VISUAL GEOMETRY MEETS DIFFUSION FOR SPARSE-VIEW NOVEL VIEW SYNTHESIS

Anonymous authors
Paper under double-blind review



Figure 1: **VGGT-*Diff* grounds a pretrained video diffusion prior with geometry-routed visual features for sparse-view novel-view synthesis.** It maintains strong target-view fidelity across varying pose difficulties while better preserving source-observed details than diffusion-based baselines.

ABSTRACT

We present VGGT-*Diff*, a geometry-routed multi-view diffusion model for sparse-view novel view synthesis. Existing novel view synthesis (NVS) methods face a fundamental trade-off: reconstruction-based approaches preserve observed geometry but struggle to synthesize unseen regions, while diffusion-based methods provide strong generative priors yet rely on implicit source-to-query correspondence. VGGT-*Diff* bridges these regimes by routing visual geometry latents from VGGT-Ω into a pretrained video diffusion model. Each visual token is associated with a 3D point and confidence, then transformed into query-aligned latent conditions through a confidence-aware Visual Geometry Router (VGR) that preserves front and back surface evidence. These conditions guide joint target-view denoising, while Point-Track Residual Consistency (PTRC) regularizes predicted-clean residuals along reliable 3D tracks, improving multi-view stability. We further introduce robust geometry conditioning, combining training-time regularization with inference-time guidance for improved robustness. Experiments show competitive or state-of-the-art performance across interpolation and extrapolation under different viewpoint difficulties. Our code will be released publicly.

1 INTRODUCTION

Novel view synthesis (NVS) aims to render previously unseen viewpoints of a scene from sparse observations. Existing generalizable methods broadly follow two complementary paradigms. Reconstruction-oriented approaches infer explicit or implicit scene representations, including neural radiance fields, Gaussian primitives, and feed-forward 3D representations (Mildenhall et al., 2021; Kerbl et al., 2023; Charatan et al., 2024; Chen et al., 2024a), while Transformer-based models such as LVSM (Jin et al., 2024), RayZer (Jiang et al., 2025a), LagerNVS (Szymanowicz et al., 2026), and SVSM (Kim et al., 2026) learn view synthesis with reduced explicit 3D inductive bias. These approaches provide strong geometric fidelity when target views are well supported by observations, but sparse inputs inevitably leave parts of the scene unobserved, making deterministic reconstruction

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.

increasingly under-constrained under large viewpoint changes. In unseen regions, predictions can degenerate into patch-like artifacts biased toward colors observed in nearby source views.

A complementary line formulates NVS as conditional generation and exploits image or video diffusion priors to complete unseen content (Watson et al., 2022; Kong et al., 2024; Zheng & Vedaldi, 2024; Müller et al., 2024; Yu et al., 2024; Zhou et al., 2025; Wu et al., 2026a). Such models provide strong generative priors, but often lack explicit geometric conditioning to anchor generation to observed scene structure. As a result, generative priors may override geometry-supported evidence and hallucinate plausible yet inconsistent content, leading to structural drift, unstable occlusions, and cross-view inconsistency under wide-baseline interpolation and extrapolation. Meanwhile, visual geometry foundation models such as VGGT (Wang et al., 2025) and VGGT-Ω (Wang et al., 2026a) learn rich multi-view representations encoding appearance, 3D structure, correspondence, and confidence, offering a natural way to ground generative NVS with explicit geometric evidence.

These limitations suggest that generative NVS should combine strong completion priors with explicit yet uncertainty-aware geometric guidance. We introduce VGGT-*Diff*, a geometry-routed multi-view diffusion framework that injects visual geometry latents from VGGT-Ω into a pretrained video diffusion model. Rather than rediscovering geometry from RGB and camera rays alone, VGGT-*Diff* associates visual tokens with 3D locations and confidence to construct spatially aligned conditions for source and query views. Since hard projection is brittle near occlusions and uncertain geometry, we propose a confidence-aware Visual Geometry Router (VGR) that preserves front and back surface evidence together with visibility uncertainty, providing query-aligned guidance without treating reconstructed geometry as a complete scene explanation.

A lightweight input projection maps the concatenated source RGB latents, Plücker ray maps, noisy query latents, and routed geometry conditions into the pretrained DiT latent space, allowing VGGT-*Diff* to reuse the video diffusion prior with minimal architectural modification. VGGT-*Diff* then jointly denoises multiple query views within the same DiT sequence, enabling direct information exchange across targets. Joint generation alone, however, does not guarantee consistency with a shared 3D scene. We therefore introduce Point-Track Residual Consistency (PTRC), which uses VGGT-derived 3D correspondences to align denoising residual errors of the same physical point across generated views, rather than forcing view-dependent features or VAE latents to match. Since geometry recovered from sparse observations is inevitably imperfect, we randomly retain, attenuate, or remove the routed geometry condition during training to prevent over-reliance on uncertain projections. This regularization also enables optional Geometry-Prior CFG at inference, which strengthens geometry-aware denoising and further improves novel-view synthesis quality. In this way, visual geometry both conditions the diffusion process and regularizes multi-view generation.

Together, VGGT-*Diff* combines geometry foundation priors with video diffusion for faithful and generative sparse-view NVS. Our contributions are threefold: **(1) Geometry-routed generation.** We introduce a multi-view diffusion framework whose confidence-aware Visual Geometry Router (VGR) transforms VGGT-Ω features and 3D locations into query-aligned conditions, grounding generative completion in observed scene structure; **(2) Geometry-grounded consistency.** We propose Point-Track Residual Consistency (PTRC), which aligns predicted-clean residuals along reliable 3D correspondences to reduce cross-view drift without suppressing valid view-dependent appearance; and **(3) Geometry-condition regularization.** We stochastically attenuate or drop routed geometry during training to prevent over-reliance on imperfect projections, improving robustness under sparse or inaccurate geometry. The dropped-condition branch also supports optional matched guidance at inference. Extensive experiments demonstrate competitive or state-of-the-art performance across pose difficulties and improved geometric reconstructability of jointly generated views.

2 Related Work

Generalizable and Feed-Forward View Synthesis. Novel view synthesis has progressed from scene-specific representations such as NeRF and Gaussian Splatting (Mildenhall et al., 2021; Barron et al., 2021; Kerbl et al., 2023; Chen et al., 2025; 2026) toward generalizable models that amortize reconstruction across scenes. Early approaches infer radiance fields or image-based representations through pixel-aligned features, multi-view stereo, epipolar reasoning, or learned ray aggregation (Yu et al., 2021; Chen et al., 2021; Wang et al., 2021; 2022). More recent methods reduce explicit 3D inductive bias and learn view synthesis with Transformers, including SRT (Sajjadi et al., 2022),

ViewFormer (Kulhánek et al., 2022), LVSM (Jin et al., 2024), Efficient-LVSM (Jia et al., 2026), and RayZer (Jiang et al., 2025a). In parallel, feed-forward models directly predict renderable 3D representations (Charatan et al., 2024; Chen et al., 2024a; Szymanowicz et al., 2024; Hong et al., 2024; Zhang et al., 2024), while recent works further improve geometry-aware scaling and efficiency through LagerNVS (Szymanowicz et al., 2026), SVSM (Kim et al., 2026), projective conditioning (Wu et al., 2026b), and SHARP (Mescheder et al., 2026). These approaches provide strong geometric fidelity and rendering, but deterministic reconstruction remains under-constrained when sparse observations do not cover content revealed by distant query views.

Generative and Geometry-Guided Novel View Synthesis. Generative NVS uses image or video priors to complete unobserved regions. Early diffusion methods perform pose-conditioned synthesis or model joint multi-view distributions (Watson et al., 2022; Liu et al., 2023; Shi et al., 2023; 2024; Liu et al., 2024; Ye et al., 2024; Kong et al., 2024; Zheng & Vedaldi, 2024), while later methods exploit video priors for stronger cross-view coherence (Kwak et al., 2024; Voleti et al., 2024; Müller et al., 2024; Yu et al., 2024; Zhou et al., 2025). Stronger pixel-space backbones improve end-to-end NVS (Elata et al., 2025), and FrameCrafter (Wu et al., 2026a) adapts pretrained video diffusion to complete unordered posed views. Related work studies hybrid deterministic and generative modeling (Le et al., 2026), test-time video completion (Xu et al., 2026), dynamic arbitrary-view generation (Van Hoorick et al., 2026), and correspondence-supervised diffusion (Kwon et al., 2026). However, source-to-query correspondence often remains implicit, limiting robustness to large viewpoint changes and cross-view consistency. Geometry foundation models such as DUSt3R (Wang et al., 2024b), MASt3R (Leroy et al., 2024), VGGT (Wang et al., 2025), and VGGT-Ω (Wang et al., 2026a) encode multi-view geometry, correspondence, and cameras. Their geometric evidence has been combined with Gaussian reconstruction and latent video diffusion (Chen et al., 2024b), sparse-view 3D optimization (Wang et al., 2024a), joint image and geometry diffusion (Kwak et al., 2026), latent generative refinement (Hirschorn et al., 2026), geometry-conditioned video diffusion (Kang et al., 2026), and wide-baseline guidance (Zhou et al., 2026). Unlike methods that construct complete Gaussian or radiance-field scenes, VGGT-*Diff* routes visual geometry features, 3D points, and confidence into query-aligned diffusion conditions and reuses their correspondences to regularize jointly generated views. Geometry therefore guides and constrains the generative prior as uncertainty-aware evidence rather than replacing it with deterministic reconstruction.

3 METHOD

Given sparse observations, novel-view synthesis must preserve the scene evidence visible in the source images while completing unobserved regions. Reconstruction models encode correspondence explicitly but have limited support for such completion; video diffusion models provide strong generative priors but leave source-to-query correspondence largely implicit. VGGT-*Diff* uses geometry to connect these two capabilities. Rather than treating the recovered geometry as a complete scene representation, we use it to route visual evidence into a pretrained multi-view diffusion model and to regularize the generated views along shared 3D point tracks. The pipeline is shown in Fig. 2.

3.1 CAMERA-CONDITIONED MULTI-VIEW DIFFUSION

We consider sparse-view novel-view synthesis from M posed source images to N prescribed query cameras. The observed views and requested cameras are represented as

$$\mathcal{S} = \{(\mathbf{I}_i^{\mathrm{s}}, \mathbf{K}_i^{\mathrm{s}}, \mathbf{T}_i^{\mathrm{s}})\}_{i=1}^{M}, \quad \mathcal{Q} = \{(\mathbf{K}_j^{\mathrm{q}}, \mathbf{T}_j^{\mathrm{q}})\}_{j=1}^{N}. \tag{1}$$

Our goal is to synthesize query views $\{\widehat{\mathbf{I}}_j^{\mathrm{q}}\}_{j=1}^{N}$. We place source and query views in one diffusion sequence, enabling information exchange across slots while preserving view-specific camera controls. A frozen VAE encodes training views into the clean latent volume $\mathbf{Z}_0 = [\mathbf{Z}_0^{\mathrm{s}}; \mathbf{Z}_0^{\mathrm{q}}]$. Following Wan's linear flow formulation (Wan et al., 2025), we sample a flow time σ and Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\mathbf{X}_\sigma = (1-\sigma)\mathbf{Z}_0 + \sigma\boldsymbol{\epsilon}, \quad \mathbf{U} = \boldsymbol{\epsilon} - \mathbf{Z}_0, \quad \mathcal{L}_{\mathrm{FM}} = \frac{1}{N}\sum_{j=1}^{N} \|\mathbf{V}_{\theta,j} - \mathbf{U}_j\|_2^2 . \tag{2}$$

Here, $\mathbf{X}_\sigma$ is the diffusion state and $\mathbf{V}_{\theta,j}$ is the predicted velocity for query view j. The loss applies only to query slots. Query images define training targets but are never exposed as conditions.

