# RoGe: Novel View Synthesis via End-to-End Implicit Reconstruction and Generation

**Anonymous authors**
Paper under double-blind review

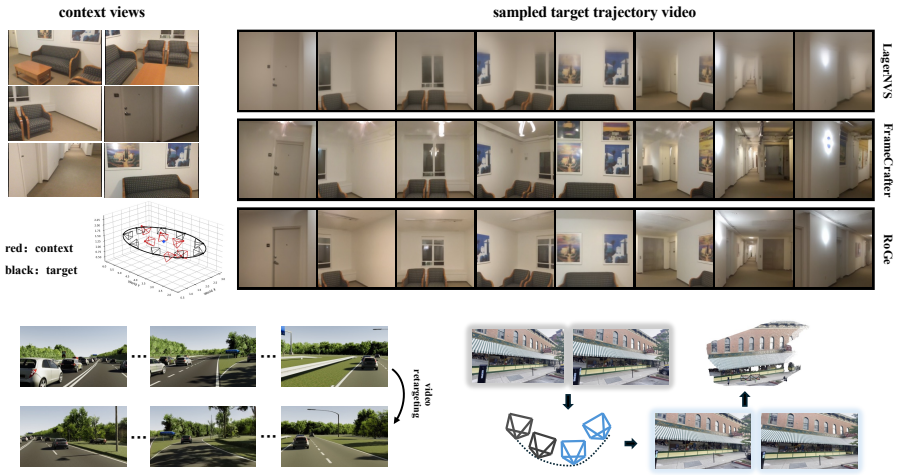



Figure 1: **RoGe**, a unified reconstruction and generation framework for sparse novel view synthesis. Given sparse views and arbitrary camera trajectories, it lifts per-view implicit geometry into a video diffusion model to synthesize temporally and geometrically coherent videos with high visual fidelity, without any explicit 3D representation. (1) Top: roaming within the scene anchored by sparse views. (2) Bottom left: extending RoGe to video retargeting in the 4D world. (3) Bottom right: reconstructing geometrically accurate point clouds from the synthesized novel views.

## Abstract

Novel view synthesis from sparse inputs requires both geometric grounding from the observed views and generative priors of unobserved regions, motivating recent hybrid methods that combine reconstruction and generation. However, existing methods bridge the two with rendered images or explicit 3D representations such as point maps or 3D Gaussians. Generation is thus conditioned on a lossy and imperfect projection of the scene, inheriting its errors, and reconstruction receives no signal from generation to correct them. We present RoGe, an end-to-end unified reconstruction and generation framework that removes this explicit bridge. It targets roaming within a scene anchored by sparse views: given a few posed images and a camera trajectory, it synthesizes a temporally coherent video along that trajectory. From the sparse input views, RoGe builds an implicit scene representation with a feed-forward reconstruction model, and queries it with camera rays to obtain per-view geometric features. These features are injected into a video diffusion model as conditioning, without any explicit 3D intermediate. Both modules are trained jointly, so the generation objective directly shapes its own geometric conditioning. We conduct extensive experiments, where RoGe outperforms reconstruction-based, generation-based, and hybrid baselines in terms of image-level quality and video-level temporal and geometric consistency. Ablations confirm that ray-queried implicit features outperform both raw reconstruction tokens and rendered RGB as conditioning, and that joint training brings further gains.

## 1 Introduction

Novel view synthesis (NVS) enables free-viewpoint exploration of captured scenes. While dense captures can support high-fidelity reconstruction, many practical settings provide only a few cali-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

brated images. In this paper, we study free-viewpoint exploration in this sparse regime, which we term ***roaming within a scene anchored by sparse views***: given a sparse set of posed observations and a user-specified camera trajectory, the goal is to generate an ordered sequence of novel views that is visually realistic, temporally coherent, and consistent with the observed scene. This setting is inherently ill-posed, since large portions of the scene may be occluded or entirely unobserved. A successful method must therefore preserve the geometric structure and appearance supported by the input views, follow the requested camera motion, and plausibly complete missing content without introducing cross-view drift.

Existing methods for NVS largely fall into two paradigms. Reconstruction-based methods, including per-scene optimization Mildenhall et al. (2021); Kerbl et al. (2023) and feed-forward models Jiang et al. (2025); Ye et al. (2026), offer precise camera control and strong multi-view consistency, but struggle with unseen regions and extrapolation of poses, often resulting in blurring, floaters, and holes. Generation-based methods Gao et al. (2024); Zhou et al. (2025); Wu et al. (2026b), in contrast, leverage diffusion priors to synthesize unobserved areas under large viewpoint changes, yet lag behind in geometric consistency and camera controllability.

The complementary strengths have inspired efforts to bridge reconstruction and generation. One line of work Wu et al. (2024); Yu et al. (2024); Wu et al. (2025a) augment per-scene optimization with diffusion-based inpainting, using generated views to extend coverage beyond observed regions. While effective, such a strategy is often limited by inconsistencies between the generated and input views. Another line, Wu et al. (2025a); Ren et al. (2025); Yang et al. (2026) first reconstructs the 3D scene, and then render or project it to condition a generative model. Reconstruction and generation are thus fully decoupled. Generation only sees a lossy and imperfect projection of the scene and inherits its errors, while reconstruction receives no signal from generation to correct them.

In this work, we present RoGe, an end-to-end unified reconstruction and generation framework for NVS. RoGe extracts per-view geometric features from the implicit scene representation built from the intermediate tokens of a feed-forward reconstruction model and injects them into a video diffusion model as conditioning, without any explicit 3D intermediate representations. Crucially, the two models are trained end-to-end, enabling direct information flow between them and allowing the generative model to guide the learning of geometric representations. Experiments across both image-level and video-level metrics demonstrate that RoGe achieves superior visual quality, geometric consistency, and camera controllability. Main contributions can be summarized as follows:

- We propose RoGe, a joint reconstruction and generation framework for NVS that connects a feed-forward reconstruction network with a video generation model in an end-to-end manner, so that the generation model can directly consume the geometric representation it is conditioned on.
- We condition the generation part on per-view geometric features obtained by querying the implicit scene representation with camera rays. Ablations show that such ray-queried features are more effective than raw reconstruction tokens and rendered or decoded RGB maps.
- We conduct comprehensive experiments, demonstrating that our method synthesizes novel trajectory videos with high visual quality, strong geometric consistency, and precise camera controllability, surpassing existing reconstruction-based, generation-based, and hybrid methods.

