolated the first 8 trajectories from the reference pose to the initial target trajectory, repeating the source video's first frame 8 times. Then we cropped the first 8 frames from the generated video. This approach allows ReCamMaster to generate different viewpoints from the initial frame at the cost of frame count. However, as shown in row 3 of Figure 5, while the viewpoint of the first frame shows slight changes, errors persist in reaching to the target trajectory during the interpolation process. TrajectoryCrafter projects the first frame's viewpoint to align with the target trajectory, but inaccurate depth estimation decreases the fidelity of the source video, such as artifacts appearing in the woman's clothing in the center. In contrast, our depth-free method is independent of such external network errors and demonstrates viewpoints aligned with the target trajectory in First-Frame Asynchronized (FF-Async) scenarios without loss of frame count.

Quantitative Results Table 2 presents quantitative evaluation results for both First-Frame Synchronized (FF-Sync) and First-Frame Asynchronized (FF-Async) trajectories. Our method demonstrates clear superiority over competing approaches in the asynchronous setting and achieves competitive or superior performance in the synchronous setting. Particularly, our method shows superior performance in the more challenging FF-Async scenario, achieving lower rotation and translation errors. These results validate that our proposed warping and augmentation techniques effectively generate videos well-aligned with camera trajectories, regardless of the given camera trajectory.

Table 3: Ablation study conducted on the WebVid dataset. The **best** and second best results are highlighted for each metric.

Components			Shared Intrinsic			Different Intrinsic			Mixed		
Aug.Traj.	Aug.Intr.	Warp	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
☐	☐	☐	19.228	0.562	0.427	18.480	0.507	0.525	18.907	0.539	0.469
✓	☐	☐	20.820	0.615	0.353	18.865	0.523	0.499	19.982	0.575	0.416
✓	✓	☐	22.807	0.680	0.250	21.866	0.649	0.293	22.404	0.667	0.268
✓	✓	✓	**24.311**	**0.720**	**0.203**	**24.412**	**0.733**	**0.198**	**24.369**	**0.727**	**0.200**

Table 4: Ablation study of our augmentation strategies on WebVid dataset. Best scores per metric are in **bold**.

Method	FF-Async			
	Rot.↓	Trans.↓	FID↓	FVD↓
w/ SynCamVideo	4.989	1.251	50.389	234.985
w/ Augmentation	**3.628**	**0.758**	**42.758**	**234.280**

5.4 Ablation Study

Analysis of Proposed Components. We perform an ablation study by progressively adding each of the proposed components in AugMVC dataset. As shown in Fig. 8 and Table 3, the baseline without warping and augmentation fails from the first frame to capture both the target trajectory and intrinsics, yielding the lowest scores under both intrinsic settings. Adding trajectory augmentation improves performance in the homogeneous intrinsic case and produces a slight rightward rotation in the first frame consistent with the target trajectory; however, it still fails to reflect the target intrinsics and offers comparable performance to the baseline under heterogeneous intrinsics. Enabling both trajectory and intrinsic augmentation yields partial adaptation to the improved intrinsics but remains inaccurate. Finally, the warping module, by explicitly warping the source latent with respect to the target pose and intrinsics, proves crucial for accurately encoding intrinsic information, leading to substantial qualitative and quantitative gains.

Effectiveness of our augmentation strategy. We evaluate the proposed trajectory-intrinsic augmentation scheme in two regimes: (i) a model with a warping module trained on MultiCamVideo and SynCamVideo, and (ii) a model trained on FreeCam using our method. SynCamVideo comprises multiple recordings of the same dynamic scene, captured simultaneously from distinct viewpoints using stationary cameras. Even under joint training, it does not cover First-Frame Asynchronized (FF-Async) case. In contrast, our augmentation yields lower camera-estimation error and improved video-quality metrics as shown in Table 4.