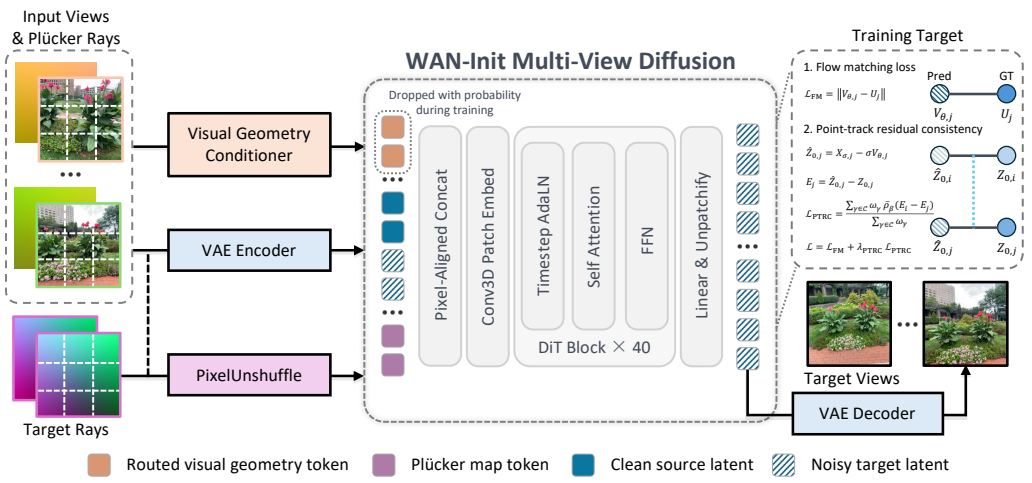

Figure 2: **Overview of VGGT-*Diff*.** The Visual Geometry Conditioner uses VGGT-Ω points and confidence to route appearance-bearing source features into query-aligned conditions. The diffusion state, clean-source image condition, Plücker rays, and routed visual condition are concatenated channel-wise, embedded independently within each view, and jointly processed by a Wan-initialized multi-view DiT. Geometry conditioning is stochastically regularized during training, while flow matching and PTRC supervise target fidelity and cross-view consistency, respectively.

The model receives three complementary conditions in addition to $\mathbf{X}_\sigma$. First, Wan's image-conditioning stream contains clean source latents and blank query slots, providing observed appearance without leaking query RGB. Second, dense Plücker ray maps encode the camera associated with every source and query pixel. Third, the geometry-routed features introduced in Sec. 3.2 provide scene-specific source-to-query correspondence. These signals are concatenated and mapped into the pretrained Wan token space by an expanded input projection. The pretrained image pathway is preserved, while newly introduced camera and geometry inputs are initialized to have no effect at the start of fine-tuning. Source and query tokens are then jointly processed by the Wan DiT. We keep spatial positional encoding but remove temporal ordering from the view axis, since the inputs form a set of camera observations rather than a video timeline. No scene-specific text is used: training and inference share the same fixed empty textual context. After joint denoising, query latents are decoded independently into the requested views.

3.2 Geometry-Routed Visual Conditioning

Feature association and projection. Camera rays specify where a view is sampled, but not which source observation supports each query location. We establish this correspondence using a frozen VGGT-Ω (Wang et al., 2026a), which associates each source feature $\mathbf{f}_n$ with a 3D point $\mathbf{p}_n$ and confidence c_n. A lightweight adapter maps these features into the diffusion conditioning space, while the points and confidence determine where each feature is routed and how strongly it contributes.

Source features retain their native image correspondence and are resampled directly onto the diffusion grid. Query views instead require geometric alignment. For query camera j, let $(\mathbf{R}_j, \mathbf{t}_j)$ denote its extrinsics and $(f_{x,j}, f_{y,j}, c_{x,j}, c_{y,j})$ its focal lengths and principal point. The camera-space coordinates of $\mathbf{p}_n$ and its projected query location are

$$\mathbf{x}^{\mathrm{c}}_{nj} = (x^{\mathrm{c}}_{nj}, y^{\mathrm{c}}_{nj}, z^{\mathrm{c}}_{nj})^\top = \mathbf{R}_j\mathbf{p}_n + \mathbf{t}_j, \quad \mathbf{u}_{nj} = \pi_j(\mathbf{p}_n) = \left(f_{x,j}\frac{x^{\mathrm{c}}_{nj}}{z^{\mathrm{c}}_{nj}} + c_{x,j}, f_{y,j}\frac{y^{\mathrm{c}}_{nj}}{z^{\mathrm{c}}_{nj}} + c_{y,j} \right)^\top. \tag{3}$$

Here, z^{c}_{nj} is the query-camera depth and π_j is perspective projection. We route $\mathbf{f}_n$ to the resulting query-grid location, transferring source-observed appearance evidence rather than pre-rendered RGB values or explicit geometric predictions.

Hard-anchor routing. The primary geometry condition is a depth-selected hard anchor. After projection, each source feature is assigned to its nearest query-grid cell. For every cell, a hard z-

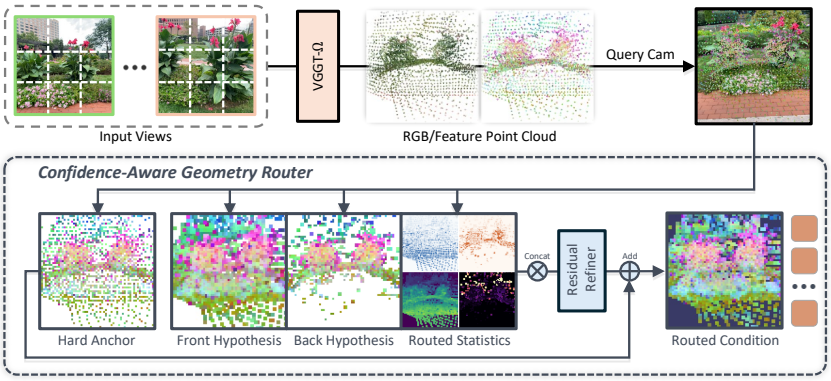

Figure 3: **Visual Geometry Conditioning.** Our conditioner constructs a query-aligned hard anchor from z-buffered near-front source features. A lightweight layered router forms confidence-weighted front and secondary hypotheses and predicts a residual correction from them and routing statistics.

buffer retains features close to the nearest valid camera-space depth and averages them according to their normalized VGGT-Ω confidence, producing $\mathbf{G}_j^{\text{hard}}$. This route is sharp, simple, and already provides a strong source-to-query correspondence.

Layered residual refinement. The hard anchor provides the primary condition, but discrete visibility may discard valid evidence near occlusion boundaries or under small geometric errors. We therefore construct confidence-weighted front and secondary hypotheses only as residual cues. For a projected feature $\mathbf{f}_n$ at continuous grid coordinate (u_n, v_n) with normalized confidence $\widetilde{c}_n$, its weight and aggregated feature at cell q, with grid coordinate $\mathbf{q} = (q_x, q_y)$, are

$$w_{nq}^{\ell} = \widetilde{c}_n \, k(u_n - q_x) \, k(v_n - q_y) \, \nu_{nq}^{\ell},$$
$$\mathbf{G}_q^{\ell} = \frac{\sum_{n \in \mathcal{A}_q^{\ell}} w_{nq}^{\ell} \mathbf{f}_n}{\sum_{n \in \mathcal{A}_q^{\ell}} w_{nq}^{\ell} + \varepsilon}, \quad k(t) = [1 - |t|]_+. \tag{4}$$

Here, k provides bilinear support and ν_{nq}^{ℓ} encodes relative-depth visibility. The front hypothesis is anchored at the nearest splatted depth z_q^{front}, with $\nu_{nq}^{\text{front}} = \exp\big(-[z_n / z_q^{\text{front}} - 1]_+ / \tau\big)$. The secondary hypothesis retains points satisfying $z_n > (1+\delta) z_q^{\text{front}}$, anchors them at their nearest retained depth, and applies the same attenuation. A zero-initialized residual refiner combines both hypotheses with statistics $\mathbf{S}_j$ describing their support, confidence, and relative depth separation:

$$\mathbf{G}_j = \mathbf{G}_j^{\text{hard}} + \mathcal{R}\big(\mathbf{G}_j^{\text{front}}, \mathbf{G}_j^{\text{back}}, \mathbf{S}_j\big). \tag{5}$$

Thus, the hard route remains the initial and primary condition, while layered evidence supplies learned corrections where useful. Unsupported locations receive no source-specific condition and are completed by the diffusion prior.

3.3 Point-Track Residual Consistency

The flow-matching objective in Eq. (2) decomposes over query views, optimizing their individual fidelity. Such per-view supervision does not enforce geometric coherence among the remaining errors: where sparse observations provide weak constraints, predictions of the same physical point can drift differently across jointly generated views even when each appears plausible and incurs a small per-view loss. A cross-view constraint is therefore needed, but directly matching RGB values, VAE latents, or velocities would be overly restrictive because viewpoint-dependent illumination, visibility, and local context legitimately alter these representations. We instead couple each prediction through its residual to its own target, enforcing consistency only in the error that should be removed.

For a sampled flow time, the predicted clean latent and its residual in query view j are

$$\widehat{\mathbf{Z}}_{0,j} = \mathbf{X}_{\sigma,j} - \sigma \mathbf{V}_{\theta,j}, \quad \mathbf{E}_j = \widehat{\mathbf{Z}}_{0,j} - \mathbf{Z}_{0,j}. \tag{6}$$

We reuse the VGGT-Ω point predictions to establish tracks across every pair of query views. A point is retained only when it projects inside both views, lies in front of both cameras, and passes a per-view z-buffer visibility test. For a valid track γ, its projections are mapped to the nearest latent-grid

locations $\mathbf{u}_{\gamma,a}$ and $\mathbf{u}_{\gamma,b}$. We define $\Delta\mathbf{E}_\gamma = \mathbf{E}_a(\mathbf{u}_{\gamma,a}) - \mathbf{E}_b(\mathbf{u}_{\gamma,b})$, and let $\bar{\rho}_\beta$ denote the channel average of the scalar Smooth-L1 penalty ρ_β. With normalized VGGT-Ω confidence w_γ, PTRC is

$$\mathcal{L}_{\mathrm{PTRC}} = \frac{\sum_{\gamma\in\mathcal{C}} w_\gamma\, \bar{\rho}_\beta(\Delta\mathbf{E}_\gamma)}{\sum_{\gamma\in\mathcal{C}} w_\gamma}, \quad \rho_\beta(x) = \begin{cases} x^2/(2\beta), & |x| < \beta, \\ |x| - \beta/2, & |x| \geq \beta. \end{cases} \tag{7}$$

The quadratic region provides precise alignment for small discrepancies, while the linear region, together with confidence weighting, limits the influence of unreliable tracks. Since PTRC aligns errors rather than predictions, corresponding views retain valid appearance changes without drifting independently from their respective targets.

3.4 Training and Inference

All camera poses are expressed in a gauge defined solely by the source cameras. Consequently, changing the number or subset of query cameras does not alter the conditioning coordinate frame. We freeze the VAE and VGGT-Ω, and fine-tune the diffusion transformer together with the new input and routing modules. The number of jointly generated query views is varied during training to support both single-view synthesis and longer view trajectories.

Training loss. The full objective combines query-wise flow matching with point-track regularization as $\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{PTRC}}\mathcal{L}_{\mathrm{PTRC}}$, where $\lambda_{\mathrm{PTRC}} = 0.1$ in all experiments. The two terms serve complementary roles: $\mathcal{L}_{\mathrm{FM}}$ learns accurate denoising for each query view, while $\mathcal{L}_{\mathrm{PTRC}}$ couples prediction errors along reliable 3D point tracks.

Geometry condition dropout. Geometry reconstructed from sparse observations is informative but inevitably imperfect. During training, we therefore randomly retain, attenuate, or remove the routed geometry condition, while leaving source appearance and camera rays unchanged. This prevents the diffusion model from treating projected geometry as an infallible scene reconstruction and improves its robustness when routed evidence is sparse, uncertain, or locally missing.

Geometry-prior CFG. Geometry-condition dropout also provides a reference for selectively strengthening the geometry prior at inference. Let $\mathbf{V}_{\mathrm{G}}$ denote the fully conditioned velocity and $\mathbf{V}_{\mathrm{ref}}$ the reference prediction obtained by removing only query-routed geometry. During early denoising, we apply $\mathbf{V}_{\mathrm{GeoCFG}} = \mathbf{V}_{\mathrm{ref}} + s_{\mathrm{g}}(\mathbf{V}_{\mathrm{G}} - \mathbf{V}_{\mathrm{ref}})$, and otherwise retain $\mathbf{V}_{\mathrm{G}}$. The two branches share the diffusion state, source appearance, source-aligned visual features, camera conditions, timestep, and empty textual context. Thus, s_{g} strengthens query geometry without removing the observations that define the scene.

