GEOVERSE: WORLD-CONSISTENT NOVEL VIEW SYNTHESIS IN GEOMETRIC LATENT SPACE

Anonymous authors
Paper under double-blind review

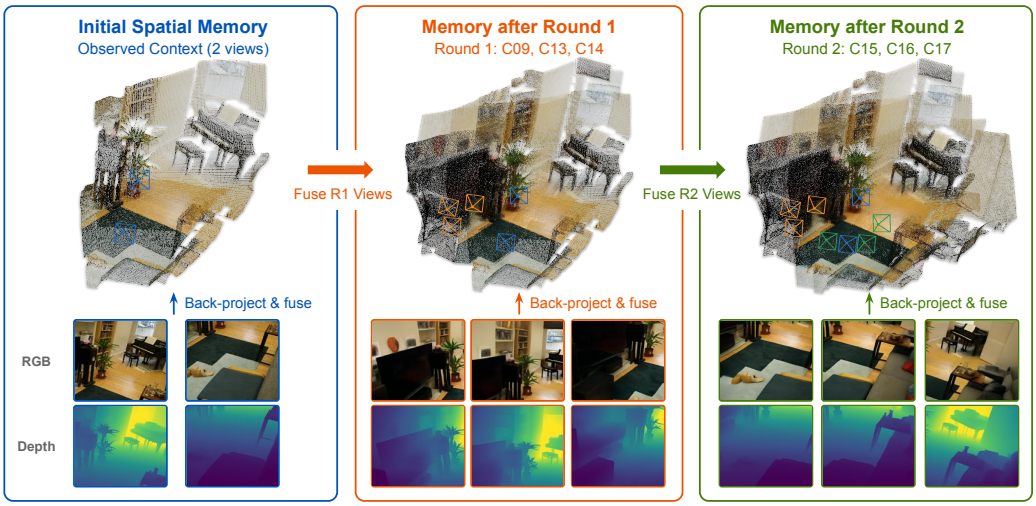


Figure 1: GeoVerse enables long-sequence novel view synthesis by iteratively integrating generated observations into a persistent spatial memory, which in turn guides subsequent view synthesis.

ABSTRACT

Novel view synthesis from sparse images must reconcile faithful reconstruction of observed regions with plausible completion of unseen content, while maintaining world consistency across viewpoints. Existing geometry-based methods preserve observed scene structure but often struggle to complete unseen regions, whereas video generative models offer rich appearance priors but accumulate inconsistencies during sequential view generation. We propose GeoVerse, a framework that synthesizes world-consistent novel views by performing generation within the geometric latent space of a pretrained 3D foundation model and injecting appearance priors from a video generative model. Specifically, GeoVerse extracts multilevel features from Wan2.2 VACE and injects them into the geometric latent diffusion model via a ControlNet-style adapter, incorporating video-learned appearance priors to enhance structural completion. To enforce cross-view coherence, a global spatial memory continuously aggregates observed and synthesized content, reprojecting target-aligned guidance to anchor subsequent predictions to a shared scene representation. Extensive experiments across diverse datasets demonstrate improved visual quality and geometric consistency, with a 2.23 dB higher PSNR on DL3DV and 32.4% lower ATE on Mip-NeRF360 compared to GLD.

1 INTRODUCTION

Novel View Synthesis (NVS) is a cornerstone task in 3D computer vision that aims to render photorealistic images from specified, previously unseen target viewpoints given one or more reference images (Mildenhall et al., 2021; Yu et al., 2024a). A fundamental challenge in NVS lies in balancing reconstruction fidelity during view interpolation with generative capability during view extrapolation, faithfully aggregating observed scene content while plausibly completing unobserved regions.

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.

To bridge faithful reconstruction with generative completion, maintaining a unified geometric representation is crucial for ensuring spatial and semantic coherence across shifting viewpoints.

Early NVS frameworks were predominantly constrained to per-scene optimization, with pioneering paradigms like Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) and 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) fitting scene-specific representations on dense captures. To bypass scene-specific training bottlenecks, recent generalizable architectures leverage scaled data to predict Gaussian primitives directly from sparse views in a feed-forward manner (Chen et al., 2024; Jiang et al., 2025a). Spurred by 3D foundation models (Wang et al., 2024; 2025a), these feed-forward methods achieve rapid feed-forward reconstruction with strong geometry grounding (Jiang et al., 2025a; Liu et al., 2025). However, bound by deterministic geometric formulations, pure reconstruction frameworks inherently lack generative imagination, frequently producing severe visual artifacts, blurriness, or hollow voids when synthesizing unobserved regions under large camera motions.

To compensate for the limited extrapolation capability of reconstruction methods, video diffusion models have recently been repurposed for view synthesis (Yu et al., 2024a), leveraging rich appearance priors learned from vast video corpora (Wan et al., 2025). Despite their expressive completion, applying video models directly or sequentially to 3D scenes often accumulates cross-frame inconsistencies, causing drift and structural breakdown over long horizons (Wu et al., 2026; Wang et al., 2026). To enforce spatial consistency, hybrid frameworks align video diffusion features with 3D representations (Wu et al., 2025; Huang et al., 2026), yet they remain burdened by the heavy computational overhead of synthesizing dense video frames. More recently, Geometry Latent Diffusion (GLD) (Jang et al., 2026) has emerged as a promising alternative that generates novel views directly within a geometric latent space. Built upon 3D foundation models such as DA3 (Lin et al., 2025), which is pretrained on large-scale data with depth and camera-pose supervision to learn cross-view attention, this space provides a robust structural foundation and a geometric prior for single-pass consistency. Furthermore, GLD's RGB head effectively decodes fine texture and appearance details from these features (Jang et al., 2026). However, restricted by a limited training distribution, GLD struggles with generative fidelity and stability on in-the-wild scenes.

In summary, existing NVS approaches struggle to seamlessly harmonize rigid 3D geometric consistency with expressively detailed generative completion. To bridge this gap, we present **GeoVerse**, a novel framework that endows geometric latent diffusion with rich video generative priors while maintaining structural consistency via a persistent spatial memory. We address these limitations through three key designs. First, we incorporate Wan2.2 VACE (Wan et al., 2025; Jiang et al., 2025b), a high-capacity video diffusion model pretrained on extensive video data, to enrich the geometric latent space with powerful appearance priors and plausible scene completion. Second, we scale the 3D training corpus from 4 to 15 diverse real and synthetic datasets to broaden scene coverage and enhance cross-domain generalization. Beyond single-pass consistency, successive view expansion requires a persistent representation across multiple inference steps. Inspired by ViewCrafter (Yu et al., 2024a), we maintain a global colored point-cloud memory that accumulates historical content. Reprojecting this memory into target views provides pixel-aligned guidance, anchoring multi-step predictions while facilitating unobserved region completion. Leveraging this structured guidance, we employ reflow distillation (Yan et al., 2024) to reduce the denoising process to 4 steps, cutting per-inference latency from over 100 seconds to under 10 seconds. Our primary contributions are summarized as follows: (1) We introduce a novel framework that seamlessly bridges geometric latent diffusion with expressive video generative priors for world-consistent novel view synthesis. (2) We propose a global spatial memory mechanism that provides target-aligned input guidance, effectively enforcing sustained long-sequence geometric consistency across extended trajectories. (3) We scale up model training across diverse real and synthetic multi-view corpora to significantly enhance cross-domain generalization. (4) We integrate a few-step reflow distillation scheme to significantly accelerate inference to under 10 seconds.

2 Related Work

2.1 Reconstructive Novel View Synthesis

Focusing on fusing existing scene content within captured view bounds, reconstructive novel view synthesis was initially dominated by scene-specific optimization. Pioneered by NeRF (Mildenhall et al., 2021), implicit radiance fields achieved novel view rendering through differentiable volume

rendering, which was later accelerated and scaled by subsequent variants (Barron et al., 2021; Müller et al., 2022). To overcome the implicit rendering bottleneck, 3DGS (Kerbl et al., 2023) introduced explicit 3D Gaussian primitives for real-time synthesis, followed by improvements in anti-aliasing and geometry modeling (Yu et al., 2024b; Lu et al., 2024; Ren et al., 2024). While capable of rendering high-fidelity views, these per-scene representations suffer from long optimization times and strictly require dense multi-view coverage. To bypass per-scene optimization, feed-forward frameworks learn generalizable mappings from sparse inputs to novel target views. Early generalizable architectures like MVSplat (Chen et al., 2024) leverage plane-sweep cost volumes for 3D Gaussian prediction, while LVSM (Jin et al., 2024) scales transformer-based synthesis with minimal explicit geometric priors. Accelerated by 3D foundation models like VGGT (Wang et al., 2025a), recent frameworks seamlessly unify camera estimation and feed-forward reconstruction. Specifically, AnySplat (Jiang et al., 2025a) jointly recovers camera poses and 3D Gaussians from unconstrained views, whereas WorldMirror (Liu et al., 2025) incorporates multi-source geometric priors. In the monocular regime, SHARP (Mescheder et al., 2026) regresses a metric 3D Gaussian field from a single image in one forward pass. Although feed-forward methods excel at interpolative synthesis within bounded views, extrapolating to unobserved regions remains inherently ambiguous.