6 Conclusion

In this paper, we present a novel camera-controlled video-to-video generation model, FreeCam. Our approach achieves accurate pose fidelity without requiring a depth prior, breaking the first-frame constraint imposed by previous methods. Our key contribution is the infinite homography warping module, which encodes 3D camera rotations directly in the 2D latent space, thereby eliminating the need for external depth estimation while improving camera-pose fidelity. Training with the proposed data augmentation scheme both contributes to performance improvements and removes prior constraints. Notably, our framework enables camera control as a post-processing step for videos, representing a significant advancement in video editing. In future work, this approach can be extended beyond the current base model's frame length limitation, enabling even longer camera-controlled video generation results.

REFERENCES

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22875–22889, 2025a.

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. *The Thirteenth International Conference on Learning Representations*, 2025b.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. 2025a.

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. *The Thirteenth International Conference on Learning Representations*, 2025b.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

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.

Richard Hartley and Andrew Zisserman. Scene planes and homographies. In *Multiple view geometry in computer vision*, pp. 325–342. Cambridge university press, 2003.

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.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

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.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10106–10116, 2024.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

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*, pp. 313–331. Springer, 2024.

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

Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Supplementary Material

A Model Architecture Details

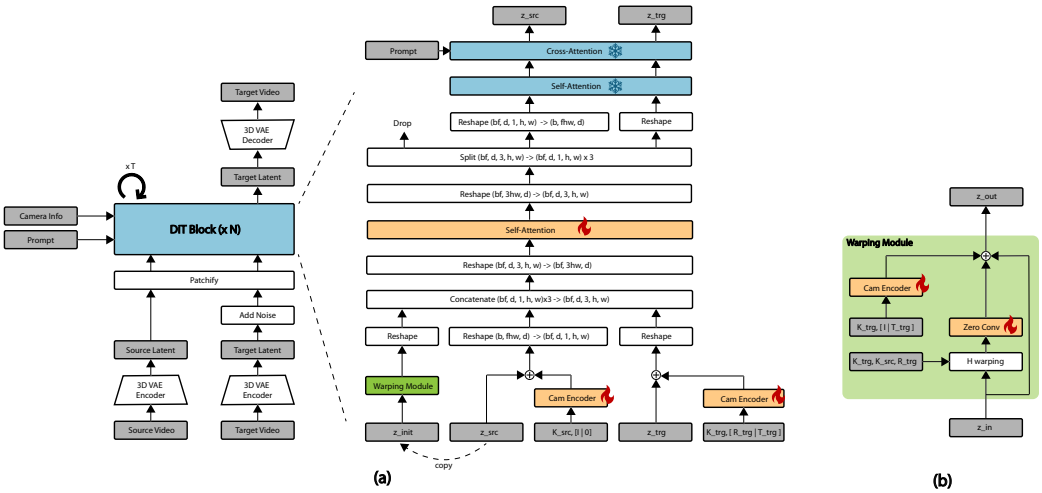

Figure 6: This figure illustrates our model overview along with the dimensions of latent feature handling. Here, d represents the feature dimension, w and h denote latent width and height, and f indicates the latent frame count. The $\mathbb{R}$ notation is omitted for simplicity.

Figure 6 presents an overview of the proposed FreeCam, along with its dimensional specifications. Further architectural details can be found in Section 4.1.

B Dataset Augmentation Details

B.1 Trajectory Augmentation

Figure 7 displays the scene example of MultiCamVideo dataset Bai et al. (2025a). Each scene contains 10 videos, including 9 with random camera trajectories and 1 with a static camera. Note that the initial frames are all identical. Two trajectories illustrated in Figure 3 are crafted using the trajectory pair (1, 9) and (2, 9)

B.2 Intrinsic Augmentation

Algorithm 1 illustrates our approach for augmenting the focal length of a given video. This process involves resizing according to the ratio between current and new focal lengths, followed by cropping to maintain the original image resolution.