4 Experiments

4.1 Experimental Setup

Data and training. We train on only the 1K-scene split of DL3DV-10K (960P) (Ling et al., 2024). Removing one unavailable scene and 19 overlapping with the 140-scene DL3DV-Benchmark leaves 980 training scenes. VGGT-*Diff* initializes from Wan2.1-I2V-14B (Wan et al., 2025). We freeze the VAE and VGGT-Ω (Wang et al., 2026a), while training the Wan DiT, expanded input projection, visual-feature adapter, and Visual Geometry Router. Training runs for 147 epochs at 192×336, then 60 at 480×832, using BF16 and a global batch size of eight. We adopt a 6-to-variable-N curriculum: each sample uses six sources and jointly denoises N targets, progressing from $N \in \{1, 2, 4\}$ to $N \in \{4, 8, 12, 16\}$. The Wan DiT uses learning rates of 10^{-5} and 5×10^{-6} across the two stages; new modules use 10^{-4} throughout. Additional half-resolution experiments on the full DL3DV-10K yield substantial gains from scaling data and optimization; see the Appendix.

Evaluation and baselines. We evaluate 6,188 DL3DV-Benchmark targets and test zero-shot transfer on Mip-NeRF 360 (Barron et al., 2022). All methods receive identical six-view sources and target cameras, with each target generated independently; VGGT-*Diff* uses 50 flow-sampling steps. Our evaluation follows FrameCrafter (Wu et al., 2026a), with native outputs resized or center-cropped to 480×480. We report PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and DreamSim (Fu et al., 2023). Baselines span regression-based and diffusion-based models in the tables. Additional diagnostics use 192 pose-stratified targets and assess eight-view geometric consistency with two frozen reconstructors, VGGT-Ω and Pi3 (Wang et al., 2026b).

Table 1: **Quantitative comparison of 6-view NVS on DL3DV and Mip-NeRF 360.** All methods follow the same protocol within each benchmark; #Scenes excludes generic foundation-model pre-training. LPIPS uses AlexNet features, and DS denotes DreamSim. Best and second-best results are highlighted by first and second.

Method	Backbone	#Scenes	DL3DV				Mip-NeRF 360			
			PSNR ↑	SSIM ↑	LPIPS ↓	DS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	DS ↓
Regression-based Models										
AnySplat (Jiang et al., 2025b)	VGGT (Wang et al., 2025)	254K	12.388	0.298	0.510	0.214	11.009	0.232	0.594	0.252
E-RayZer (Zhao et al., 2026)	–	10K	16.850	0.442	0.455	0.254	16.560	0.343	0.621	0.340
LVSM (Jin et al., 2024)	–	67.5K	17.090	0.478	0.333	0.204	15.250	0.317	0.609	0.577
DepthSplat (Xu et al., 2025)	Depth Anything V2 (Yang et al., 2024)	77.5K	17.284	0.566	0.307	0.158	15.977	0.370	0.410	0.215
Diffusion-based Models										
EscherNet (Kong et al., 2024)	SD1.5 (Rombach et al., 2022)	10K	12.070	0.251	0.484	0.227	11.140	0.126	0.540	0.315
Aether (Zhu et al., 2025)	CogVideoX-5B (Yang et al., 2025)	–	12.660	0.258	0.469	0.140	12.600	0.220	0.651	0.334
MVSplat360 (Chen et al., 2024b)	SVD (Blattmann et al., 2023)	69.5K	14.150	0.358	0.513	0.174	13.859	0.292	0.636	0.278
SEVA (Zhou et al., 2025)	SD2.1 (Rombach et al., 2022)	80K	16.150	0.470	0.253	0.088	14.590	0.294	0.372	0.137
FrameCrafter (Wu et al., 2026a)	Wan2.1-I2V-14B (Wan et al., 2025)	1K	17.180	0.445	0.223	0.066	15.640	0.279	0.365	0.111
VGGT-*Diff*	Wan2.1-I2V-14B (Wan et al., 2025)	1K	18.104	0.508	0.222	0.062	16.296	0.317	0.315	0.091

Table 2: **PSNR across target-pose difficulty on DL3DV.** All methods use independent 6-to-1 inference over 32 targets per bin.

Method	Interpolation			Extrapolation		
	Near	Mid	Far	Near	Mid	Far
Regression-based Models						
AnySplat (Jiang et al., 2025b)	14.196	12.415	10.862	13.977	12.566	11.827
E-RayZer (Zhao et al., 2026)	17.388	14.878	13.824	16.833	14.477	14.712
LVSM (Jin et al., 2024)	20.276	16.599	13.969	20.790	16.633	15.579
DepthSplat (Xu et al., 2025)	20.376	17.424	14.820	20.653	17.943	16.374
Diffusion-based Models						
Aether (Zhu et al., 2025)	14.030	12.167	11.065	15.739	12.295	11.795
GEN3C (Ren et al., 2025)	13.620	12.356	11.548	14.288	13.050	12.288
MVSplat360 (Chen et al., 2024b)	15.282	14.420	13.389	15.542	14.479	14.176
FrameCrafter (Wu et al., 2026a)	20.338	17.485	14.649	21.168	17.888	16.594
SEVA (Zhou et al., 2025)	20.940	17.861	14.703	22.311	18.050	16.575
VGGT-*Diff*	20.920	18.391	15.853	22.396	18.793	17.793

Table 3: **Ablations on DL3DV.** All variants use independent 6-to-1 inference on the 192 pose-stratified targets; LPIPS uses VGG features.

Configuration	PSNR ↑	SSIM ↑	LPIPS ↓
Model design			
w/o VGGT-Ω prior	17.294	0.501	0.357
w/ point-rendered RGB	18.553	0.536	0.316
w/o PTRC	18.586	0.530	0.312
w/o layered residual refinement	18.895	0.548	0.303
Geometry-conditioning strategy			
w/o condition regularization	18.722	0.538	0.309
w/o CFG (full method)	19.024	0.557	0.301
w/ geometry-prior CFG	19.139	0.563	0.296

4.2 Comparison with Baselines

(1) Overall performance. Table 1 compares all methods under the common protocol. On 6,188 DL3DV targets, VGGT-*Diff* ranks first in PSNR, LPIPS, and DreamSim and second in SSIM. Notably, these results are achieved using only ∼1k training scenes, substantially fewer than many competing methods, demonstrating strong data efficiency. Against FrameCrafter with the same Wan2.1-I2V-14B prior and a comparable training scale, VGGT-*Diff* gains 0.924 dB PSNR and 0.063 SSIM while retaining comparable perceptual quality, demonstrating the benefit of geometry-routed adaptation. On zero-shot Mip-NeRF 360, VGGT-*Diff* ranks first in LPIPS and DreamSim and second in PSNR. Although E-RayZer and DepthSplat lead PSNR and SSIM, respectively, their substantially higher perceptual errors indicate that these gains come at the cost of perceptual quality.

(2) Qualitative comparison. Fig. 4 and Fig. 5 localize these gains. On DL3DV, VGGT-*Diff* preserves chair legs, desk boundaries, facade geometry, and legible signage while recovering target layout and occlusion order; competing methods blur thin structures, distort geometry, or drift from the requested camera. On zero-shot Mip-NeRF 360, VGGT-*Diff* likewise better preserves object boundaries and source-observed appearance, indicating scene-grounded rather than unsupported synthesis.

(3) Performance across pose difficulties. To expose failures hidden by aggregate metrics, we stratify 192 DL3DV targets by interpolation/extrapolation and near/mid/far difficulty (32 per bin). Table 2 shows that VGGT-*Diff* leads five bins and trails SEVA by only 0.020 dB in interpolation-near. Its margin over FrameCrafter ranges from 0.582 to 1.228 dB and exceeds 0.9 dB in five bins, confirming broad effectiveness across pose difficulties.

(4) Cross-view consistency. Single-target metrics cannot determine whether jointly generated views support a coherent scene. We therefore reconstruct the same eight views with two frozen geometry models: VGGT-Ω as the primary probe and Pi3, unused in training or conditioning, as an independent check for shared-backbone evaluator bias. ***If generated views are mutually consistent and match GT content and detail, each model should recover a coherent point cloud close to its GT-view reconstruction.*** Across 52 DL3DV scenes, all methods use the same six sources and eight

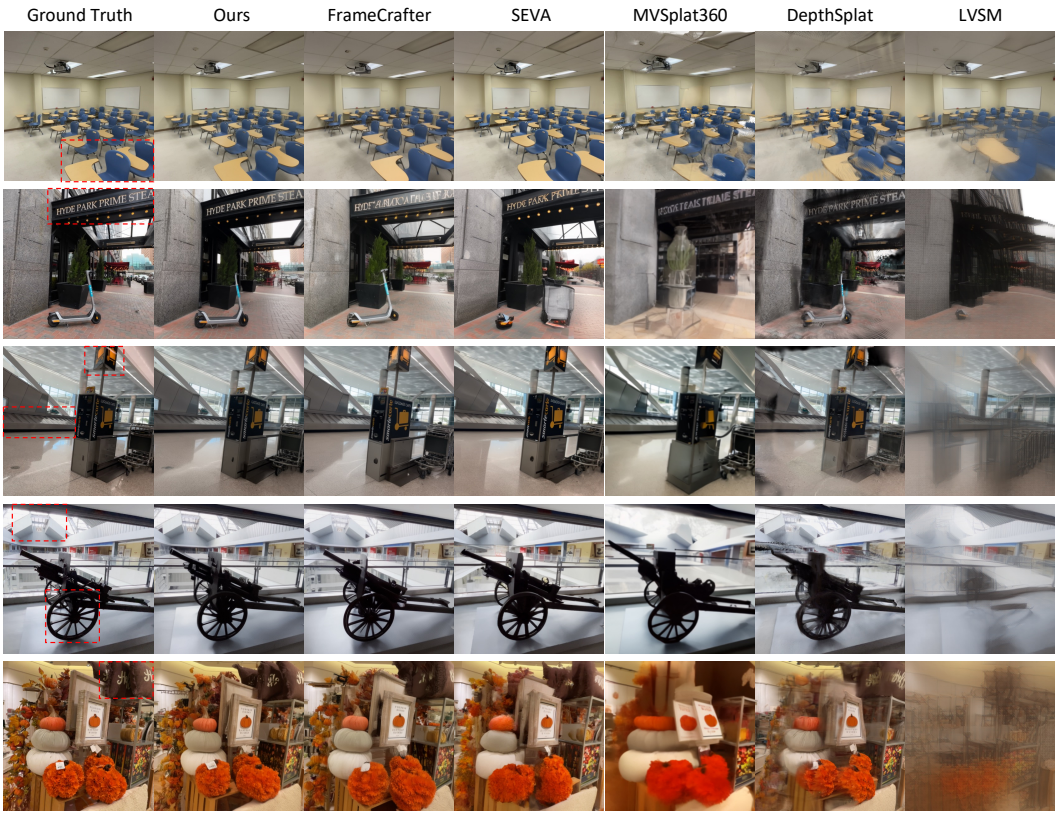

Figure 4: **Qualitative comparison on diverse DL3DV scenes.** VGGT-*Diff* better matches the requested viewpoints while preserving fine geometry, legible text, and object identity; previous methods exhibit blur, structural distortion, or content drift.

Table 4: **Cross-reconstructor evaluation on DL3DV.** Eight jointly generated views are reconstructed with frozen VGGT-Ω and Pi3 and compared with GT-view references produced by the same reconstructor. Pi3 is not used for training or conditioning VGGT-*Diff*. Metrics are trajectory-normalized and should be compared within each reconstructor.