2.2 Generative Novel View Synthesis

Generative novel view synthesis introduces generative priors to synthesize content beyond the observed scene coverage. ViewCrafter combines point-based 3D clues with a pretrained video diffusion model and iteratively expands both its camera trajectory and reconstructed content (Yu et al., 2024a). NeoVerse couples feed-forward 4D reconstruction with novel-trajectory video generation to model scenes from in-the-wild monocular videos (Yang et al., 2026). To reduce drift over longer horizons, spatial-memory approaches explicitly store and retrieve previously generated scene content: geometry-grounded long-term memory caches 3D history, while Mirage lifts diffusion latents into a persistent 3D cache and queries it through latent-space warping (Wu et al., 2026; Wang et al., 2026). Lyra 2.0 combines per frame geometric matching for history retrieval with self augmented history training to suppress cumulative drift during long sequence generation (Shen et al., 2026). These methods improve long-horizon consistency while retaining a video-based generation pipeline.

Geometric foundation models offer a complementary route to geometry-aware extrapolation. Geometry Forcing supervises intermediate video-diffusion representations with features from a geometric foundation model, while Gen3R adapts VGGT tokens into geometric latents and aligns them with pretrained video appearance latents for joint RGB and 3D generation (Wu et al., 2025; Huang et al., 2026). Although these methods improve the 3D awareness of video diffusion, their generation process remains organized as a video sequence. Geometric Latent Diffusion (GLD) instead repurposes the geometric feature space of DA3 as the native latent space for multi-view diffusion, enabling direct generation at specified target cameras with strong cross-view correspondence (Lin et al., 2025; Jang et al., 2026). Unlike prior sequence-based approaches, GeoVerse adopts the geometric-latent formulation, augmented by video generative priors and a persistent 3D memory, enabling direct target-view synthesis free of dense video interpolation.

3 Method

Fig. 2 illustrates the overall pipeline of GeoVerse, a framework that integrates geometric latent diffusion with video generative priors and a global spatial memory for world-consistent novel view synthesis. Let $\mathcal{C}$ and $\mathcal{T}$ denote the context and target view sets, with sizes $N_{\mathcal{C}}$ and $N_{\mathcal{T}}$, respectively. Given context images $\mathbf{I}_{\mathcal{C}}$ and target-aligned guidance $\mathbf{I}_{\mathcal{T}}^{\text{proj}}$, GeoVerse predicts target RGB images $\hat{\mathbf{I}}_{\mathcal{T}}$ and geometry $\hat{\mathbf{G}}_{\mathcal{T}} = \{\hat{\mathbf{D}}_{\mathcal{T}}, \hat{\mathbf{R}}_{\mathcal{T}}\}$, comprising depth and raymaps, and then updates the spatial memory $\mathbf{M}$. Specifically, Sec. 3.1 details the integration of geometric latent diffusion with video generative priors, Sec. 3.2 describes the spatial-memory aggregation and target-aligned projection, and Sec. 3.3 outlines the model distillation strategy for efficient inference.

3.1 Enhancing Geometric Latent Diffusion with Video Priors

Geometry Latent Space. Standard image-video latent spaces provide robust appearance priors but struggle to maintain explicit 3D correspondence across novel viewpoints (Rombach et al., 2022).

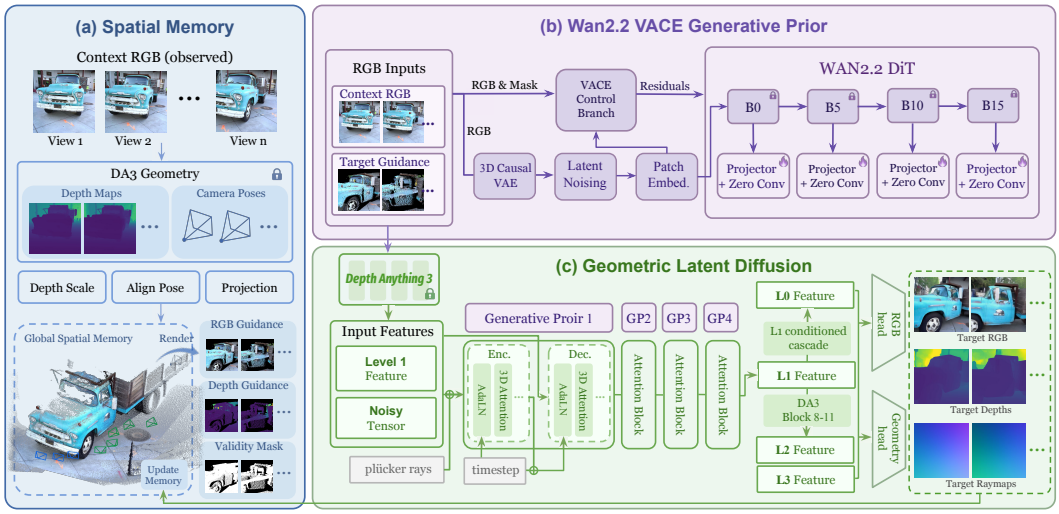


Figure 2: **Overview of GeoVerse.** Context observations initialize a global spatial memory that provides target-aligned RGB-D hints. Guided by the hints and Wan2.2 VACE features injected through a ControlNet-style adapter, geometric latent diffusion synthesizes target views in four denoising updates. The decoded RGB and geometry update the memory for subsequent view expansion.

In contrast, GLD combines the geometry head of Depth Anything 3 (Lin et al., 2025) for spatial layouts (depth and raymaps) with an auxiliary RGB head for high-frequency textures (Jang et al., 2026). Following GLD (Jang et al., 2026), we perform diffusion generation directly within the multi-level geometric latent space $\mathcal{F} = \{\mathbf{F}^i\}_{i=0}^3$ of a frozen DA3 encoder, where feature levels correspond to Transformer blocks $(b_0, b_1, b_2, b_3) = (5, 7, 9, 11)$, and designate Level-1 as the synthesis boundary to balance geometric accuracy and visual fidelity. Unlike GLD's zero-padding target conditions, GeoVerse constructs the input sequence $\bar{\mathbf{I}} = [\mathbf{I}_{\mathcal{C}}, \mathbf{I}_{\mathcal{T}}^{\text{proj}}]$ by merging context images with spatial memory projections, which are then processed by the frozen DA3 encoder up to the synthesis boundary:

$$\overline{\mathbf{F}} = \mathbf{V}^{b_1} \odot \mathcal{E}_{\text{geo}}^{1:b_1}(\bar{\mathbf{I}}), \tag{1}$$

where $\mathcal{E}_{\text{geo}}^{1:b_1}$ encodes up to block b_1, and $\mathbf{V}^{b_1}$ is the validity mask aligned to its feature resolution.

Let $\mathbf{F}^1$ denote the clean Level 1 features of the ground-truth multi-view images. Given Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $t \sim \mathcal{U}(0, 1)$, our multi-view flow-matching model (Lipman et al., 2023) adopts the linear probability path $\mathbf{X}_t = (1-t)\mathbf{F}^1 + t\boldsymbol{\epsilon}$ with target velocity $\mathbf{u}_t = \boldsymbol{\epsilon} - \mathbf{F}^1$. The denoiser jointly predicts the velocity of all view tokens via:

$$\hat{\mathbf{u}}_t = v_\theta\left(\mathbf{X}_t, t; \overline{\mathbf{F}}, \boldsymbol{\Gamma}, \mathbf{F}^{\text{W}}, \mathbf{D}^{\text{proj}}, \mathbf{V}\right), \tag{2}$$

where $\boldsymbol{\Gamma}$ denotes camera Plücker-ray embeddings, $\mathbf{F}^{\text{W}}$ represents Wan2.2 features, and $\mathbf{D}^{\text{proj}}$ and $\mathbf{V} = [\mathbf{1}_{\mathcal{C}}, \mathbf{V}_{\mathcal{T}}]$ refer to projected depth guidance and its validity mask. Under these fixed conditions, we integrate the predicted velocity field from $t = 1$ to $t = 0$ to transform Gaussian noise into the Level-1 features $\hat{\mathbf{F}}^1$. These features then drive a conditional cascade, alongside context features and camera embeddings, to synthesize the Level-0 features $\hat{\mathbf{F}}^0$:

