

FreeOrbit4D: Training-Free Arbitrary Camera Redirection for Monocular Videos via Foreground-Complete 4D Reconstruction

WEI CAO, University of Illinois Urbana-Champaign, USA
 HAO ZHANG, University of Illinois Urbana-Champaign, USA
 FENGRUI TIAN, University of Pennsylvania, USA
 YULUN WU, University of Illinois Urbana-Champaign, USA
 YINGYING LI, University of Illinois Urbana-Champaign, USA
 SHENLONG WANG, University of Illinois Urbana-Champaign, USA
 NING YU, Eyeline Labs, USA and Netflix, USA
 YAOYAO LIU*, University of Illinois Urbana-Champaign, USA

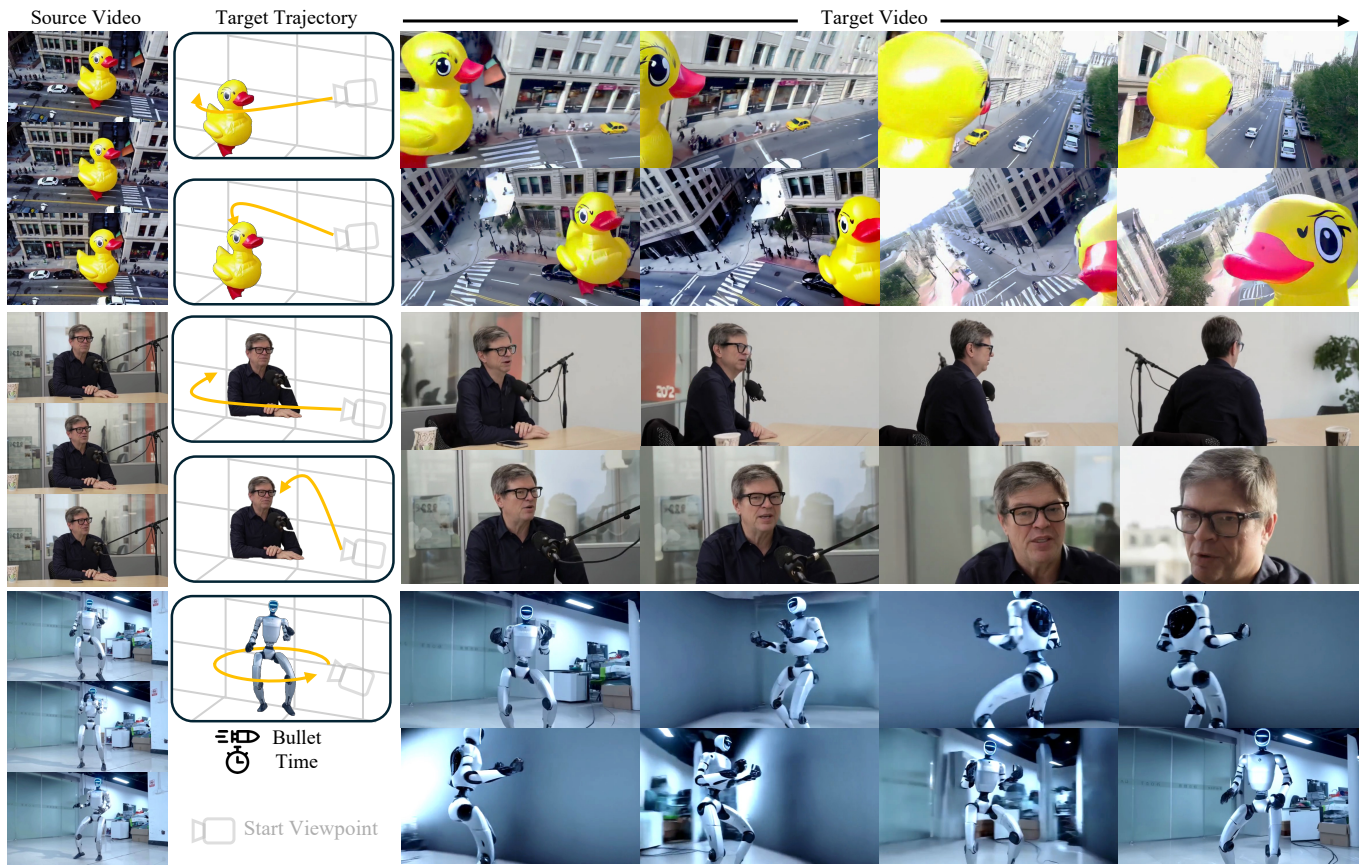


Fig. 1. FreeOrbit4D enables **training-free** camera redirection from a single monocular video to arbitrary target camera trajectories. Given a source video (left) and a target trajectory (middle), our method produces a redirected video (right) with faithful appearance and strong temporal coherence under large-angle redirected camera motions, even including bullet-time orbits, demonstrated on diverse scenes and subjects.

*Corresponding author.

Authors' Contact Information: Wei Cao, University of Illinois Urbana-Champaign, Champaign, Illinois, USA, weicao3@illinois.edu; Hao Zhang, University of Illinois Urbana-Champaign, Urbana, Illinois, USA, haoz19@illinois.edu; Fengrui Tian, University of Pennsylvania, Philadelphia, Pennsylvania, USA, tianfr@upenn.edu; Yulun Wu, University of Illinois Urbana-Champaign, Champaign, Illinois, USA, yulun5@illinois.edu; Yingying Li, University of Illinois Urbana-Champaign, Urbana, Illinois, USA,

yl101@illinois.edu; Shenlong Wang, University of Illinois Urbana-Champaign, Urbana, Illinois, USA, shenlong@illinois.edu; Ning Yu, Eyeline Labs, Los Angeles, California, USA and Netflix, Los Angeles, California, USA, ningyu.hust@gmail.com; Yaoyao Liu, University of Illinois Urbana-Champaign, Champaign, Illinois, USA, lyy@illinois.edu.



This work is licensed under a Creative Commons Attribution 4.0 International License.

SIGGRAPH Conference Papers '26, Los Angeles, CA, USA

SIGGRAPH Conference Papers '26, July 19–23, 2026, Los Angeles, CA, USA.

Camera redirection aims to replay a dynamic scene from a single monocular video under a user-specified camera trajectory. However, large-angle redirection is inherently ill-posed: a monocular video captures only a narrow spatio-temporal view of a dynamic 3D scene, providing severely limited observations of the underlying 4D world. The key challenge is therefore to recover a complete and coherent plenoptic representation from this limited input, with consistent geometry and coherent motion. While recent diffusion-based methods achieve impressive visual generation quality, they often break down under large-angle viewpoint changes far from the original trajectory, where missing visual grounding leads to severe geometric ambiguity and temporal inconsistency. To address this, we present **FreeOrbit4D**, an effective training-free framework that tackles this geometric ambiguity by recovering a foreground-complete 4D proxy as structural grounding for video generation. We obtain this proxy by decoupling foreground and background reconstructions: we unproject the monocular video into a static background and partial foreground point clouds in a unified global space, then leverage an object-centric multi-view diffusion model to synthesize multi-view images and reconstruct complete foreground point clouds in canonical object space. By aligning the canonical foreground point cloud to the global scene space via dense pixel-synchronized 3D–3D correspondences and projecting the foreground-complete 4D proxy onto target camera viewpoints, we provide geometric scaffolds (e.g., depth/visibility cues) that guide a conditional video diffusion model. Extensive experiments show that FreeOrbit4D produces more faithful and temporally coherent redirected videos under challenging large-angle trajectories, and our foreground-complete 4D proxy further opens a potential avenue for practical applications such as edit propagation and 4D data generation. **Project page:** <https://freeorbit4d.vision.ischool.illinois.edu/>

CCS Concepts: • **Computing methodologies** → **Image-based rendering**; **Rendering**; **Reconstruction**.

ACM Reference Format:

Wei Cao, Hao Zhang, Fengrui Tian, Yulun Wu, Yingying Li, Shenlong Wang, Ning Yu, and Yaoyao Liu. 2026. FreeOrbit4D: Training-Free Arbitrary Camera Redirection for Monocular Videos via Foreground-Complete 4D Reconstruction. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '26)*, July 19–23, 2026, Los Angeles, CA, USA. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3799902.3811122>

1 Introduction

Camera redirection seeks to endow machines with the ability to synthesize novel video sequences from a source video along a user-specified camera trajectory [Bai et al. 2025; He et al. 2024; Hou and Chen 2024; Yu et al. 2024]. For example, consider the video interview footage in Fig. 1. Although recorded from a fixed viewpoint, we can easily imagine the same scene from the side, back, or top-down, by mentally reconstructing a 3D world and replaying its dynamics from a virtual perspective. This capability is central to applications such as Autonomous Driving [Cao et al. 2025; Zhou et al. 2023, 2025a], AR/VR [Azuma 1997], and cinematic effects like bullet-time replay from limited camera setups [Gao et al. 2021; Tian et al. 2023; Wang et al. 2025d; Zhang et al. 2025d]. Today, in industry, such effects rely on costly multi-camera rigs and multi-view reconstruction techniques [Charatan et al. 2024; Chen et al. 2024; Jiang et al. 2025b; Kerbl et al. 2023; Mildenhall et al. 2021; Tancik et al.