Method	VGGT-Ω (Wang et al., 2026a)			Pi3 (Wang et al., 2026b)		
	Chamfer ↓	F@1% ↑	F@2% ↑	Chamfer ↓	F@1% ↑	F@2% ↑
LVSM (Jin et al., 2024)	0.2387	0.1116	0.2218	0.3610	0.0673	0.1525
SEVA (Zhou et al., 2025)	0.1944	0.2050	0.3244	0.2514	0.1640	0.2751
FrameCrafter (Wu et al., 2026a)	0.1000	0.3086	0.4793	0.1404	0.2140	0.3880
VGGT-*Diff*	0.0733	0.4231	0.6041	0.1174	0.3089	0.5021

target cameras, with generated and GT views processed identically by each reconstructor. Table 4 shows that VGGT-*Diff* ranks first under both models. Relative to FrameCrafter, it reduces Chamfer-L1 by 26.7% with VGGT-Ω and 16.4% with Pi3, while improving F-score@1% and F-score@2% by 11.45/12.48 and 9.49/11.41 percentage points, respectively. This agreement indicates that the gains do not arise solely from reusing VGGT-Ω as the evaluator. Appendix Fig. 10 provides corresponding visualizations. Since the references are reconstructed from target RGB rather than physical scans, these metrics measure relative consistency and reconstructability, not absolute 3D accuracy.

4.3 Ablation Studies

(1) Protocol. All variants reported in Table 3 follow the same full-resolution controlled protocol and use independent 6-to-1 inference on 192 pose-stratified targets. Training-time variants share the same Wan initialization, change only the listed component, and are evaluated without Geometry-Prior CFG. The final row applies CFG to the full checkpoint and changes only inference. Unless stated otherwise, differences are measured against the full checkpoint without CFG.

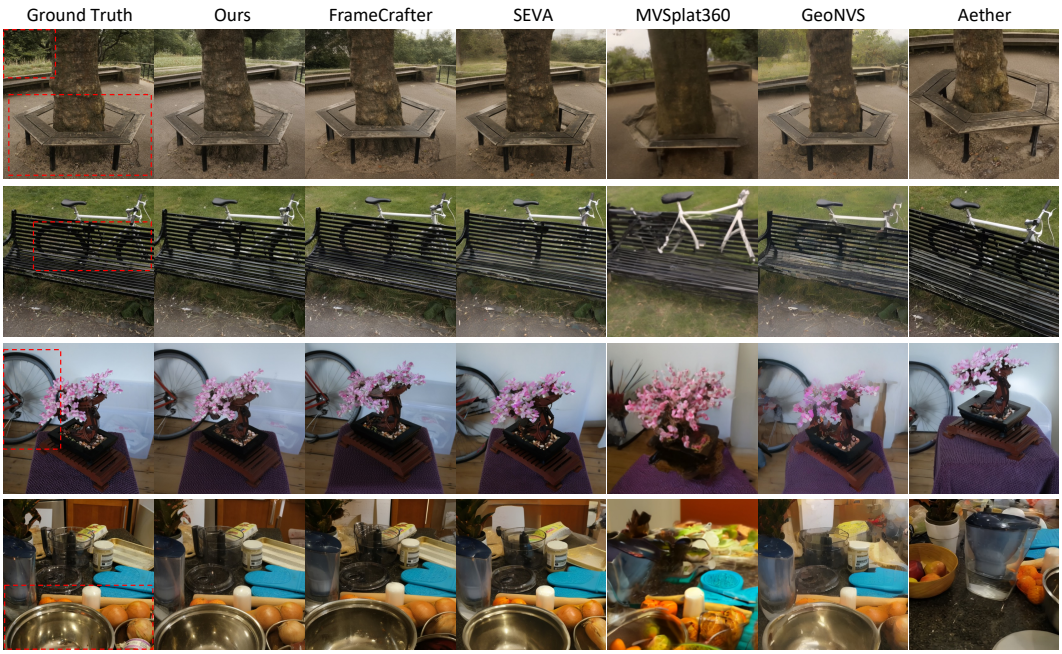

Figure 5: **Qualitative comparison on Mip-NeRF 360.** Without dataset-specific fine-tuning, VGGT-*Diff* better preserves fine structures, occlusion boundaries, and source-observed appearance while more accurately matching the target view.

(2) Model design. Table 3 shows that removing all VGGT-Ω-dependent components causes the largest degradation: PSNR falls by 1.730 dB and SSIM by 0.056, while LPIPS rises by 0.056. Camera rays and the Wan prior alone therefore cannot replace visual geometry. Replacing routed features with point-rendered RGB, while retaining the geometry, router, and PTRC, costs 0.471 dB PSNR and raises LPIPS by 0.015, confirming that appearance-bearing features provide richer evidence than incomplete RGB renderings. Removing PTRC costs 0.438 dB PSNR and 0.027 SSIM, the largest SSIM loss among single-component ablations, supporting residual alignment over shared 3D tracks. Removing layered refinement while retaining the hard anchor causes smaller drops of 0.129 dB PSNR and 0.009 SSIM, indicating that hard routing captures most of the benefit while layered refinement corrects ambiguous locations near occlusions and imperfect projections.

(3) Geometry-conditioning strategy. Removing condition regularization costs 0.302 dB PSNR and 0.019 SSIM while increasing LPIPS by 0.008, confirming that stochastic attenuation and dropout reduce over-reliance on imperfect geometry. The no-CFG row evaluates the full checkpoint without guidance; applying Geometry-Prior CFG to the same checkpoint improves PSNR by 0.115 dB and SSIM by 0.006 while reducing LPIPS by 0.005. The larger gain from regularization identifies it as the primary mechanism, while matched CFG provides a modest complementary improvement at inference without retraining or changing the learned model parameters or training objective.

Additional experiments, protocol details, results and analyses are provided in the Appendix.

5 Conclusion

We presented VGGT-*Diff*, a geometry-routed multi-view framework combining visual geometry with pretrained video diffusion for sparse-view NVS. Its confidence-aware VGR maps source features into query-aligned conditions using 3D points and confidence, while PTRC aligns predicted-clean residuals along reliable correspondences. Stochastic training-time weakening and selective inference-time strengthening improve robustness to imperfect geometry. Across diverse pose difficulties, VGGT-*Diff* achieves strong target-view fidelity and cross-view coherence, combining geometry-grounded scene structure with the generative prior's ability to complete unseen content.

AI USE STATEMENT

We used generative AI tools to assist with language polishing, improving clarity and concision, and refining overleaf formatting and presentation. We also used these tools to support literature discovery and track recent developments relevant to our work. All suggested references were manually verified against their original sources, and all AI-assisted text and formatting changes were reviewed and revised by us. Generative AI was not used to conduct experiments, or determine the scientific claims and conclusions of this work. We take full responsibility for the final content, including all text, citations, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed data construction, view sampling, training schedules, conditioning hyperparameters, evaluation protocols, and additional controlled experiments in the Appendix. All reported comparisons use fixed source views, target cameras, evaluation cases, and metric backbones as specified in the paper. We will release the training and evaluation code, model checkpoints, and configuration files to support reproduction of our results.

REFERENCES

Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5835–5844. IEEE, 2021. 2

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5460–5469. IEEE, 2022. 6

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. 7

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. 1, 3

Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 14104–14113. IEEE, 2021. 2

Kangjie Chen, BingQuan Dai, Minghan Qin, Dongbin Zhang, Peihao Li, Yingshuang Zou, and Haoqian Wang. Slgaussian: Fast language gaussian splatting in sparse views. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3047–3056, 2025. 2

Kangjie Chen, Yingji Zhong, Zhihao Li, Jiaqi Lin, Youyu Chen, Minghan Qin, and Haoqian Wang. Quantifying and alleviating co-adaptation in sparse-view 3d gaussian splatting. *Advances in Neural Information Processing Systems*, 38:115939–115968, 2026. 2

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, 2024a. 1, 3

Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. *Advances in Neural Information Processing Systems*, 37:107064–107086, 2024b. 3, 7, 20, 21

Noam Elata, Bahjat Kawar, Yaron Ostrovsky-Berman, Miriam Farber, and Ron Sokolovsky. Novel view synthesis with pixel-space diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26756–26766. IEEE, 2025. 3

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. 6

Or Hirschorn, Omer Sela, Inbar Huberman-Spiegelglas, Netalee Efrat, Eli Alshan, Ianir Ideses, Frederic Devernay, Yochai Zvik, and Lior Fritz. Splatent: Splatting diffusion latents for novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8319–8330, 2026. 3

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *International Conference on Learning Representations*, volume 2024, pp. 50678–50702, 2024. 3

Xiaosong Jia, Yihang Sun, Junqi You, Zichen Zou, Junchi Yan, Zuxuan Wu, Yu-Gang Jiang, et al. Efficient-lvsm: Faster, cheaper, and better large view synthesis model via decoupled co-refinement attention. In *International Conference on Learning Representations*, volume 2026, pp. 40029–40046, 2026. 3

Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4918–4929. IEEE, 2025a. 1, 3

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, 2025b. 7, 20, 21

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. *arXiv preprint arXiv:2410.17242*, 2024. 1, 3, 7, 8, 20, 21

Minjun Kang, Inkyu Shin, Taeyeop Lee, Myungchul Kim, In So Kweon, and Kuk-Jin Yoon. Geonvs: Geometry grounded video diffusion for novel view synthesis. *arXiv preprint arXiv:2603.14965*, 2026. 3

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. 1, 2

Evan Kim, Hyunwoo Ryu, Thomas W Mitchel, and Vincent Sitzmann. Scaling view synthesis transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28893–28902, 2026. 1, 3

Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J Davison. Eschernet: A generative model for scalable view synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9503–9513. IEEE, 2024. 2, 3, 7

Jonáš Kulhánek, Erik Derner, Torsten Sattler, and Robert Babuška. Viewformer: Nerf-free neural rendering from few images using transformers. In *European Conference on Computer Vision*, pp. 198–216. Springer, 2022. 3

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

Min-Seop Kwak, Junho Kim, Sangdoo Yun, Dongyoon Han, Taekyung Kim, Seungryong Kim, and Jin-Hwa Kim. Aligned novel view image and geometry synthesis via cross-modal attention instillation. In *International Conference on Learning Representations*, volume 2026, pp. 16029–16050, 2026. 3

Minkyung Kwon, Jinhyeok Choi, Jiho Park, Seonghu Jeon, Jinhyuk Jang, Junyoung Seo, Minseop Kwak, Jin-Hwa Kim, and Seungryong Kim. Correspondence-attention alignment for multi-view diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2316–2326, 2026. 3

Thanh-Tung Le, Tuan Pham, Tung Nguyen, Deying Kong, Xiaohui Xie, and Stephan Mandt. Umami: Unifying masked autoregressive models and deterministic rendering for view synthesis. *Advances in Neural Information Processing Systems*, 38:175038–175065, 2026. 3

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. 3

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. 6, 15

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9264–9275. IEEE, 2023. 3

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *International conference on learning representations*, volume 2024, pp. 27676–27697, 2024. 3

Lars Mescheder, Wei Dong, Shiwei Li, Xuyang Bai, Marcel Santos, Peiyun Hu, Bruno Lecouat, Mingmin Zhen, Amaël Delaunoy, Tian Fang, et al. Sharp monocular view synthesis in less than a second. In *International Conference on Learning Representations*, volume 2026, pp. 24192–24230, 2026. 3

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. 1, 2

Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10258–10268. IEEE, 2024. 2, 3

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. 7, 20, 21

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022. 7

Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6219–6228. IEEE, 2022. 2

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *International conference on learning representations*, volume 2024, pp. 39838–39859, 2024. 3

Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10208–10217, 2024. 3

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. 1, 3

Basile Van Hoorick, Dian Chen, Shun Iwase, Pavel Tokmakov, Muhammad Zubair Irshad, Igor Vasiljevic, Swati Gupta, Fangzhou Cheng, Sergey Zakharov, and Vitor Campagnolo Guizilini. Anyview: Synthesizing any novel view in dynamic scenes. *arXiv preprint arXiv:2601.16982*, 2026. 3

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pp. 439–457. Springer, 2024. 3

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. 3, 6, 7, 16

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, 2025. 2, 3, 7

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. 2, 3, 4, 6, 8, 15

Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, Zhangyang Wang, et al. Is attention all that nerf needs? *arXiv preprint arXiv:2207.13298*, 2022. 2

Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2021. 2

Qisen Wang, Yifan Zhao, Jiawei Ma, and Jia Li. How to use diffusion priors under sparse views? *Advances in Neural Information Processing Systems*, 37:30394–30424, 2024a. 3

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, 2024b. 3

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, volume 2026, pp. 10481–10497, 2026b. 6, 8, 21