C Implementation Details

We use Wan2.1(Wan et al. (2025)) as our backbone model and employ LLaVA(Liu et al. (2024)) for text extraction from the source video. During inference, we estimate the intrinsic parameters of the source video using UniDepth(Piccinelli et al. (2024)). We conduct ablation studies using low-resolution training (F=41, H=320, W=544) for 20k iterations on 4 H100 GPUs with 81GB VRAM each, using a batch size of 32. For quantitative evaluation, we train our model at high resolution (F=81, H=480, W=832) for 15k iterations. For RotErr(degree) and TransErr(meter) computation, ViPE(Huang et al. (2025)) is used for camera trajectory extraction of the generated video and the extracted trajectory is compared with the ground truth trajectory. To ensure that camera trajectory extraction with ViPE(Huang et al. (2025)) references the initial camera pose of the source video,

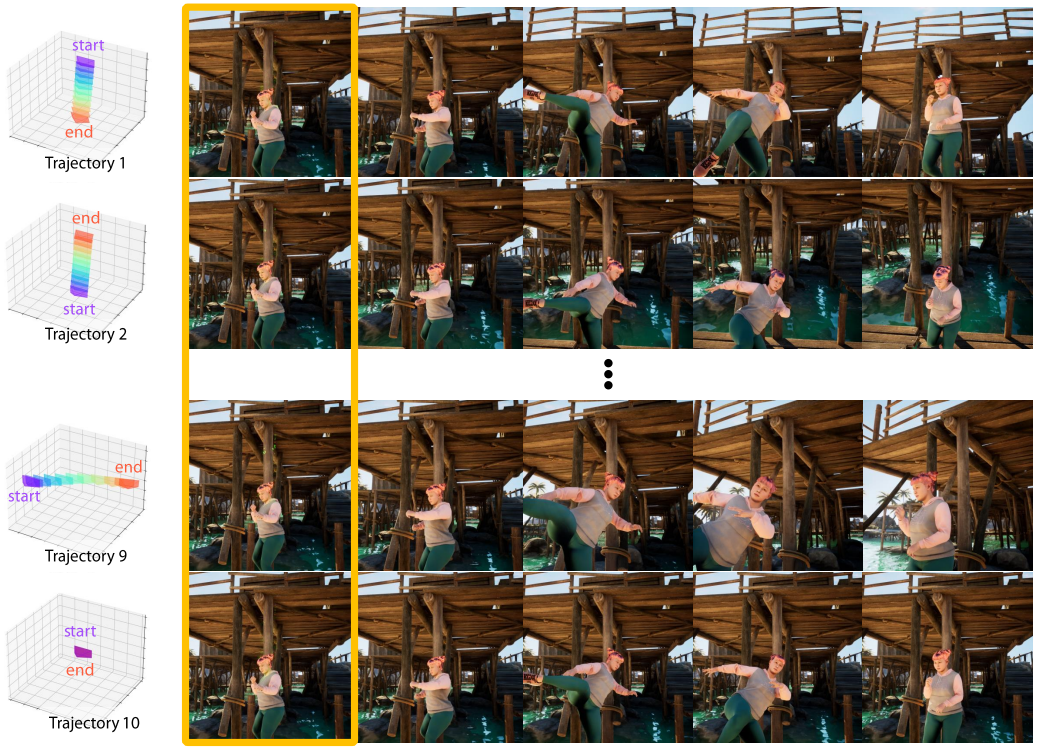

Figure 7: Example scene from the MultiCamVideo Dataset. Each scene contains 10 videos, including 9 with random camera trajectories and 1 with a static camera. Each row displays frames from one camera trajectory. All videos share the same initial frame.

Algorithm 1 Video Focal Length Augmentation

Input: Scene path P_{scene}, focal length f
Output: Augmented video dataset with modified focal lengths
procedure PROCESSVIDEOS(P^{in}_{scene}, P^{out}_{scene}, f_{now}, f_{new})
  for $i = 1$ to 10 **do** ▷ Process 10 cameras
    $video \leftarrow$ Load(P^{in}_{scene}/cam_i.mp4)
    $(W_{orig}, H_{orig}) \leftarrow$ GetDimensions($video$)
    $(W_{new}, H_{new}) \leftarrow (\frac{f_{new}}{f_{now}} \cdot W_{orig}, \frac{f_{new}}{f_{now}} \cdot H_{orig})$
    while frame exists in $video$ **do**
      $frame \leftarrow$ ReadFrame($video$)
      $frame \leftarrow$ Resize($frame$, (W_{new}, H_{new}))
      $frame \leftarrow$ CenterCrop($frame$, (W_{orig}, H_{orig}))
      WriteFrame(P^{out}_{scene}/cam_i.mp4, $frame$)
    end while
  end for
end procedure

the first frame of each source video is concatenated at the beginning of every generated video. ViPE then estimates relative poses with respect to this concatenated frame. Pose estimation results from the concatenated first frame are excluded before evaluation.