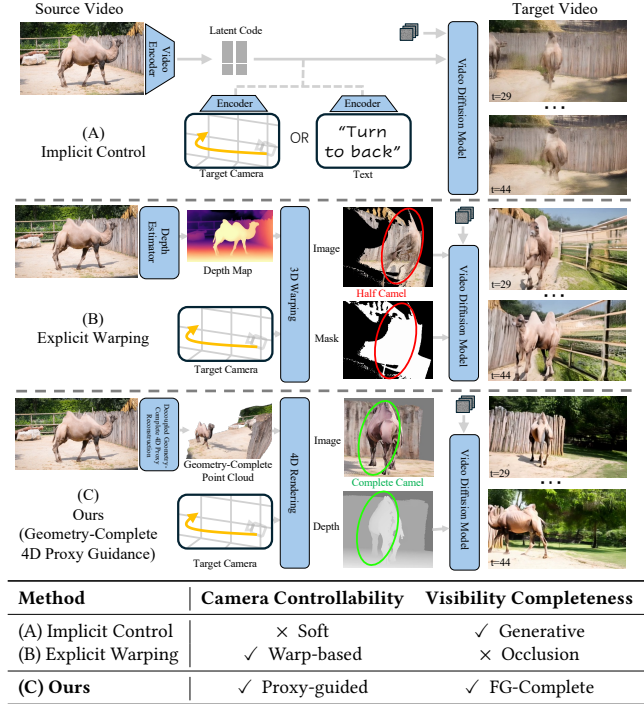


Fig. 2. **Comparison of video camera redirection paradigms.** We compare 3 representative approaches for camera redirection from monocular video. **(A) Implicit Control:** Camera motion is specified via learned embeddings. Such implicit representations provide only soft controllability: text cannot precisely describe complex trajectories, and learned conditions often fail to follow the intended path (e.g., the “turn to back” instruction). Moreover, training requires paired data, which is expensive and scarce. **(B) Explicit Warping:** These methods warp observed pixels to target viewpoints using estimated depth. However, occluded regions remain unfilled, producing visible holes (red circles show “Half Camel”). **(C) Ours:** We reconstruct a foreground-complete 4D proxy that recovers both visible and occluded foreground surfaces, then render it from target viewpoints as geometric guidance for video generation. This enables precise camera control with complete visibility from arbitrary viewpoints (green circle shows “Complete Camel”). The table summarizes the key trade-offs. ✓ and × denote full and no support, respectively.

2023]; enabling free-viewpoint camera redirection from a single video would democratize these experiences for everyday capture.

However, camera redirection from a single monocular video is inherently ill-posed. A monocular video captures only a narrow spatio-temporal view of a dynamic 3D scene, providing highly partial observations of the 4D world. The key challenge is to recover a complete and coherent plenoptic representation from this limited input, with consistent geometry and coherent motion. While prior methods can interpolate nearby novel views by reconstructing observed surfaces [Kerbl et al. 2023; Mildenhall et al. 2021; Tian et al. 2023], they break down under large-angle viewpoint changes far from the original camera trajectory. In these regimes, the absence of visual grounding introduces severe geometric ambiguity, fundamentally limiting replay from arbitrary viewpoints.

To tackle the challenge, recent studies [Bai et al. 2025; Gu et al. 2025; Hou and Chen 2024; Hu et al. 2025; Huang et al. 2025a; Li et al. 2025e; Luo et al. 2025b; Ren et al. 2025; You et al. 2024; Yu et al. 2025, 2024; Zhang et al. 2025a; Zheng et al. 2026] try to explore the generative power of video foundation models for camera redirection. As shown in Fig. 2, these methods generally follow two paradigms: implicit control and explicit warping. Implicit control methods [Bai et al. 2025; He et al. 2024; Hou and Chen 2024] encode camera trajectories as learned embeddings [Bai et al. 2025; He et al. 2024; Hou and Chen 2024] or text prompts [Wang et al. 2025a] and inject them into video diffusion models; however, such implicit representations offer only soft controllability and require expensive 4D training data. Explicit warping methods [Hu et al. 2025; Ren et al. 2025; Seo et al. 2024; Yu et al. 2025] instead estimate depth and warp observed pixels to target viewpoints, but since only visible surfaces are available, occluded regions produce unfilled holes that must be hallucinated by downstream generators [Wang et al. 2025a; Yang et al. 2024]. Neither paradigm achieves both precise camera control and complete visibility of unseen regions, both essential for faithful large-angle redirection.

In this paper, we propose FreeOrbit4D, an effective training-free framework for arbitrary camera redirection from a single monocular video via foreground-complete 4D reconstruction. Our key idea is to explicitly reason about and recover a complete 4D geometry from partial monocular observations, and to use this geometry as structural guidance for generative rendering at novel viewpoints. By completing unseen geometry, our method resolves the ambiguity that causes artifacts in prior work, enabling consistent synthesis even when the target trajectory deviates significantly from the original viewpoints.

Specifically, we observe that reconstructing dynamic scenes from monocular video and completing occluded object geometry are fundamentally different tasks: the former requires temporally consistent scene-level reasoning, while the latter demands multi-view understanding of object shape. This motivates handling them in distinct representation spaces: we unproject the source video into a global scene space with geometrically incomplete foreground, and complete the foreground geometry via multi-view synthesis in canonical object space. While this separation yields tractable sub-tasks addressable by existing techniques [Wang et al. 2025b; Xie et al. 2024; Zhou et al. 2025b], the key challenge lies in unifying their outputs into a coherent whole. Our correspondence-aware alignment addresses this by bridging canonical and global spaces through dense pixel-synchronized 3D–3D correspondences, producing a unified foreground-complete 4D proxy without task-specific training. To interface this proxy with a video generator, we distill it into view-dependent depth maps as conditioning [Wang et al. 2025a]. While depth is not full geometry, it compactly captures the metric layout and visibility cues that enforce cross-view and temporal consistency. This conditioning enables faithful large-angle view synthesis and paves the way for potential 4D data generation.

Our contributions are summarized as follows:

- We present a training-free method that achieves foreground-complete 4D reconstruction from monocular video by combining global scene lifting with object geometry completion through dense 3D correspondences.
- Building on this reconstruction, we propose a camera redirection framework tailored for large-angle view synthesis, enabling expansive camera motions with strong geometric consistency across time and viewpoints.
- Extensive experiments, including a user study, validate state-of-the-art performance under challenging large-angle trajectories. The explicit 4D representation further supports downstream applications by providing consistent geometry for video editing, editable structure for 4D manipulation, and dense annotations for potential 4D data generation.

2 Related Work

Camera-controlled video generation. With the rapid development of video foundation models [Blattmann et al. 2023; Kong et al. 2024], research has progressed from pure text-to-video generation toward controllable video synthesis with multiple conditioning signals [Bahmani et al. 2024; Wang et al. 2024; Zhang et al. 2023]. Among these directions, camera-controlled video generation has attracted increasing attention due to its relevance to downstream video creation applications. Early approaches encode camera intrinsics and extrinsics as control signals [Van Hoorick et al. 2024; Wang et al. 2024], but subsequent studies show that such low-dimensional parameters are difficult to exploit effectively in high-dimensional attention-based diffusion models. This has motivated alternative camera representations, including Plücker coordinates [Fan et al. 2025; He et al. 2024; Zhang et al. 2025c], PRope [Huang et al. 2025b; Li et al. 2025b], and per-token trajectory attention [Xiao et al. 2024], together with diverse control injection mechanisms [Bai et al. 2025; He et al. 2024; Huang et al. 2025b; Luo et al. 2025b; Wang et al. 2024] and decoupled time–camera conditioning [Wang et al. 2025d]. This paradigm has been further extended to user-provided source videos: ReCapture [Zhang et al. 2025a], SpaceTimePilot [Huang et al. 2025a], and CAT4D [Wu et al. 2025a] re-render an input video along new camera trajectories via per-video fine-tuning or learned camera-motion disentanglement. Despite promising results enabled by powerful diffusion priors, methods in this family often struggle to faithfully follow prescribed camera trajectories (Fig. 3), especially under large-angle redirection, due to the absence of explicit, dense 4D geometric conditioning. In contrast, our training-free approach achieves precise camera redirection by reconstructing and completing 4D scene geometry.

