

MoCam: Unified Novel View Synthesis via Structured Denoising Dynamics

Haofeng Liu^{*,1}, Yang Zhou^{*,1,2,5}, Ziheng Wang^{*,1,3}, Zhengbo Xu¹, Zhan Peng^{1,4}, Jie Ma¹, Jun Liang^{1,✉}, Shengfeng He^{5,✉}, and Jing Li¹

¹Orange Team, Moku Lab, HUJING Digital Media & Entertainment Group

²South China University of Technology ³Columbia University ⁴Huazhong

University of Science and Technology ⁵Singapore Management University

orange-3dv-team.github.io/MoCam

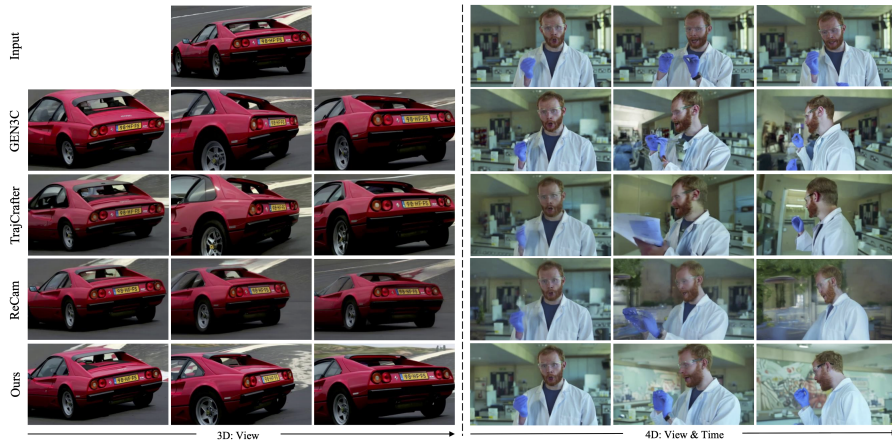


Fig. 1: We propose MoCam, a method that unifies novel view synthesis through structured denoising dynamics. Existing methods rely on static guidance that entangles geometry and appearance, often resulting in geometric collapse and visual artifacts. MoCam introduces structured denoising dynamics that guide generation from motion alignment to appearance refinement, producing coherent and photorealistic results.

Abstract. Generative novel view synthesis faces a fundamental dilemma: geometric priors provide spatial alignment but become sparse and inaccurate under view changes, while appearance priors offer visual fidelity but lack geometric correspondence. Existing methods either propagate geometric errors throughout generation or suffer from signal conflicts when fusing both statically. We introduce MoCam, which employs structured denoising dynamics to orchestrate a coordinated progression from geometry to appearance within the diffusion process. MoCam first leverages geometric priors in early stages to anchor coarse structures and tolerate their incompleteness, then switches to appearance priors in later stages to actively correct geometric errors and refine details. This design naturally unifies static and dynamic view synthesis by temporally decoupling

*Equal contribution.

✉Corresponding authors: junliang.lj@alibaba-inc.com, shengfenghe@smu.edu.sg

geometric alignment and appearance refinement within the diffusion process. Experiments demonstrate that MoCam significantly outperforms prior methods, particularly when point clouds contain severe holes or distortions, achieving robust geometry-appearance disentanglement.

Keywords: View Synthesis · Video Re-camera · Visual Generation

1 Introduction

Novel view synthesis aims to create photorealistic views from arbitrary camera trajectories given limited input, and it remains a fundamental challenge in computer vision with broad applications in virtual production, immersive reality, and content creation. This encompasses two closely related problems: single-image 3D reconstruction, where a static scene is reconstructed from one photograph, and video 4D re-camera, where dynamic scenes are rendered along new camera paths given a monocular video. Success in both settings requires reconciling precise geometric control with high-fidelity appearance synthesis, particularly when the target viewpoint significantly deviates from the input.

Recent advances in diffusion models [3, 18, 34, 40] have enabled impressive progress in both domains. For 3D reconstruction, methods like ViewCrafter [45], VistaDream [35], and SpatialCrafter [49] leverage reconstructed geometry to guide novel view synthesis. For 4D re-camera, approaches such as Gen3C [30] and TrajectoryCrafter [44] employ 3D scaffolds (*e.g.*, point clouds) to render target-view videos with explicit camera control. However, these methods share a critical vulnerability: they rely on geometric priors (depth maps or point clouds reconstructed from monocular input) that inevitably become sparse, incomplete, and erroneous under large view changes. Existing pipelines either propagate these geometric flaws throughout generation [30, 44] or attempt to fuse geometry and appearance statically, causing signal conflicts that degrade both structure and texture (see Fig. 1), limiting their applicability in settings that demand high-fidelity and precise cinematic control.

We argue that this bottleneck stems from a fundamental tension between two complementary yet incompatible signal sources. Rendered geometric scaffolds provide essential spatial alignment with target trajectories but suffer from holes and distortions due to disocclusion and depth inaccuracy. Conversely, source images/videos offer rich, high-fidelity appearance but are geometrically misaligned with novel views. Crucially, these signals cannot be effectively combined simultaneously: early in generation, strong appearance cues dominate and cause geometric drift; late in generation, flawed geometry permanently bakes structural errors into the output.