$$\hat{\mathbf{F}}^0 = \mathcal{S}_\phi(\boldsymbol{\epsilon} \mid \hat{\mathbf{F}}^1, \mathbf{F}_{\mathcal{C}}^0, \boldsymbol{\Gamma}), \tag{3}$$

where $\mathcal{S}_\phi$ denotes the cascade sampler initialized from Gaussian noise $\boldsymbol{\epsilon}$. Subsequently, the remaining frozen DA3 blocks forward-process $\hat{\mathbf{F}}^1$ to compute $(\hat{\mathbf{F}}^2, \hat{\mathbf{F}}^3) = \mathcal{E}_{\text{geo}}^{b_1+1:b_3}(\hat{\mathbf{F}}^1)$. Specialized RGB and geometry heads then utilize the multi-scale hierarchy $\hat{\mathcal{F}} = \{\hat{\mathbf{F}}^i\}_{i=0}^3$ to decode target appearance $\hat{\mathbf{I}}_{\mathcal{T}} = \mathcal{D}_{\text{rgb}}(\hat{\mathcal{F}})_{\mathcal{T}}$ and target geometry $\hat{\mathbf{G}}_{\mathcal{T}} = \mathcal{D}_{\text{geo}}(\hat{\mathcal{F}})_{\mathcal{T}}$.

Video-Prior Injection. To complement GLD's geometric representation with learned visual priors, we use a frozen Wan2.2 VACE model (Wan et al., 2025; Jiang et al., 2025b). Rather than sampling a complete video, we extract its intermediate features once and reuse them throughout geometric denoising. Further details of Wan2.2 VACE are provided in Appendix A.

Specifically, its feature extractor processes the RGB sequence $\bar{\mathbf{I}}$. Upon normalization and resizing, we inject Gaussian noise ($\sigma = 0.1$) exclusively into valid target projections, producing $\widetilde{\mathbf{I}}$ while leaving context RGB unperturbed. We encode $\widetilde{\mathbf{I}}$ with the frozen video VAE to obtain $\mathbf{z}_{\text{proxy}} = E_{\text{VAE}}(\widetilde{\mathbf{I}})$ and construct $\mathbf{x}_t = (1-\sigma_t)\mathbf{z}_{\text{proxy}} + \sigma_t\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ and σ_t is the noise level associated with the Wan timestep t. Using $\mathbf{x}_t$ as the backbone input and $[\mathbf{z}_{\text{proxy}}, \mathbf{V}]$ as the VACE condition, we extract features from blocks $(k_0, k_1, k_2, k_3) = (0, 5, 10, 15)$ in a single forward pass:

$$\mathbf{F}^{\text{W},\ell} = \mathcal{W}^{0:k_\ell}(\mathbf{x}_t, t; \mathcal{E}_{\text{VACE}}([\mathbf{z}_{\text{proxy}}, \mathbf{V}]))\,, \tag{4}$$

where $\ell \in \{0, 1, 2, 3\}$, $\mathcal{E}_{\text{VACE}}$ denotes the VACE conditioning unit, $\mathcal{W}^{0:k_\ell}$ denotes the Wan backbone up to the selected block, and $\mathbf{F}^{\text{W}}$ collectively denotes the extracted features.

To transfer these video features into the geometric latent space, a convolutional module $\mathcal{A}_\ell$ matches each feature $\mathbf{F}^{\text{W},\ell}$ to the spatial resolution and channel dimension of its corresponding injection layer. A ControlNet-style branch (Zhang et al., 2023) processes the aligned features under the same diffusion timestep and camera conditions as the main denoiser, then supplies an additive residual:

$$\mathbf{c}_{\ell+1} = \mathcal{B}_\ell^{\text{ctrl}}(\mathbf{c}_\ell + \mathcal{A}_\ell(\mathbf{F}^{\text{W},\ell}); t, \boldsymbol{\Gamma}), \tag{5}$$

$$\mathbf{h}_{\ell+1} = \mathcal{B}_\ell(\mathbf{h}_\ell; t, \boldsymbol{\Gamma}) + \mathcal{Z}_\ell(\mathbf{c}_{\ell+1}), \tag{6}$$

where $\mathbf{c}_\ell$ and $\mathbf{h}_\ell$ are the control and main-branch features. The zero-initialized projections $\mathcal{Z}_\ell$ ensure that the auxiliary branch keeps the pretrained denoiser initially unmodified. During training, these residuals learn to adapt frozen video features for geometric generation, boosting appearance fidelity and completion while preserving explicit camera control.

Loss Function and Training. We supervise both latent generation and decoded target predictions. The Level 1 denoiser uses a weighted flow-matching objective:

$$\mathcal{L}_{\text{FM}}^1 = \mathbb{E}_{t,\boldsymbol{\epsilon}}\Big[\sum_{i\in\mathcal{C}\cup\mathcal{T}} \alpha_i \Big\| v_\theta\big(\mathbf{X}_t, t; \overline{\mathbf{F}}, \boldsymbol{\Gamma}, \mathbf{F}^{\text{W}}, \mathbf{D}^{\text{proj}}, \mathbf{V}\big)_i - \mathbf{u}_{t,i}\Big\|_2^2\Big], \tag{7}$$

where $\alpha_i = 0.25$ for context views and 1 for target views, prioritizing novel-view synthesis while retaining context reconstruction. The Level 0 cascade uses an analogous objective $\mathcal{L}_{\text{FM}}^0$.

For decoded target RGB, $\mathcal{L}_{\text{rgb}}$ represents the mean valid-pixel ℓ_1 reconstruction error, whereas $\mathcal{L}_{\text{hf}}$ applies the same loss to high-gradient regions. Following DA3 (Lin et al., 2025), we estimate a camera-center similarity transform and leverage its scale $s^\star$ to align predicted depth, yielding $\mathcal{L}_{\text{depth}}$ as the mean valid-pixel ℓ_1 error between $s^\star\hat{\mathbf{D}}_\mathcal{T}$ and the reference depth. The full objective is

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{FM}}^1 + \lambda_0\mathcal{L}_{\text{FM}}^0 + \lambda_{\text{rgb}}\mathcal{L}_{\text{rgb}} + \lambda_{\text{hf}}\mathcal{L}_{\text{hf}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}. \tag{8}$$

3.2 Persistent Spatial Memory for Long-Sequence Consistency

Spatial Memory Aggregation. To maintain multi-round scene consistency, GeoVerse incrementally updates a colored point-cloud memory $\mathbf{M}^r$ in a shared coordinate system (Yu et al., 2024a; Ren et al., 2025). Context observations initialize the memory as $\mathbf{M}^0 = \mathbf{P}^0$ using DA3-predicted (Lin et al., 2025) or metric depth if available. For generated rounds, predicted depth is scale-aligned via $\widetilde{\mathbf{D}}_i^r = s_r^\star\hat{\mathbf{D}}_i^r$ using a camera-center similarity transform. Back-projecting valid pixels with $\widetilde{\mathbf{D}}_i^r$ and camera parameters $\mathbf{C}_i$ constructs a point set $\mathbf{P}^r$ containing 3D positions, colors, and confidence weights. Fusing each new point set $\mathbf{P}^r$ into $\mathbf{M}^{r-1}$ produces $\mathbf{M}^r$, placing accumulated content in a unified coordinate frame while mitigating scale drift.

Target-Aligned Projection. For each target camera $\mathbf{C}_j$ ($j \in \mathcal{T}$), depth-aware point splatting renders the spatial memory $\mathbf{M}^r$ into RGB-D guidance ($\mathbf{I}_j^{\text{proj}}, \mathbf{D}_j^{\text{proj}}$) and a validity mask $\mathbf{V}_j$, accounting for point confidence and z-buffer visibility. The mask $\mathbf{V}_j$ equals one for pixels supported by valid point projections and zero elsewhere. These projected hints and validity mask are subsequently encoded into target-side conditioning features via Eq. (1). Crucially, valid projections anchor known scene structures, while masked regions guide the generative prior to synthesize unobserved areas.

3.3 Few-Step Inference via Reflow Distillation

Hierarchical Few-Step Sampling. We allocate three denoising steps to Level-1 for primary multi-view generation and one step to the frozen Level-0 cascade for shallow feature recovery. This budget

Table 1: **Quantitative comparison across multiple datasets** in terms of visual quality, geometric accuracy, and average inference time. **Bold** and underline denote the best and second-best results.