## 2 Related Work

**Scene Reconstruction.** Classical Structure-from-Motion (SfM) recovers sparse geometry through per-scene optimization, which is computationally expensive and brittle under sparse inputs Schonberger & Frahm (2016). Feed-forward methods enable efficient multi-view reconstruction in a single pass Wang et al. (2024a); Leroy et al. (2024); Wang et al. (2025a; 2026a;b); Keetha et al. (2026); Lin et al. (2025). However, both of these approaches primarily output point clouds, which are insufficient for high-fidelity NVS. Neural Radiance Fields (NeRF) Mildenhall et al. (2021) and 3D Gaussian Splatting (3DGS) Kerbl et al. (2023) instead provide renderable scene representations. 3DGS in particular offers fast optimization and real-time rendering at competitive visual quality, and has become a dominant choice for reconstruction-based NVS. Recently, feed-forward 3DGS has shown increasingly promising results, often outperforming optimization-based 3DGS under sparse inputs Charatan et al. (2024); Chen et al. (2024); Xu et al. (2025); Ziwen et al. (2025); Ye et al. (2025); Jiang et al. (2025); Ye et al. (2026). Departing from explicit representations, LagerNVS Szymanowicz et al. (2026) shows that the intermediate tokens of VGGT Wang et al. (2025a) already encode sufficient scene geometry to render novel views directly. Despite strong multi-view consis-

tency and precise camera control, these reconstruction-based methods, including LagerNVS, lack a generative prior and thus produce blur and artifacts in regions unobserved by the sparse inputs.

**Video Generation.** Video generation models have demonstrated impressive capabilities in generating open-domain videos Blattmann et al. (2023); Zheng et al. (2024); Yang et al. (2025); Kong et al. (2024); Wan et al. (2025); Agarwal et al. (2025); Lin et al. (2026). On top of these models, camera-controllable approaches generate videos from a single image or text prompt along user-specified trajectories Wang et al. (2024b); He et al. (2024; 2025); Zhang et al. (2026); Zhao et al. (2026). Another line of work focuses re-rendering existing videos along new trajectories Yu et al. (2025); Jeong et al. (2025); Bai et al. (2025a;b). More closely related to classical NVS, CAT3D Gao et al. (2024), SEVA Zhou et al. (2025) and FrameCrafter Wu et al. (2026b) leverage multi-view or video diffusion models to synthesize novel views from sparse observations. While generation-based NVS methods excel at extrapolating unobserved content from sparse inputs, they generally lag behind reconstruction-based approaches in geometric consistency and precise camera control.

**Reconstruction and Generation.** Given the complementary nature of reconstruction and generation, an increasing number of hybrid approaches have emerged. One family aligns video diffusion with feed-forward 3D models in the feature space to generate geometrically consistent videos and 3D quantities Huang et al. (2025); Wu et al. (2026a); Dai et al. (2026); Huang et al. (2026). These methods target world modeling from a single image or text prompt, where the geometry is imagined rather than observed. Our task instead starts from sparse posed views, whose geometry can be recovered and used to ground video generation. Several works Wu et al. (2024); Yu et al. (2024); Liu et al. (2024); Wu et al. (2025b) retain per-scene optimization and use generated views as additional supervision, but inconsistencies between the generated and input views limit further gains. Two-stage methods Wu et al. (2025a); Ren et al. (2025); Yang et al. (2026); Kang et al. (2026) instead recover an explicit 3D representation first and then use it as a condition in generative models, so that reconstruction and generation remain decoupled. Our method requires neither per-scene optimization nor an explicit 3D intermediate representation. Geometric features derived from a feed-forward reconstruction network are directly injected into the video generation model, and these two modules are trained end-to-end so that the generation objective shapes its own geometric conditioning.

## 3 METHOD

Given a sparse set of posed context observations $\mathcal{S} = \{(\mathbf{I}_i, \mathbf{K}_i, \mathbf{T}_i)\}_{i=1}^{M}$ of a static scene, our goal is to synthesize a video along a target camera trajectory $\mathcal{T} = \{(\mathbf{K}_j, \mathbf{T}_j)\}_{j=1}^{N}$. Here, $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ denotes an image, while $\mathbf{K}$ and $\mathbf{T} = [\mathbf{R} \mid \mathbf{t}] \in SE(3)$ denotes its camera intrinsics and camera-to-world pose, respectively, where $\mathbf{T}$ consists of a rotation $\mathbf{R} \in SO(3)$ and a translation $\mathbf{t} \in \mathbb{R}^3$. Without loss of generality, the first view in $\mathcal{S}$ is set as the world frame and the translations of both $\mathcal{S}$ and $\mathcal{T}$ are normalized by the scale factor derived from the maximum translation magnitude in $\mathcal{S}$.

Sparse inputs leave large portions of the scene unobserved, making the problem inherently both geometric and generative. Neither reconstruction nor generation alone is sufficient to achieve optimal performance. We therefore bring these complementary paradigms together in a unified, differentiable framework shown in Fig. 2. Specifically, we build an implicit scene representation from the sparse context views with a feed-forward 3D reconstruction model, and query it with per-view rays to output geometry-aware features for the target trajectory. We then inject these features into a pre-trained video generation model as conditions, and optimize both modules jointly in an end-to-end manner. In the following, we first introduce two preliminaries, the Plücker ray representation and latent flow matching, and then elaborate the above pipeline in Sec. 3.1–3.3.

**Plücker Ray Representation.** We adopt Plücker ray map Plucker (1865); Zhang et al. (2024) as the common camera representation for both reconstruction and generation. For a pixel $\boldsymbol{\rho} = (u, v)$ from a view, let $\tilde{\boldsymbol{\rho}} = (u, v, 1)^\top$ denote its homogeneous coordinate. Given the camera-to-world pose $\mathbf{T} = [\mathbf{R} \mid \mathbf{t}]$, the corresponding Plücker coordinate of the pixel ray is defined as:

$$\mathbf{p} = \big[\mathbf{d},\, \mathbf{o} \times \mathbf{d}\big] \in \mathbb{R}^6, \quad \mathbf{o} = \mathbf{t}, \quad \mathbf{d} = \frac{\mathbf{R}\mathbf{K}^{-1}\tilde{\boldsymbol{\rho}}}{\|\mathbf{R}\mathbf{K}^{-1}\tilde{\boldsymbol{\rho}}\|_2}, \tag{1}$$

where $\mathbf{o}$ and $\mathbf{d}$ are the ray's origin and the normalized direction in the world frame. Stacking Plücker coordinates $\mathbf{p}$ over all pixels of the view yields a Plücker ray map $\mathbf{P} \in \mathbb{R}^{6\times H\times W}$. We retain two complementary encodings for such a ray representation: a tokenized ray map for query-