To resolve this, we introduce MoCam, a framework that exploits *structured denoising dynamics* to temporally decouple geometry and appearance priors within the diffusion process. Our key insight is that diffusion models exhibit distinct representational needs across denoising phases: early stages require coarse structural anchoring, while later stages demand high-frequency refinement. Mo-

Cam orchestrates a coordinated progression: in early timesteps, the model conditions solely on rendered scaffolds to establish global structure and motion coherence, deliberately tolerating geometric incompleteness. As denoising progresses and the latent stabilizes, MoCam transitions to conditioning on the source appearance. At this stage, the established geometry enables the model to use appearance not merely for texture transfer, but to actively correct geometric errors and fill disoccluded regions without destabilizing the overall structure.

Notably, this mechanism naturally provides a unified solution for both static and dynamic view synthesis. By structuring the denoising process to first establish geometry and then refine appearance, MoCam separates geometric alignment from appearance synthesis in a manner that is independent of the input modality. As a result, the same generation principle applies to both single-image 3D view synthesis and video 4D re-camera, highlighting that our approach addresses the underlying challenge of synthesizing views under unreliable geometry.

By transforming denoising into a structured progression from alignment to realism, MoCam achieves robust geometry-appearance disentanglement. Even when point clouds contain severe holes or distortions, our method generates geometrically coherent and photorealistic results, significantly outperforming static conditioning approaches (Fig. 1).

In summary, our contributions are threefold:

- We identify the fundamental conflict between geometric and appearance priors in generative view synthesis, and propose *structured denoising dynamics* as a principled solution that temporally decouples these signals.
- We present a unified framework for both single-image 3D reconstruction and video 4D re-camera, demonstrating that stage-wise conditioning generalizes across input modalities.
- We show that active geometric error correction through late-stage appearance signal achieves state-of-the-art robustness under sparse and inaccurate geometry, setting new standards for controllable view synthesis.

2 Related Works

Optimization-Based Novel View Synthesis. A classical approach to novel view synthesis involves reconstructing a 3D or 4D representation from posed images. Neural Radiance Fields (NeRF) [26] transformed this field by representing a static scene as a continuous volumetric function, enabling unprecedented photorealism. More recently, 3D Gaussian Splatting (3DGS) [17] has achieved comparable or superior quality with real-time rendering by modeling the scene as a set of explicit 3D Gaussians. Extending such methods to dynamic scenes, which is essential for video re-camera, requires modeling temporal evolution [51]. One strategy learns 4D representations that map spacetime coordinates to scene properties [4, 7, 9, 10, 21, 22, 25, 39, 50], while another explicitly models motion through deformation fields [8, 20, 23, 24, 29, 32, 36, 41]. Although powerful, these approaches typically require dense multi-view video and involve costly per-scene

optimization. When limited to monocular input, both reconstruction quality and appearance fidelity degrade significantly. In contrast, our method avoids per-scene optimization entirely, instead leveraging the generative priors of large-scale video models to synthesize photorealistic and geometrically consistent results from a single input video.

Generative Novel View Synthesis. Recent single-view 3D reconstruction methods [6, 31, 47] leverage pretrained image diffusion models to enable view synthesis from single images. However, generating smooth camera trajectories rather than isolated views requires temporal consistency, motivating the shift to video generative models [3, 18, 34, 40]. For instance, ViewCrafter [45] harnesses video diffusion to synthesize high-fidelity view sequences along camera paths. Extending these methods to dynamic scenes for 4D video re-camera introduces further complexity, as the generation process must simultaneously handle temporal dynamics and viewpoint changes. Existing approaches fall into two categories. The first injects camera pose information directly into the model’s conditioning mechanism [1, 2, 19, 33, 38], offering end-to-end generation but often lacking geometric accuracy, especially for complex or large-scale trajectories. The second category follows a render-then-inpaint strategy [5, 16, 30, 43, 44, 46], where a 3D scaffold (e.g., a point cloud) is reconstructed from the source video, rendered along the target path, and then refined using a video inpainting model. Gen3C [30] constructs a spatiotemporal 3D cache to guide generation, while TrajectoryCrafter [44] introduces a Ref-DiT block for reference-based conditioning. Although these methods better enforce target-view geometry, they suffer from a key bottleneck: the rendered scaffold is built on sparse and inaccurate geometry, which permanently bakes errors into the generation process. The inpainting stage inherits these flaws and lacks the capacity to correct them. Our method addresses this limitation by introducing a temporally structured guidance strategy. By decoupling geometry and appearance over the denoising process, it mitigates error propagation and improves stability under large camera motions.

Conditioning Mechanisms in Diffusion Models. Conditioning is the core mechanism for controllability in diffusion models. Techniques such as ControlNet [48] and T2I-Adapter [27] allow spatial control using depth maps or other signals, while IP-Adapter [42] enables lightweight image-prompt conditioning. These approaches typically apply static guidance, using the same control signal across all timesteps. More recent work has begun to explore dynamic conditioning. TSM [52] and DMP [12] demonstrate that adjusting or switching control inputs over time can significantly improve generation quality. Building on this idea, our method introduces a dynamic conditioning scheme tailored to video re-camera. We design a stage-wise handover between two complementary but conflicting inputs: a geometrically aligned yet flawed scaffold, and a view-disaligned but visually rich reference video. This design specifically resolves the error propagation problem by aligning each guidance signal with the stage of denoising where it is most effective.