Figure 8: Qualitative ablation study of proposed components. From top to bottom, each row incrementally adds proposed components, showing cumulative improvements toward the target video. Best viewed in zoom

D BASELINES

D.1 RECAMMASTER

We use the official code and checkpoints to generate videos with resolution 480×832 and length 81 frames. In the ReCamMaster MultiCamVideo dataset, all videos of a given scene share the same initial frame. To realize the dynamic first-frame setting in Section 5.2, we construct eight short camera trajectories that move from the reference pose to the target sequence's first pose. We then repeat the first frame of the source video eight times and render it along these trajectories, assigning the results to the first eight frames of each generated video. This procedure yields a dynamic initial segment while preserving content consistency.

D.2 TRAJECTORYCRAFTER

YU et al. (2025) utilize CogVideoX as a baseline, with the input resolution fixed at 384×672 and sequence length constrained to 49 frames. For fair comparison under the 81-frame setting, we extended inference by generating outputs in 49-frame segments, reusing the last frame of each segment as the first frame of the subsequent segment, thereby producing continuous sequences. This inference-level extension alleviates the architectural limitation, enabling 81-frame sequence generation with preserved temporal continuity. All other experimental settings are kept identical to those in the original paper, ensuring a fair comparison with our proposed approach.

E ADDITIONAL QUALITATIVE RESULTS

E.1 ABLATION STUDY

Figure 8 visualizes the results of our ablation study. Training on the AugMCV Dataset, augmented with the proposed data augmentation scheme, produces generation results that more closely resemble the target video. When the warping module proposed in this paper is further incorporated, the generated output achieves the highest alignment with the target video.

Figure 9: Additional qualitative results of AugMVC dataset. Best viewed in zoom.

E.2 Additional Qualitative Results

Figure 9, Figure 10, and Figure 11 present additional qualitative results for AugMVC dataset cases as well as WebVid under both synchronous and asynchronous settings. Figure 12 further demonstrates our method's performance across diverse camera trajectories. All corresponding video results are included in the supplementary material as video files.