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 2, 3

Qi Wu, Khiem Vuong, Minsik Jeon, Srinivasa Narasimhan, and Deva Ramanan. Novel view synthesis as video completion. *arXiv preprint arXiv:2604.08500*, 2026a. 2, 3, 6, 7, 8, 16, 20, 21, 23

Zirui Wu, Zeren Jiang, Martin R Oswald, and Jie Song. From rays to projections: Better inputs for feed-forward view synthesis. *arXiv preprint arXiv:2601.05116*, 2026b. 3

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. 7, 20, 21

Yan Xu, Yixing Wang, and Stella X Yu. Novel view synthesis from a few glimpses via test-time natural video completion. *Advances in Neural Information Processing Systems*, 38:46541–46561, 2026. 3

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024. 7

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. 7

Jianglong Ye, Peng Wang, Kejie Li, Yichun Shi, and Heng Wang. Consistent-1-to-3: Consistent image to 3d view synthesis via geometry-aware diffusion models. In *2024 International Conference on 3D Vision (3DV)*, pp. 664–674. IEEE, 2024. 3

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4576–4585. IEEE, 2021. 2

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. 2, 3

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2024. 3

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018. 6

Qitao Zhao, Hao Tan, Qianqian Wang, Sai Bi, Kai Zhang, Kalyan Sunkavalli, Shubham Tulsiani, and Hanwen Jiang. E-rayzer: Self-supervised 3d reconstruction as spatial visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7525–7535, 2026. 7, 20, 21

Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9720–9731. IEEE, 2024. 2, 3

Haiyang Zhou, Wangbo Yu, Chaoran Feng, Xunyu Zhou, Yonghong Tian, and Li Yuan. Uniworld-view: Large-baseline view synthesis via video diffusion models. *arXiv preprint arXiv:2608.04701*, 2026. 3

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. 2, 3, 7, 8, 20, 21

Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8535–8546. IEEE, 2025. 7, 20, 21

A APPENDIX

Metric convention. The cross-method comparison in Table 1 uses LPIPS-AlexNet to follow the evaluation protocol of prior work. Controlled ablations use LPIPS-VGG consistently within the fixed 192-case diagnostic protocol. We label the backbone explicitly and do not compare absolute LPIPS values across these protocols.

Appendix organization. The supplementary material is organized as follows.
Secs. A.1 and A.2. We specify the view-sampling protocol, camera gauge, pose-difficulty definition, and fixed training and conditioning hyperparameters.
Secs. A.3, A.4, and A.5. We analyze data and optimization scaling, joint multi-target generation, and the feature depth and optimization scope of VGGT-Ω (Wang et al., 2026a).
Secs. A.6 and A.7. We report complete pose-difficulty results and evaluate cross-view consistency using two independent point-cloud reconstructors.
Secs. A.8 and A.9. We isolate the residual formulation of PTRC and provide additional optimization and scene-level statistical diagnostics.
Secs. A.10 and A.11. We examine robustness to fewer source views and clarify the VAE protocols used for quantitative evaluation and continuous-trajectory generation.
Sec. A.12. We conclude with promising directions for extending geometry-routed multi-view generation.

A.1 VIEW SAMPLING AND POSE-DIFFICULTY PROTOCOL

Training view sampling. For each training scene, we precompute 16 reproducible six-view source bundles that rotate deterministically across epochs. Thirteen bundles use global sampling: the ordered camera trajectory is divided into six contiguous segments, and one frame is sampled from each segment. The remaining three use the same stratification within a randomly selected local window. The six selected views are shuffled before entering the model, exposing it to both broad trajectory coverage and locally concentrated observations.

Targets are sampled online after excluding the active source frames. Local bundles preferentially draw targets from their window and fall back to the full scene when necessary. The target-count curriculum progresses through $\{1,2,4\}\to\{2,4,8\}\to\{4,8,12\}\to\{4,8,12,16\}$ in the first training stage and $\{4,8\}\to\{4,8,12\}\to\{4,8,12,16\}$ in the second. For eight or more targets, we always sample an ordered trajectory by selecting evenly spaced frames within a random temporal span. Smaller target sets use this mode with probability 0.5 and otherwise sample views uniformly without replacement. Training does not explicitly balance the pose-difficulty bins used for evaluation.

Source-defined camera gauge. Camera normalization depends only on the six source views. We choose the source camera nearest their centroid as the anchor, align all cameras to its rotation and origin, and normalize translation by the mean source-to-anchor distance. Query cameras do not affect this transformation, so single-target and joint inference share the same coordinate frame for a fixed source set.

Pose-difficulty definition. We construct the diagnostic set from the official six-view DL3DV-Benchmark (Ling et al., 2024) split, using its fixed source views and candidate test frames. Let $i_{\min}$ and $i_{\max}$ be the minimum and maximum source frame indices. A target is trajectory interpolation when $i_{\min} < i_{\mathrm{t}} < i_{\max}$ and trajectory extrapolation when it lies outside this interval. This definition follows the ordered capture trajectory and does not imply containment within the 3D convex hull of the source cameras.

Difficulty is measured by the target's distance from its nearest source pose. Let $\mathbf{c}_{\mathrm{t}}, \mathbf{f}_{\mathrm{t}}$ denote the target camera center and unit forward direction, and $\mathbf{c}_j, \mathbf{f}_j$ those of source j. We define the source-trajectory diameter as $D_{\mathcal{S}} = \max_{j,k\in\mathcal{S}} \|\mathbf{c}_j - \mathbf{c}_k\|_2$ and its stabilized value as $\bar{D}_{\mathcal{S}} = \max(D_{\mathcal{S}}, 10^{-6})$. The angular difference to source j is $\theta_j = \arccos\left[\mathrm{clip}(\mathbf{f}_{\mathrm{t}}^{\top}\mathbf{f}_j, -1, 1)\right]$. Pose novelty is then

$$\nu_{\mathrm{t}} = \min_{j\in\mathcal{S}} \sqrt{\frac{\|\mathbf{c}_{\mathrm{t}} - \mathbf{c}_j\|_2^2}{\bar{D}_{\mathcal{S}}^2} + \frac{\theta_j^2}{\pi^2}}. \tag{8}$$

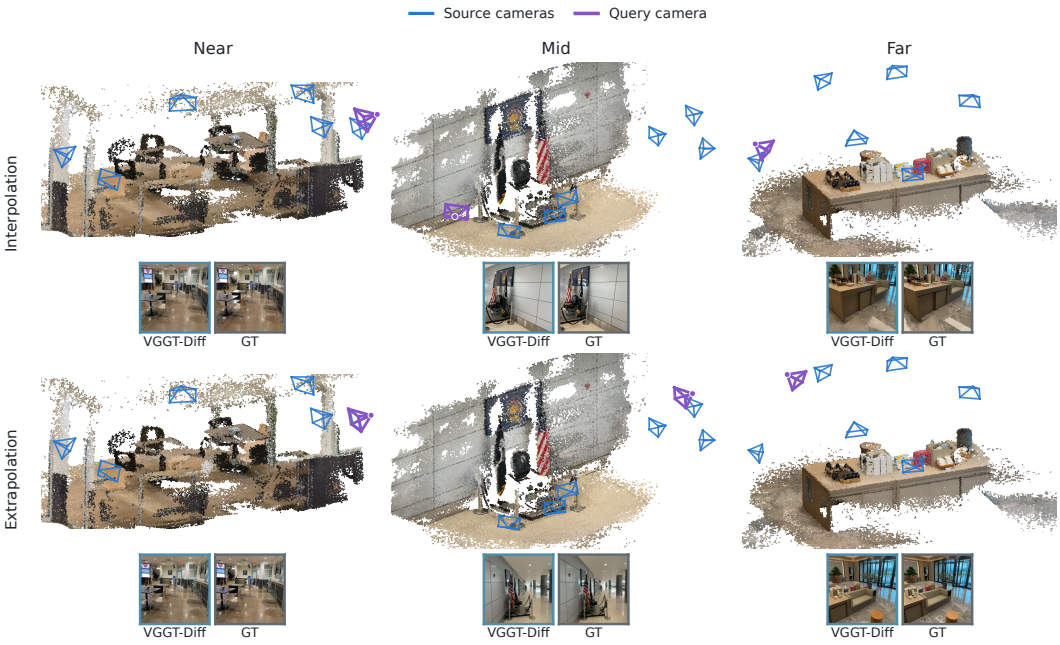

Figure 6: **Pose-difficulty protocol and representative predictions.** Blue frustums denote the six source cameras and purple denotes the query. Rows distinguish trajectory interpolation from extrapolation; columns increase pose novelty from near to far. Frozen VGGT-Ω point clouds, fused from multiple continuous GT-view windows, provide scene context only. Insets compare VGGT-*Diff* with GT and do not determine the bins.

This combines scene-normalized translation and viewing-direction change. Near, mid, and far bins are defined by global pose-novelty tertiles computed separately for interpolation and extrapolation. Their two thresholds are $(0.1165, 0.2147)$ for interpolation and $(0.0560, 0.1272)$ for extrapolation.

Figure 6 visualizes the two complementary criteria. The row depends on the target's position along the ordered capture trajectory, whereas the column reflects its translation-and-rotation novelty ν_{t}. A target can therefore be extrapolative but pose-near, or interpolative yet pose-far.

Frozen evaluation cases. The candidate pool contains 5,549 interpolation and 638 extrapolation targets. From each of the six bins, we deterministically select 32 targets from 32 different scenes. The resulting protocol contains 192 unique targets from 109 scenes; scenes may recur across bins, but no scene-target pair is duplicated. Every method receives the same six source views and predicts each target independently using fixed per-case noise, without additional target slots, rollout, or prediction feedback. This pose-stratified diagnostic is separate from the aggregate FrameCrafter-aligned (Wu et al., 2026a) comparison in Table 1, and its cases remain fixed across model resolutions.

A.2 Fixed Hyperparameters and Conditioning Regularization

Table 5 collects the constants shared by the main experiments. The router thresholds operate on relative depth and are therefore dimensionless; PTRC is evaluated directly in the Wan (Wan et al., 2025) latent space.

Visual Geometry Router. We retain the notation of Eq. 4. The raw confidence c_n is normalized per scene as $\widetilde{c}_n = \mathrm{clip}(c_n / \max_m c_m, 0, 1)$. For query-grid cell q, the front anchor is the nearest splatted depth z_q^{front}, and $\nu_{nq}^{\mathrm{front}} = \exp[-(z_n/z_q^{\mathrm{front}} - 1)_+/\tau]$, with $\tau = 0.02$. The back (secondary) hypothesis retains points satisfying $z_n > (1+\delta) z_q^{\mathrm{front}}$, where $\delta = 0.04$; their nearest depth defines z_q^{back}, around which the same attenuation is applied. Besides $\mathbf{G}_q^{\mathrm{front}}$ and $\mathbf{G}_q^{\mathrm{back}}$, the router outputs four per-cell statistics: clipped front and back support, bilinear-weighted mean confidence, and the clipped relative gap $(z_q^{\mathrm{back}}/z_q^{\mathrm{front}} - 1)$.

PTRC optimization. Following Eq. 6, query view j uses $\widehat{\mathbf{Z}}_{0,j} = \mathbf{X}_{\sigma,j} - \sigma \mathbf{V}_{\theta,j}$ and residual $\mathbf{E}_j = \widehat{\mathbf{Z}}_{0,j} - \mathbf{Z}_{0,j}$. We compute Eq. 7 in the Wan latent space, using the confidence-weighted

Smooth-L1 penalty with $\beta = 0.1$ and no VAE decoding. Within each training stage, $u_{\mathrm{tr}} = e/(E-1)$ denotes normalized epoch progress and

$$\lambda_{\mathrm{eff}}(u_{\mathrm{tr}}) = \lambda_{\mathrm{PTRC}} \begin{cases} 0, & u_{\mathrm{tr}} < 0.1, \\ (u_{\mathrm{tr}} - 0.1)/0.2, & 0.1 \le u_{\mathrm{tr}} < 0.3, \\ 1, & u_{\mathrm{tr}} \ge 0.3, \end{cases} \qquad \lambda_{\mathrm{PTRC}} = 0.1. \tag{9}$$

Thus, PTRC is disabled for the first 10% of each stage, linearly activated during the next 20%, and fully weighted thereafter. The latent-space implementation uses no additional timestep gate.