	Method	2D Metrics			3D Metrics					Efficiency
		PSNR↑	SSIM↑	LPIPS↓	ATE↓	RPE_r ↓	RPE_t ↓	Reproj.↓	MEt3R↓	Time (s)↓
DL3DV	ViewCrafter	15.98	0.516	0.494	0.367	10.214	0.728	0.738	0.313	286.75
	NeoVerse	12.14	0.353	0.622	0.164	7.934	0.382	**0.638**	0.278	249.19
	GEN3C	17.06	0.561	0.435	0.114	5.788	0.238	0.692	0.301	343.10
	MVGenMaster	17.28	0.573	0.384	0.098	5.235	0.205	0.669	0.282	37.64
	Matrix3D	13.47	0.416	0.494	0.153	6.127	0.323	0.715	0.288	53.39
	CAMEO	11.14	0.383	0.677	0.906	44.157	1.987	0.887	0.425	14.78
	GLD	17.38	0.546	0.383	0.058	1.538	0.125	0.652	0.262	169.29
	Ours	**19.61**	**0.598**	**0.339**	**0.028**	**0.942**	**0.060**	0.649	**0.261**	**9.24**
RealEstate10K	ViewCrafter	16.84	0.646	0.407	0.091	1.715	0.160	0.685	0.208	287.87
	NeoVerse	11.89	0.450	0.598	0.044	0.858	0.079	0.728	0.184	250.05
	GEN3C	18.51	0.712	0.324	0.055	0.512	**0.060**	0.686	0.182	314.18
	MVGenMaster	20.54	0.740	0.295	0.039	0.991	0.067	0.653	0.196	26.78
	Matrix3D	15.71	0.562	0.408	0.041	0.688	0.072	0.671	0.218	53.39
	CAMEO	13.68	0.515	0.529	0.140	4.853	0.329	0.771	0.303	14.28
	GLD	19.66	0.709	0.299	0.037	1.338	0.067	**0.609**	0.183	163.50
	Ours	**21.29**	**0.756**	**0.275**	**0.035**	**0.437**	0.065	0.631	**0.179**	**9.29**
Mip-NeRF 360	ViewCrafter	15.84	0.417	0.553	0.446	19.046	1.102	0.706	0.346	295.81
	NeoVerse	12.17	0.297	0.652	0.394	10.574	0.584	0.626	0.301	204.14
	GEN3C	17.13	0.465	0.476	0.420	11.549	0.679	0.674	0.304	339.81
	MVGenMaster	16.75	0.446	0.427	0.093	2.507	0.217	0.645	0.284	24.71
	Matrix3D	15.06	0.360	0.489	0.118	2.819	0.197	0.658	0.287	52.93
	CAMEO	10.82	0.275	0.689	0.964	33.244	2.183	0.885	0.436	14.95
	GLD	17.57	0.440	0.406	0.105	2.337	0.202	0.621	0.251	156.01
	Ours	**20.02**	**0.532**	**0.351**	**0.071**	**1.629**	**0.168**	**0.594**	**0.244**	**9.12**
ScanNetV2	ViewCrafter	12.82	0.654	0.599	0.068	2.861	0.101	0.658	0.126	515.81
	NeoVerse	10.53	0.422	0.671	0.010	0.804	0.022	**0.494**	0.100	605.04
	GEN3C	12.54	0.548	0.567	0.058	2.478	0.071	0.665	0.169	1257.07
	GLD	15.33	0.664	0.491	0.011	0.396	0.019	0.598	0.108	476.27
	Ours	**19.65**	**0.781**	**0.398**	**0.009**	**0.233**	**0.016**	0.589	**0.088**	**33.11**

allocation is empirically grounded in Table 2 and Fig. 4: three Level-1 steps recover noticeably finer details than a single step, whereas a single L0 cascade step suffices to match the quality of 49 steps. Consequently, distillation is applied solely to the Level-1 denoiser, leaving the cascade and decoding heads unchanged.

Level 1 Reflow Distillation. To achieve Level-1 generation within just three steps, we initialize a student model from the multi-step teacher and perform piecewise reflow (Yan et al., 2024) over three time intervals bounded by $1 = \tau_0 > \tau_1 > \tau_2 > \tau_3 = 0$. For notation brevity, the conditioning inputs from Eq. (2) remain fixed throughout each trajectory and are omitted below. For a given interval a, the starting state is constructed by corrupting clean ground-truth features $\mathbf{F}^1$ with Gaussian noise: $\mathbf{X}_a^+ = (1-\tau_a)\mathbf{F}^1 + \tau_a\boldsymbol{\epsilon}$. The frozen teacher then integrates the ODE from τ_a to τ_{a+1} to obtain the target endpoint $\mathbf{X}_a^-$. The effective velocity vector connecting these endpoints is computed as $\mathbf{u}_a^{\mathrm{T}} = (\mathbf{X}_a^+ - \mathbf{X}_a^-)/(\tau_a - \tau_{a+1})$. By learning to match these straight-line endpoint displacements, the student effectively shortcuts multiple teacher evaluations into a single update per interval.

For any timestep $t \sim \mathcal{U}(\tau_{a+1}, \tau_a)$, we form the linearly interpolated state $\widetilde{\mathbf{X}}_t = \eta\mathbf{X}_a^+ + (1-\eta)\mathbf{X}_a^-$, where $\eta = (t - \tau_{a+1})/(\tau_a - \tau_{a+1})$. The student learns this constant reflow velocity via

$$\mathcal{L}^1_{\mathrm{PRF}} = \mathbb{E}\Big[\sum_{j\in\mathcal{C}\cup\mathcal{T}} \alpha_j \big\| v_{\theta_S}(\widetilde{\mathbf{X}}_t, t)_j - \mathbf{u}^{\mathrm{T}}_{a,j}\big\|_2^2\Big], \tag{9}$$

where the expectation is taken over training samples, noise vectors, intervals, and timesteps, and α_j retains the view-dependent weights from Sec. 3.1. To stabilize optimization, we regularize the student with the original flow-matching objective: $\mathcal{L}^1_{\mathrm{fast}} = \rho\mathcal{L}^1_{\mathrm{PRF}} + (1-\rho)\mathcal{L}^1_{\mathrm{FM}}$. During inference, a single Euler step per interval traverses the trajectory before the cascade recovers Level-0 features.

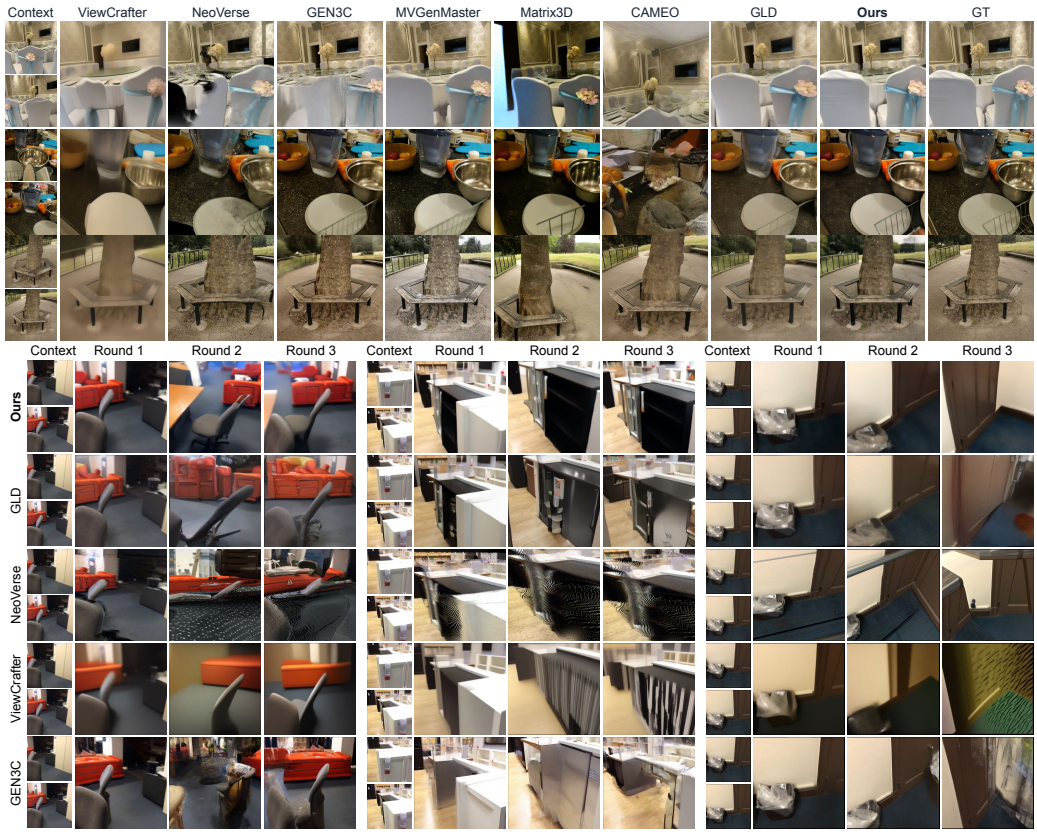


Figure 3: **Qualitative comparisons across multiple datasets.** Top: comparisons of results from a single inference pass. Bottom: comparisons of long-sequence generation over three rounds.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We evaluate on two in-domain benchmarks, RealEstate10K (Zhou et al., 2018) and DL3DV (Ling et al., 2024), and two out-of-domain benchmarks, Mip-NeRF 360 (Barron et al., 2022) and ScanNetv2 (Dai et al., 2017), with ScanNetv2 used for long-sequence evaluation. Specifically, each generation round uses $N_{\mathcal{C}} = 2$ context views and $N_{\mathcal{T}} = 6$ target views. We report PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) for visual quality, ATE and relative pose errors ($\mathrm{RPE_r}$/$\mathrm{RPE_t}$) (Sturm et al., 2012) from VGGT-estimated poses (Wang et al., 2025a) for target-camera fidelity, reprojection error and MEt3R (Asim et al., 2025) for cross-view geometric and feature consistency, and average inference time for efficiency.

