SEG: LEVERAGING SELF-EXPANDING GEOMETRIC STATE FOR SUCCESSIVE VIEW SYNTHESIS

Anonymous authors
Paper under double-blind review

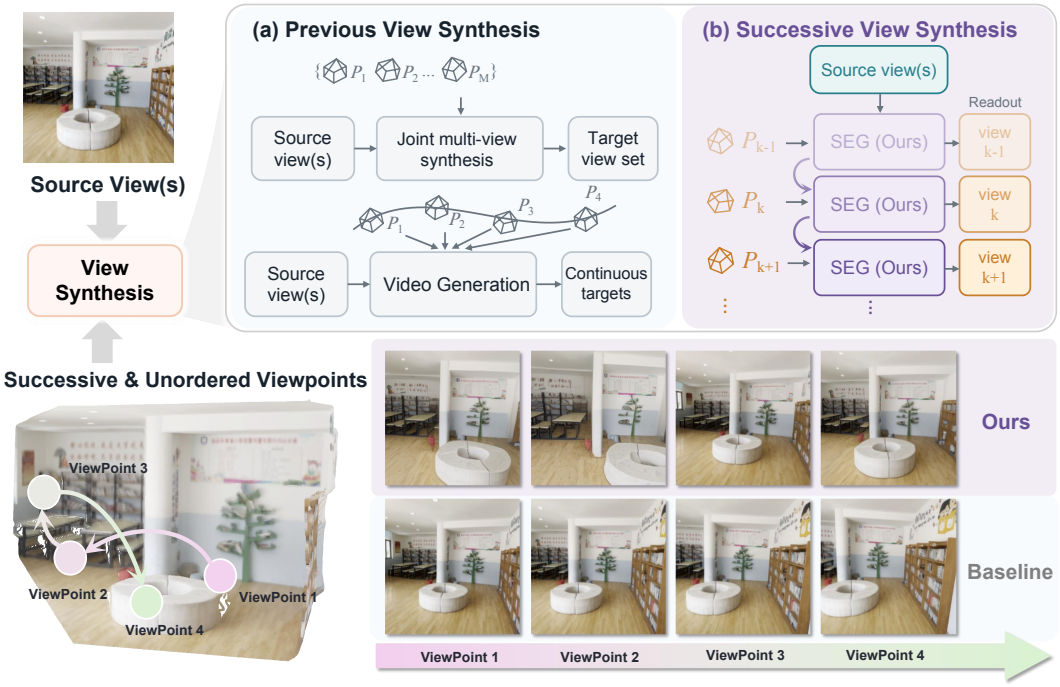

Figure 1: **Successive view synthesis with SEG.** Starting from sparse source views, SEG supports successiveviewpoint queries in arbitrary spatial order through expanding geometric state, which retains historical context for subsequent synthesis.

ABSTRACT

Recent advances in both reconstruction and multi-view diffusion have greatly improved the quality and geometric consistency of novel view synthesis (NVS). However, most existing methods either adopt joint multi-view synthesis for predetermined target views, or rely on video generation models for continuous synthesis, which limits their ability in successive view synthesis where incremental and unordered view queries are required. To address this limitation, we present **SEG**, a successive view synthesis framework built around a Self-Expanding Geometric state. The key insight of SEG is to turn context-dependent geometric latent into self-expanding states with consistency during expansion and causality under successive queries. Therefore, we construct our SEG state from early geometric encoder level based on consistency analysis, and design causal-prefix ground-truth state for supervision. We pair these latents with KV reference memory to form the final SEG state for efficient successive queries. Based on SEG state, we design our pipeline with a causal geometric diffusion model for state synthesis and expanding, and introduce a geometric adapter to map generated latents to multi-level features for RGB readout. We show that SEG state provides a well-suited solution for successive NVS, including extensive experiments on RealEstate10K and DL3DV with strong view synthesis quality, viewpoint accuracy, and computational efficiency.

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.

1 INTRODUCTION

Novel view synthesis (NVS) makes it possible to explore a scene from viewpoints beyond those captured by the input images (Mildenhall et al., 2020; Kerbl et al., 2023; Jin et al., 2025). Recently, advances in latent diffusion and representation-based generation (Rombach et al., 2022; Zheng et al., 2026) have improved the quality and efficiency of image synthesis, with tremendous progress in both joint multi-view synthesis (Gao et al., 2024; Zhou et al., 2025; Jang et al., 2026; Çelen et al., 2026) and video generation based view synthesis (Huang et al., 2026; Ren et al., 2025). As illustrated in Figure 1, these methods typically synthesize a fixed set of target views known in advance simultaneously, or use continuous target queries for video generation.