Reconstruction-grounded 4D generation. A second line of work uses explicit 3D or 4D geometry to ground camera-redirection video generation. Methods either inject lightweight cues such as rendered geometry buffers [Gu et al. 2025; Zheng et al. 2026] or sparse 3D point tracks [Lee et al. 2025] into video diffusion, or reconstruct a full 4D proxy of the scene from the source video [Cao et al. 2024; Chen et al. 2025; Hou and Chen 2024; Li et al. 2025e; Liu et al. 2025; Song et al. 2025; Tang et al. 2026; Tian et al. 2025; Yu et al. 2025, 2024; Zhang et al. 2025b]. Pipelines that reconstruct a full 4D proxy rely on varying reconstruction backbones—3D representations such

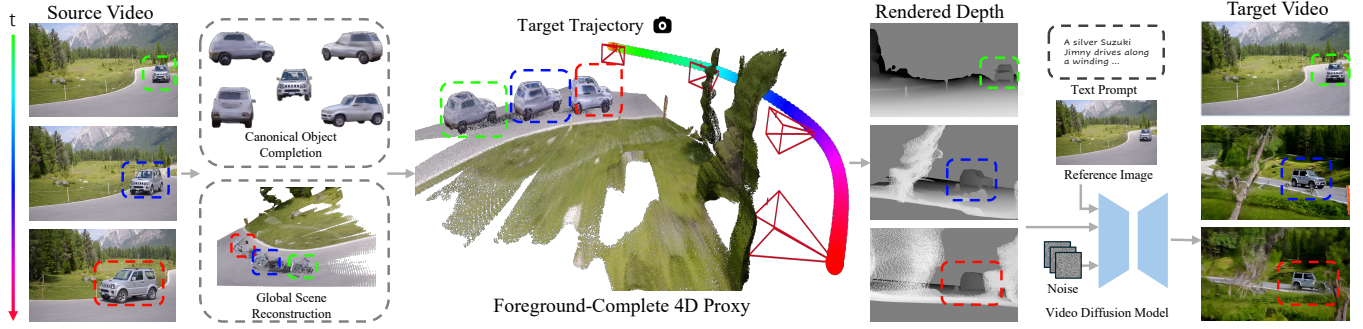


Fig. 3. **Overview of FreeOrbit4D.** Our framework redirects a monocular video to a target camera trajectory via a foreground-complete 4D proxy. This proxy is constructed through two branches: *Global Scene Reconstruction* recovers background and partial foreground in global space, while *Canonical Object Completion* reconstructs complete foreground geometry via multi-view synthesis. After alignment, we render depth maps from the unified 4D proxy to condition a video diffusion model, along with a reference image and text prompt, enabling faithful novel-view synthesis under large viewpoint changes.

as NeRF [Mildenhall et al. 2021; Tian et al. 2023, 2024] and Gaussian Splatting [Kerbl et al. 2023], or 4D representations including motion scaffolds and dynamic Gaussians [He et al. 2025; Lei et al. 2025; Liao et al. 2025; Luo et al. 2025a; Wang et al. 2025c], video-native geometric estimators [Huang et al. 2025c; Li et al. 2025a; Sucar et al. 2026; Wang et al. 2025e; Yao et al. 2025b; Zhang et al. 2024; Zhu et al. 2026], and feed-forward 4D predictors [Jiang et al. 2025c; Lin et al. 2025]. Partial novel views are then rendered from the reconstructed geometry along the target trajectory, and the resulting unfilled holes are hallucinated by a downstream video diffusion model. However, the reconstructed geometry captures only visible surfaces from the source video, leaving the occluded foreground surfaces unrecovered in 3D. Under large-angle camera redirection, these unrecovered regions dominate the target view, and pixel-level hallucination without 3D support cannot synthesize them consistently, producing incomplete foreground and visible holes (Fig. 3). A complementary thread of 4D generative models instead synthesizes 4D content end-to-end. Object-only variants [Chen et al. 2026; Li et al. 2024b; Ren et al. 2023; Zeng et al. 2024] are confined to a single dynamic object, while scene-level variants, including multi-view video diffusion [Li et al. 2024a; Lu et al. 2025; Mi et al. 2025; Yuan et al. 2025], native 4D synthesis [Li et al. 2025d; Yang et al. 2026], and action-conditioned scene generation [Li et al. 2025c; Liu et al. 2026; Zhan et al. 2026], do not ingest a user-provided source video and therefore cannot redirect existing footage along a prescribed camera trajectory. In contrast, our method reconstructs a foreground-complete 4D proxy via correspondence-aware alignment, recovering both visible and occluded foreground surfaces to enable precise, geometry-grounded redirection of user-provided monocular videos.

3 Method

In this work, we study the camera redirection task: Given a monocular source video $\mathcal{V}^{src}$ and a target camera trajectory, our goal is to synthesize a visually faithful and temporally consistent target video $\mathcal{V}^{tgt}$ that depicts the same dynamic scene as $\mathcal{V}^{src}$ but viewed from the target trajectory.

Fig. 3 shows our three-stage framework. First, we reconstruct a static background and geometrically incomplete foreground in

global scene space, and complete foreground geometry in canonical object space (Sec. 3.1). Second, we align the canonical foreground to the global scene via dense pixel-synchronized 3D–3D correspondences, yielding a unified foreground-complete 4D proxy (Sec. 3.2), from which we render depth scaffolds along the target trajectory and synthesize the output with a video diffusion model (Sec. 3.3).

3.1 Decoupled 4D Reconstruction

Since the static scene and the moving object exhibit fundamentally different geometric and motion characteristics, it is natural to decouple their reconstructions. We therefore reconstruct geometry in two complementary coordinate spaces. In **global scene space**, we recover the static background point cloud $\mathcal{P}^{bg}$ and partial foreground point clouds $\tilde{\mathcal{P}}_t^{fg}$, which are lifted from the source frames and thus capture only visible surfaces. In **canonical object space**, we reconstruct complete foreground point clouds $\hat{\mathcal{P}}_t^{fg}$ that capture the full 3D structure but lack alignment to the global scene. Throughout, we use tilde ($\tilde{\cdot}$) to denote geometrically incomplete representations in global scene space, and hat ($\hat{\cdot}$) to denote geometrically complete but unaligned representations in canonical object space. These are unified in Sec. 3.2.

Global Scene Reconstruction. VGGT [Wang et al. 2025b] is a feed-forward model that predicts 3D point maps from images, but assumes a static scene. To handle dynamic videos, we adopt its temporally-aware extension [Zhou et al. 2025b], which processes the entire monocular sequence $\mathcal{V}^{src}$ and predicts temporally consistent point maps $\tilde{\mathcal{P}}_t \in \mathbb{R}^{H \times W \times 3}$ for each frame, all registered in a unified global coordinate system.

We then use semantic masks $\mathbf{M}_t$ from SAM2 [Ravi et al. 2024] to separate background and foreground, initiated by a user-provided point prompt on the first frame, or fully automated via detector-based prompting [Ren et al. 2024]. The foreground-background decomposition is:

$$\mathcal{P}^{bg} = \bigcup_{t=1}^T \{\tilde{\mathcal{P}}_t(\mathbf{u}) \mid \mathbf{M}_t(\mathbf{u}) = 0\}, \quad \tilde{\mathcal{P}}_t^{fg} = \{\tilde{\mathcal{P}}_t(\mathbf{u}) \mid \mathbf{M}_t(\mathbf{u}) = 1\}. \quad (1)$$

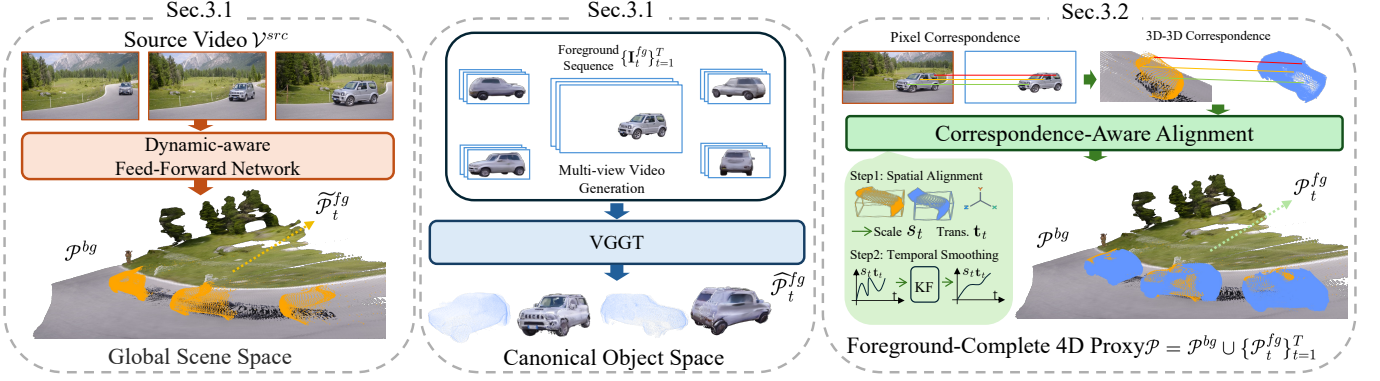


Fig. 4. **Decoupled 4D reconstruction and alignment pipeline.** **Left (Sec. 3.1):** A dynamic-aware feed-forward network lifts the source video $\mathcal{V}^{src}$ into *global scene space*, producing the static background $\mathcal{P}^{bg}$ and geometrically incomplete foreground $\tilde{\mathcal{P}}_t^{fg}$ (orange). **Middle (Sec. 3.1):** The foreground sequence $\{\mathbf{I}_t^{fg}\}_{t=1}^T$ is fed into an object-centric video diffusion model to synthesize multi-view images, from which VGGT reconstructs complete foreground geometry $\hat{\mathcal{P}}_t^{fg}$ in *canonical object space*. **Right (Sec. 3.2):** Dense 3D-3D correspondences derived from pixel-synchronized point maps enable *correspondence-aware alignment*: per-frame spatial alignment estimates scale s_t and translation $\mathbf{t}_t$, followed by temporal smoothing via Kalman filtering. The result is the unified foreground-complete 4D proxy $\mathcal{P} = \mathcal{P}^{bg} \cup \{\mathcal{P}_t^{fg}\}_{t=1}^T$.