Geometry-condition regularization. Each training sample draws one shared geometry-condition scale

$$a \sim \begin{cases} 0, & \Pr = 0.10, \\ \mathcal{U}(0.2, 0.8), & \Pr = 0.20, \\ 1, & \Pr = 0.70. \end{cases} \tag{10}$$

The same a scales the global Ω context and dense routed geometry over both source and query slots. Source-image latents, Plücker rays, diffusion state, and timestep remain unchanged; this operation is not text dropout.

Geometry-Prior CFG. The formal guided result uses a target-only reference branch. The conditional prediction $\mathbf{V}_{\mathrm{G}}$ retains all routed geometry, whereas $\mathbf{V}_{\mathrm{ref}}$ zeros only query-slot dense geometry while preserving source-aligned geometry, global Ω context, image latents, Plücker rays, diffusion state, timestep, and empty-text context. For denoising step k, with $u_{\mathrm{den}} = k/(K-1)$, we use

$$\mathbf{V}_{\mathrm{GeoCFG}} = \mathbf{V}_{\mathrm{ref}} + s_{\mathrm{g}}^{\mathrm{eff}}(u_{\mathrm{den}})(\mathbf{V}_{\mathrm{G}} - \mathbf{V}_{\mathrm{ref}}), \qquad s_{\mathrm{g}}^{\mathrm{eff}}(u_{\mathrm{den}}) = 1 + (s_{\mathrm{g}} - 1)\cos^2\left[\frac{\pi}{2}\min\left(\frac{u_{\mathrm{den}}}{0.6}, 1\right)\right]. \tag{11}$$

Hence guidance decays smoothly from 2 to 1 over the first 60% of the 50-step trajectory and is inactive thereafter; it is neither a hard switch nor constant-scale guidance. The full-Ω CFG ablation instead removes the global context and all source/query routed geometry from its reference branch.

Table 5: **Fixed geometry-conditioning hyperparameters.** The relative-depth router constants are shared across scenes and resolutions.

Component	Setting	Value
Visual Geometry Router	Secondary-layer relative margin	$\delta = 0.04$
Visual Geometry Router	Relative-depth visibility temperature	$\tau = 0.02$
PTRC	Full loss weight; Smooth-L1 transition	0.1; 0.1
PTRC schedule	Stage-wise off / linear ramp / full-weight fractions	10% / 20% / 70%
Condition regularization	Dropped / weakened / full-condition probabilities	0.10 / 0.20 / 0.70
Geometry-Prior CFG	Guidance scale; denoising steps; guided interval	2.0; 50; early 60%
Text conditioning	Prompt; text-CFG scale	empty; 1

A.3 Scaling with Data and Optimization Budget

Protocol. We extend the full VGGT-Diff model trained on DL3DV clean-10K, containing 9,510 available scenes, from the original 18K fixed-compute budget to 100K effective steps. Every checkpoint is evaluated on the same 192 targets, with 32 cases in each interpolation/extrapolation near, mid, and far bin. All evaluations use six source views, 192×336 inference, a 192×192 center crop for metric computation, seed 20260823, and 50 sampling steps. We report LPIPS-VGG throughout this diagnostic; these values should not be compared directly with the LPIPS-AlexNet results in Table 1.

The losses in Table 6 average the ten log entries from the 1K effective steps preceding each checkpoint. Flow loss is the original flow-matching objective, while PTRC denotes the raw *unweighted* point-track loss before applying its training weight or activation scale.

Evaluation scaling. Table 6 and Fig. 7 show that the clean-10K model remains substantially under-optimized at 18K steps. Extending training to 100K raises PSNR from 18.7861 to 20.2780 dB and

Table 6: **Checkpoint scaling on DL3DV clean-10K.** Image metrics use the fixed 192-case half-resolution protocol. Losses average the preceding 1K effective steps; PTRC is reported before weighting.

Effective steps	PSNR ↑	SSIM ↑	LPIPS ↓	Flow loss ↓	PTRC loss ↓
18K	18.7861	0.5272	0.2793	0.074406	0.087436
24K	19.1529	0.5509	0.2621	0.057234	0.086353
30K	19.1980	0.5579	0.2593	0.056245	0.085368
36K	19.6120	0.5770	0.2507	0.055986	0.084214
42K	19.5829	0.5762	0.2541	0.053832	0.080950
48K	19.6875	0.5830	0.2490	0.053431	0.080535
54K	19.8138	0.5889	0.2467	0.053542	0.079932
60K	19.4351	0.5725	0.2534	0.052149	0.077931
72K	19.9368	0.5973	0.2411	0.052154	0.077268
84K	20.1706	0.6088	0.2373	0.052251	0.076402
88K	20.1350	0.6059	0.2388	0.052218	0.076123
96K	20.2644	0.6146	0.2342	0.050399	0.074175
100K	20.2780	0.6166	0.2335	0.050209	0.073912

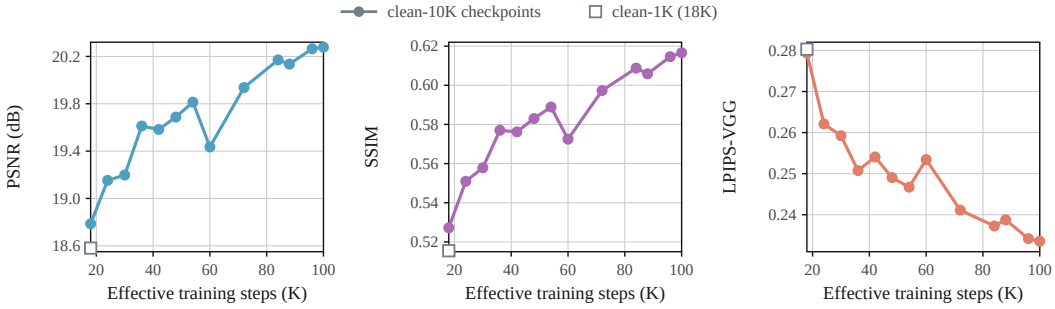

Figure 7: **Evaluation metrics across the clean-10K optimization budget.** Under the 192-case half-resolution protocol, PSNR, SSIM, and LPIPS-VGG are measured as training extends from 18K to 100K effective steps. Despite local fluctuations, all metrics improve substantially, with the best results at 100K. The open square denotes the clean-1K model evaluated at the matched 18K budget.

SSIM from 0.5272 to 0.6166, while reducing LPIPS-VGG from 0.2793 to 0.2335. These correspond to gains of 1.4918 dB PSNR and 0.0894 SSIM, together with a 0.0458 reduction in LPIPS-VGG.

Checkpoint quality is not strictly monotonic: 42K, 60K, and 88K exhibit local regressions, but each is followed by recovery and the long-term trend remains consistently favorable. The 100K checkpoint is best on all three image metrics. Its PSNR gain over 96K narrows to 0.0135 dB, suggesting that optimization is approaching a plateau; continued SSIM and LPIPS-VGG improvements prevent concluding that performance has fully saturated.

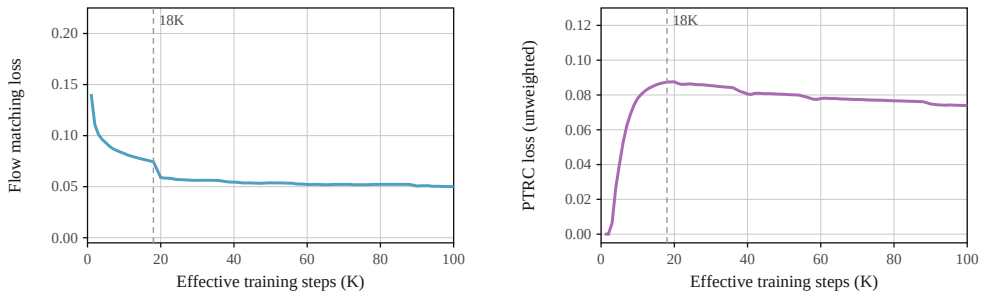

Figure 8: **Training dynamics on DL3DV clean-10K.** Curves connect non-overlapping 1K-step averages across the full training trajectory. The dashed line marks the original 18K fixed-compute budget. PTRC is reported as the raw, unweighted consistency loss. Its initial rise reflects scheduled activation and the increasingly difficult target curriculum rather than optimization divergence.

Training dynamics. Figure 8 shows that optimization continues well beyond the original fixed-compute boundary. Between the 18K and 100K checkpoints, the preceding-1K Flow loss decreases from 0.074406 to 0.050209, a 32.5% reduction, while the unweighted PTRC loss decreases from

0.087436 to 0.073912, a 15.5% reduction. PTRC rises early because the objective is progressively activated as the target curriculum introduces more difficult multi-view correspondences. Once fully active, its sustained decline accompanies the continued improvement in image quality. Together, the evaluation and loss curves show that scaling the dataset requires a commensurate optimization budget to realize its benefit.

A.4 Benefits of Joint Multi-Target Generation

We compare generating eight targets in one diffusion sequence (Joint-8) with eight strictly independent 6-to-1 runs. Across 52 DL3DV scenes and 416 targets, both modes use the same checkpoint, six sources, eight target cameras, seed, per-view noise, and 50-step sampler. Independent runs cannot access other target cameras, latents, or predictions.

Table 7: **Joint versus independent eight-view generation.** Image metrics use 416 targets; geometry proxies are computed from the same outputs with frozen VGGT-Ω. LPIPS uses VGG features, and Chamfer-L1 is trajectory-normalized.

Mode	PSNR ↑	SSIM ↑	LPIPS ↓	Chamfer ↓	F@1% ↑	F@2% ↑
Joint-8	17.6401	0.4787	0.3568	0.0483	0.3438	0.5341
Independent-8	17.1784	0.4453	0.3744	0.0741	0.2432	0.4047

Joint generation improves PSNR by 0.462 dB and SSIM by 0.0334, reduces LPIPS-VGG by 0.0176, lowers normalized Chamfer-L1 by 0.0259, and raises F-score@1% and F-score@2% by 0.1006 and 0.1294. Scene-bootstrap 95% confidence intervals exclude zero for all three image metrics and all three geometry metrics. Joint target slots therefore provide useful mutual context rather than merely batching otherwise independent predictions.

A.5 VGGT-Ω Design Choices for Geometry Conditioning

Setup. We examine feature depth and optimization scope under the same half-resolution 192-case protocol. First, we vary the extraction layer while keeping VGGT-Ω frozen. Second, using Layer 23, we jointly optimize its visual aggregator and encoder at a learning rate of 10^{-6} while retaining frozen camera and depth heads. The Wan DiT, adapter, and Visual Geometry Router are trained normally in all variants; all remaining inputs and objectives are fixed.

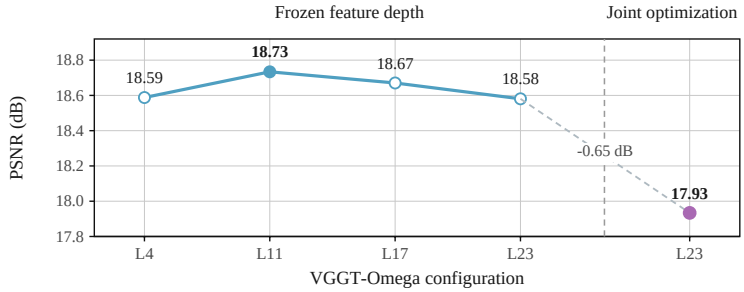

Figure 9: **VGGT-Ω feature depth and optimization.** Intermediate frozen features perform best, while jointly optimizing the Layer 23 visual pathway degrades PSNR.

Table 8: **VGGT-Ω feature and optimization diagnostics.** All variants use the same half-resolution protocol; LPIPS uses VGG features.

Configuration	PSNR ↑	SSIM ↑	LPIPS ↓
Frozen VGGT-Ω: feature depth			
Layer 4	18.588	0.5160	0.2828
Layer 11	18.734	0.5253	0.2763
Layer 17	18.670	0.5197	0.2788
Layer 23	18.581	0.5155	0.2803
Additional optimization diagnostic			
Layer 23, jointly optimized	17.934	0.4803	0.3021

Feature depth. Table 8 shows a modest preference for intermediate features. Layer 11 performs best across all metrics, improving over Layer 23 by 0.153 dB PSNR and 0.0098 SSIM while reducing LPIPS-VGG by 0.0040. Intermediate representations may better balance cross-view correspondence with local appearance: shallower features contain less aggregated multi-view context, whereas deeper features may trade fine detail for more abstract scene information. The limited variation among frozen layers also indicates robustness to feature depth.