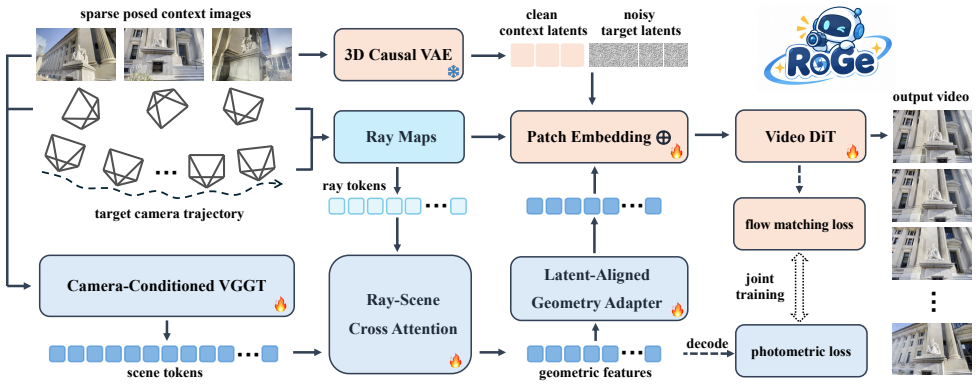



Figure 2: Overview of RoGe. Given sparse posed images and a target camera trajectory, the camera-conditioned VGGT builds an implicit scene representation. Camera rays query this representation to extract geometric cues, which are injected into the video diffusion model to synthesize a photorealistic novel trajectory video. The reconstruction and generation models are trained end-to-end.

ing the implicit scene and a lossless packed ray map directly supplied to the diffusion transformer (DiT) Peebles & Xie (2023).

**Latent Flow Matching.** We employ a pretrained video diffusion model operating in the latent space of a 3D causal Variational Autoencoder (VAE) as in Wan et al. (2025). Let $\mathbf{X}_0$ denote the clean video latents and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. Given a noise level $\sigma \in [0, 1]$ sampled from the noise schedule, we construct the rectified flow path Liu et al. (2022); Lipman et al. (2022):

$$\mathbf{X}_\sigma = (1-\sigma)\mathbf{X}_0 + \sigma\boldsymbol{\epsilon}, \qquad \mathbf{V}^* = \boldsymbol{\epsilon} - \mathbf{X}_0, \tag{2}$$

where the target velocity $\mathbf{V}^*$ is constant along the path. The diffusion transformer $\mathcal{D}_\psi$ predicts the velocity field that transports samples from Gaussian noise towards the data distribution:

$$\hat{\mathbf{V}}_\sigma = \mathcal{D}_\psi\left(\mathbf{X}_\sigma, \sigma; \mathcal{C}\right), \tag{3}$$

where $\mathcal{C}$ denotes the conditioning signal detailed in Sec. 3.2. The model $\mathcal{D}_\psi$ is trained by regressing $\hat{\mathbf{V}}_\sigma$ to $\mathbf{V}^*$. During inference, the learned velocity field is integrated from $\sigma$=1 to $\sigma$=0, and the denoised latents are then decoded into output videos.

### 3.1 Implicit Reconstruction

Our implicit reconstruction module builds on LagerNVS Szymanowicz et al. (2026), a feed-forward NVS framework based on VGGT Wang et al. (2025a) without explicit 3D reconstruction.

**Camera-Conditioned VGGT Reconstruction.** VGGT is pretrained for 3D reconstruction tasks, estimating camera poses, depths, and point maps. Its intermediate tokens thus carry strong geometric priors and implicitly represent the 3D scene, from which we extract geometric cues for conditioning subsequent video generation. Specifically, we first use a two-layer Multi-Layer Perceptron (MLP) to project the intrinsics $\mathbf{K}$ and extrinsics $\mathbf{T}$ of each context frame in $\mathcal{S}$ into a 1024-dimensional token, which is added to the camera token that VGGT prepends to each view. The camera-conditioned VGGT network $\mathcal{R}_\theta$ then aggregates multi-view information through alternating local and global attention: $\{\mathbf{Z}_i\}_{i=1}^M = \mathcal{R}_\theta\left(\{\mathbf{I}_i, \mathbf{K}_i, \mathbf{T}_i\}_{i=1}^M\right)$. For each context view, we concatenate the tokens from the last local and global attention layers along the channel dimension, yielding 2048-dimensional scene tokens $\mathbf{Z}_i$. Unlike explicit methods, we do not further instantiate depth maps, point clouds, or 3D Gaussians. Instead, the tokens $\mathcal{Z} = \{\mathbf{Z}_i\}_{i=1}^M$ form an implicit, appearance-preserving scene representation that remains directly 'renderable' or queryable from arbitrary camera rays.

**Ray Map Encoder.** We encode the Plücker ray map of each view in $\mathcal{S}$ and $\mathcal{T}$ as $\{\mathbf{Q}_i\}_{i=1}^{M+N} = \{\mathcal{E}_{\text{ray}}(\mathbf{P}_i)\}_{i=1}^{M+N}$. The 2D convolution $\mathcal{E}_{\text{ray}}$ projects each $8 \times 8$ spatial patch into a 768-dimensional ray token, yielding a set of ray tokens $\mathbf{Q}_i \in \mathbb{R}^{768 \times \frac{H}{8} \times \frac{W}{8}}$ for view $i$, with four learnable register tokens prepended to the ray tokens Darcet et al. (2024). Such a dense representation preserves the spatial correspondence between each latent location and its associated camera ray.

**Ray–Scene Cross-Attention.** Before interacting with the ray tokens, we linearly project the 2048-dimensional scene tokens $\mathcal{Z}$ to 768 dimensions. Then the ray tokens query the scene tokens through

a stack of ray–scene transformer blocks. Each block first performs self-attention among the ray tokens of view $i$ and then exchanges information with the shared scene representation:

$$\tilde{\mathbf{Q}}_i^\ell = \mathbf{Q}_i^\ell + \mathrm{SelfAttn}\big(\mathbf{Q}_i^\ell\big), \tag{4}$$

$$\mathbf{Q}_i^{\ell+1} = \tilde{\mathbf{Q}}_i^\ell + \mathrm{CrossAttn}\big(\tilde{\mathbf{Q}}_i^\ell, \mathcal{Z}^\ell, \mathcal{Z}^\ell\big), \tag{5}$$

$$\mathcal{Z}^{\ell+1} = \mathcal{Z}^\ell + \mathrm{CrossAttn}\big(\mathcal{Z}^\ell, \tilde{\mathbf{Q}}_i^\ell, \tilde{\mathbf{Q}}_i^\ell\big). \tag{6}$$

$\ell = 1, \ldots, L$. The first $L-1$ blocks use bidirectional cross-attention, allowing the scene tokens to be refined by the queried rays, while the final block performs ray-to-scene cross-attention only. After discarding the register tokens, the resulting ray tokens form a feature map $\mathbf{\Phi}_i \in \mathbb{R}^{768\times\frac{H}{8}\times\frac{W}{8}}$[1], which can be decoded to an RGB map to supervise the reconstruction module during joint training.