Here, $\mathbf{u}$ denotes pixel coordinates, and $\tilde{\mathcal{P}}_t^{fg}$ indicates partial foreground point clouds in global space, as only visible surfaces are captured from the source view (orange in Fig. 4).

Canonical Object Completion. To resolve the inherent single-view ambiguity in foreground geometry, we leverage a multi-view video diffusion model [Yao et al. 2025a] to synthesize novel views. Since this model is object-centric and requires isolated object inputs, we first extract the foreground from each frame $\mathbf{I}_t$ of the source video $\mathcal{V}^{src}$ using the semantic masks, yielding the masked foreground sequence $\{\mathbf{I}_t^{fg}\}_{t=1}^T$. The model then synthesizes four temporally synchronized novel-view videos at 90° azimuthal intervals, denoted as $\{\mathbf{I}_t^{(k)}\}_{t=1}^T$ for $k = 1, \dots, 4$.

We then use VGGT [Wang et al. 2025b] to reconstruct multi-view point maps from all five views (source + four synthesized) per frame:

$$\hat{\mathbf{P}}_t^{fg} = \Phi_{\text{VGGT}}(\mathbf{I}_t^{fg}, \mathbf{I}_t^{(1)}, \dots, \mathbf{I}_t^{(4)}), \quad \hat{\mathbf{P}}_t^{fg} \in \mathbb{R}^{5 \times H \times W \times 3} \quad (2)$$

where the source view $\mathbf{I}_t^{fg}$ serves as the reference. $\hat{\mathbf{P}}_t^{fg}$ resides in canonical object space and requires alignment to the global scene.

Since the synthesized multi-view images contain white backgrounds, the raw point maps include redundant points. We filter them using binary masks: the SAM2 mask $\mathbf{M}_t$ for the source view, and color thresholding masks $\{\mathbf{M}_t^{(k)}\}_{k=1}^4$ for the novel views. Let $\hat{\mathbf{P}}_t$ and $\hat{\mathbf{P}}_t^{(k)}$ denote the point maps for the source and k -th novel view, respectively. The complete canonical foreground point cloud is:

$$\hat{\mathcal{P}}_t^{fg} = \{\hat{\mathbf{P}}_t(\mathbf{u}) \mid \mathbf{M}_t(\mathbf{u}) = 1\} \cup \bigcup_{k=1}^4 \{\hat{\mathbf{P}}_t^{(k)}(\mathbf{u}) \mid \mathbf{M}_t^{(k)}(\mathbf{u}) = 1\} \quad (3)$$

3.2 Correspondence-Aware Alignment

We now have partial foreground point clouds $\tilde{\mathcal{P}}_t^{fg}$ in global scene space, and complete canonical foreground point clouds $\hat{\mathcal{P}}_t^{fg}$ in

canonical object space. Our goal is to align $\hat{\mathcal{P}}_t^{fg}$ to global scene space, yielding the aligned foreground point clouds $\mathcal{P}_t^{fg}$, which together with the static background $\mathcal{P}^{bg}$ form the unified foreground-complete 4D proxy $\mathcal{P} = \mathcal{P}^{bg} \cup \{\mathcal{P}_t^{fg}\}_{t=1}^T$.

Per-frame Spatial Alignment. We estimate a per-frame transformation $\mathcal{T}_t$ such that $\mathcal{P}_t^{fg} = \mathcal{T}_t(\hat{\mathcal{P}}_t^{fg})$, using $\hat{\mathcal{P}}_t^{fg}$ as the spatial anchor. A key observation is that both $\hat{\mathbf{P}}_t$ and $\hat{\mathbf{P}}_t^{(k)}$ originate from the same source image $\mathbf{I}_t$, so points at the same pixel $\mathbf{u}$ correspond to the same surface point. This yields dense 3D-3D correspondences:

$$C_t = \{(\hat{\mathbf{P}}_t(\mathbf{u}), \hat{\mathbf{P}}_t^{(k)}(\mathbf{u})) \mid \mathbf{M}_t(\mathbf{u}) = 1\}, \quad (4)$$

which we filter by confidence and outlier removal.

Given C_t , a natural approach would be to minimize point-wise distances. However, monocular lifting cannot determine absolute depth scale, which may vary across frames, causing per-point inconsistencies in $\tilde{\mathcal{P}}_t^{fg}$. While point-wise fitting would propagate these errors, the overall spatial extent and location of $\tilde{\mathcal{P}}_t^{fg}$ within the scene remain reliable. In contrast, $\hat{\mathcal{P}}_t^{fg}$ provides accurate object geometry from multi-view reconstruction but lacks scene-level positioning. We therefore use $\hat{\mathcal{P}}_t^{fg}$ only to determine the global placement of the object (position and scale), while preserving the geometry of $\hat{\mathcal{P}}_t^{fg}$. Specifically, we parameterize $\mathcal{T}_t$ as: $s_t \mathbf{x} + \mathbf{t}_t$, where $(s_t, \mathbf{t}_t)$ are estimated from C_t such that the transformed geometry matches the scale and location of $\tilde{\mathcal{P}}_t^{fg}$. Rotation is fixed since both reconstructions use the source view as coordinate reference.

Temporal Smoothing. The geometrically incomplete foreground $\tilde{\mathcal{P}}_t^{fg}$ may exhibit frame-to-frame depth inconsistency due to the inherent ambiguity of monocular lifting. Since we use it to anchor global placement, this inconsistency propagates into the aligned



Fig. 5. **Qualitative comparison on the “Swing” sequence.** The top-left shows the target camera trajectory orbiting around the scene. This challenging scenario involves rapid foreground motion and thin structures (swing ropes). ReCamMaster [Bai et al. 2025] and EX-4D [Hu et al. 2025] fail to preserve human body structure, producing blurred limbs and ghosting artifacts. TrajectoryCrafter [Yu et al. 2025], GEN3C [Ren et al. 2025], and CogNVS [Chen et al. 2025] exhibit noticeable geometric distortions and semantic drift as the camera rotates. In contrast, our method maintains sharp details and stable geometry throughout the sequence, benefiting from the foreground-complete 4D proxy. More visualizations are provided in the supplementary material.

Table 1. **Quantitative comparison and User Study.** We report VBench for perceptual video quality, DINO/CLIP-SIM for semantic similarity, FID-V/FVD-V for distributional fidelity, and user ratings (1–5 scale). **Bold**: best; underline: second-best.

Method	VBench $\uparrow$						Similarity & Fidelity				User Study		
	Subject Consis.	BG Consis.	Motion Smooth.	Overall Consis.	Aesth. Qual.	Imaging Qual.	DINO-SIM ($\uparrow$)	CLIP-SIM ($\uparrow$)	FID-V ($\downarrow \times 10^2$)	FVD-V ($\downarrow \times 10^3$)	Overall ($\uparrow$)	Motion ($\uparrow$)	Stab. ($\uparrow$)
ReCamMaster	<u>0.84</u>	<u>0.92</u>	0.98	0.16	0.39	43	0.37	0.75	2.6	3.9	2.0	2.5	2.0
TrajectoryCrafter	0.80	0.91	0.94	<u>0.19</u>	<u>0.47</u>	<u>53</u>	<u>0.47</u>	<u>0.79</u>	<u>2.0</u>	3.6	<u>2.8</u>	3.2	<u>2.9</u>
EX-4D	0.76	0.89	0.94	0.16	0.42	46	0.28	0.69	3.2	3.8	2.0	2.5	2.0
GEN3C	0.79	0.88	0.95	0.18	0.42	49	0.43	0.75	2.3	3.3	2.4	<u>3.5</u>	2.3
CogNVS	0.82	0.91	<u>0.97</u>	0.15	0.42	<u>53</u>	0.34	0.78	3.0	<u>3.4</u>	2.0	3.2	1.9
Ours	0.88	0.94	0.96	0.24	0.52	64	0.65	0.84	1.7	3.6	4.6	4.5	4.5

foreground trajectory. To compensate, we smooth the centroid trajectory of aligned point clouds using a bidirectional Kalman filter with a constant-velocity motion model, with stronger regularization along depth. The per-frame scale is preserved for geometric fidelity, yielding a temporally coherent sequence $\{\mathcal{P}_t^{fg}\}_{t=1}^T$.