Baselines. We compare with ViewCrafter (Yu et al., 2024a), NeoVerse (Yang et al., 2026), GEN3C (Ren et al., 2025), MVGenMaster (Cao et al., 2025), Matrix3D (Lu et al., 2025), CAMEO (Kwon et al., 2025), and GLD (Jang et al., 2026), covering diverse approaches to generative novel-view synthesis. For long-sequence evaluation, we compare with ViewCrafter, NeoVerse, GEN3C, and GLD across successive rounds to assess visual quality and cross-round consistency.

Implementation Details. Starting from a pretrained GLD (Jang et al., 2026), GeoVerse is trained on 15 real and synthetic multi-view datasets using DA3-Base (Lin et al., 2025) at 504×504 resolution, with eight ordered views per sample (one to four selected as context views). Training proceeds in two stages: a 300k-iteration adaptation phase with a learning rate of 3×10^{-5}, followed by up to 50k iterations of reflow distillation at 1×10^{-5}. Both stages are executed on 32 NVIDIA A800 GPUs using AdamW (Loshchilov & Hutter, 2017) with a global batch size of 32. Adaptation hyperparameters are set to $(\beta_1, \beta_2) = (0.9, 0.95)$, gradient-norm clipping at 1.0, and an EMA decay of 0.9995. Further details are provided in Appendix D.

Table 2: **Quantitative comparisons on DL3DV (Ling et al., 2024) with different denoising budgets.** Three L1 updates and one Cascade L0 update provide the best overall trade-off between synthesis quality and inference speed among the compared configurations.

Steps	L1 sweep (Cascade L0 fixed to 1)						Cascade L0 sweep (L1 fixed to 3)					
	Time (s)↓	PSNR↑	SSIM↑	LPIPS↓	ATE↓	MEt3R↓	Time (s)↓	PSNR↑	SSIM↑	LPIPS↓	ATE↓	MEt3R↓
49	80.04	17.72	0.535	0.376	0.037	0.290	52.55	19.48	0.597	0.342	0.025	0.260
24	41.33	17.87	0.541	0.370	0.035	0.286	29.88	19.50	0.598	0.342	0.027	0.261
12	23.02	18.15	0.550	0.362	0.033	0.282	19.29	19.53	0.599	0.342	0.029	0.261
6	13.84	18.61	0.567	0.349	0.032	0.272	13.75	19.60	0.600	0.342	0.031	0.271
3	**9.24**	**19.61**	**0.598**	**0.339**	**0.028**	**0.261**	11.02	19.66	0.601	0.341	0.029	0.261
1	6.07	20.55	0.630	0.367	0.028	0.269	**9.24**	**19.61**	**0.598**	**0.339**	**0.028**	**0.261**



Figure 4: **Qualitative comparisons on DL3DV (Ling et al., 2024) with different denoising budgets.** One L1 update produces blurrier fine details than three L1 updates, while one Cascade L0 update preserves detail comparable to 49 Cascade L0 updates.

4.2 Comparison

Visual Quality. GeoVerse achieves state-of-the-art performance across all four benchmarks in Tab. 1, surpassing GLD (Jang et al., 2026) in PSNR by 2.23 dB on DL3DV and 2.45 dB on Mip-NeRF 360. These gains translate into clear visual improvements (Fig. 3), effectively suppressing smearing and ghosting while recovering crisp object boundaries and faithful surface colors relative to context views. Unlike baselines that introduce misplaced structures or color shifts, GeoVerse preserves scene identity during synthesis. On long-sequence ScanNetv2 testing across three rounds, GeoVerse boosts PSNR from 15.33 dB (GLD) to 19.65 dB, maintaining temporal stability across large indoor structures.

Geometric Consistency. GeoVerse lowers ATE on Mip-NeRF 360 from 0.105 (GLD (Jang et al., 2026)) to 0.071, representing a 32.4% improvement. As reflected in Tab. 1 and Fig. 3, lower pose errors manifest as enhanced 3D structural fidelity, preserving intricate geometry like outdoor furniture without distortion. Over long sequences, key scene elements retain cross-round consistency, avoiding the structural repetitions and layout shifts prevalent in baselines. This validates the role of persistent spatial memory in anchoring new updates to accumulated context. While performance varies slightly across specific metrics, where some baselines achieve lower reprojection or translation errors on individual datasets, GeoVerse maintains significantly better global scene coherence.

Inference Efficiency. GeoVerse synthesizes six target views in roughly nine seconds across the three main benchmarks, operating about $18\times$ faster than GLD (Jang et al., 2026). This acceleration stems from using only four geometric denoising steps paired with single-pass video feature extraction. Even under this constrained sampling budget, generated outputs maintain crisp object contours and recognizable 3D geometry. This efficiency advantage naturally carries over to sequential view expansion, as each additional round employs the same compact pipeline. Per-pass runtimes for the main benchmarks and cumulative three-round times on ScanNetv2 are summarized in Tab. 1.

4.3 Few-Step Analysis

We analyze the denoising budget within GLD's hierarchical sampling framework (Jang et al., 2026), with Level-1 acceleration based on reflow (Yan et al., 2024). Table 2 evaluates denoising budget

Table 3: **Ablation study on ScanNetV2 (Dai et al., 2017).** We evaluate the contributions of key components, training losses, and alternative architectures for video-prior injection.

Variant	2D Metrics			3D Metrics					Scaled time
	PSNR↑	SSIM↑	LPIPS↓	ATE↓	RPE_r ↓	RPE_t ↓	Reproj.↓	MEt3R↓	Reference (s)↓
GeoVerse	19.65	0.781	0.398	0.009	0.223	0.016	0.589	0.088	33.11
w/o Wan2.2 prior	16.83	0.696	0.463	0.022	0.388	0.034	0.591	0.092	29.19
w/o spatial memory	17.87	0.727	0.434	0.014	0.268	0.023	0.591	0.087	33.48
MoT video-KV	19.24	0.769	0.403	0.009	0.238	0.016	0.587	0.089	36.29
w/o RGB loss	18.58	0.756	0.415	0.014	0.360	0.023	0.581	0.091	33.21
w/o depth loss	19.05	0.764	0.413	0.010	0.282	0.015	0.588	0.092	33.29

trade offs on DL3DV (Ling et al., 2024). Reducing Level-1 generation from three steps to one lowers runtime from 9.24 to 6.07 s, yet worsens LPIPS (Zhang et al., 2018) from 0.339 to 0.367. As shown in Fig. 4, this setting causes a loss of fine brick textures and object boundaries despite higher PSNR, showing that PSNR alone fails to reflect perceptual quality. On the other hand, cutting Level 0 cascade steps from 49 to one preserves comparable local details with an LPIPS of 0.342 versus 0.339 at much lower latency. These complementary trends validate our $3 + 1$ budget allocation.

4.4 Ablation Studies

Table 3 summarizes diagnostic ablations on ScanNetv2 Dai et al. (2017) across model components and supervision losses. (1) Removing the Wan2.2 Wan et al. (2025) video prior causes the most severe visual degradation, proving that video appearance priors are essential for synthesizing faithful content. (2) Disabling spatial memory substantially degrades both visual fidelity and cross view alignment, highlighting the importance of accumulated scene context for multi round consistency. (3) The Mixture of Transformers variant explores key value attention as an alternative to additive residual injection, though isolating its relative advantages requires comparison under matched training settings. (4) Omitting RGB supervision leads to a moderate decline in visual quality, showing that direct appearance supervision complements the video prior. (5) Removing depth supervision weakens overall geometric consistency and confirms the value of explicit structural constraints, though non uniform metric variations indicate nuanced geometric trade offs.