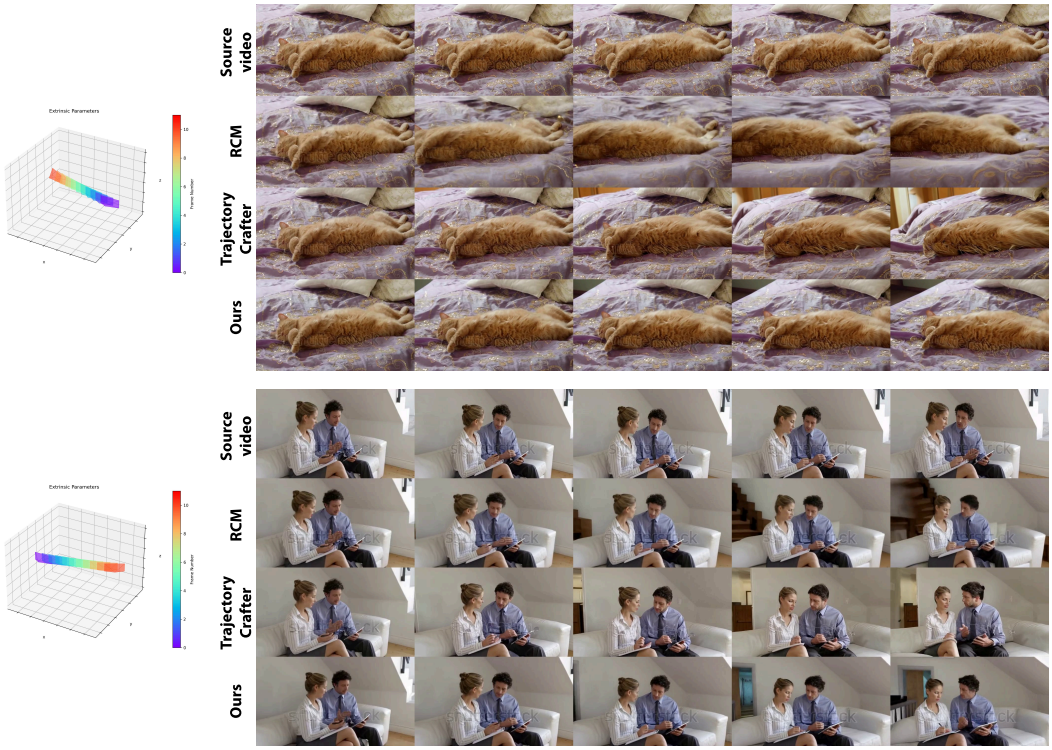

Figure 10: Additional qualitative results of WebVid dataset with First-Frame Sync (FF-Sync). Best viewed in zoom.

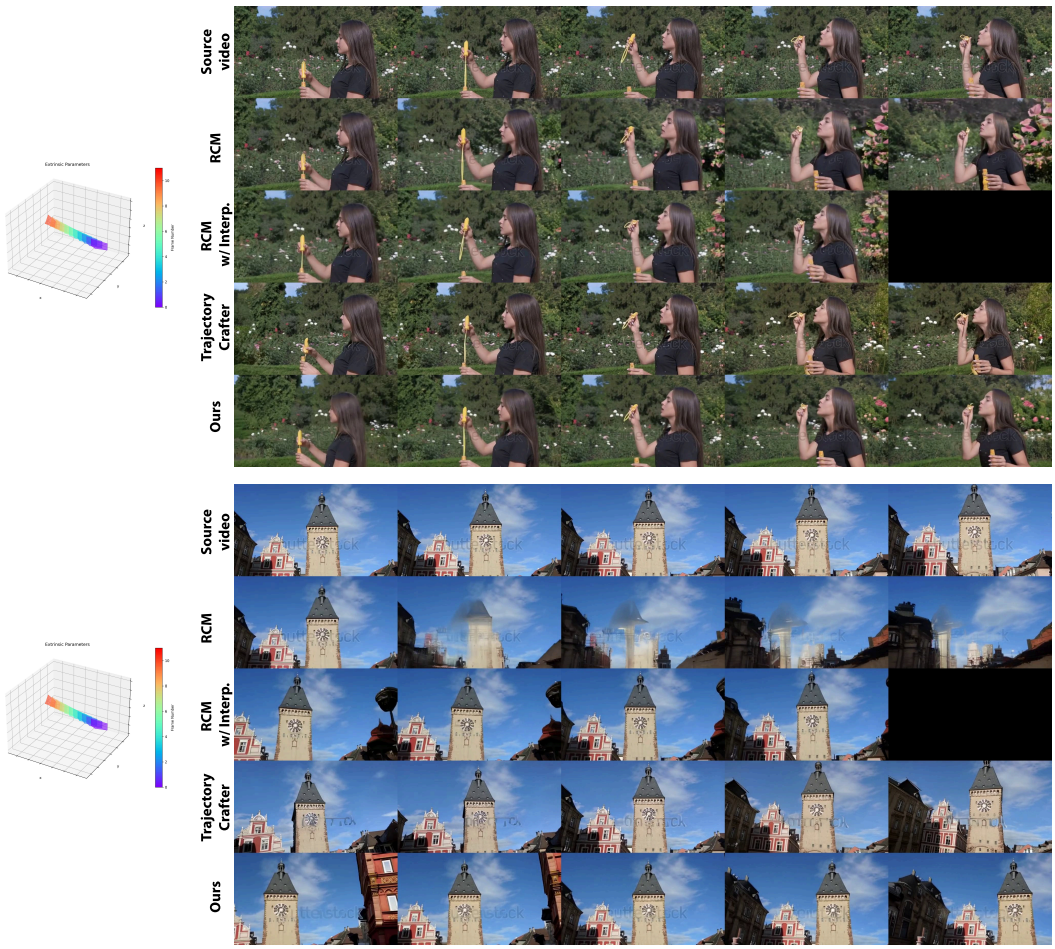

Figure 11: Additional qualitative results of WebVid dataset with First-Frame Async (FF-Async). Best viewed in zoom.