3.3 Geometry-conditioned Video Synthesis

Given the foreground-complete 4D proxy $\mathcal{P} = \mathcal{P}^{bg} \cup \{\mathcal{P}_t^{fg}\}_{t=1}^T$ and a target camera trajectory $\{\pi_t^{tgt}\}_{t=1}^T$, we render depth scaffolds to guide video synthesis:

$$\mathcal{V}^{tgt} = \Phi_{\text{VDM}} \left(\mathbf{I}_1, \left\{ \text{Render} \left(\mathcal{P}, \pi_t^{tgt} \right) \right\}_{t=1}^T, \mathbf{c} \right), \quad (5)$$

where Φ_{VDM} is a depth-conditioned video diffusion model [Jiang et al. 2025a; Wang et al. 2025a], $\mathbf{I}_1$ is the first frame of $\mathcal{V}^{src}$ as appearance reference, and $\mathbf{c}$ is a text prompt. Depth maps form a compact yet informative geometric scaffold: they encode the 3D layout from the target viewpoint, allowing video diffusion to generate spatially coherent content that faithfully follows the prescribed camera motion.

4 Experiments and Applications

We evaluate FreeOrbit4D on diverse real-world and synthetic videos under challenging large-angle camera trajectories. After describing implementation details, datasets, and metrics, we compare against

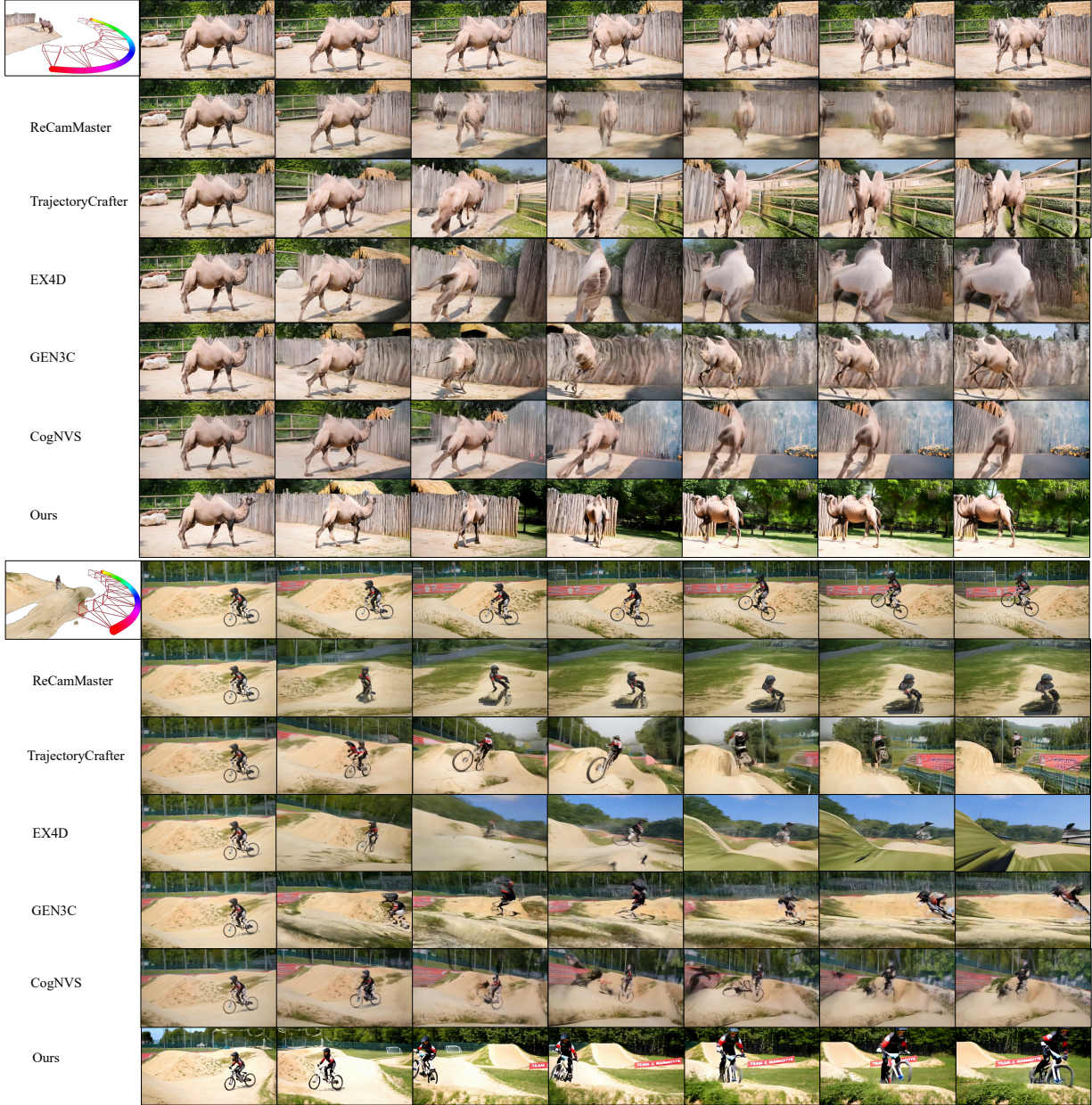


Fig. 6. **Additional qualitative comparisons on “Camel” (top) and “BMX” (bottom).** Compared to baselines that suffer from geometric distortions, motion blur, and semantic drift under large viewpoint changes, our method produces consistently sharp textures and stable geometry across diverse scenes, demonstrating the effectiveness of our foreground-complete 4D proxy.

state-of-the-art methods (Sec. 4.1), demonstrate practical applications (Sec. 4.2), and provide ablation studies (Sec. 5).

Implementation Details. Our pipeline is training-free and relies on off-the-shelf pretrained models: PAGE-4D [Zhou et al. 2025b] for global scene reconstruction, SAM2 [Ravi et al. 2024] for foreground segmentation, SV4D2.0 [Yao et al. 2025a] for multi-view

video synthesis, VGGT [Wang et al. 2025b] for multi-view point-map reconstruction, and Wan2.2-VACE [Jiang et al. 2025a; Wang et al. 2025a] for depth-conditioned video synthesis. All experiments use a single NVIDIA A40 GPU, processing each 45-frame clip at 832×480 in roughly 50 minutes end-to-end.

Evaluation Datasets. We evaluate on a diverse set of real-world and synthetic videos. For real-world data, we use sequences from