5 Limitations

While GeoVerse achieves strong multi-view consistency and high inference efficiency, two main limitations remain. First, its peak visual fidelity is constrained by the appearance capacity of the DA3 feature space. By prioritizing speed through compact feature conditioning, GeoVerse may not fully match the fine texture richness of computationally heavy, full-scale video diffusion models. Second, multi-round consistency depends on both the underlying geometry backbone and synthesized RGB coherence. In textureless or complex regions, initial depth errors and generated appearance drift can accumulate through spatial memory updates, occasionally compromising cross-view alignment across long trajectories. Improving joint 3D geometric and photometric stability remains a key objective for future research.

6 Conclusion

We introduced GeoVerse, a framework designed to resolve the fundamental trade off between geometric consistency and generative completion in novel view synthesis. By embedding pretrained video appearance priors into a 3D geometric latent diffusion model, GeoVerse generates high fidelity views while maintaining rigid spatial structure. Our global spatial memory further prevents cumulative drift across extended trajectories by grounding ongoing synthesis in accumulated 3D scene context. Accelerated by a few step reflow distillation scheme, GeoVerse achieves over 17 times speedup compared to GLD while setting new state of the art benchmarks in visual quality and pose accuracy. By demonstrating the effectiveness of combining video learned priors with geometric latent spaces, GeoVerse establishes a robust foundation and offers valuable insights toward building multi view consistent video world models.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation. We have not used generative AI tools for designing research methods and experiments, implementing methodologies, interpreting results, proposing or refining hypotheses, cleaning and reformatting datasets, or supporting qualitative and thematic data analysis, and generating synthetic datasets, proposing mathematical claims, providing key elements for proving mathematical claims, and assisting in writing proofs are not applicable to this work. Additionally, we used generative AI tools for summarizing or analyzing existing literature, and editing the manuscript to enhance readability. We have reviewed all AI-assisted work: translated or polished text was manually cross-checked sentence-by-sentence to ensure that the original intent remained uncompromised. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision*, pp. 690–708. Springer, 2022.

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *CVPR*, 2025.

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.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, pp. 5470–5479, 2022.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

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

Chenjie Cao, Chaohui Yu, Shang Liu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Mvgenmaster: Scaling multi-view generation from any image via 3d priors enhanced diffusion model. In *CVPR*, pp. 6045–6056, 2025.

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

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pp. 5828–5839, 2017.

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

Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2821–2830. IEEE, 2018.

Wooseok Jang, Seonghu Jeon, Jisang Han, Jinhyeok Choi, Minkyung Kwon, Seungryong Kim, Saining Xie, and Sainan Liu. Repurposing geometric foundation models for multi-view diffusion. *arXiv preprint arXiv:2603.22275*, 2026.

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, 2025a.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202. IEEE, 2025b.

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.

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

Minkyung Kwon, Jinhyeok Choi, Jiho Park, Seonghu Jeon, Jinhyuk Jang, Junyoung Seo, Minseop Kwak, Jin-Hwa Kim, and Seungryong Kim. Cameo: Correspondence-attention alignment for multi-view diffusion models. *arXiv preprint arXiv:2512.03045*, 2025.

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

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 *CVPR*, pp. 22160–22169, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20654–20664. IEEE, 2024.

Yuanxun Lu, Jingyang Zhang, Tian Fang, Jean-Daniel Nahmias, Yanghai Tsin, Long Quan, Xun Cao, Yao Yao, and Shiwei Li. Matrix3d: Large photogrammetry model all-in-one. In *CVPR*, 2025.

Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4981–4991. IEEE, 2023.

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.

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.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20076–20086. IEEE, 2023.

Kerui Ren, Lihan Jiang, Tao Lu, Mulin Yu, Linning Xu, Zhangkai Ni, and Bo Dai. Octree-gs: Towards consistent real-time rendering with lod-structured 3d gaussians. *arXiv preprint arXiv:2403.17898*, 2024.

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 *CVPR*, 2025.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, Sanja Fidler, Jiahui Huang, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra 2.0: Explorable generative 3D worlds. *arXiv preprint arXiv:2604.13036*, 2026. URL https://arxiv.org/abs/2604.13036.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 573–580. IEEE, 2012.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2443–2451. IEEE, 2020.

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

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, pp. 5294–5306, 2025a.

Kaixuan Wang and Shaojie Shen. Flow-motion and depth network for monocular stereo and beyond. *IEEE Robotics and Automation Letters*, 5(2):3307–3314, 2020.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, pp. 20697–20709, 2024.

Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y Chen, Yuqing Yang, and Bohan Zhuang. Latent spatial memory for video world models. *arXiv preprint arXiv:2606.09828*, 2026.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *IROS*, 2020.

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. *arXiv preprint arXiv:2507.13347*, 2025b.

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.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *Advances in Neural Information Processing Systems*, 38:49371–49393, 2026.

Hanshu Yan, Xingchao Liu, Jiachun Pan, Jun Hao Liew, Qiang Liu, and Jiashi Feng. Perflow: Piecewise rectified flow as universal plug-and-play accelerator. *Advances in Neural Information Processing Systems*, 37:78630–78652, 2024.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. *arXiv preprint arXiv:2601.00393*, 2026.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22. IEEE, 2023.

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*, 2024a.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19447–19456. IEEE, 2024b.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026. URL https://arxiv.org/abs/2603.16666.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM TOG*, 37, 2018. URL https://arxiv.org/abs/1805.09817.

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, et al. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. *arXiv preprint arXiv:2509.12201*, 2025.

This supplementary material provides additional technical details and experimental results to complement the main paper. Section A introduces Wan2.2 VACE, including its inference pipeline, pretraining scale, and use as a frozen feature extractor in GeoVerse. Section B defines the training losses and details their computation, including high-frequency weighting and depth-scale alignment. Section C describes the key/value-based MoT variant used in the architectural comparison. Section D summarizes the training data, with dataset types, scene counts, and image counts. Section E presents additional qualitative comparisons for novel-view synthesis and successive view expansion.

A WAN2.2 VACE DETAILS

Video generation and conditioning. Wan is a video diffusion framework with a causal video VAE, a text encoder, and a diffusion Transformer (Wan et al., 2025). In the standard generation pipeline, the Transformer iteratively denoises a video latent under text and optional visual conditions, and the VAE decoder converts the final latent into RGB frames. Visual inputs requiring latent encoding are processed by the VAE encoder before entering the corresponding conditioning path. Wan2.2-A14B uses separate high- and low-noise experts for different portions of the denoising trajectory, with approximately 14 billion active parameters per step.

VACE (Jiang et al., 2025b) adds a unified interface for reference images, source videos, and spatiotemporal masks. Its Video Condition Unit organizes these inputs for reference-guided generation, video editing, inpainting, outpainting, and temporal extension. The standard VACE pipeline tokenizes visual conditions through context encoding and introduces them into the video backbone through context adapters. Its training curriculum progresses from foundational completion tasks to multiple references and task combinations, followed by quality refinement. These components provide visual completion cues that complement the geometric representation used by GeoVerse.

Pretraining data scale. The Wan technical report describes pretraining on billions of images and videos (Wan et al., 2025). The Wan2.2 release reports 65.6% more images and 83.2% more videos than Wan2.1; these are relative increases, rather than disclosed absolute counts for the Wan2.2 corpus. VACE constructs task-specific conditions from filtered videos, including segmentation masks, reference crops, depth, pose, and optical flow (Jiang et al., 2025b). The available model card does not specify an absolute training-set size for the particular Wan2.2-VACE-Fun checkpoint used here. These external pretraining corpora are separate from GeoVerse's multi-view training data in Sec. D.

Feature extraction in GeoVerse. Our checkpoint is derived from the low-noise component of Wan2.2-VACE-Fun-A14B (Wan et al., 2025). We retain backbone blocks 0–15 and the corresponding VACE blocks, extracting 5,120-dimensional features at blocks 0, 5, 10, and 15. The frozen extractor runs once per multi-view prediction, and the geometric denoiser reuses these features across its denoising steps.

The two input branches share an RGB proxy assembled from observed context images and projected target images. After normalization and resizing, Gaussian perturbations with standard deviation 0.1 are added only to valid target projections; context RGB remains unchanged and invalid target regions are zero-filled. Target ground-truth RGB is not used to construct this proxy.