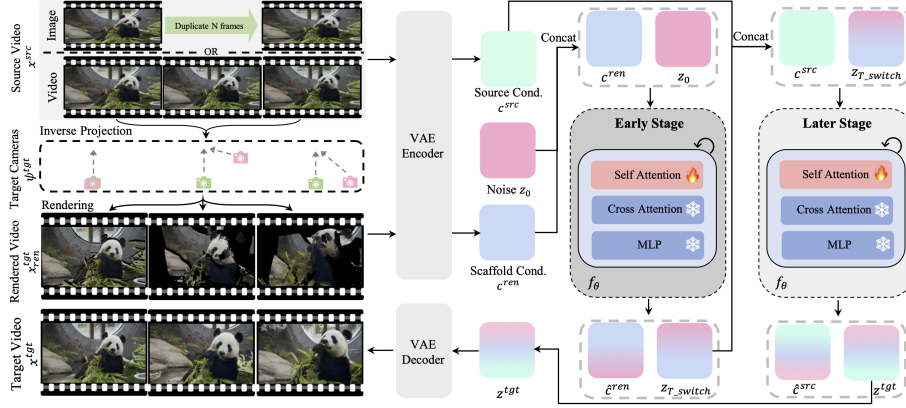


Fig. 2: Overview of the MoCam Framework. Given a source video x^{src} (or a single image repeated for N frames), a geometrically-aligned but imperfect scaffold video x^{tgt} is first rendered along the target trajectory ψ^{tgt} . After encoding these conditions into latent space, the model processes the initial noise z_0 via the proposed structured denoising dynamics. Specifically, in the early stage, the denoising is guided by the scaffold condition c^{ren} to establish geometrically-aligned noise z_{T_switch} . Subsequently, the signal switches to the original source condition c^{src} to obtain the clean latent z^{tgt} with refined appearance and decode it as the target video x^{tgt} . This temporal decoupling of conditions prevents the propagation of scaffold errors, enabling stable and photorealistic synthesis.

3 MoCam

In this section, we present MoCam, a novel framework to generate novel views based on video model. The core challenge is maintaining geometric and temporal consistency, especially under complex camera movements. Our approach is built on the key insight that different types of conditions are optimal at different stages of the generation process. Specifically, our method consists of three main stages, as illustrated in Fig. 2: (1) We first construct a dynamic point cloud from the monocular input video (or a single image replicated to N frames, *i.e.*, a stationary video) and render it along the target trajectory to create a coarse scaffold video. (2) We then use this scaffold video and the original source video as dual conditioning inputs to our novel stage-Wise generation model. (3) The model first enforces the coarse structure using the scaffold, then switches to the source video to perfect the appearance and geometry.

3.1 Preliminary: Video Generative Models

Since our method builds upon a video generative model, we first provide a brief overview of its fundamental principles. For computational efficiency, modern video generative models [3, 18, 34, 40] operate not in the high-dimensional pixel space but in a compressed latent space. This space is constructed by a pre-trained Variational Autoencoders (VAEs) [37]. The VAE consists of an encoder $\mathcal{E}$ that compresses an input video $x \in \mathbb{R}^{N \times H \times W \times 3}$ into a compact latent representation $z = \mathcal{E}(x) \in \mathbb{R}^{n \times h \times w \times c}$, and a decoder $\mathcal{D}$ that reconstructs the video $\hat{x} = \mathcal{D}(z)$

from this latent representation. Upon this latent space, a generative model f_θ is trained to model the data distribution. This is typically achieved through one of two primary training paradigms: a denoising diffusion objective or a flow matching objective. Under the denoising diffusion schema, the model learns to reverse a process that gradually adds noise to the data. The objective is to predict the noise added to a latent representation:

$$\min_{f_\theta} \mathbb{E}_{z_0, z_1, t, c} \|f_\theta(z_t, t, c) - z_0\|_2^2. \quad (1)$$

Alternatively, under the flow matching schema, the model learns a vector field that transports samples from a simple prior distribution to the data distribution:

$$\min_{f_\theta} \mathbb{E}_{z_0, z_1, t, c} \|f_\theta(z_t, t, c) - v_t\|_2^2, \quad (2)$$

where $z_1 = \mathcal{E}(x)$ is the latent encoding of a real video sampled from the data distribution p_{data} , and $z_0 \sim \mathcal{N}(0, \mathbf{I})$ is a random latent sampled from a standard Gaussian prior. The variable $t \in [0, 1]$ is a continuous time step, and c represents optional conditioning information (such as text prompts or image frames). For the denoising objective, z_t is a noisy latent created by interpolating between z_1 and z_0 according to a noise schedule (e.g., $z_t = \alpha_t z_1 + \sigma_t z_0$). For the flow matching objective, z_t is typically a linear interpolation $z_t = (1-t)z_0 + tz_1$, and the target velocity is $v_t = z_1 - z_0$. Crucially, the timestep t represents not merely a noise level, but a progression from global structure to local detail—a property we exploit in our stage-wise conditioning strategy.

3.2 Scaffold Generation

The first step of our pipeline is to generate a coarse video draft that is spatially and temporally aligned with the target camera trajectory. This *scaffold video*, denoted as x_{ren}^{tgt} , serves as the initial structural guide for our diffusion model.