However, real-world interactions such as virtual tours and scene inspection turn NVS into a successive exploration process: users may reveal any unseen regions unexpectedly instead of following a predetermined path, and jump between unordered viewpoints. We refer to this setting as *successive view synthesis* containing two core properties: **incremental queries** which call for efficient accumulation and reuse of generated context, and **unordered viewpoints** which require spatial reasoning across potentially discontinuous and unordered camera poses. Existing joint multi-view formulation does not accumulate generated history and support incremental queries, while video-based synthesis typically organizes generation around continuous camera trajectories or clips. Although autoregressive pipeline supports incremental queries (Kong et al., 2025), the natural 2D representation still leaves it hard to jointly encode geometric features and precisely follow viewpoint queries (Jang et al., 2026). The central question still leaves: Can we efficiently reuse synthesized context to expand a geometrical-guaranteed state for unordered and incremental queries?

In this paper, we address this question by first constructing a **Self-Expanding Geometric (SEG) state** as the core representation of our successive view synthesis framework. Although it seems natural to apply geometric features as our core state from frozen foundation models (Wang et al., 2025a; Lin et al., 2025), we find that ideal reusable geometric state for successive NVS still demands two extra properties: consistent representation under expansion, and causal target construction under successive queries. Therefore, we first analyze geometric feature stability under the expansion setting, and select latent from early geometric encoder level that is most stable to newly added views. To ensure causal conditioning, we then construct each target ground-truth state from its own available image prefix to prevent knowledge from future views. In order for causal conditioning queries, we pair each geometric latent with KV reference memory, which groups as our final state. Each completed prediction extends both components by appending new entries while preserving earlier ones, allowing generated content to accumulate as reusable geometric context.

Based on SEG state, we further design our full SEG pipeline. After the initialization of source views, we use a causal geometric diffusion model that efficiently generates one view at a time via KV reference memory from SEG state. A clean reference pass through the shared backbone converts the completed prediction into new reusable KV memory for subsequent queries, allowing historical features to be reused. The model is trained with causal-prefix supervision, which uses the aforementioned ground-truth state. During RGB readout period, we notice that there still exists a gap between successively generated latents and the ground-truth features, so we design a geometric adapter that maps the archived latents to multi-level features, which are finally used for ViT-based RGB readout.

We extensively evaluate our method on real-world view synthesis datasets including RealEstate10K (Zhou et al., 2018) and DL3DV (Ling et al., 2024). Through comparisons with SOTA baselines (Jin et al., 2025; Jang et al., 2026; Ren et al., 2025; Huang et al., 2026; Çelen et al., 2026), we find that our method achieves superior performance in terms of 2D view synthesis quality, viewpoint accuracy, and computational efficiency. We also conduct ablation studies to validate the effectiveness of our SEG and causal designs.

2 RELATED WORK

Reconstruction-based View Synthesis. Reconstruction-based novel view synthesis answers camera queries through a scene representation estimated from observed images. NeRF (Mildenhall et al., 2020) and 3D Gaussian Splatting (Kerbl et al., 2023) optimizes implicit radiance field and explicit Gaussian primitives for NVS view rendering. Many following generalizable approaches (Yu et al., 2021; Charatan et al., 2024; Fei et al., 2024; Chen et al., 2024; Xu et al., 2025) infer NeRF or

3DGS representations and performing NVS via a feed-forward network. More recently, LVSM (Jin et al., 2025) together with following works (Szymanowicz et al., 2026) use implicit queries for NVS, and feed-forward geometry methods such as CUT3R (Wang et al., 2025b) also support the implicit query paradigm. These approaches make scene information reusable across camera queries or observations. However, as illustrated in Table 1, these methods are not good at rendering novel view in unseen parts due to the absence of generation ability.