For the Wan backbone, the frozen video VAE encodes the proxy into a 16-channel latent $\mathbf{z}_{\text{proxy}} = E_{\text{VAE}}(\widetilde{\mathbf{I}})$, using the checkpoint's latent normalization. We then form $\mathbf{x}_t = (1-\sigma_t)\mathbf{z}_{\text{proxy}} + \sigma_t\boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, and apply the backbone's patch embedding. The extraction timestep is fixed at $t = 500$, and its noise level σ_t is obtained from the corresponding scheduler mapping. This latent-space perturbation is distinct from the 0.1 RGB perturbation above and is applied across the latent, rather than only in invalid regions.

For the VACE branch, we convert visibility $\mathbf{V}$ into the generation mask $\mathbf{m} = 1 - \mathbf{V}$. Following VACE's condition encoding (Jiang et al., 2025b), the inactive and reactive RGB inputs, $\widetilde{\mathbf{I}} \odot (1-\mathbf{m})$ and $\widetilde{\mathbf{I}} \odot \mathbf{m}$, are separately VAE-encoded into two 16-channel latents. Each spatial 8×8 mask block is rearranged into 64 channels and temporally aligned to the latent grid. Concatenating the two latents and the packed mask produces the 96-channel VACE condition; the mask itself is not VAE-encoded. The VACE blocks combine the encoded condition with backbone tokens and supply control residuals to the corresponding Wan blocks. Unlike the backbone input, the condition latents

receive no additional timestep-dependent diffusion noise. This single-pass feature extraction does not run a complete video-sampling trajectory or invoke the VAE decoder; its features are reused by the ControlNet-style adapters (Zhang et al., 2023) in the geometric denoiser.

B Loss Definitions and Computation

We expand the five loss terms in Eq. (8). The formulas below describe a single multi-view training sample; losses are averaged over the mini-batch. Latent flow matching supervises context and target views, whereas decoded RGB and depth losses supervise target views only.

Two-level flow matching. For feature level $i \in \{0, 1\}$, we sample $t \sim \mathcal{U}(0, 1)$ and Gaussian noise with the same shape as the clean reference features $\mathbf{F}^i$. The noisy state is $\mathbf{X}_t^i = (1-t)\mathbf{F}^i + t\boldsymbol{\epsilon}$, and its velocity target is $\boldsymbol{\epsilon} - \mathbf{F}^i$. Writing $\hat{\mathbf{u}}_{t,j}^i$ for the predicted velocity of view j, we obtain

$$\mathcal{L}_{\mathrm{FM}}^i = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[\sum_{j \in \mathcal{C}\cup\mathcal{T}} \alpha_j \left\|\hat{\mathbf{u}}_{t,j}^i - (\boldsymbol{\epsilon}_j - \mathbf{F}_j^i)\right\|_2^2\right], \qquad i \in \{0, 1\}. \tag{10}$$

The squared norm sums errors over the feature grid and channels of each view. We use $\alpha_j = 0.25$ for context views and 1 for target views. At Level 1, the velocity is predicted by v_θ under the conditions in Eq. (2); at Level 0, it is predicted by the denoiser underlying the conditional cascade in Eq. (3).

RGB supervision. Let Ω_j^{rgb} denote the pixels with valid reference RGB in target view j, and define the channel-averaged error $e_j(p) = \|\hat{\mathbf{I}}_j(p) - \mathbf{I}_j(p)\|_1/3$. The reconstruction loss is

$$\mathcal{L}_{\mathrm{rgb}} = \frac{1}{N_{\mathrm{rgb}}} \sum_{j \in \mathcal{T}} \sum_{p \in \Omega_j^{\mathrm{rgb}}} e_j(p). \tag{11}$$

Here, $N_{\mathrm{rgb}} = \max(1, \sum_{j\in\mathcal{T}} |\Omega_j^{\mathrm{rgb}}|)$ is the valid-pixel count clamped to at least one. Predictions and references use the same RGB normalization. These supervision pixels are distinct from the projection-visibility mask $\mathbf{V}$: valid reference pixels remain supervised even when they are not visible in the memory projection.

To further emphasize image boundaries and texture, we weight the same RGB error using reference-image gradients. We first convert reference RGB to grayscale and compute its gradient magnitude $g_j(p)$ using horizontal and vertical 3×3 Sobel filters. Let $q_{0.50}$ and $q_{0.98}$ be the corresponding gradient-magnitude quantiles over valid target pixels in the sample. We form $w_j(p) = \mathrm{clip}_{[0,1]}((g_j(p) - q_{0.50})/(q_{0.98} - q_{0.50} + \delta))$, with $\delta = 10^{-6}$, and compute

$$\mathcal{L}_{\mathrm{hf}} = \frac{1}{Z_{\mathrm{hf}}} \sum_{j \in \mathcal{T}} \sum_{p \in \Omega_j^{\mathrm{rgb}}} w_j(p) e_j(p). \tag{12}$$

The normalizer is $Z_{\mathrm{hf}} = \max(\delta, \sum_{j\in\mathcal{T}} \sum_{p\in\Omega_j^{\mathrm{rgb}}} w_j(p))$. The weights are computed from the reference images, so the model cannot reduce this term by suppressing its own image gradients. The loss is zero when all weights are zero.

Scale-aligned depth supervision. Following the camera-center alignment described in the main paper, let $\hat{\mathbf{o}}_j$ and $\mathbf{o}_j$ denote predicted and reference camera centers for views with valid camera estimates, indexed by $\mathcal{J} \subseteq \mathcal{C} \cup \mathcal{T}$. We fit a single similarity transform per sample:

$$(s^\star, \mathbf{Q}^\star, \mathbf{b}^\star) = \underset{s>0,\, \mathbf{Q}\in\mathrm{SO}(3),\, \mathbf{b}\in\mathbb{R}^3}{\arg\min} \sum_{j\in\mathcal{J}} \|s\mathbf{Q}\hat{\mathbf{o}}_j + \mathbf{b} - \mathbf{o}_j\|_2^2 . \tag{13}$$

We solve this least-squares problem by centering the camera centers and applying singular value decomposition to their cross-covariance. Only the recovered scale $s^\star$ is applied to camera-space depth; rotation and translation align the camera coordinate systems and do not enter the depth residual. The

Table 4: **Training datasets used by GeoVerse.** We summarize the data types, scene counts, and image counts of the real and synthetic multi-view datasets in our training mixture.

Dataset	Data type	#Scenes	#Images
Aria Digital Twin (Pan et al., 2023)	Real / digital twin	188	232,115
ARKitScenes (Baruch et al., 2021)	Real / RGB-D	644	125,134
DL3DV (Ling et al., 2024)	Real / multi-view video	10,475	3,594,809
MapFree (Arnold et al., 2022)	Real / multi-view video	460	515,113
ScanNet++ v2 (Yeshwanth et al., 2023)	Real / RGB + scans	954	1,032,198
Waymo (Sun et al., 2020)	Real / driving	3,990	790,405
RealEstate10K (Zhou et al., 2018)	Real / real-estate video	66,033	8,832,823
GTA-SfM (Wang & Shen, 2020)	Synthetic / outdoor	200	17,649
Hypersim (Roberts et al., 2021)	Synthetic / indoor	107	17,348
MVSSynth (Huang et al., 2018)	Synthetic / driving	120	12,000
OmniWorld (Zhou et al., 2025)	Synthetic / diverse scenes	4,486	794,596
TartanAir (Wang et al., 2020)	Synthetic / diverse scenes	18	613,274
TartanAir v2 (Wang et al., 2020)	Synthetic / diverse scenes	74	7,135,955
Virtual KITTI (Cabon et al., 2020)	Synthetic / driving	100	42,520
Spring (Mehl et al., 2023)	Synthetic / dynamic scenes	37	5,000
Total	15 datasets	87,886	23,760,939

same scale is shared by all target views, rather than fitted separately to each depth map. With $\Omega_j^{\mathrm{depth}}$ denoting pixels whose reference depth is finite and positive, the loss is

$$\mathcal{L}_{\mathrm{depth}} = \frac{1}{N_{\mathrm{depth}}} \sum_{j \in \mathcal{T}} \sum_{p \in \Omega_j^{\mathrm{depth}}} \left| s^\star \hat{\mathbf{D}}_j(p) - \mathbf{D}_j(p) \right|. \tag{14}$$

Here, $N_{\mathrm{depth}} = \max(1, \sum_{j \in \mathcal{T}} |\Omega_j^{\mathrm{depth}}|)$ normalizes by the valid depth-pixel count.

In each training iteration, we compute the latent velocity residuals, decode target RGB and depth from the predicted feature hierarchy, construct the RGB gradient weights and camera-based scale, and combine the resulting terms using the coefficients in Eq. (8). The reflow objective in Sec. 3.3 is a separate distillation objective.