### 3.2 Reconstruction-Generation Fusion

After implicit reconstruction, we have the sparse context images $\{\mathbf{I}_i\}_{i=1}^{M}$, along with the Plücker ray maps $\{\mathbf{P}_i\}_{i=1}^{M+N}$ and the ray-queried geometric features $\{\mathbf{\Phi}_i\}_{i=1}^{M+N}$ for the context and target views. They respectively provide the appearance, the camera pose, and the implicit geometry that conditions the video DiT, forming $\mathcal{C}$ in Eq. (3). We describe their construction below.

**Hybrid VAE Representation.** Both the context images and the target video have to be encoded into the latent space of the VAE before entering the DiT. However, sparse context views form an unordered set, whereas the target views constitute a temporally coherent video. Encoding both with the same temporal strategy would either impose an artificial ordering on the context views or break temporal continuity in the output. We therefore adopt a hybrid latent representation: each context image $\mathbf{I}_i \in \mathcal{S}$ is independently encoded as a single-frame video following Wu et al. (2026b), while the target video is jointly encoded from the ground-truth frames $\{\mathbf{I}_j^{gt}\}_{j=1}^{N}$ available during training:

$$\mathbf{X}^{\mathcal{S}} = [\mathcal{E}_{\mathrm{VAE}}(\mathbf{I}_1), \ldots, \mathcal{E}_{\mathrm{VAE}}(\mathbf{I}_M)] \in \mathbb{R}^{16\times M\times\frac{H}{8}\times\frac{W}{8}}, \tag{7}$$

$$\mathbf{X}^{\mathcal{T}} = \mathcal{E}_{\mathrm{VAE}}\left([\mathbf{I}_1^{gt}, \ldots, \mathbf{I}_N^{gt}]\right) \in \mathbb{R}^{16\times(1+\frac{N-1}{4})\times\frac{H}{8}\times\frac{W}{8}}, \tag{8}$$

where $[\cdot,\cdot]$ denotes concatenation along the temporal dimension. The clean training sequence is then formulated as $\mathbf{X}_0 = [\mathbf{X}^{\mathcal{S}}, \mathbf{X}^{\mathcal{T}}]$. This hybrid representation preserves permutation invariance among the sparse context views while retaining the temporal structure of the causal video VAE for the target video. We additionally construct the appearance condition $\mathcal{Y} = \left[\mathbf{B}; \left[\mathbf{X}^{\mathcal{S}}, \mathbf{0}^{\mathcal{T}}\right]\right]$, where $[\cdot;\cdot]$ denotes concatenation along the channel dimension. $\mathbf{B}$ is a binary mask distinguishing context and target latent slots, and $\mathbf{0}^{\mathcal{T}}$ denotes zero-filled target latents.

**Latent-Aligned Geometry Adapter.** The implicit geometric features $\{\mathbf{\Phi}_i\}_{i=1}^{M+N}$ and VAE latents $\mathbf{X}_0$ differ in channel width, spatial resolution, and temporal organization. Hence we introduce a lightweight geometry adapter to bridge the two representations. For view $i$, $\mathbf{\Phi}_i$ is first projected from 768 to $C$ channels using a $1\times1$ convolution $\mathcal{E}_{\mathrm{geom}}$, and then bilinearly resampled to the latent resolution: $\mathbf{\Phi}_i' = \mathrm{Resize}\left(\mathcal{E}_{\mathrm{geom}}(\mathbf{\Phi}_i)\right) \in \mathbb{R}^{C\times\frac{H}{8}\times\frac{W}{8}}$. In terms of temporal alignment, each of the $M$ context frames and the first target frame is replicated four times, and the remaining $N-1$ target frames are kept unchanged. Finally, every four consecutive frames are concatenated along the channel dimension, which yields the implicit geometry condition $\mathcal{G} \in \mathbb{R}^{4C\times(M+1+\frac{N-1}{4})\times\frac{H}{8}\times\frac{W}{8}}$.

**Geometry-Grounded Diffusion Transformer.** We ground a pretrained video diffusion transformer Wan et al. (2025) $\mathcal{D}_\psi$ with both explicit camera rays and ray-queried geometric features.

Similar to Sec. 3.2, the Plücker ray maps of all views $\{\mathbf{P}_i\}_{i=1}^{M+N} \in \mathbb{R}^{6\times(M+N)\times H\times W}$ can be duplicated into $\mathbb{R}^{6\times(4M+4+N-1)\times H\times W}$, and then spatially rearranged using pixel unshuffle: $\mathcal{P} \in \mathbb{R}^{1536\times(M+1+\frac{N-1}{4})\times\frac{H}{8}\times\frac{W}{8}}$, where $6\times4\times8\times8 = 1536$. Pixel unshuffle preserves the complete per-pixel ray information while aligning with VAE latents.

Together with $\mathcal{Y}$ and $\mathcal{G}$, $\mathcal{P}$ completes the conditioning signal $\mathcal{C} = \{\mathcal{Y}, \mathcal{P}, \mathcal{G}\}$ of the video diffusion transformer. We then patchify the noisy latents and the conditions, and fuse them together:

$$\tilde{\mathbf{X}}_\sigma = \mathcal{E}_{\mathrm{patch}}^1\big([\mathbf{X}_\sigma; \mathcal{Y}]\big) + \mathcal{E}_{\mathrm{patch}}^2(\mathcal{P}) + \mathcal{E}_{\mathrm{patch}}^3(\mathcal{G}), \tag{9}$$

[1] In practice, the implicit reconstruction module operates on resized images. However, for simplicity of notation, we still use $H\times W$ to denote the image resolution in Sec. 3.1.

where $\tilde{\mathbf{X}}_\sigma \in \mathbb{R}^{5120\times(M+1+\frac{N-1}{4})\times\frac{H}{16}\times\frac{W}{16}}$ and 5120 is the hidden dimension of the DiT. $\mathcal{E}^1_{\text{patch}}$, $\mathcal{E}^2_{\text{patch}}$, and $\mathcal{E}^3_{\text{patch}}$ denote 3D convolutional patch embeddings. The fused tokens $\tilde{\mathbf{X}}_\sigma$ are then processed by the subsequent DiT blocks. Note that temporal rotary embeddings are disabled on context latent slots for permutation equivariance, and are retained only along the target trajectory.