Joint Multi-View Synthesis. Generative NVS complements reconstruction by inferring plausible content outside the observed regions. CAT3D, MVGenMaster, and SEVA coordinate target views through joint multi-view diffusion (Gao et al., 2024; Cao et al., 2025; Zhou et al., 2025). Many approaches include geometry guidance through different representations, such as 3D-aware diffusion using a volumetric feature field (Chan et al., 2023), jointly denoising RGB and scene-coordinate maps with observation-dependent supervision (Çelen et al., 2026), aligning latent with geometric representation (Huang et al., 2026) and adopting latent features from geometric foundation models (Wang et al., 2025a; Lin et al., 2025) for diffusion (Jang et al., 2026). Though these methods can generate multiple unseen views, they do not retain generated content for incremental and successive queries. Each new query is conditioned only on the source observations, and the generated content can not be reused. Few works (Kong et al., 2025) support auto-regressive formulation, but the heavy reliance on 2D representations have been identified as a limitation (Jang et al., 2026).

Table 1: **Different NVS capabilities.** △: varies by method.

View Synthesis Type	Unordered Query	Incremental Query	Unseen Generation	Gen.History Reuse
Reconstruction NVS	✓	✓	×	×
Joint multi-view NVS	✓	×	✓	×
Video-based NVS	△	△	✓	△
SEG (ours)	✓	✓	✓	✓

Video Generation for View Synthesis. Video-based view synthesis uses learned video priors to generate image sequences conditioned on camera motion. CameraCtrl (He et al., 2024) introduces camera control into video diffusion, while ViewCrafter (Yu et al., 2024) incorporates geometric guidance to synthesize novel views from sparse observations. Auto-regressive approaches such as ARSS (Teng et al., 2025) generate view sequences along specified camera trajectories. Memory-based methods (Ren et al., 2025; Li et al., 2025b) further retain generated scene content to guide subsequent synthesis. Self Forcing trains on self-generated contexts with K/V caching, while Self Gradient Forcing lets future prediction losses supervise the encoding of historical latents into cached features (Huang et al., 2025; Zhuang et al., 2026). VGGT-World predicts the temporal evolution, and decodes future geometry (Sun et al., 2026) for reconstruction. However, video-based methods typically organize synthesis around continuous camera trajectories or video clips, whereas our setting considers individual camera queries that arrive successively in arbitrary spatial order. SEG instead conditions on specified camera queries of a static scene and expanding synthesized geometric state without new visual observations, making it more suitable and flexible for interactive exploration.

3 METHOD

Overview Our goal is to enable successive scene synthesis from sparse observations via self-expanding geometric states. We first formulate successive view synthesis (Section 3.1) and introduce the self-expanding geometric state (Section 3.2). We then describe the causal view synthesis pipeline and state readout (Section 3.3), followed by causal-prefix supervision and the training objective (Section 3.4).

3.1 PROBLEM FORMULATION

Given a set of N posed source images $\mathcal{O} = \{(I_i^{\text{src}}, P_i^{\text{src}})\}_{i=1}^{N}$ of a static scene, conventional novel view synthesis predicts an image at each target camera P_k^{tgt} as

$$\widehat{I}_k^{\text{tgt}} = f_\theta\big(\mathcal{O}, P_k^{\text{tgt}}\big)\,. \tag{1}$$