C MoT Variant for Video-Prior Injection

Shared-attention conditioning. Following the shared-attention design of Fast-WAM (Yuan et al., 2026), we couple video-prior tokens and geometric latent tokens through attention while retaining separate processing streams. At each coupled layer, modality-specific projections map the two representations into a common attention space. The geometric stream produces queries, keys, and values from its current denoising state, while the video stream supplies visual keys and values. We concatenate the video and geometric keys and values along the token dimension, allowing each geometric query to attend jointly to scene features and video priors. The resulting attention output passes through the geometric stream's output projection, residual connection, and feed-forward layers to update its tokens.

Feature reuse during denoising. The interaction is asymmetric: geometric tokens attend to the video stream, while video features remain independent of the evolving geometric state. We therefore compute and cache the video keys and values once, reusing them across denoising updates. Geometric queries, keys, and values are recomputed at each step. This variant introduces video information within the attention computation, whereas the ControlNet-style design injects it through an auxiliary residual branch.

D Training Data and Statistics

Dataset composition. Our training mixture contains 15 datasets spanning real indoor captures, outdoor videos, driving sequences, and synthetic environments. TartanAir and TartanAir v2 are

listed separately as distinct sampling sources. Table 4 follows the dataset/type/count presentation of DA3 (Lin et al., 2025), but reports our own training-manifest and RGB-inventory statistics rather than DA3's differently selected subsets. MegaDepth is excluded from this mixture.

Sampling and geometric inputs. We sample dataset sources with probabilities proportional to the square root of their RGB frame counts, balancing coverage against the dominance of the largest collections. RealEstate10K and Waymo use Pi3-estimated poses and depth (Wang et al., 2025b). For the remaining sources, available dataset geometry and Pi3-estimated geometry are selected at a nominal ratio of $2:1$; samples with missing or incomplete depth fall back to Pi3. Samples that fail loading or validity checks are resampled. The inventory counts therefore describe the available training pool, not the number of distinct images consumed by a completed run.

E Additional Qualitative Results

E.1 Novel-View Synthesis

Figure 5 extends the main-paper comparison with two target views from each of three scenes from RE10K datasets and DL3DV datasets.

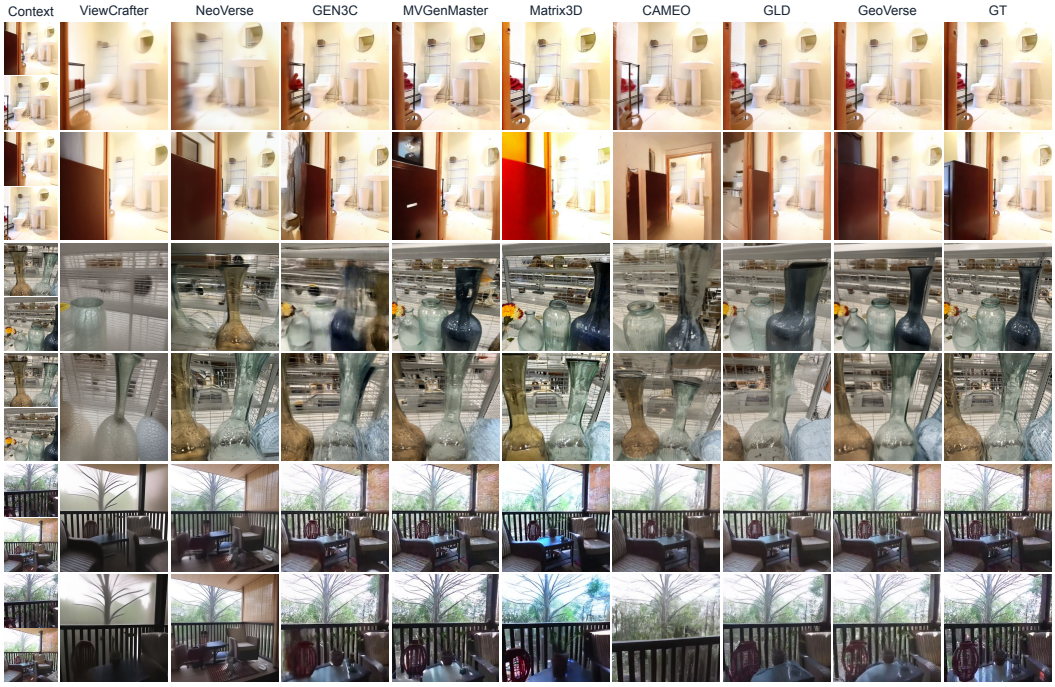


Figure 5: **Additional qualitative comparisons on RE10K and DL3DV.** Two target views are shown for each of three scenes, alongside context images and ground truth.

E.2 SUCCESSIVE VIEW EXPANSION

Figure 6 presents three successive generation rounds on two ScanNetv2 scenes. We display target frames 1, 3, and 5 from each round to illustrate how scene appearance evolves as viewpoints expand.

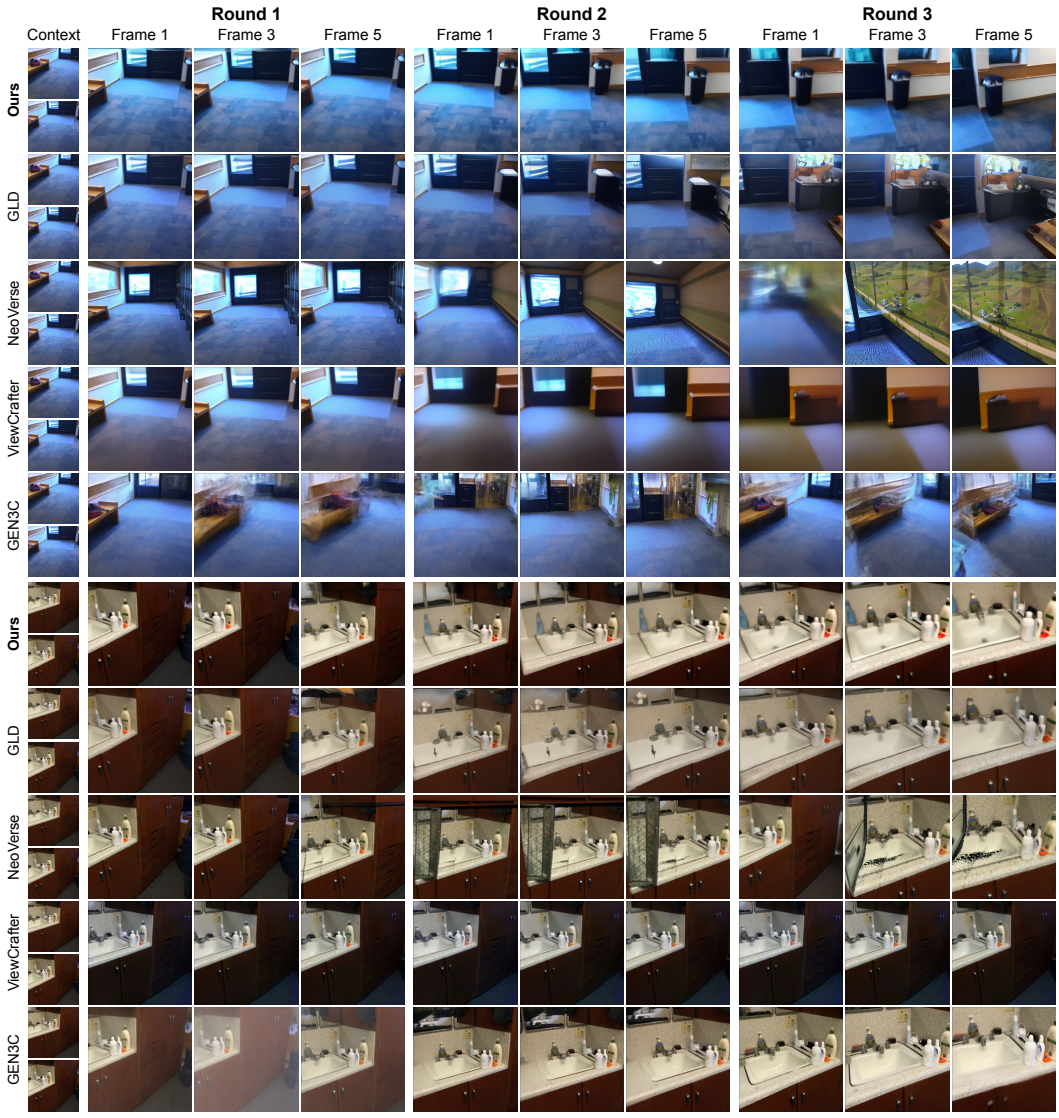


Figure 6: **Additional long-sequence comparisons on ScanNetv2.** Each row contains two context images followed by target frames 1, 3, and 5 from each of three rounds.