### 3.3 End-to-End Joint Training

Our model is trained with reconstruction and generation supervision in a single optimization step:

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \lambda\mathcal{L}_{\text{photometric}}. \tag{10}$$

**Flow Matching Loss.** Given clean latents $\mathbf{X}_0 = [\mathbf{X}^{\mathcal{S}}, \mathbf{X}^{\mathcal{T}}]$, we sample a noise level $\sigma$ and Gaussian noise $\boldsymbol{\epsilon}$ to construct noisy latents $\mathbf{X}_\sigma$ and the ground-truth velocity $\mathbf{V}^*$ as in Eq. (2). Since the context images are observed inputs, the loss is applied only to the target latent slots:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\mathbf{X}_0,\boldsymbol{\epsilon},\sigma}\left[w(\sigma)\cdot\left\|\mathcal{M}_t\odot\left(\hat{\mathbf{V}}_\sigma - \mathbf{V}^*\right)\right\|_2^2\right], \tag{11}$$

where $\hat{\mathbf{V}}_\sigma = \mathcal{D}_\psi(\mathbf{X}_\sigma, \sigma; \mathcal{C})$ as in Eq. (3), $\mathcal{M}_t$ selects the target latent slots with the operator $\odot$, and $w(\sigma)$ is the noise-level-dependent training weight.

**Photometric Loss.** The reconstruction module projects the geometric features $\{\mathbf{\Phi}_i\}_{i=1}^{M+N}$ into RGB maps according to LagerNVS, supervised by a pixel-wise mean-squared error $\mathcal{L}_{\text{photometric}}$ against the corresponding ground-truth images. This auxiliary loss works jointly with the generator to refine the reconstruction module while preventing it from forgetting its learned geometric priors. No depth, point map, or any other explicit 3D supervision is used in our training.

## 4 Experiments

### 4.1 Experimental Setup

**Implementation Details.** The camera-conditioned VGGT $\mathcal{R}_\theta$, ray map encoder $\mathcal{E}_{\text{ray}}$, and $L = 12$ ray–scene cross-attention blocks are initialized from LagerNVS Szymanowicz et al. (2026). We adopt Wan2.1-I2V-14B-480P Wan et al. (2025) as our video generation backbone, containing the VAE $\mathcal{E}_{\text{VAE}}$, the patch embedding $\mathcal{E}^1_{\text{patch}}$ for video latents and RGB conditions, and the DiT $\mathcal{D}_\psi$. The geometry adapter $\mathcal{E}_{\text{geom}}$, which projects 768-dimensional ray-queried features to $C$=384 channels, and the patch embeddings $\mathcal{E}^2_{\text{patch}}$ for ray maps and $\mathcal{E}^3_{\text{patch}}$ for implicit geometry are randomly initialized. During training, we apply Low-Rank Adaptation (LoRA) Hu et al. (2021) with rank 32 to $\mathcal{D}_\psi$ and rank 8 to $\mathcal{R}_\theta$, $\mathcal{E}_{\text{ray}}$, and the ray-scene cross-attention blocks. All parameters except the LoRA modules, $\mathcal{E}_{\text{geom}}$, and the three patch embeddings are frozen. We randomly sample $M \in 3, \ldots, 9$ context views and use $N = 81$ temporally ordered target frames. The same views are fed to the feed-forward reconstruction module at $288 \times 504$ and to the video generation module at $192 \times 336$ or $480 \times 832$. We first train the model at $192 \times 336$ for 15,000 steps on 192 H20 GPUs with a batch size of 1 and $\lambda_{\text{render}} = 0.1$. We then further fine-tune the model at $480 \times 832$ for 500 steps on 192 H20 GPUs, with the entire reconstruction module fully frozen.

**Datasets.** We train on RealEstate10K (RE10K) Zhou et al. (2018) and DL3DV Ling et al. (2024), which cover indoor and outdoor scenes with diverse camera trajectories and have been widely utilized for the training of feed-forward 3D models and camera-controllable video generation models. We follow CUT3R Wang et al. (2025b) for data preprocessing and official splits. After filtering, the training set contains 6,527 sequences from RE10K and 6,264 sequences from DL3DV. For quantitative evaluation, we use the 10 RE10K test sequences and 10 DL3DV test sequences selected by SEVA Zhou et al. (2025). For cross-dataset generalization, we conduct experiments on ACID Liu et al. (2021), a large-scale aerial coastline dataset frequently used by prior works. We select 10 sequences with substantial viewpoint variation for evaluation, with an average of 474 frames per sequence, and follow SEVA to obtain the context and target splits. To cover the scene as comprehensively as possible with a minimal number of context views, we use 3, 6, and 9 context views for RE10K, DL3DV, and ACID, respectively. In addition, we evaluate NVS along extrapolated camera trajectories on ScanNet Dai et al. (2017) to assess the capability of roaming.

**Baselines.** We compare against representative state-of-the-art methods on NVS: 1) *reconstruction-based*: 3DGS Kerbl et al. (2023), AnySplat Jiang et al. (2025), YoNoSplat Ye et al. (2026),

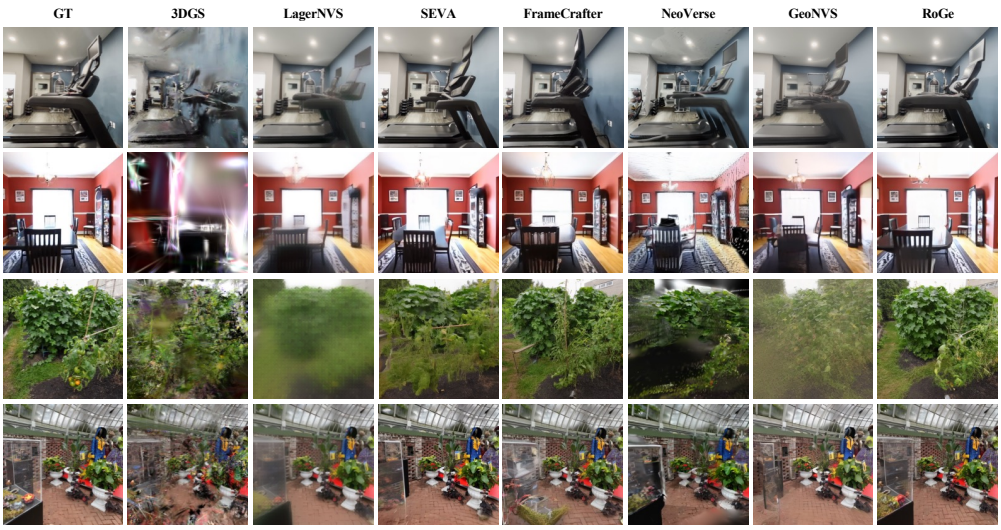



Figure 3: Qualitative NVS results across different methods with 3 or 6 sparse inputs.