Given a source video $x^{src} = \{I_i^{src}\}_{i=1}^N \in \mathbb{R}^{N \times H \times W \times 3}$, we first leverage a depth estimator to acquire its depth $d^{src} = \{D_i^{src}\}_{i=1}^N$. We then follow the inverse perspective projection Φ^{-1} to construct a dynamic point cloud $p = \{P_i\}_{i=1}^N$:

$$p = \Phi^{-1}(x^{src}, d^{src}, K), \quad (3)$$

where $K \in \mathbb{R}^{3 \times 3}$ denotes the camera intrinsic. We refer this dynamic point cloud p as the 3D scaffold, which provide us a way to precisely control the camera trajectory. Specifically, conditioned by a target camera trajectory $\psi^{tgt} = \{\Psi_i^{tgt}\}_{i=1}^N$, we render the target video $x^{tgt} = \{I_i^{tgt}\}_{i=1}^N \in \mathbb{R}^{N \times H \times W \times 3}$ from p following the perspective projection Φ :

$$x_{ren}^{tgt} = \Phi(p, \psi^{tgt}, K). \quad (4)$$

As shown in Fig. 2, the rendered scaffold video spatially aligns with the target camera motion. However, due to the inherent limitations of monocular input,

this video suffers from significant artifacts: holes from disocclusion, and geometric distortions, particularly in views far from the original camera path. While unsuitable as a final output, it provides an invaluable, spatially-aligned motion prior for the initial stages of generation.

3.3 Stage-Wise Dual-Conditioning Diffusion

The proposed latent video generative model integrates conditions from two distinct sources—the scaffold video x_{ren}^{tgt} and the source video x^{src} —at different phases of the generation process. We build upon a pretrained latent video diffusion architecture [34], which is trained to denoise a noisy latent variable z_t at timestep t . **Our innovation lies in how we formulate the conditioning term c .**

We design a stage-wise dual-conditioning architecture. Each stage is responsible for processing one of our condition signals:

Spatial Scaffold Condition. To inject the strong motion and structural prior from the scaffold video x_{ren}^{tgt} , we follow the frame dimension conditioning to retain temporal synchronization [2]. Particularly, x_{ren}^{tgt} is first projected into the latent space by the VAE encoder $\mathcal{E}$, $z_{ren}^{tgt} = \mathcal{E}(x_{ren}^{tgt})$, *i.e.*, conditioning term c^{ren} . After that we concatenate the c^{ren} with the initial noise z_0 along the frame dimension as the input of the video model. This provides direct, spatially-explicit guidance, forcing the generated output to conform to the layout and motion defined by the scaffold.

Reference Appearance Condition. Unlike the scaffold video x_{ren}^{tgt} that contain spatial-aligned information, the source video x^{src} emphasizes high-fidelity appearance and object dynamics of the scene. It forms a complement relationship with the x_{ren}^{tgt} during the video generation process, in which x_{ren}^{tgt} provide geometry signal and x^{src} supplement the appearance signal. The conditioning of x^{src} is the same as x_{ren}^{tgt} : $\mathcal{E}$, $z^{src} = \mathcal{E}(x^{src})$, *i.e.*, conditioning term c^{src} , then concatenated along the frame dimension. This mechanism is effective at transferring content and texture, making it ideal for our view-disaligned source video.

Why Stage-wise Conditioning is Necessary. An intuitive way to leverage these two kinds of condition (*i.e.*, c^{ren} and c^{src}) is to concatenate them together with the initial noise z_0 and let the model to learn the combinative condition by itself. Though c^{ren} and c^{src} exhibit the mentioned-above complement relationship, they also contain conflicting signal to each other, *i.e.*, the different camera movements. Since the camera movement of c^{src} is different with c^{ren} , it introduces interference against the guidance of c^{ren} , which may confuse the model learning and decrease the final effectiveness. Fig. 3 illustrates the re-



Fig. 3: Results of different guiding methods.

sults of such conditioning. Besides, persistent exposure to c^{ren} ’s geometric errors causes irreversible structural artifacts. See experiment (Sec. 4.3) for detail discussion.

To circumvent this conflict, our approach MoCam is motivated by the inherent behavior of diffusion models. We align our conditioning strategy with the progressive denoising process, prioritizing the establishment of global structure in early stages before refining high-frequency details in later ones. The central novelty of MoCam is the *structured denoising dynamics*, where we temporally align these two conditioning signals with respective denoising timesteps t . The model’s prediction is conditioned on a time-dependent context $c(t)$:

$$f_{\theta}(z_t, t, c(t))$$

where $c(t)$ is defined by a switch at a pre-defined timestep threshold T_{switch} :

$$c(t) = \begin{cases} c^{ren} & \text{if } t > T_{switch} \\ c^{src} & \text{if } t \leq T_{switch} \end{cases} \quad (5)$$

The intuition is as follows:

- Early Stage ($t > T_{switch}$): Geometry Anchoring. The latent z_t is mostly noise. The model’s primary task is to establish the global structure and motion of the video. By using c^{ren} , we force the generation to adhere to the target camera trajectory from the very beginning.
- Later Stage ($t \leq T_{switch}$): Active Error Correction & Refinement. The latent z_t already contains a coherent, low-frequency structure that aligns with the target structure. The task now shifts to synthesizing high-frequency details, refining appearance, and correcting geometric inaccuracies. We switch to c^{src} , which provides a rich source of clean textures and consistent object appearance. Because the coarse structure is already established, the model can use this high-fidelity reference to “inpaint” and “correct” the structure inherited from the first stage, without being corrupted by the scaffold’s persistent errors.

This deliberate handover prevents the scaffold’s flaws from being “baked in” during the final, high-fidelity synthesis steps, effectively resolving the core limitation of static pipelines. As shown in Fig. 2, the clean latent z^{tgt} is put into the decoder for the final output video x^{tgt} :

$$x^{tgt} = \mathcal{D}(z^{tgt}) \quad (6)$$

4 Experiments

We implement MoCam by building upon the pretrained Wan2.2 video diffusion model [34] and train it using 20,000 data pairs from the MultiCamVideo dataset [2]. Each training sample consists of a reference video, the resulting scaffold video, and the ground-truth target video. For scaffold generation, we use

Table 1: Quantitative 3D reconstruction comparisons on the OpenVid dataset. BC: Background Consistency, SC: Subject Consistency, IQ: Imaging Quality, RotErr: Rotation Error, TransErr: Translation Error. Cells highlighted in red and yellow denote the best and second-best performance.