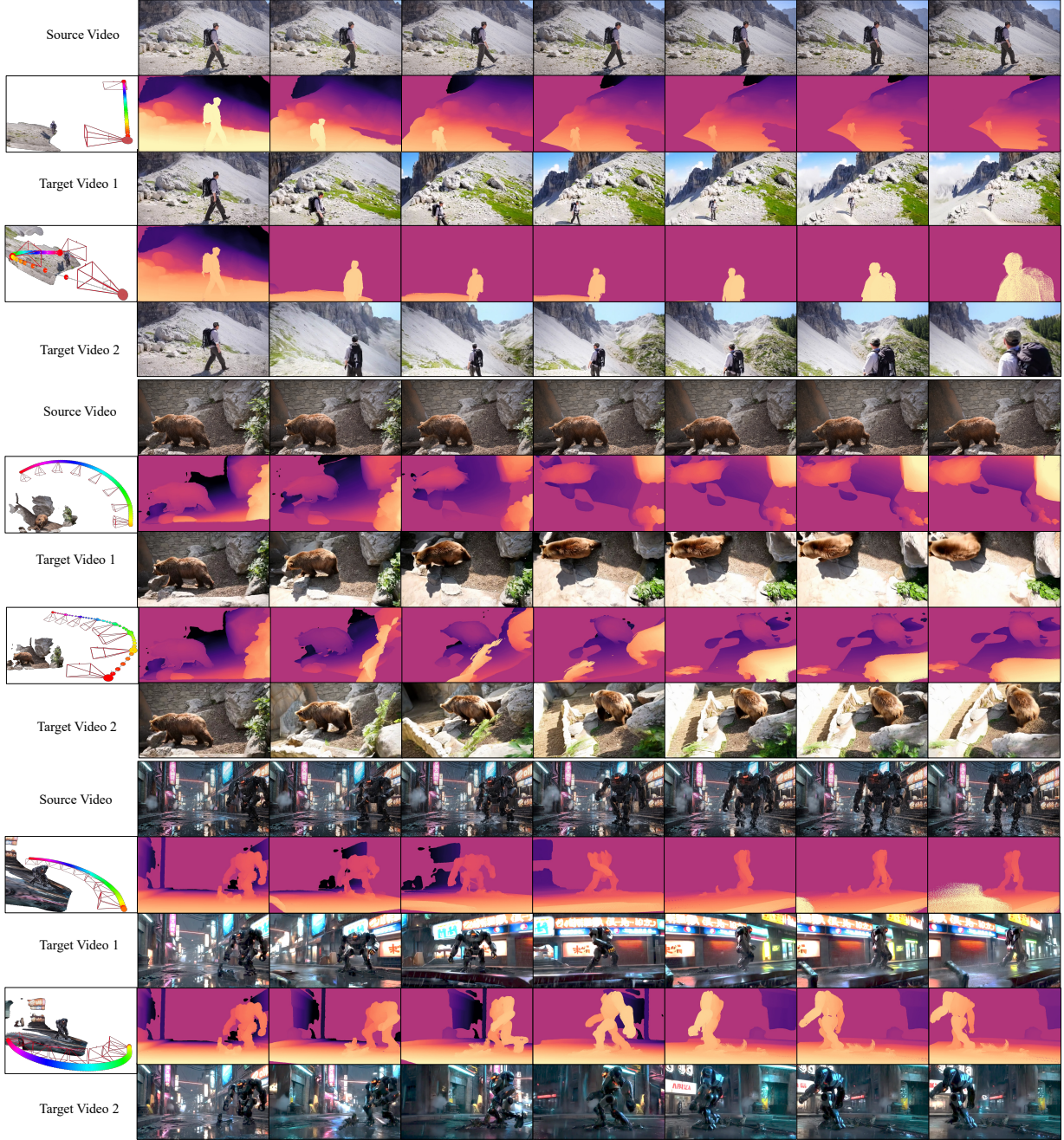


Fig. 7. **Multi-trajectory video synthesis results on “Hike” (top), “Bear” (middle), and “Robot” (bottom).** We show that our method generates temporally and geometrically consistent videos along diverse target camera trajectories. By leveraging the foreground-complete 4D proxy, our approach effectively handles articulated motion, complex lighting, and thin structures across all sequences.

DAVIS [Perazzi et al. 2016], which provides high-quality monocular videos with varied dynamic foregrounds, as well as publicly available online videos (e.g., Unitree robot demos, LeCun interview footage) that feature complex real-world motion and cluttered backgrounds. For synthetic data, we include sequences generated by VEO [Google

DeepMind 2024] and Sora [OpenAI 2025], covering diverse scenes and motion patterns. To evaluate large-angle camera redirection, we define target trajectories with extreme yaw or pitch rotations (e.g., 120° or 180°) from the initial viewpoint.

Evaluation Metrics. We measure visual quality with FID-V [Heusel et al. 2017] and FVD-V [Unterthiner et al. 2019], semantic consistency with CLIP-SIM [Radford et al. 2021] and DINO-SIM [Oquab et al. 2023], and perceptual video quality with VBench [Huang et al. 2024]. A user study further evaluates overall preference, camera accuracy, and temporal stability.

4.1 Comparison with State-of-the-Art Methods

Baselines. We compare our method with state-of-the-art camera-controlled video-to-video generation methods, including ReCamMaster [Bai et al. 2025], TrajectoryCrafter [Yu et al. 2025], EX-4D [Hu et al. 2025], GEN3C [Ren et al. 2025], and CogNVS [Chen et al. 2025]. ReCamMaster [Bai et al. 2025] conditions a video generator on target camera poses for single-video re-rendering. TrajectoryCrafter [Yu et al. 2025] uses diffusion with point-cloud rendering guidance to follow a user-specified camera trajectory. EX-4D [Hu et al. 2025] leverages a depth-watertight mesh to improve extreme-viewpoint synthesis. GEN3C [Ren et al. 2025] maintains and renders a 3D point-cloud cache to enforce 3D-consistent camera control. CogNVS [Chen et al. 2025] completes disoccluded regions in rendered novel views via test-time fine-tuned video diffusion.

Qualitative Results. We present qualitative comparisons in Fig. 5, Fig. 6, and Fig. 7. In the “Swing” scenario (Fig. 5), which involves rapid foreground motion and thin structures (swing ropes), TrajectoryCrafter, GEN3C, and CogNVS [Chen et al. 2025] exhibit geometric distortions and semantic drift as the camera rotates. ReCamMaster and EX-4D fail to preserve the structural integrity of the human body, producing blurred limbs or ghosting artifacts under large viewpoint changes. In contrast, our method generates high-fidelity frames with sharp details and stable geometry by leveraging the foreground-complete 4D proxy to resolve single-view ambiguity. Fig. 6 provides additional comparisons across diverse scenes, and Fig. 7 showcases results under multiple user-specified trajectories, highlighting the flexibility of our trajectory control.

Quantitative Results. Tab. 1 summarizes the quantitative comparison across automatic metrics and user ratings. On VBench, our method ranks first in five out of six dimensions, achieving the highest subject consistency (0.88), background consistency (0.94), overall consistency (0.24), aesthetic quality (0.52), and imaging quality (64). Motion smoothness (0.96) ranks behind ReCamMaster (0.98) and CogNVS (0.97), both of which tend to produce over-smoothed results at the cost of geometric detail. For semantic similarity and distributional fidelity, we achieve the best DINO-SIM (0.65), CLIP-SIM (0.84), and FID-V (1.7×10^2), with competitive FVD-V (3.6×10^3). Despite these strong results, we observe that automatic metrics do not fully capture trajectory control quality. As shown in Fig. 5, methods such as ReCamMaster and TrajectoryCrafter achieve reasonable scores while still deviating from the prescribed path or losing geometric coherence, motivating our user study.

User Study. We conducted a user study with 20 participants across 10 diverse sequences. Participants rated results on a 1–5 scale along three axes (see the questionnaire interface in Fig. 8): (1) overall preference, (2) camera motion accuracy (adherence to the prescribed trajectory), and (3) temporal stability and source consistency (flicker

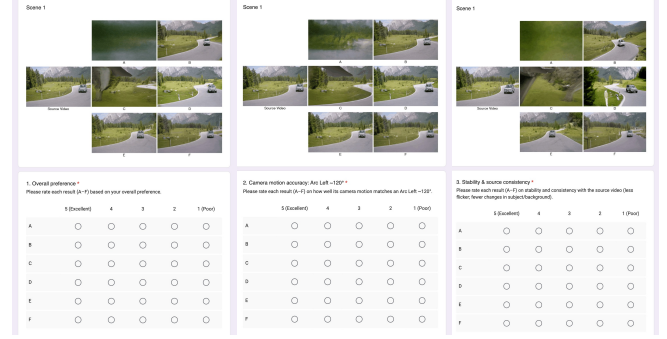


Fig. 8. **User study interface.** We present participants with two anonymized videos (ours vs. baseline) and ask them to select the one with superior temporal stability and geometric fidelity.



Fig. 9. **Applications enabled by FreeOrbit4D.** Our explicit 4D representation enables various downstream applications. **Top:** Appearance editing—given a single edited reference frame (e.g., zebra pattern or anime style), our foreground-complete proxy propagates the edit consistently across all novel viewpoints. **Bottom:** Geometry editing—by directly manipulating the point cloud (scaling or compositing objects from different sources), we synthesize plausible redirected videos from the modified 4D geometry.

and identity preservation). As shown in Tab. 1, our method outperforms all baselines across all axes, with particularly large margins in motion accuracy (4.5 vs. 3.5 for the next best method). This validates that our foreground-complete 4D proxy provides precise camera control and structural stability that traditional metrics fail to capture, yielding an overall preference of 4.6.

4.2 Applications

As shown in Fig. 9, FreeOrbit4D also enables several practical and interesting applications beyond camera redirection, thanks to our explicit and foreground-complete 4D representation.

Appearance Propagation. Our framework propagates appearance edits from a single reference frame to the entire redirected video. We

Table 2. **Ablation study.** We progressively add multi-view generation (MVG) and Kalman filter smoothing (KF) to evaluate each component’s contribution. **Bold:** best.