Frozen versus jointly optimized features. Joint optimization reduces PSNR by 0.647 dB and SSIM by 0.0352, while increasing LPIPS-VGG by 0.0218 relative to the frozen Layer 23 setting. Under the current clean-980 data, learning rate, and frozen-head configuration, additional trainable capacity therefore does not improve NVS. One plausible explanation is that flow-matching gradients alter the pretrained visual representation while the frozen geometry heads retain their original feature organization, producing internal drift.

These half-resolution results are not directly comparable with the full-resolution main results. All main-paper experiments retain the pre-specified frozen Layer 23 configuration to avoid retrospective model selection on the fixed evaluation set. The complementary behavior across depths motivates multi-level feature fusion as a structured alternative that preserves the pretrained geometry representation while combining local appearance with broader multi-view context.

A.6 Complete Pose-Difficulty Results

We provide complete PSNR and LPIPS results across interpolation and extrapolation difficulty. All methods use six fixed source views and independent 6-to-1 inference on a common 480×480 image plane, with bins determined solely by camera pose. Each DL3DV bin contains 32 targets; Mip-NeRF 360 contains 66 targets per interpolation bin and 16 per extrapolation bin. Ranking highlights follow Table 1, and dashes denote unavailable results.

Unlike the aggregate comparison in Table 1, this diagnostic separates trajectory relation from pose novelty. Interpolation and extrapolation are each divided into near, mid, and far bins, revealing whether performance degrades smoothly as target cameras move away from the available observations. The same frozen cases are used for every method, so differences across bins reflect pose robustness rather than sampling changes. Figure 6 provides the corresponding camera layouts and representative VGGT-*Diff* predictions for the six categories.

Table 9: **PSNR across target-pose difficulty (↑).** Results are reported separately for DL3DV and Mip-NeRF 360.

Method	DL3DV						Mip-NeRF 360					
	Interpolation			Extrapolation			Interpolation			Extrapolation		
	Near	Mid	Far	Near	Mid	Far	Near	Mid	Far	Near	Mid	Far
Regression-based Models												
AnySplat (Jiang et al., 2025b)	14.196	12.415	10.862	13.977	12.566	11.827	12.424	11.040	9.851	12.594	10.273	8.972
E-RayZer (Zhao et al., 2026)	17.388	14.878	13.824	16.833	14.477	14.712	15.964	15.085	14.028	15.404	14.760	12.609
LVSM (Jin et al., 2024)	20.276	16.599	13.969	20.790	16.633	15.579	16.117	14.488	13.722	15.296	14.040	13.016
DepthSplat (Xu et al., 2025)	20.376	17.424	14.820	20.653	17.943	16.374	17.478	16.046	15.303	16.208	15.518	12.511
Diffusion-based Models												
Aether (Zhu et al., 2025)	14.030	12.167	11.065	15.739	12.295	11.795	12.977	12.994	12.867	11.841	10.873	10.894
GEN3C (Ren et al., 2025)	13.620	12.356	11.548	14.288	13.050	12.288	–	–	–	–	–	–
MVSplat360 (Chen et al., 2024b)	15.282	14.420	13.389	15.542	14.479	14.176	14.283	13.928	13.585	14.407	13.463	12.794
FrameCrafter (Wu et al., 2026a)	20.338	17.485	14.649	21.168	17.888	16.594	16.933	15.700	15.016	16.222	15.330	13.017
SEVA (Zhou et al., 2025)	20.940	17.861	14.703	22.311	18.050	16.575	17.911	16.467	15.427	17.448	15.925	13.933
VGGT-*Diff*	20.920	18.391	15.853	22.396	18.793	17.793	17.254	16.630	15.611	16.336	16.061	13.989

Analysis. On DL3DV, VGGT-*Diff* achieves the best PSNR in five of six bins and the best LPIPS in all six, while trailing SEVA by only 0.020 dB in interpolation-near. Its margin over FrameCrafter ranges from 0.582 to 1.228 dB and exceeds 0.9 dB in five of six bins. On zero-shot Mip-NeRF 360, VGGT-*Diff* obtains the best LPIPS in every bin and the best PSNR in four of six bins. The consistent perceptual gains across both datasets suggest that the geometry-routed diffusion prior remains scene-grounded under increasing viewpoint change.

A.7 Point-Cloud Reconstruction

Figure 10 complements the quantitative geometry evaluation with point clouds reconstructed from jointly generated views. We use two frozen geometry models: VGGT-Ω as the primary probe and

Table 10: **LPIPS across target-pose difficulty (↓).** We use VGG features on DL3DV and AlexNet features on Mip-NeRF 360.

Method	DL3DV						Mip-NeRF 360					
	Interpolation			Extrapolation			Interpolation			Extrapolation		
	Near	Mid	Far	Near	Mid	Far	Near	Mid	Far	Near	Mid	Far
Regression-based Models												
AnySplat (Jiang et al., 2025b)	0.5052	0.5609	0.6214	0.5057	0.5483	0.5828	0.5513	0.5854	0.6135	0.5920	0.6349	0.6830
E-RayZer (Zhao et al., 2026)	0.5496	0.6357	0.6925	0.5527	0.6433	0.6508	0.7270	0.7978	0.8284	0.7494	0.7975	0.7859
LVSM (Jin et al., 2024)	0.2911	0.4413	0.5656	0.2375	0.3913	0.4835	0.4980	0.6461	0.6897	0.5315	0.6589	0.7058
DepthSplat (Xu et al., 2025)	0.2688	0.3592	0.4643	0.2600	0.3141	0.4077	0.3164	0.4058	0.4555	0.3965	0.4842	0.5703
Diffusion-based Models												
Aether (Zhu et al., 2025)	0.4949	0.5891	0.6405	0.3952	0.5579	0.6098	0.5731	0.5855	0.5978	0.5965	0.6857	0.6980
GEN3C (Ren et al., 2025)	0.5194	0.5808	0.6269	0.4773	0.5578	0.5990	–	–	–	–	–	–
MVSplat360 (Chen et al., 2024b)	0.5009	0.5374	0.5952	0.4855	0.5223	0.5724	0.5820	0.6453	0.6610	0.6404	0.6683	0.6743
FrameCrafter (Wu et al., 2026a)	0.2489	0.3558	0.4689	0.2150	0.3011	0.3996	0.2712	0.3639	0.3996	0.3148	0.4204	0.5152
SEVA (Zhou et al., 2025)	0.2460	0.3390	0.4546	0.2122	0.2959	0.3929	0.2564	0.3267	0.3854	0.2983	0.4013	0.4744
VGGT-*Diff*	0.2423	0.3227	0.4163	0.1922	0.2732	0.3572	0.2474	0.3085	0.3504	0.2922	0.3685	0.4424

Pi3 (Wang et al., 2026b) as an independent reconstructor that is not used by VGGT-*Diff* for training or conditioning. Their agreement helps distinguish genuine cross-view improvements from potential shared-backbone evaluator bias.

A coherent set of generated views provides mutually compatible evidence and should produce a stable point cloud close to the corresponding GT-view reconstruction. Cross-view drift instead produces duplicated, displaced, or fragmented structures. All methods use the same six source views and eight target cameras. Within each reconstructor, generated and GT views follow identical reconstruction settings and the same downstream alignment, filtering, and sampling procedure. Because VGGT-Ω and Pi3 produce different point-cloud representations, comparisons are made within each reconstructor rather than across them. Their GT-view reconstructions serve as model-specific references, not absolute 3D scans.

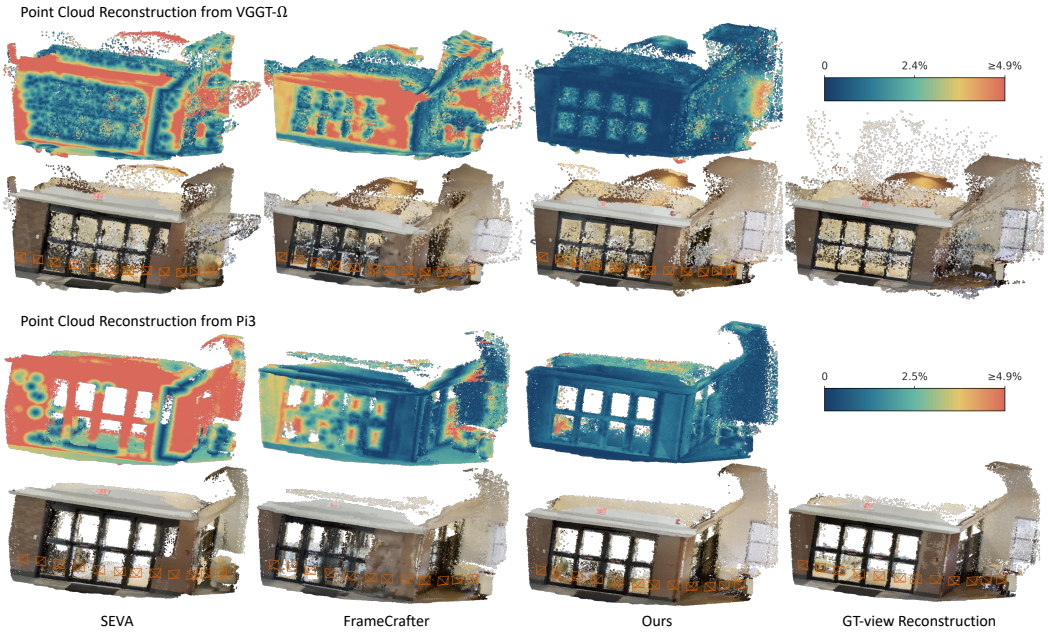

Figure 10: **Point-cloud reconstruction from jointly generated views.** The upper and lower blocks use VGGT-Ω and Pi3, respectively. Within each block, normalized point-to-reference errors are shown above the corresponding reconstructions using the displayed color scale. Both reconstructors show fewer displaced points and more coherent geometry for VGGT-*Diff*, indicating stronger cross-view consistency.

Both reconstructors produce the same overall ordering summarized in Table 4. SEVA (Zhou et al., 2025) and FrameCrafter exhibit larger high-error regions, displaced foreground points, and less stable facade structure. In contrast, VGGT-*Diff* produces more compact reconstructions with fewer structural offsets and better preserves the shared layout across views. The agreement between

VGGT-Ω and Pi3 supports the Chamfer-L1 and F-score gains while reducing dependence on any single geometry evaluator.

A.8 Residual versus Raw-Velocity Consistency

Motivation. PTRC regularizes predicted-clean residuals rather than directly tying raw model outputs across views. To isolate this design choice, we replace PTRC with a naive objective that applies the same point tracks, confidence weights, and robust penalty directly to predicted velocities:

$$\mathcal{L}_{\text{Vel}} = \frac{\sum_{\gamma \in \mathcal{C}} w_\gamma \, \bar{\rho}_\beta \left(\mathbf{V}_{\theta,a}(\mathbf{u}_{\gamma,a}) - \mathbf{V}_{\theta,b}(\mathbf{u}_{\gamma,b}) \right)}{\sum_{\gamma \in \mathcal{C}} w_\gamma}. \tag{12}$$

This alternative assumes that corresponding locations should share the same velocity. Under linear flow matching, however, each view has its own target $\mathbf{U}_j = \boldsymbol{\epsilon}_j - \mathbf{Z}_{0,j}$; view-specific clean latents and noise states therefore make raw velocity equality generally invalid. Since $\mathbf{E}_j = -\sigma(\mathbf{V}_{\theta,j} - \mathbf{U}_j)$, PTRC instead aligns errors relative to each view's own flow target, preserving valid differences between their velocities.

This comparison is stricter than simply removing PTRC. The direct-matching baseline retains the same 3D correspondences and explicitly couples the same query locations, but changes what is required to agree. It therefore tests whether the benefit comes from generic point-track smoothing or specifically from aligning residual errors relative to each view's supervision target.