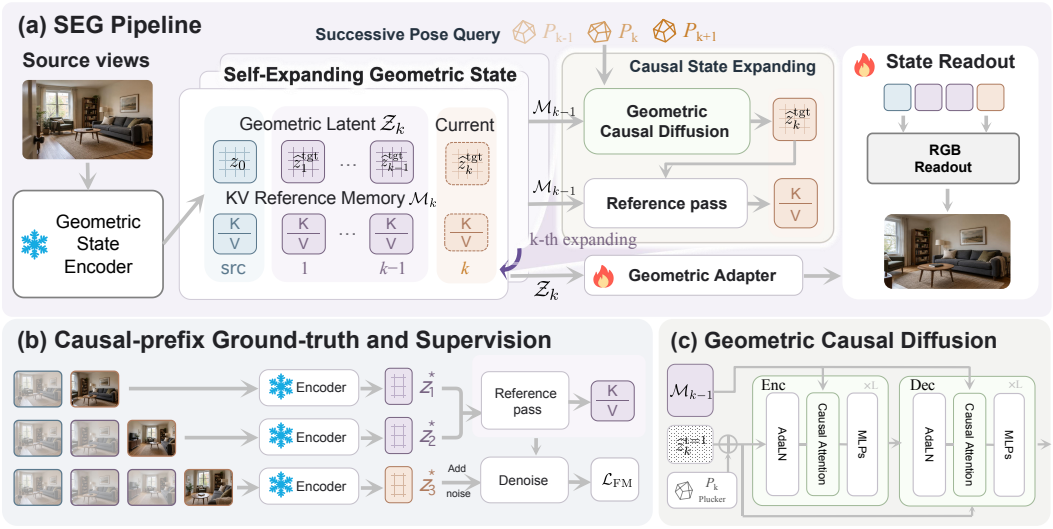

Figure 2: **Overview of SEG.** (a) Source views initialize a geometric latent archive and KV reference memory. Generated latents and KV entries from a clean reference pass are appended without changing earlier entries; the Geometric Adapter enables RGB readout. (b) The frozen encoder constructs each ground-truth target from its own image prefix. Earlier clean targets build reference memory for supervising the noisy current target. (c) The Geometric Causal Diffusion combines camera conditioning with causal attention in both its encoder and decoder.

Here P contains camera intrinsics and extrinsics, and f_θ denotes a neural network parameterized by θ. This fixed-context formulation does not specify how synthesized information is retained for subsequent queries. We consider *successive view synthesis*, where camera queries arrive incrementally in arbitrary spatial order and each prediction is conditioned on the source observations and context retained from earlier queries:

$$\widehat{I}_k^{\mathrm{tgt}} = f_\theta\left(\mathcal{O}, \mathcal{H}_{k-1}, P_k^{\mathrm{tgt}}\right), k = 1, 2, \ldots, \tag{2}$$

where $\mathcal{H}_{k-1}$ denotes the retained synthesis context, containing information about previously generated content. The formulation leaves its representation open: it may consist of images, latent features, or a learned scene representation. The index k denotes query order rather than scene time, and viewpoints may arrive in any spatial order without following a smooth trajectory.

Successive view synthesis requires causal conditioning and poses two core requirements for $\mathcal{H}_k$: (1) **efficient incremental updates**, which require a state that accumulates generated content and reuses reference features across queries; (2) **spatial reasoning**, which relates each query to geometrically relevant earlier views regardless of their query order.

3.2 Self-Expanding Geometric State

Geometric State and Challenges. We begin our method by instantiating the retained context $\mathcal{H}_k$ as a persistent latent state $\mathcal{S}_k$ containing information from source observations and completed predictions. Existing view synthesis methods often use image-based features or latents (Ren et al., 2025; Zhou et al., 2025; Jin et al., 2025). To retain spatial information across discontinuous and unordered camera queries, a natural choice is to adopt intermediate geometric features from frozen foundation models (Lin et al., 2025; Wang et al., 2025a) and expand them through generation. However, directly adopting these features does not inherently ensure two properties required for successive NVS: **consistent representation under expansion** and **causal target construction under multi-view encoding**. These encoders jointly aggregate information across input views, so adding a new view can alter earlier representations, while encoding a complete training sequence can introduce future-view information into earlier geometric targets. We therefore construct a self-expanding state that addresses both requirements.

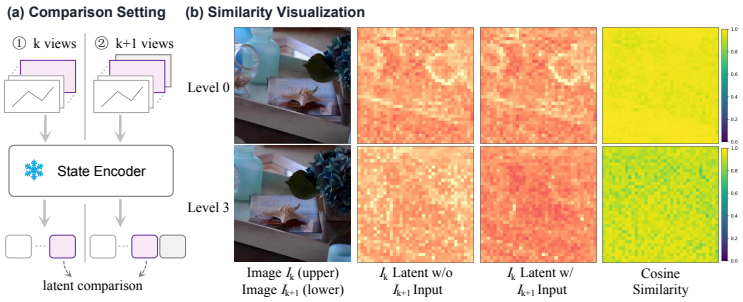

Figure 3: **Feature stability under expansion.** (a) Comparing features of I_k with and without I_{k+1}. (b) DA3 features and cosine similarity.

Table 2: **Stability across DA3 levels.** Cosine similarity of I_k's features as the input grows from k to $k+1$ views.

Level	$k=1$	$k=5$
Level 0	**0.988**	**0.991**
Level 1	0.943	0.963
Level 2	0.921	0.947
Level 3	0.924	0.944

Representation Stability under Expansion. Stability under expansion guides our choice of retained features. We compare four intermediate levels[1] (Levels 0–3) of the Depth Anything 3 encoder (Lin et al., 2025), measuring how the representation of I_k changes when I_{k+1} is added (Figure 3). As shown in Table 2, Level 0 achieves the highest cosine similarity at both evaluated prefix lengths. Its lower sensitivity to newly added views motivates our choice of Level 0 as the retained representation. We denote the corresponding frozen feature extractor by $\mathcal{E}$.