Table 1: Quantitative NVS results on RE10K and DL3DV. *R* and *G* denote reconstruction-only methods and generation-only methods, respectively. *R+G* means reconstruction-generation-combined methods, where *1* and *2* represent generation-aided reconstruction and decoupled reconstruction-then-generation. *3* tightly couples two modules end-to-end without any explicit 3D. Difix3D and Difix apply generative priors during 3DGS training and after rendering, respectively.

| Method | Type | RE10K 3v | | | | DL3DV 6v | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | PSNR↑ | SSIM↑ | LPIPS↓ | DreamSim↓ | PSNR↑ | SSIM↑ | LPIPS↓ | DreamSim↓ |
| 3DGS | R | 14.76 | 0.449 | 0.469 | 0.380 | 13.75 | 0.308 | 0.426 | 0.438 |
| AnySplat | R | 16.77 | 0.572 | 0.190 | 0.069 | 13.21 | 0.279 | 0.315 | 0.176 |
| YoNoSplat | R | 18.83 | 0.641 | 0.142 | 0.063 | 17.04 | 0.459 | 0.277 | 0.184 |
| LagerNVS | R | <u>32.09</u> | <u>0.955</u> | 0.031 | 0.036 | <u>20.90</u> | <u>0.639</u> | 0.193 | 0.122 |
| SEVA | G | 29.26 | 0.929 | 0.036 | 0.025 | 18.63 | 0.603 | 0.158 | 0.080 |
| FrameCrafter | G | 29.03 | 0.925 | 0.037 | <u>0.022</u> | 19.85 | 0.609 | <u>0.121</u> | <u>0.057</u> |
| Difix3D | R+G, 1 | 14.85 | 0.452 | 0.425 | 0.325 | 13.89 | 0.313 | 0.388 | 0.375 |
| 3DGS + Difix | R+G, 2 | 14.42 | 0.427 | 0.461 | 0.358 | 13.45 | 0.276 | 0.421 | 0.413 |
| LagerNVS + Difix | R+G, 2 | 30.98 | 0.895 | <u>0.027</u> | 0.034 | 19.74 | 0.574 | 0.181 | 0.119 |
| GEN3C | R+G, 2 | 23.55 | 0.816 | 0.096 | 0.043 | 12.78 | 0.256 | 0.434 | 0.269 |
| NeoVerse | R+G, 2 | 16.20 | 0.538 | 0.264 | 0.102 | 16.45 | 0.432 | 0.237 | 0.166 |
| GeoNVS | R+G, 2 | 27.89 | 0.913 | 0.049 | 0.047 | 18.78 | 0.619 | 0.155 | 0.077 |
| RoGe | R+G, 3 | **33.35** | **0.965** | **0.021** | **0.020** | **21.50** | **0.704** | **0.104** | **0.052** |

LagerNVS Szymanowicz et al. (2026); 2) *generation-based*: SEVA Zhou et al. (2025), FrameCrafter Wu et al. (2026b); 3) *reconstruction-generation-combined*: Difix3D+ Wu et al. (2025a) (facilitating 3DGS training or denoising rendered images), GEN3C Ren et al. (2025), NeoVerse Yang et al. (2026), GeoNVS Kang et al. (2026). Since FrameCrafter is trained on only 1K DL3DV samples, we fine-tune it on the same data for the same number of steps as RoGe for a fair comparison. Note that NeoVerse is originally designed for synchronized video retargeting. We adapt it to sparse NVS by aggregating the predicted Gaussians from all frames for rendering.

**Metrics.** Since we focus on trajectory NVS Zhou et al. (2025) rather than merely isolated novel views, we consider video-level metrics in addition to image-based ones. Image-level metrics include PSNR, SSIM, LPIPS, and DreamSim Fu et al. (2023). While for video-level evaluation, we adopt FID Heusel et al. (2017), FVD Unterthiner et al. (2018), TSED Yu et al. (2023), and MEt3R Asim et al. (2025). Specifically, FID and FVD assess the visual realism and temporal coherence of the generated video, while TSED and MEt3R focus on the geometric consistency. We further use VGGT-Ω Wang et al. (2026a) to estimate the camera trajectory from the generated video and evaluate the RMSE of the translational and rotational Absolute Pose Error (APE), which simultaneously reflects visual quality, multi-view consistency, and camera controllability. We note that the ground-truth camera trajectories in RE10K, DL3DV, and ACID do not have an absolute scale. Therefore, for translational APE, we normalize the length of each ground-truth trajectory before computing the error. The ACID sequences contain nearly straight-line trajectories, for which the rotational component of $Sim(3)$ alignment can be ambiguous, making rotational APE unreliable. We therefore report rotational Relative Pose Error (RPE) instead in ACID.

| | RE10K 3v | | | | | | DL3DV 6v | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FID↓ | FVD↓ | TSED↑ | MEt3R↓ | APE-T↓ | APE-R↓ | FID↓ | FVD↓ | TSED↑ | MEt3R↓ | APE-T↓ | APE-R↓ |
| LagerNVS | 21 | 76 | 0.994 | **0.014** | **0.007** | 12.51 | 33 | 206 | 0.977 | **0.062** | **0.005** | 2.42 |
| SEVA | 15 | 78 | 0.987 | 0.019 | 0.018 | 20.99 | 23 | 85 | 0.972 | 0.076 | 0.011 | 5.38 |
| FrameCrafter | 14 | 83 | 0.969 | 0.018 | 0.030 | 28.37 | 16 | 89 | 0.720 | 0.080 | 0.018 | 6.18 |
| GEN3C | 31 | 154 | **0.997** | 0.020 | 0.027 | 23.68 | 61 | 281 | 0.854 | 0.089 | 0.057 | 35.24 |
| NeoVerse | 69 | 324 | **0.997** | 0.023 | 0.009 | 10.69 | 59 | 184 | 0.919 | 0.079 | 0.008 | 2.91 |
| GeoNVS | 24 | 92 | 0.986 | 0.021 | 0.018 | 23.63 | 29 | 136 | 0.928 | 0.083 | 0.021 | 11.63 |
| RoGe | **11** | **45** | **0.997** | 0.015 | 0.009 | **9.45** | **15** | **50** | **0.991** | 0.071 | **0.005** | **2.31** |

Table 2: Quantitative comparison on RE10K and DL3DV with video-level metrics.