	VBench			Perceptual Quality		Pose Accuracy	
Methods	BC $\uparrow$	SC $\uparrow$	IQ $\uparrow$	FVD-V $\downarrow$	CLIP-V $\uparrow$	RotErr $\downarrow$	TransErr $\downarrow$
GEN3C	0.9295	0.9076	0.6909	289.37	0.80	1.36	5.12
TrajCrafter	0.9241	0.9070	0.6700	313.65	0.79	1.36	5.11
ReCam	0.9016	0.8818	0.5842	355.26	0.77	2.13	5.79
Ours	0.9334	0.9250	0.6961	255.16	0.87	1.35	5.11



Fig. 4: Qualitative results for single-view 3D reconstruction.

ViPE [14] for depth and camera estimation. The model is trained for 20,000 steps on eight GPUs with a learning rate of 1e-5 and batch size of 8. T_{switch} is set as 0.85 empirically.

4.1 Evaluation on In-the-wild Benchmark

To provide a broad quantitative assessment, we collected 100 monocular videos from OpenVid-1M [28] and generated outputs for 9 distinct camera trajectories per video, including orbital, translational, and zoom motions. These monocular videos serve as direct input for the 4D re-camera experiments. For single-view 3D reconstruction, we randomly sample one frame from each video and replicate it to N frames. Our evaluation metrics include: (1) background consistency, subject consistency and imaging quality from VBench metrics [15], (2) FVD-V and CLIP-V that calculate FVD and CLIP scores between different viewpoints, (3) pose accuracy: rotation error and translation error [13].

3D Reconstruction Qualitative Results. Fig. 4 visualizes single-view synthesis results. GEN3C and TrajCrafter struggle with the extreme sparsity of single-image point clouds, leading to structural distortions. ReCamMaster fails to infer correct 3D layouts without explicit geometry. In contrast, MoCam leverages structured denoising dynamics to overcome this sparsity: we first anchor



Fig. 5: Qualitative results from in-the-wild videos. The first example illustrates an ‘orbit-to-left’ trajectory, while the second example demonstrates a camera motion that initially moves to the top-left with zoom-in, followed by a transition to the bottom-right with a corresponding zoom-out.

plausible geometry using the limited scaffold, then refine appearance, yielding coherent and detailed results.

3D Reconstruction Quantitative Results. Tab. 1 demonstrates our superiority. MoCam significantly outperforms competitors in perceptual quality (e.g., FVD-V 255.16 vs. 289.37) and achieves the lowest pose errors. This confirms that our structured denoising strategy effectively maintains both high perceptual fidelity and precise camera control, effectively handling the geometric ambiguity inherent in single-view 3D reconstruction.

4D Re-Camera Qualitative Results. Fig. 5 presents a qualitative comparison against state-of-the-art methods across diverse scenes and camera trajectories. Methods based on 3D scaffolds, like GEN3C [30] and TrajectoryCrafter [44], successfully follow the target movement but suffer from severe geometric degradation. For instance, in the first example, the cat’s body becomes distorted as the camera orbits. Similarly, in the second example, the human’s arm collapses unrealistically. These artifacts are direct consequences of error propagation from the sparse and inaccurate point cloud reconstruction. The implicit conditioning method, ReCamMaster [2], struggles to maintain geometric consistency and fails to follow the complex trajectory, resulting in chaotic and unusable outputs. In contrast, MoCam generates results that are both geometrically coherent and photorealistic. Our method correctly preserves the 3D structure of the subjects (the cat’s volume, the person’s limbs) while rendering high-fidelity textures, even under significant view changes. We provide more dynamic results in the supplementary video.

4D Re-Camera Quantitative Results. As shown in Tab. 2, MoCam achieves the highest scores across the majority of metrics, notably in background consistency, subject consistency, and imaging quality, outperforming all competitors by significant margins. This confirms that our method not only maintains the iden-

Table 2: Quantitative 4D re-camera comparisons on the OpenVid dataset. BC: Background Consistency, SC: Subject Consistency, IQ: Imaging Quality, RotErr: Rotation Error, TransErr: Translation Error. Cells highlighted in red and yellow denote the best and second-best performance.

Methods	VBench			Perceptual Quality		Pose Accuracy	
	BC $\uparrow$	SC $\uparrow$	IQ $\uparrow$	FVD-V $\downarrow$	CLIP-V $\uparrow$	RotErr $\downarrow$	TransErr $\downarrow$
GEN3C	0.9270	0.9067	0.6908	291.13	0.79	1.36	5.13
TrajCrafter	0.9235	0.9062	0.6697	317.08	0.80	1.38	5.12
ReCam	0.8977	0.8801	0.5837	361.98	0.76	2.15	5.82
Scaffold-Only	0.8898	0.8448	0.4807	359.38	0.76	1.37	5.10
Scaffold-Early	0.9139	0.9053	0.6172	273.19	0.83	1.37	5.13
Static-Both	0.9190	0.9203	0.6740	242.81	0.87	2.71	11.01
Ours (Wan2.1)	0.9330	0.9248	0.6931	253.13	0.84	1.37	5.11
Ours	0.9332	0.9247	0.6932	260.05	0.85	1.36	5.12

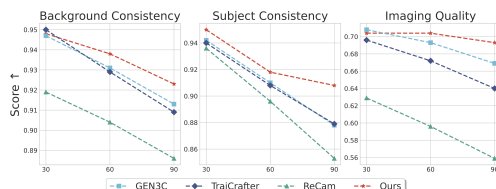


Fig. 6: Quantitative results of VBench metrics on various motion magnitudes.