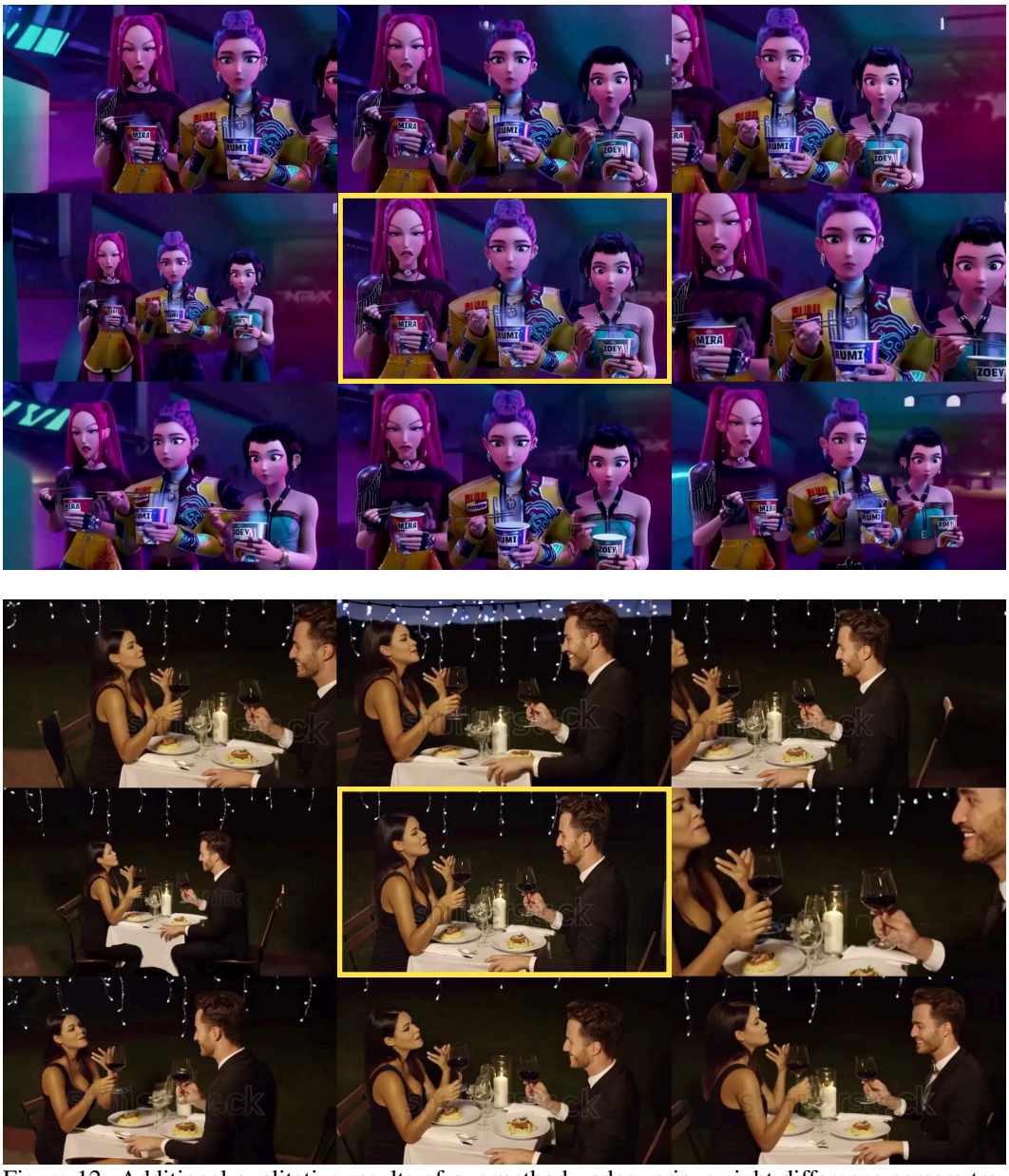

Figure 12: Additional qualitative results of our method under various eight difference camera trajectories. Best viewed in zoom.