Table 3: Quantitative comparison on ACID with both image-level and video-level metrics.

| | ACID 9v | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | DreamSim↓ | FID↓ | FVD↓ | TSED↑ | MEt3R↓ | APE-T↓ | RPE-R↓ |
| LagerNVS | 24.11 | 0.708 | 0.139 | 0.088 | 29 | 133 | 0.994 | 0.028 | 0.032 | 0.21 |
| SEVA | 22.97 | 0.696 | 0.094 | 0.048 | 12 | 78 | 0.969 | 0.048 | 0.043 | 0.37 |
| FrameCrafter | 24.58 | 0.734 | 0.081 | 0.040 | 12 | 88 | 0.973 | 0.031 | 0.046 | 0.38 |
| GEN3C | 16.39 | 0.427 | 0.288 | 0.153 | 43 | 313 | 0.989 | 0.031 | 0.057 | 0.22 |
| NeoVerse | 15.99 | 0.401 | 0.322 | 0.209 | 89 | 399 | **0.995** | 0.032 | 0.036 | 0.27 |
| GeoNVS | 22.07 | 0.656 | 0.134 | 0.082 | 24 | 285 | 0.934 | 0.057 | 0.058 | 0.70 |
| RoGe | **26.30** | **0.786** | **0.064** | **0.032** | **10** | **36** | **0.995** | **0.027** | **0.028** | **0.20** |

### 4.2 Experimental Results and Analysis

**Novel View Synthesis.** For each test sequence, image-based methods render, predict, or generate each non-context frame individually, whereas video-based methods generate non-context frames in 81-frame chunks, with each chunk produced in a single forward pass following their default inference settings. Since different methods produce images with different aspect ratios, we adopt a similar strategy to Wu et al. (2026b) by cropping and resizing them to $192 \times 192$ to ensure identical FoV for final evaluation. As in Tab. 1 and Fig. 3, per-scene optimization-based 3DGS struggles with sparse inputs, and Difix3D+ provides limited improvement by incorporating generative priors during optimization or directly refining the renderings. Feed-forward reconstruction methods such as YoNoSplat achieve stronger performance, but still suffer from holes and artifacts in regions not covered by the input views. Despite its high PSNR, LagerNVS tends to produce overly blurred results, especially under large viewpoint changes. On the other hand, generation-only methods such as SEVA and FrameCrafter exhibit strong capabilities of extrapolation and completion, but sometimes suffer from unstable camera control and geometric distortions. Hybrid methods, including GEN3C, NeoVerse, and GeoNVS, first reconstruct an explicit 3D representation and then obtain generation conditions from it through post-processing. The relatively complex post-processing pipeline plays a significant role in determining the final performance, while the two stages remain fully decoupled. As a result, they achieve only suboptimal performance. RoGe efficiently extracts geometric features from the implicit 3D representation to condition the video generation, while jointly training the reconstruction and generation modules end-to-end, achieving state-of-the-art visual quality across all the image-level metrics even on the generalization dataset (shown in Tab. 3).

We also evaluate on the output 81-frame videos of each method. As reported in Tab. 2 and Tab. 3, FrameCrafter delivers good visual quality but struggles with geometric consistency. In contrast, reconstruction-based methods like LagerNVS display strong multi-view consistency and accurate camera control. By fully fusing reconstruction and generation, RoGe produces temporally coherent videos with high visual fidelity, consistent scene geometry, and accurate camera controllability.

**Roaming in the Scene.** We further evaluate scene roaming along extrapolated trajectories that deviate from the trajectory along which the context views were captured. Tab. 4 and Fig. 1 present the results on ScanNet, showcasing RoGe's ability to extrapolate beyond the observed views while preserving the known scene geometry and faithfully following the prescribed camera trajectory.

**Extension for Video Retargeting.** We extend RoGe to video retargeting for 4D world modeling. Specifically, given a video with a known camera trajectory and a target trajectory for retargeting, we independently encode each posed context frame into scene tokens. We then apply ray–scene cross-attention to obtain ray-queried geometric features for each context frame and its corresponding target frame. The context video is instead encoded in the same way as the target video in Eq. (8). We initialize the adapted model from the pretrained RoGe and further fine-tune it on VKitti Cabon et al. (2020) for 5,000 steps using 16 H20 GPUs with a batch size of 1. Fig. 1 shows an example on a test sequence, where a video captured from $15^\circ$ left is retargeted to $15^\circ$ right.

| | MEt3R↓ | APE-T↓ | APE-R↓ |
|---|---|---|---|
| LagerNVS | **0.059** | <u>0.20</u> | <u>3.52</u> |
| SEVA | 0.087 | 0.23 | 5.21 |
| FrameCrafter | 0.153 | 0.57 | 7.96 |
| GEN3C | 0.114 | 1.48 | 10.61 |
| NeoVerse | <u>0.069</u> | 0.25 | 4.92 |
| GeoNVS | 0.085 | 0.27 | 5.09 |
| RoGe | **0.059** | **0.15** | **3.12** |

Table 4: Quantitative comparison on the synthesized videos when roaming.

| | Geometric Condition | Joint Train | Resolution | PSNR↑ | SSIM↑ | LPIPS↓ |
|---|---|---|---|---|---|---|
| #1 | None (generation only) | ✗ | $192 \times 336$ | 18.49 | 0.517 | 0.156 |
| #2 | VGGT Patch tokens | ✗ | $192 \times 336$ | 17.98 | 0.477 | 0.168 |
| #3 | VGGT Patch tokens with camera-conditioned | ✗ | $192 \times 336$ | 18.83 | 0.541 | 0.147 |
| #4 | Rendered RGB | ✗ | $192 \times 336$ | 19.36 | 0.564 | 0.141 |
| #5 | Predicted RGB | ✗ | $192 \times 336$ | 19.49 | 0.573 | 0.139 |
| #6 | Ray-queried geometric features | ✗ | $192 \times 336$ | 20.07 | 0.608 | 0.133 |
| #7 | Ray-queried geometric features | ✓ | $192 \times 336$ | 20.40 | 0.624 | 0.130 |
| #8 | Ray-queried geometric features | ✓ | $480 \times 832$ | **21.50** | **0.704** | **0.104** |

Table 5: Ablation on geometric conditioning and joint training on DL3DV. #1–#6 freeze the reconstruction branch and differ only in what is injected into the video generation module. #7 additionally trains both modules end-to-end. #8 further fine-tunes the model at a higher resolution of $480 \times 832$.

### 4.3 Ablation Studies

We ablate on DL3DV to answer two questions: what geometric condition to inject into the video generation branch, and whether to train the two modules jointly. Results are reported in Tab. 5.

**Geometric Condition.** First of all, naively injecting raw VGGT patch tokens (#2) is worse than no geometry at all (#1). The world frame and scale underlying VGGT tokens are generally inconsistent with those of the input context and target poses. Although one could attempt to align them via $Sim(3)$, VGGT pose estimation or the alignment itself might fail, and even successful alignment cannot re-express the VGGT tokens in the new world frame and scale. As a result, the diffusion model receives geometry and camera control that disagree with each other. We instead condition VGGT with the input cameras (#3), so that its features are built directly in the given frame. This resolves the mismatch and brings a modest gain over #1.

We query per-view geometric features (#6) from the tokens output by the camera-conditioned VGGT and further decode them into RGB maps (#5) following LagerNVS. We also experiment with connecting the camera-conditioned VGGT to the GS head of World-Mirror Liu et al. (2025), training the model on RE10K and DL3DV to predict Gaussians and render them into RGB maps (#4). In comparison, using RGB maps from either implicit or explicit reconstruction as the conditioning signal provides less benefit than directly using the implicit geometric features.