Table 3: Quantitative comparisons on iPhone Dataset. Cells highlighted in red and yellow denote the best and second-best performance.

Methods	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$
GEN3C	12.36	0.4028	0.5112	260.15
TrajCrafter	13.74	0.4555	0.4819	273.36
ReCam	11.44	0.3768	0.5622	301.41
Ours	14.60	0.4581	0.4213	180.35

tity and structure of the main subject but also produces more visually pleasing and realistic images. Besides, MoCam achieves this superior perceptual quality without sacrificing geometric precision, maintaining the lowest rotation error and competitive translation error.

Robustness to Geometric Degradation. Geometric degradation under large view changes poses a fundamental challenge to all view synthesis paradigms. For scaffold-based methods, large camera motions induce severe geometric sparsity (disocclusion holes, depth inaccuracy); for scaffold-free implicit methods like ReCamMaster, the same motions cause geometric drift due to the lack of explicit spatial constraints. We design an experiment to measure robustness to this unified challenge, using motion magnitude as a controlled proxy to induce progressive geometric degradation. Starting from a modest 30-degree orbit (minimal geometric stress), we progressively increase motion magnitude to a challenging 90-degree trajectory. At 90 degrees, scaffold-based methods face extremely sparse geometry, while implicit methods face severe misalignment between source and target views. Fig. 6 plots performance as geometric degradation intensifies with increasing motion. All competitors deteriorate under this stress test: GEN3C and TrajectoryCrafter propagate errors from sparse, hole-ridden point clouds; ReCamMaster, despite being scaffold-free, suffers catastrophic geometric drift without early-stage structural anchoring. MoCam maintains consistently high scores because its stage-wise decoupling addresses both failure modes: the early



Fig. 7: Qualitative results on various motion scales. The models are inferred under camera trajectories with three different scales of orbit degree.

geometry-anchoring stage prevents drift (solving ReCamMaster’s vulnerability), while the late appearance stage corrects sparse geometric errors (solving scaffold-based methods’ vulnerability). The qualitative results in Fig. 7 confirm this: as geometric degradation increases, previous methods produce warped, broken, or drifted figures, while MoCam renders coherent subjects by leveraging geometry for structure without being corrupted by its sparsity.

4.2 Evaluation on Multi-view Video Benchmark

While our primary focus is on in-the-wild monocular videos, we also conduct experiments on a multi-view dataset to enable evaluation with pixel-wise metrics. Following the setup of TrajectoryCrafter [44], we use the iPhone dataset [11], treating one moving camera view as the monocular input and a static camera view as the ground-truth target.



Fig. 8: Qualitative results on iPhone Dataset.



Fig. 9: Ablation results on structured denoising dynamics.

Tab. 3 shows that MoCam significantly outperforms other methods on PSNR, SSIM, LPIPS and FVD. The strong improvement in LPIPS and FVD, perceptual metrics for perceptual quality, is particularly noteworthy, indicating that our generated views are perceptually closer to the ground truth. This demonstrates that our timestep-gated conditioning not only improves general coherence but also preserves geometric and appearance details with high fidelity. The qualitative results in Fig. 8 corroborate this; for instance, notice the finer details on the subject’s clothing and the more accurate facial structure rendered by our method compared to the blurry or distorted results from others.

4.3 Ablation Studies

We conduct a series of ablation studies to dissect the contributions of our key design choices.

Effectiveness of Structured Denoising Dynamics. To validate our core hypothesis that a temporally-aligned structured denoising generation is essential, we compare MoCam against several variants. The variants are: (1) *Scaffold-Only*: The model is conditioned only on the scaffold video c^{ren} for all timesteps. (2) *Scaffold-Early*: The model is conditioned on the scaffold video c^{ren} only during the early timesteps ($t > T_{switch}$), with no explicit conditioning in the later stages. This variant tests the hypothesis that simply removing the flawed scaffold signal is sufficient to mitigate artifact propagation. (3) *Static-Both*: Both scaffold and

reference conditions (*i.e.*, c^{ren} and c^{src}) are provided simultaneously throughout the entire denoising process.

As shown in Tab. 2, MoCam significantly outperforms all variants, demonstrating the advantage of the proposed structured denoising dynamics, and Fig. 9 qualitatively illustrate the results. The ‘Scaffold-Only’ baseline collapses across all metrics (IQ merely 0.4807), proving that persistent geometric conditioning permanently bakes scaffold errors into the output. The ‘Scaffold-Early’ model successfully avoids inheriting the worst point cloud artifacts, but without the reference signal in the later stages, it struggles to synthesize fine-grained, scene-consistent textures and often produces blurry or generic details in disoccluded regions. The ‘Static-Both’ model suffers catastrophic geometric instability (Rotation Error 2.71, Translation Error 11.01) despite competitive perceptual scores, validating that simultaneous conditioning creates signal interference. In contrast, MoCam leverages the strengths of both signals in sequence: it first establishes correct geometry with the scaffold and then actively corrects its flaws and refines photorealism using the reference video. These results validate that merely removing imperfect geometry (Scaffold-Early) or adding appearance statically (Static-Both) is insufficient; the deliberate stage-wise handover is essential to resolve the geometry-appearance conflict.