Method	DINO-SIM ($\uparrow$)	CLIP-SIM ($\uparrow$)	FID-V ($\downarrow \times 10^2$)	FVD-V ($\downarrow \times 10^3$)
Baseline	0.58	0.81	1.9	4.1
+ MVG	0.60	0.82	1.9	4.1
+ KF (Full)	0.65	0.84	1.7	3.6

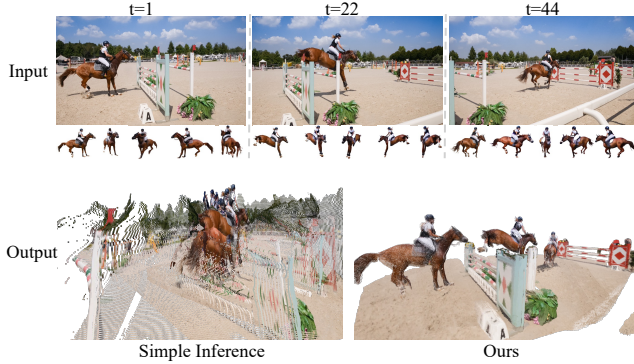


Fig. 10. **Ablation: Simple Inference vs. Ours.** Directly combining multi-view images with source video leads to ghosting artifacts (left). Our decoupled strategy produces coherent 4D reconstruction (right).

edit the reference frame with Qwen-Image-Edit [Wu et al. 2025b], applying color or style changes. The foreground-complete 4D proxy then provides consistent depth scaffolds, enabling faithful appearance transfer across all viewpoints.

4D Geometry Manipulation. Our explicit point-based 4D representation enables direct geometric manipulation in space-time. By modifying the point cloud, such as scaling the foreground or compositing point clouds from different sources into a shared scene, we can synthesize plausible redirected videos from the edited 4D geometry. This highlights the flexibility of explicit 4D representations over purely implicit approaches for scene manipulation.

5 Ablation Study

We conduct ablation studies on each key component under the same setting as our main experiments. Tab. 2 and Fig. 10 report the quantitative and qualitative results, respectively.

Simple Inference. A straightforward baseline directly feeds the multi-view images together with the source video into a dynamic-aware feed-forward reconstruction network, bypassing our decoupled strategy. However, as shown in Fig. 10, this naive combination fails to produce a coherent 4D reconstruction: since the same object across frames exhibits similar appearance, the network collapses correspondences across time, producing misalignment and ghosting.

Without Multi-View Generation (MVG). We remove the multi-view generation step and rely solely on partial foreground point clouds for guidance. As shown in Tab. 2, performance degrades across all metrics, indicating that multi-view generation provides

complete foreground geometry that is critical for stable synthesis under large viewpoint changes.

Without Kalman Filter (KF). We disable temporal smoothing. As discussed in Sec. 3.2, monocular lifting produces frame-to-frame depth inconsistency that propagates into the aligned foreground trajectory. The performance drop upon removing the Kalman filter confirms its necessity for coherence.

6 Limitations

Our method has three main limitations. First, the pipeline assumes a single dominant foreground object and a mostly static background. It can handle multiple objects when each is sufficiently observed over time, but complex interactions with heavy mutual occlusions or dynamic backgrounds remain challenging for current object-centric models. Second, as a modular system built on off-the-shelf components, upstream errors in segmentation or multi-view synthesis can propagate to downstream stages. The modular design, however, enables independent component upgrades as stronger models emerge. Third, the multi-stage pipeline prioritizes redirection quality over runtime efficiency, incurring substantial computational cost.

7 Conclusion

We present **FreeOrbit4D**, a training-free framework for camera redirection from a single monocular video. By reconstructing a foreground-complete 4D proxy as geometric scaffolds for video generation, our method achieves SOTA fidelity, temporal coherence, and camera control under large viewpoint changes, while enabling applications such as appearance propagation and scene manipulation. Future work may explore integrating learned priors for improved robustness and leveraging the recovered 4D proxy for downstream tasks such as 4D asset creation and synthetic data generation.

Acknowledgments

This research is supported by the National Artificial Intelligence Research Resource (NAIRR) Pilot under award NAIRR250199. Computational resources are also provided by Delta and DeltaAI at the National Center for Supercomputing Applications (NCSA) through ACCESS allocations CIS250012, CIS250816, and CIS251188. S. W. is also supported by NSF Awards #2525287, #2404385, #2414227, #2340254, #2312102, and #2331878, and research grants from IBM, Meta, NVIDIA, and Intel.

References

- Ronald T Azuma. 1997. A survey of augmented reality. *Presence: teleoperators & virtual environments* 6, 4 (1997), 355–385.
- Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. 2024. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781* (2024).
- Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. 2025. Recammaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14834–14844.
- Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).

- Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, et al. 2025. Pseudo-simulation for autonomous driving. *arXiv preprint arXiv:2506.04218* (2025).
- Wei Cao, Chang Luo, Biao Zhang, Matthias Nießner, and Jiapeng Tang. 2024. Motion2vecsets: 4d latent vector set diffusion for non-rigid shape reconstruction and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20496–20506.
- David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. 2024. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 19457–19467.
- Hongyuan Chen, Xingyu Chen, Youjia Zhang, Zexiang Xu, and Anpei Chen. 2026. Motion 3-to-4: 3D Motion Reconstruction for 4D Synthesis. *arXiv preprint arXiv:2601.14253* (2026).
- Kaihua Chen, Tarasha Khurana, and Deva Ramanan. 2025. Reconstruct, Inpaint, Test-Time Finetune: Dynamic Novel-view Synthesis from Monocular Videos. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Yuedong Chen, Haoifei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. 2024. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European conference on computer vision*. Springer, 370–386.
- Xiang Fan, Sharath Girish, Vivek Ramanujan, Chaoyang Wang, Ashkan Mirzaei, Petr Sushko, Aliaksandr Siarohin, Sergey Tulyakov, and Ranjay Krishna. 2025. OmniView: An All-Seeing Diffusion Model for 3D and 4D View Synthesis. *arXiv preprint arXiv:2512.10940* (2025).
- Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. 2021. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5712–5721.
- Google DeepMind. 2024. Veo. <https://deepmind.google/models/veo/>. Accessed: 2024-11-05.
- Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. 2025. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Papers*. 1–12.
- Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. 2024. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101* (2024).
- Jixuan He, Chieh Hubert Lin, Lu Qi, and Ming-Hsuan Yang. 2025. Restage4D: Reanimating Deformable 3D Reconstruction from a Single Video. *arXiv preprint arXiv:2508.06715* (2025).
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).
- Chen Hou and Zhibo Chen. 2024. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126* (2024).
- Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. 2025. Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh. *arXiv preprint arXiv:2506.05554* (2025).
- Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. 2025c. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934* (2025).
- Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. 2025b. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)* 44, 6 (2025), 1–15.
- Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21807–21818.
- Zhenning Huang, Hyeonho Jeong, Xuelin Chen, Yulia Gryaditskaya, Tuanfeng Y Wang, Joan Lasenby, and Chun-Hao Huang. 2025a. SpaceTimePilot: Generative Rendering of Dynamic Scenes Across Space and Time. *arXiv preprint arXiv:2512.25075* (2025).
- Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. 2025b. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *ACM Transactions on Graphics (TOG)* 44, 6 (2025), 1–16.
- Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. 2025a. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 17191–17202.
- Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. 2025c. Geo4d: Leveraging video generators for geometric 4d scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20658–20671.
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* 42, 4 (2023), 139–1.
- Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603* (2024).
- Yao-Chih Lee, Zhoutong Zhang, Jiahui Huang, Jui-Hsien Wang, Joon-Young Lee, Jia-Bin Huang, Eli Shechtman, and Zhengqi Li. 2025. Generative Video Motion Editing with 3D Point Tracks. *arXiv preprint arXiv:2512.02015* (2025).
- Jiahui Lei, Yijia Weng, Adam W Harley, Leonidas Guibas, and Kostas Daniilidis. 2025. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 6165–6177.
- Bing Li, Cheng Zheng, Wenxuan Zhu, Jinjie Mai, Biao Zhang, Peter Wonka, and Bernard Ghanem. 2024b. Vivid-zoo: Multi-view video generation with diffusion model. *Advances in Neural Information Processing Systems* 37 (2024), 62189–62222.
- Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, Zhengzhong Tu, et al. 2024a. 4k4dgen: Panoramic 4d generation at 4k resolution. *arXiv preprint arXiv:2406.13527* (2024).
- Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. 2025b. Cameras as relative positional encoding. *arXiv preprint arXiv:2507.10496* (2025).
- Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yehan Xiong, Min Chen, et al. 2025e. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 28785–28796.
- Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. 2025a. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10486–10496.
- Zizhang Li, Hong-Xing Yu, Wei Liu, Yin Yang, Charles Herrmann, Gordon Wetzstein, and Jiajun Wu. 2025c. Wonderplay: Dynamic 3d scene generation from a single image and actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9080–9090.
- Zhibing Li, Mengchen Zhang, Tong Wu, Jing Tan, Jiaqi Wang, and Dahua Lin. 2025d. SS4D: Native 4D Generative Model via Structured Spacetime Latents. *ACM Transactions on Graphics (TOG)* 44, 6 (2025), 1–12.
- Ting-Hsuan Liao, Haowen Liu, Yiran Xu, Songwei Ge, Gengshan Yang, and Jia-Bin Huang. 2025. Pad3r: Pose-aware dynamic 3d reconstruction from casual videos. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. 1–11.
- Chenguo Lin, Yuchen Lin, Panwang Pan, Yifan Yu, Tao Hu, Honglei Yan, Katerina Fragkiadaki, and Yadong Mu. 2025. Movies: Motion-aware 4d dynamic view synthesis in one second. *arXiv preprint arXiv:2507.10065* (2025).
- Tianqi Liu, Zhaoxi Chen, Zihao Huang, Shaocong Xu, Saining Zhang, Chongjie Ye, Bohan Li, Zhiguo Cao, Wei Li, Hao Zhao, et al. 2025. Light-X: Generative 4D Video Rendering with Camera and Illumination Control. *arXiv preprint arXiv:2512.05115* (2025).
- Wei Liu, Ziyu Chen, Zizhang Li, Yue Wang, Hong-Xing Yu, and Jiajun Wu. 2026. RealWonder: Real-Time Physical Action-Conditioned Video Generation. *arXiv preprint arXiv:2603.05449* (2026).
- Dongyue Lu, Ao Liang, Tianxin Huang, Xiao Fu, Yuyang Zhao, Baorui Ma, Liang Pan, Wei Yin, Lingdong Kong, Wei Tsang Ooi, et al. 2025. SEE4D: Pose-Free 4D Generation via Auto-Regressive Video Inpainting. *arXiv preprint arXiv:2510.26796* (2025).
- Yawen Luo, Xiaoyu Shi, Jianhong Bai, Menghan Xia, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. 2025b. Camclonemaster: Enabling reference-based camera control for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. 1–10.
- Zhanpeng Luo, Haoxi Ran, and Li Lu. 2025a. Instant4D: 4D Gaussian Splatting in Minutes. *arXiv preprint arXiv:2510.01119* (2025).
- Zhenxing Mi, Yuxin Wang, and Dan Xu. 2025. One4D: Unified 4D Generation and Reconstruction via Decoupled LoRA Control. *arXiv preprint arXiv:2511.18922* (2025).
- Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.
- OpenAI. 2025. Sora: Creating video from text. <https://openai.com/index/sora/>. Accessed: 2024-11-05.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).
- Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 724–732.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam

- 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024).
- Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. 2023. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142* (2023).
- Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159* (2024).
- Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. 2025. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6121–6132.
- Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungyong Kim, and Yuki Mitsufuji. 2024. Genwarp: Single image to novel views with semantic-preserving generative warping. *Advances in Neural Information Processing Systems* 37 (2024), 80220–80243.
- Chenxi Song, Yanming Yang, Tong Zhao, Ruibo Li, and Chi Zhang. 2025. Worldforge: Unlocking emergent 3d/4d generation in video diffusion model via training-free guidance. *arXiv preprint arXiv:2509.15130* (2025).
- Edgar Sucar, Eldar Insafutdinov, Zihang Lai, and Andrea Vedaldi. 2026. V-DPM: 4D Video Reconstruction with Dynamic Point Maps. *arXiv preprint arXiv:2601.09499* (2026).
- Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. 2023. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 conference proceedings*. 1–12.
- Jiapeng Tang, Wei Cao, Biao Zhang, Chang Luo, Yaoyao Liu, and Matthias Nießner. 2026. Motion2VecSets: Non-Rigid Shape Reconstruction and Tracking with 4D Latent Set Diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2026).
- Fengrui Tian, Tianjiao Ding, Jinqi Luo, Hancheng Min, and Rene Vidal. 2025. Voyaging into Perpetual Dynamic Scenes from a Single View. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7698–7708.
- Fengrui Tian, Shaoyi Du, and Yueqi Duan. 2023. Mononerf: Learning a generalizable dynamic radiance field from monocular videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 17903–17913.
- Fengrui Tian, Yueqi Duan, Angtian Wang, Jianfei Guo, and Shaoyi Du. 2024. Semantic flow: Learning semantic field of dynamic scenes from monocular videos. *arXiv preprint arXiv:2404.05163* (2024).
- Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2019. FVD: A new metric for video generation. (2019).
- Basile Van Hoorick, Rundt Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. 2024. Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *European Conference on Computer Vision*. Springer, 313–331.
- Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Fei Wu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. 2025a. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314* 3, 4 (2025), 6.
- Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. 2025b. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5294–5306.
- Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. 2025c. Shape of motion: 4d reconstruction from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9660–9672.
- Yiming Wang, Qihang Zhang, Shengqu Cai, Tong Wu, Jan Ackermann, Zhengfei Kuang, Yang Zheng, Frano Rajić, Siyu Tang, and Gordon Wetzstein. 2025d. BulletTime: Decoupled Control of Time and Camera Pose for Video Generation. *arXiv preprint arXiv:2512.05076* (2025).
- Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. 2025e. π^3 : Scalable Permutation-Equivariant Visual Geometry Learning. *arXiv e-prints* (2025), arXiv–2507.
- Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. 2024. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.
- Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. 2025b. Qwen-image technical report. *arXiv preprint arXiv:2508.02324* (2025).
- Rundt Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. 2025a. Cat4d: Create anything in 4d with multi-view video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26057–26068.
- Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. 2024. Trajectory attention for fine-grained video motion control. *arXiv preprint arXiv:2411.19324* (2024).
- Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. 2024. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470* (2024).
- Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. 2026. NeoVerse: Enhancing 4D World Model with in-the-wild Monocular Videos. *arXiv preprint arXiv:2601.00393* (2026).
- Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072* (2024).
- Chun-Han Yao, Yiming Xie, Vikram Voleti, Huaizu Jiang, and Varun Jampani. 2025a. Sv4d 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4d generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13248–13258.
- David Yifan Yao, Albert J Zhai, and Shenlong Wang. 2025b. Uni4d: Unifying visual foundation models for 4d modeling from a single video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 1116–1126.
- Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. 2024. Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. *arXiv preprint arXiv:2405.15364* (2024).
- Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. 2025. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 100–111.
- Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. 2024. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048* (2024).
- Yu Yuan, Tharindu Wickremasinghe, Zeeshan Nadir, Xijun Wang, Yiheng Chi, and Stanley H Chan. 2025. SeeU: Seeing the Unseen World via 4D Dynamics-aware Generation. *arXiv preprint arXiv:2512.03350* (2025).
- Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. 2024. Stag4d: Spatial-temporal anchored generative 4d gaussians. In *European Conference on Computer Vision*. Springer, 163–179.
- Jiahao Zhan, Zizhang Li, Hong-Xing Yu, and Jiajun Wu. 2026. PerpetualWonder: Long-Horizon Action-Conditioned 4D Scene Generation. *arXiv preprint arXiv:2602.04876* (2026).
- David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. 2025a. Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2050–2062.
- Hao Zhang, Chun-Han Yao, Simon Donné, Narendra Ahuja, and Varun Jampani. 2025d. Stable Part Diffusion 4D: Multi-View RGB and Kinematic Parts Video Generation. *arXiv preprint arXiv:2509.10687* (2025).
- Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. 2024. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825* (2024).
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3836–3847.
- Songchun Zhang, Huiyao Xu, Sitong Guo, Zhongwei Xie, Hujun Bao, Weiwei Xu, and Changqing Zou. 2025c. Spatialcrafter: Unleashing the imagination of video diffusion models for scene reconstruction from limited observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 27794–27805.
- Yanran Zhang, Ziyi Wang, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. 2025b. Joint 3D Geometry Reconstruction and Motion Generation for 4D Synthesis from a Single Image. *arXiv preprint arXiv:2512.05044* (2025).
- Sixiao Zheng, Minghao Yin, Wenbo Hu, Xiaoyu Li, Ying Shan, and Yanwei Fu. 2026. VerseCrafter: Dynamic Realistic Video World Model with 4D Geometric Control. *arXiv preprint arXiv:2601.05138* (2026).
- Hongkuan Zhou, Wei Cao, Aifen Sui, and Zhenshan Bing. 2023. What Matters to Enhance Traffic Rule Compliance of Imitation Learning for End-to-End Autonomous Driving. *arXiv preprint arXiv:2309.07808* (2023).
- Hongkuan Zhou, Stefan Schmid, Yicong Li, Lavdim Halilaj, Xiangtong Yao, and Wei Cao. 2025a. Predicting the road ahead: A knowledge graph based foundation model for scene understanding in autonomous driving. In *European Semantic Web Conference*. Springer, 116–132.
- Kaichen Zhou, Yuhang Wang, Grace Chen, Xinhai Chang, Gaspard Beaudouin, Fangneng Zhan, Paul Pu Liang, and Mengyu Wang. 2025b. Page-4d: Disentangled pose and geometry estimation for 4d perception. *arXiv e-prints* (2025), arXiv–2510.
- Ruijie Zhu, Jiahao Lu, Wenbo Hu, Xiaoguang Han, Jianfei Cai, Ying Shan, and Chuanxia Zheng. 2026. MotionCrafter: Dense Geometry and Motion Reconstruction with a 4D VAE. *arXiv preprint arXiv:2602.08961* (2026).