Controlled comparison. Both variants share the data, initialization, geometry conditioning, point-track construction, and optimization, differing only in the consistency objective. For efficiency, they follow the same half-resolution protocol and are evaluated on the same 192 pose-stratified targets. Geometry conditioning, condition regularization, target curriculum, and robust penalty remain unchanged; both objectives use the stage-wise schedule in Eq. 9, with full weight 0.1.

Table 11: **PTRC versus direct velocity matching across pose difficulty.** Both use the same half-resolution protocol; PTRC values are highlighted and LPIPS uses VGG features.

Interpolation						
Difficulty	PSNR ↑		SSIM ↑		LPIPS ↓	
	Direct	PTRC	Direct	PTRC	Direct	PTRC
Near	19.008	20.133	0.531	0.602	0.274	0.221
Mid	17.520	18.103	0.435	0.467	0.341	0.300
Far	15.210	15.670	0.308	0.352	0.440	0.397

Extrapolation						
Difficulty	PSNR ↑		SSIM ↑		LPIPS ↓	
	Direct	PTRC	Direct	PTRC	Direct	PTRC
Near	20.021	21.326	0.591	0.674	0.238	0.173
Mid	17.863	18.613	0.478	0.535	0.308	0.260
Far	16.994	17.641	0.415	0.463	0.377	0.331

Analysis. Across all 192 targets, PTRC improves PSNR from 17.769 to 18.581 dB and SSIM from 0.460 to 0.516, while reducing LPIPS from 0.330 to 0.280. Table 11 further shows consistent gains on every metric across all six pose bins, with PSNR improvements ranging from 0.460 to 1.305 dB. Because both variants use identical correspondences and weighting, these gains isolate the residual-relative formulation rather than the point tracks themselves. Direct velocity matching penalizes valid differences between view-specific flow targets, whereas PTRC couples only the prediction errors that should be removed. The improvement appears in both interpolation and extrapolation, showing that the residual formulation is not tied to one camera relation or difficulty range.

A.9 Additional Optimization and Statistical Diagnostics

Wan adaptation. We compare full DiT fine-tuning with a rank-32 LoRA adaptation while retaining the same geometry conditioner, data, initialization, objectives, and half-resolution 192-case evaluation. Full fine-tuning improves PSNR by 1.324 dB and SSIM by 0.0478, while reducing LPIPS-VGG by 0.0751. The large gap indicates that sparse low-rank updates are insufficient to integrate routed geometry into the pretrained video prior under this setup.

Table 12: **Effect of Wan adaptation strategy.** Both variants use the same half-resolution controlled protocol; LPIPS uses VGG features.

Wan adaptation	PSNR ↑	SSIM ↑	LPIPS ↓
LoRA, rank 32	17.2563	0.4677	0.3554
Full fine-tuning	18.5807	0.5155	0.2803

Paired scene-level statistics. We additionally compare locally generated VGGT-*Diff* and FrameCrafter outputs under the same 140-scene, 6,188-target protocol. The scene-mean PSNR gain

is 0.7893 dB, with a 92.14% scene win rate and a scene-bootstrap 95% confidence interval of $[0.7003, 0.8769]$ dB. The corresponding scene-mean changes are $+0.0330$ SSIM, -0.0365 LPIPS-AlexNet, and -0.0118 DreamSim. These paired results show that the aggregate improvement is distributed across scenes rather than driven by a small subset. They are reported separately from Table 1, whose baseline rows follow the published cross-method protocol.

A.10 Robustness to Sparse Input Views

Protocol. We evaluate whether the frozen full-resolution checkpoint trained with six source views transfers to reduced source coverage without retraining or test-time optimization. Both VGGT-*Diff* and FrameCrafter use independent M-to-1 inference on the same 192 pose-stratified DL3DV targets, with $M \in \{3, 4\}$. Inference runs at 480×832 for 50 sampling steps, and metrics are computed on the center 480×480 crop. Both methods use identical source images, target cameras, and per-case noise seeds. For each target, we select M views from the six sources fixed by the main protocol using camera poses only. A source s is ranked by $\sqrt{(\|\mathbf{c}_t - \mathbf{c}_s\|_2/D)^2 + (\theta_{ts}/\pi)^2}$, where D is the maximum pairwise distance among the six source cameras and θ_{ts} is their viewing direction difference. The selected views are restored to their registered trajectory order, producing nested source subsets independent of image content, predictions, and method identity.

Table 13: **Robustness to sparse input views on DL3DV.** Both methods use full-resolution independent M-to-1 inference on the same 192 targets. Source subsets are selected identically using only camera poses; LPIPS uses VGG features.

Sources	Method	PSNR ↑	SSIM ↑	LPIPS ↓	DreamSim ↓
3	FrameCrafter (Wu et al., 2026a)	16.9559	0.4647	0.3709	0.0807
	VGGT-*Diff*	17.7372	0.4966	0.3391	0.0680
4	FrameCrafter (Wu et al., 2026a)	17.8545	0.5076	0.3390	0.0685
	VGGT-*Diff*	18.4380	0.5307	0.3135	0.0592

Results. Despite being trained with six sources, VGGT-*Diff* remains stronger under both reduced-view settings. With three inputs, it improves over FrameCrafter by 0.781 dB PSNR and 0.0319 SSIM while reducing LPIPS-VGG by 0.0318 and DreamSim by 0.0127. With four inputs, the corresponding gains are 0.584 dB, 0.0231, 0.0255, and 0.0093. The consistent improvement across all four metrics shows that geometry-routed conditioning remains effective when the available observations are reduced to three or four views, without adapting the six-view-trained checkpoint.

A.11 VAE Encoding Protocols

All VGGT-*Diff* variants use the same frozen Wan2.1 Video VAE, which produces 16-channel latents with an $8\times$ spatial downsampling factor ($192\times336 \to 16\times24\times42$ and $480\times832 \to 16\times60\times104$). All quantitative experiments adopt an independent-view latent layout. During training, each source image and ground-truth target image is independently encoded, with the latter providing the clean latent for denoising supervision. During inference, only source images are encoded; target latent slots are initialized from noise, and no target RGB image is provided to the model. Consequently, six sources and N targets produce $6 + N$ view-time latent slots, each aligned one-to-one with a physical view, camera, Plücker map, and routed geometry condition. Thus, none of our reported quantitative results uses temporal VAE compression.

To accelerate experimentation and reduce the training cost of long sequences, we use a hybrid protocol for the continuous 80-frame trajectory demo and fine-tune it at the half resolution of 192×336. The six sources remain independently encoded, while the ordered target sequence is jointly encoded through the causal temporal pathway of the same frozen VAE, yielding $L_T = 1 + \lceil (T-1)/4 \rceil$ target slots. For $T = 80$, we repeat the final frame once to form an 81-frame sequence, obtaining $L_T = 21$ target slots; the repeated output is discarded after decoding. This reduces the number of view-time latent slots processed by the DiT from $6 + 80 = 86$ to $6 + 21 = 27$. Camera conditions bypass the VAE. Each six-channel Plücker map is spatially rearranged into 384 channels at latent resolution using an $8\times$ PixelUnshuffle. To avoid temporally discarding intermediate target poses, we group these maps according to the causal VAE windows: $[0]$, $[1, 2, 3, 4]$, …, and $[77, 78, 79, 79]$. Within each four-frame window, the ordered 384-channel maps are concatenated into 1536 channels and projected back to 384 by a lightweight $1\times1\times1$ Plücker adapter, making all four poses available to the conditioning pathway while preserving the DiT interface. The first target slot retains its

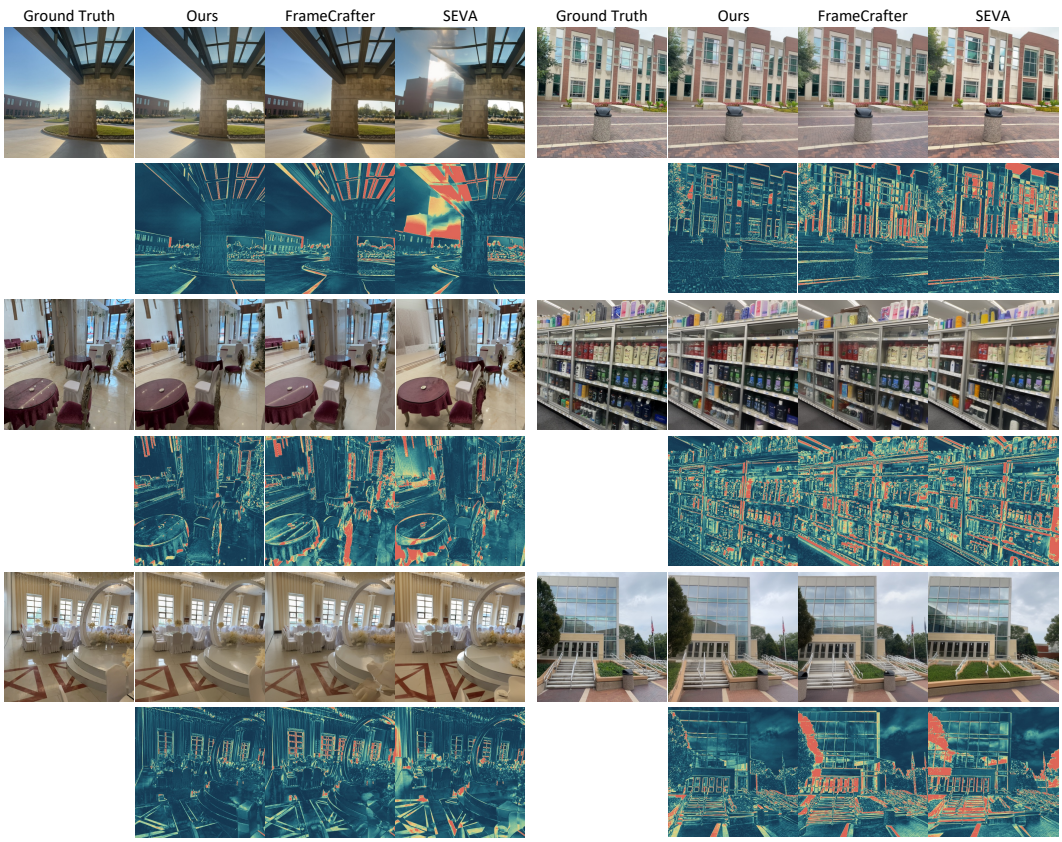

Figure 11: **Target-view alignment and pixel error.** Predicted RGB views are shown above their absolute pixel differences from GT under a shared error scale. VGGT-*Diff* produces lower and more localized errors than the baselines, indicating closer alignment with the GT viewpoint and fewer spatial or structural mismatches.

single-frame conditioning path. We initialize the adapter to select only the fourth map, making the augmented model exactly equivalent to the previous endpoint-only checkpoint at initialization and enabling checkpoint-compatible fine-tuning without modifying the DiT backbone. In the current adaptation, target camera matrices and routed geometry conditions remain aligned with the endpoint indices $\mathcal{A}_T = \{0, 4, 8, \ldots, 76, 79\}$; only the Plücker pathway aggregates all poses within each temporal window. Because independently encoded and temporally compressed target latents have different temporal semantics, we fine-tune the initialized model for this 80-frame regime rather than treating the two representations as interchangeable. This half-resolution extension is used only for the continuous demo and does not affect any reported quantitative result.

A.12 Future Directions

Our results suggest several directions for extending geometry-routed multi-view generation. First, the complementary behavior observed across VGGT-Ω feature depths motivates adaptive multi-level fusion that combines fine appearance cues with increasingly global geometric context. Second, although directly optimizing the VGGT-Ω visual pathway is ineffective under our current setting, parameter-efficient adaptation, geometry-preserving objectives, or staged optimization may better specialize the encoder without disrupting its pretrained representations. The Visual Geometry Router could also be extended beyond its hard anchor and layered residual refinement to model richer visibility distributions, multiple occlusion layers, and better-calibrated geometric uncertainty. At the generation level, memory-efficient joint denoising and chunked generation may scale the benefits of shared multi-view context to more views and longer camera trajectories while maintaining consistency across chunks. Finally, extending 3D point tracks to spatiotemporal correspondences could generalize the framework from static scenes to dynamic and non-rigid content with moving objects and time-varying occlusions.