Generalization across Different Backbones. To further validate the robustness of our approach, we extend our ablation studies to different video generation backbones. Specifically, we further evaluate our method on both Wan2.1. As illustrated in Fig. 9, the qualitative comparison demonstrates that our method achieves comparable high-fidelity results on Wan2.1, similar to the performance observed on Wan2.2. This consistency across different architectures confirms that our method is robust to the choice of backbone. More importantly, it provides strong evidence that the performance gains stem from our proposed structured denoising dynamics, rather than depending on a specific video model architecture. Quantitative results in Tab. 2 also demonstrate its generalization ability.

Robustness of Depth Estimation. Monocular depth estimation inevitably introduces inaccuracies that propagate into point cloud reconstructions. Our stage-wise mechanism tolerates these imperfections: early geometry anchoring prevents structural drift, while late appearance correction rectifies artifacts before they are baked into the final output. Fig. 10 validates this robustness under depth errors, where MoCam maintains coherence despite perturbed inputs: lights distorted by inaccurate depth in the scaffold are successfully corrected in the final output, demonstrating that MoCam explicitly compensates for reconstruction noise rather than relying on perfect geometry.



Fig. 10: Depth Robustness.

4.4 Comparison with 3D-based method

We further compare our method with ViewCrafter [45] on single-view 3D reconstruction both qualitatively and quantitatively. Fig. 11 shows that ViewCrafter introduce geometry distortion and view coherence (*e.g.*, the tire in the first sample and the text in the second sample), while ours maintain correct shape due to its structured denoising dynamic mechanism. The quantitative results in Tab. 4 also demonstrates our method achieve better visual quality and structure accuracy as well.

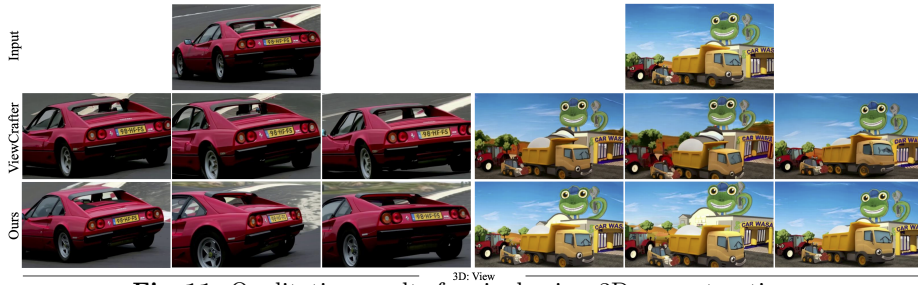


Fig. 11: Qualitative results for single-view 3D reconstruction.

Table 4: Quantitative 3D reconstruction comparisons on the OpenVid dataset. BC: Background Consistency, SC: Subject Consistency, IQ: Imaging Quality, RotErr: Rotation Error, TransErr: Translation Error.

Methods	VBench			Perceptual Quality		Pose Accuracy	
	BC $\uparrow$	SC $\uparrow$	IQ $\uparrow$	FVD-V $\downarrow$	CLIP-V $\uparrow$	RotErr $\downarrow$	TransErr $\downarrow$
ViewCrafter	0.9268	0.9082	0.6836	308.24	0.79	1.39	5.14
Ours	0.9334	0.9250	0.6961	255.16	0.87	1.35	5.11

5 Conclusion

We presented MoCam, a unified 3D/4D view synthesis framework that addresses sparse and erroneous geometry through structured denoising dynamics. By temporally decoupling geometry and appearance, early scaffold anchoring followed by late-stage error correction, our method prevents propagation of point cloud flaws without sacrificing geometric control. This stage-wise design achieves robustness to the imperfect reconstruction inevitable in monocular settings. Future work may explore joint scaffold-video refinement.

Acknowledgements

This research is supported by the Guangdong Natural Science Funds for Distinguished Young Scholars (Grant 2023B1515020097), the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG4-TC-2025-018-SGKR), and the Lee Kong Chian Fellowships.

References

1. Bahmani, S., Skorokhodov, I., Siarohin, A., Menapace, W., Qian, G., Vasilkovsky, M., Lee, H.Y., Wang, C., Zou, J., Tagliasacchi, A., et al.: Vd3d: Taming large video diffusion transformers for 3d camera control. arXiv preprint arXiv:2407.12781 (2024)
2. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: Recammaster: Camera-controlled generative rendering from a single video. arXiv preprint arXiv:2503.11647 (2025)
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
4. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 130–141 (2023)
5. Chen, K., Khurana, T., Ramanan, D.: Reconstruct, inpaint, finetune: Dynamic novel-view synthesis from monocular videos. In: Advances in Neural Information Processing Systems (NeurIPS) (2025)
6. Chung, J., Lee, S., Nam, H., Lee, J., Lee, K.M.: Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. IEEE Transactions on Visualization and Computer Graphics (2025)
7. Duan, Y., Wei, F., Dai, Q., He, Y., Chen, W., Chen, B.: 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
8. Fan, C.D., Chang, C.W., Liu, Y.R., Lee, J.Y., Huang, J.L., Tseng, Y.C., Liu, Y.L.: Spectromotion: Dynamic 3d reconstruction of specular scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21328–21338 (June 2025)
9. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12479–12488 (2023)
10. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5712–5721 (2021)
11. Gao, H., Li, R., Tulsiani, S., Russell, B., Kanazawa, A.: Dynamic novel-view synthesis: A reality check. In: NeurIPS (2022)
12. Ham, S., Woo, S., Kim, J.Y., Go, H., Park, B., Kim, C.: Diffusion model patching via mixture-of-prompts. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 17023–17031 (2025)
13. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024)
14. Huang, J., Zhou, Q., Rabeti, H., Korovko, A., Ling, H., Ren, X., Shen, T., Gao, J., Slepichev, D., Lin, C.H., et al.: Vipe: Video pose engine for 3d geometric perception. arXiv preprint arXiv:2508.10934 (2025)
15. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: VBench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)