**Joint Training.** With ray-queried features as the condition, training the reconstruction branch together with the generation branch (#7) further improves all metrics. The generation objective can now shape the geometric features it is conditioned on, which the frozen reconstruction branch in #1–#6 cannot benefit from. After fine-tuning at a higher resolution (#8), the visual quality is improved.

## 5 Conclusion

In this work, we present RoGe, a unified reconstruction and generation framework for novel view synthesis. RoGe extracts per-view implicit geometric features from a feed-forward reconstruction model and injects them into a video generation model, tightly coupling geometric reconstruction with generation without an intermediate explicit 3D representation. End-to-end joint training enables effective information exchange between the two modules, allowing RoGe to synthesize temporally and coherent videos along arbitrary camera trajectories with strong geometric consistency and visual fidelity, including in unseen regions. Extensive experiments on public dataset demonstrate that RoGe surpasses state-of-the-art reconstruction-based, generation-based, and hybrid methods.

### AI USE STATEMENT

In this work, we used generative AI tools for tasks of required disclosure: implementing our proposed method, aiding or polishing writing. Generative AI was not used to generate synthetic datasets, and other mandatory disclosure tasks do not apply to this work. For recommended disclosure tasks, generative AI helped revise scientific figures, identify research gaps, and draft partial manuscript paragraphs.We have manually reviewed all AI-assisted outputs. We cross-checked cited literature, verified AI-assisted code, and examined the factual accuracy of text descriptions and experimental observations. The authors take full responsibility for all final content of this paper, including text, claims and research artifacts generated with generative AI assistance.

### ETHICS STATEMENT

This work adheres to the academic ethics policy of ICLR regarding generative AI usage. All generative AI assistance adopted in this research does not compromise the originality of our core contributions. We have screened AI-generated content to rule out plagiarism, factual hallucinations and misleading claims.All key mathematical derivations, experimental designs and conclusions are conceived and validated by the authors. No undisclosed AI assistance exists in this manuscript. We confirm that all AI usage follows community ethical guidelines, and we bear full ethical accountability for every claim presented in this work.

### REPRODUCIBILITY STATEMENT

The experimental pipeline, mathematical derivations and implementation code of this work are fully reproducible. Although generative AI assisted in code writing and experiment planning, all hyperparameters, mathematical steps and implementation details were finalized, debugged and documented manually by the authors.We will release the source code and complete experimental configurations. The AI-assisted parts of code have been independently tested to ensure consistent experimental outcomes. The mathematical proofs aided by generative AI have been manually re-derived and verified to guarantee correctness for independent reproduction.

## REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6034–6044. IEEE, 2025.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844. IEEE, 2025a.

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. In *International Conference on Learning Representations*, volume 2025, pp. 58038–58060, 2025b.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19457–19467. IEEE, 2024.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European conference on computer vision*, pp. 370–386. Springer, 2024.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2432–2443. IEEE, 2017.

Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. In *International Conference on Learning Representations*, volume 2026, pp. 103603–103622, 2026.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

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

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

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13416–13426. IEEE, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Chun-Hao Paul Huang, Niloy Mitra, Hyeonho Jeong, Jae Shin Yoon, and Duygu Ceylan. Jog3r: Towards 3d-consistent video generators. *arXiv preprint arXiv:2501.01409*, 2025.

Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3r: 3d scene generation meets feed-forward reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25358–25369, 2026.

Hyeonho Jeong, Suhyeon Lee, and Jong Chul Ye. Reangle-a-video: 4d video generation as video-to-video translation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11164–11175. IEEE, 2025.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *ACM Transactions on Graphics (TOG)*, 44(6):1–16, 2025.

Minjun Kang, Inkyu Shin, Taeyeop Lee, Myungchul Kim, In So Kweon, and Kuk-Jin Yoon. Geonvs: Geometry grounded video diffusion for novel view synthesis. In *European Conference on Computer Vision*, pp. 565–586. Springer, 2026.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction; map-anything. github. io. In *2026 International Conference on 3D Vision (3DV)*, pp. 499–509. IEEE, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Yu Tong Tiffany Ling, Yuhan Huang, Rushikesh Zawar, Xue Bai, et al. Towards understanding camera motions in any video. *Advances in Neural Information Processing Systems*, 38, 2026.

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

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 14458–14467, 2021.

Xi Liu, Chaoyi Zhou, and Siyu Huang. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. *Advances in Neural Information Processing Systems*, 37:133305–133327, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. Worldmirror: Universal 3d world reconstruction with any-prior prompting. *arXiv preprint arXiv:2510.10726*, 2025.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Julius Plucker. I. on a new geometry of space. *Proceedings of the Royal Society of London*, (14): 53–58, 1865.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132. IEEE, 2025.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Stanislaw Szymanowicz, Minghao Chen, Jianyuan Wang, Christian Rupprecht, and Andrea Vedaldi. Lagernvs: Latent geometry for fully neural real-time novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15443–15453, 2026.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-omega. *arXiv preprint arXiv:2605.15195*, 2026a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522. IEEE, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024a.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. pi3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, volume 2026, pp. 10481–10497, 2026b.

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

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. In *International Conference on Learning Representations*, volume 2026, pp. 119186–119209, 2026a.

Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26024–26035. IEEE, 2025a.

Qi Wu, Khiem Vuong, Minsik Jeon, Srinivasa Narasimhan, and Deva Ramanan. Novel view synthesis as video completion. *arXiv preprint arXiv:2604.08500*, 2026b.

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21551–21561, 2024.

Sibo Wu, Congrong Xu, Binbin Huang, Andreas Geiger, and Anpei Chen. Genfusion: Closing the loop between reconstruction and generation via videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6078–6088. IEEE, 2025b.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16453–16463. IEEE, 2025.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. 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. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In *International Conference on Learning Representations*, volume 2025, pp. 54009–54033, 2025.

Botao Ye, Boqi Chen, Haofei Xu, Daniel Barath, and Marc Pollefeys. Yonosplat: You only need one model for feedforward 3d gaussian splatting. In *International Conference on Learning Representations*, volume 2026, pp. 39852–39871, 2026.

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

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 100–111. IEEE, 2025.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38027–38037, 2026.

Jason Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *International conference on learning representations*, volume 2024, pp. 23345–23366, 2024.

Haoyu Zhao, Jiaxi Gu, Haoran Chen, Qingping Zheng, Yeying Jin, Hongyi Yang, Junqi Cheng, Yuang Zhang, Zenghui Lu, Huan Yu, et al. Cameranoise: Enabling faithful camera control in video diffusion through geometry-flow-guided noise warping. *arXiv preprint arXiv:2605.30774*, 2026.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12405–12414. IEEE, 2025.

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

Chen Ziwen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Li Fuxin, and Zexiang Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4349–4359. IEEE, 2025.