16. Jeong, H., Lee, S., Ye, J.C.: Reangle-a-video: 4d video generation as video-to-video translation. arXiv preprint arXiv:2503.09151 (2025)
17. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **42**(4), 139–1 (2023)
18. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
19. Lei, G., Wang, C., Wang, Y., Li, H., Song, Y., Xu, W.: Motionflow: Learning implicit motion flow for complex camera trajectory control in video generation. arXiv preprint arXiv:2509.21119 (2025)
20. Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Newcombe, R., et al.: Neural 3d video synthesis from multi-view video. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 5521–5531 (2022)
21. Li, Z., Chen, Z., Li, Z., Xu, Y.: Spacetime gaussian feature splatting for real-time dynamic view synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8508–8520 (2024)
22. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6498–6508 (2021)
23. Lin, Y., Dai, Z., Zhu, S., Yao, Y.: Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 21136–21145 (2024)
24. LIU, Q., Liu, Y., Wang, J., Lyu, X., Wang, P., Wang, W., Hou, J.: MoDGS: Dynamic gaussian splatting from casually-captured monocular videos with depth priors. In: *The Thirteenth International Conference on Learning Representations* (2025), <https://openreview.net/forum?id=2prShxdLkX>
25. Luo, Z., Ran, H., Lu, L.: Instant4d: 4d gaussian splatting in minutes. *Advances in neural information processing systems* (2025)
26. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: *European Conference on Computer Vision*. pp. 405–421. Springer (2020)
27. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 38, pp. 4296–4304 (2024)
28. Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., Tai, Y.: Openvid-1m: A large-scale high-quality dataset for text-to-video generation. arXiv preprint arXiv:2407.02371 (2024)
29. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. 2021 ieee. In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 10313–10322 (2020)
30. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 6121–6132 (2025)
31. Shriram, J., Trevithick, A., Liu, L., Ramamoorthi, R.: Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. In: *2025 International Conference on 3D Vision (3DV)*. pp. 882–892. IEEE (2025)

32. Song, R., Liang, C., Xia, Y., Zimmer, W., Cao, H., Caesar, H., Festag, A., Knoll, A.: Coda-4dgs: Dynamic gaussian splatting with context and deformation awareness for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 28031–28041 (October 2025)
33. Van Hoorick, B., Wu, R., Ozguroglu, E., Sargent, K., Liu, R., Tokmakov, P., Dave, A., Zheng, C., Vondrick, C.: Generative camera dolly: Extreme monocular dynamic novel view synthesis. In: European Conference on Computer Vision. pp. 313–331. Springer (2024)
34. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
35. Wang, H., Liu, Y., Liu, Z., Wang, W., Dong, Z., Yang, B.: Vistadream: Sampling multiview consistent images for single-view scene reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26772–26782 (2025)
36. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20310–20320 (2024)
37. Wu, P., Zhu, K., Liu, Y., Zhao, L., Zhai, W., Cao, Y., Zha, Z.J.: Improved video vae for latent video diffusion model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18124–18133 (2025)
38. Wu, R., Gao, R., Poole, B., Trevithick, A., Zheng, C., Barron, J.T., Holynski, A.: Cat4d: Create anything in 4d with multi-view video diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26057–26068 (2025)
39. Yang, Z., Yang, H., Pan, Z., Zhang, L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. arXiv preprint arXiv:2310.10642 (2023)
40. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
41. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20331–20341 (2024)
42. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
43. You, M., Zhu, Z., Liu, H., Hou, J.: Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. arXiv preprint arXiv:2405.15364 (2024)
44. YU, M., Hu, W., Xing, J., Shan, Y.: Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. arXiv preprint arXiv:2503.05638 (2025)

45. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)
46. Zhang, D.J., Paiss, R., Zada, S., Karnad, N., Jacobs, D.E., Pritch, Y., Mosseri, I., Shou, M.Z., Wadhwa, N., Ruiz, N.: Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 2050–2062 (2025)
47. Zhang, J., Li, X., Wan, Z., Wang, C., Liao, J.: Text2nerf: Text-driven 3d scene generation with neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics* **30**(12), 7749–7762 (2024)
48. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
49. Zhang, S., Xu, H., Guo, S., Xie, Z., Bao, H., Xu, W., Zou, C.: 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*. pp. 27794–27805 (2025)
50. Zhang, X., Liu, Z., Zhang, Y., Ge, X., He, D., Xu, T., Wang, Y., Lin, Z., Yan, S., Zhang, J.: Mega: Memory-efficient 4d gaussian splatting for dynamic scenes. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 27828–27838 (October 2025)
51. Zhu, J., Tang, H.: Dynamic scene reconstruction: Recent advance in real-time rendering and streaming. *arXiv preprint arXiv:2503.08166* (2025)
52. Zhuang, S., Guo, Y., Ding, Y., Li, K., Chen, X., Wang, Y., Wang, F., Zhang, Y., Li, C., Wang, Y.: Timestep master: Asymmetrical mixture of timestep lora experts for versatile and efficient diffusion models in vision. *arXiv preprint arXiv:2503.07416* (2025)