Causal-prefix Geometric Target Construction. To prevent future-view information from entering the training targets, we encode each ground-truth geometric target from its own image prefix, as illustrated in Figure 2(b). For query k, the encoder $\mathcal{E}$ processes the source images and the first k ground-truth target images, and we retain only the current target's feature:

$$z_k^{\mathrm{tgt},\star} = \left[\mathcal{E}\left(I_{1:N}^{\mathrm{src}}, I_{1:k}^{\mathrm{tgt}}\right)\right]_{N+k}, \tag{3}$$

where $\star$ denotes an encoded ground-truth target and $[\cdot]_{N+k}$ selects the final view. Each earlier target retains the feature extracted from its own image prefix as later views are added. Thus, extending the prefix defines a new target without revising earlier ones or introducing future information into their representations.

Self-Expanding State Formulation. To make these geometric representations reusable for successive causal querying, we pair a latent archive with KV reference memory, formulating the complete state as:

$$\mathcal{S}_k = (\mathcal{Z}_k, \mathcal{M}_k), \qquad k = 0, 1, \ldots, \tag{4}$$

where $\mathcal{Z}_k$ stores source and generated geometric latents, and $\mathcal{M}_k$ stores their reference keys and values for reuse by subsequent queries. We initialize $\mathcal{Z}_0 = \mathcal{E}(I_{1:N}^{\mathrm{src}})$ and obtain $\mathcal{M}_0$ through the reference pass described in Section 3.3. After generating $\hat{z}_k^{\mathrm{tgt}}$, the same reference pathway computes its new memory entries $\Delta\mathcal{M}_k$. Both components then expand by appending these entries:

$$\mathcal{Z}_k = \mathcal{Z}_{k-1} \,\|\, \hat{z}_k^{\mathrm{tgt}}, \qquad \mathcal{M}_k = \mathcal{M}_{k-1} \,\|\, \Delta\mathcal{M}_k, \tag{5}$$

where $\|$ denotes appending along the view sequence, applied separately at each layer for the KV memory. These updates preserve all earlier latent and memory entries, allowing generated content to accumulate as reusable geometric context for successive queries.

3.3 Causal View Synthesis Pipeline

Based on the self-expanding geometric state, our pipeline successively generates geometric latents to expand the state and synthesizes RGB images. As illustrated in Figure 2(a), it comprises a Geometric Causal Diffusion, a Geometric Adapter and an RGB readout head.

Geometric Causal Diffusion. Building on latent diffusion in (Gao et al., 2024; Jang et al., 2026), we introduce causal conditioning for successive queries. As shown in Figure 2(c), given a camera query P_k^{tgt}, we initialize the current latent with Gaussian noise and denoise it from $t=1$ to $t=0$.

[1] See Appendix A.1 for details.

Table 3: **Successive View Synthesis on RE10K and DL3DV** with 1 source and 7 query views. Best and second-best results are **bold** and underlined, respectively. Time (s) and GPU memory (GB) values are shared due to the common evaluation setup.

Method	Type	Total			Observed		Unobserved		Time↓	Memory↓
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	PSNR↑	SSIM↑		
RE10K (Zhou et al., 2018)										
LVSM (Jin et al., 2025)	R	11.85	0.212	0.800	12.27	0.237	11.71	0.195	**0.82**	**1.52**
Gen3R (Huang et al., 2026)	V	12.99	0.472	0.543	13.13	0.475	11.89	0.377	36.66	23.10
Gen3C (Ren et al., 2025)	V	13.16	0.451	0.484	13.47	0.455	12.26	0.397	1140.37	42.64
GLD (Jang et al., 2026)	J	12.56	0.489	0.543	12.71	0.493	12.21	0.460	40.20	15.90
GenRec (Çelen et al., 2026)	J	13.81	0.508	0.483	13.96	0.514	12.95	0.439	18.70	40.79
SEG (ours)	S	**14.68**	**0.542**	**0.459**	**14.83**	**0.547**	**14.01**	**0.476**	8.02	10.92
DL3DV (Ling et al., 2024)										
LVSM (Jin et al., 2025)	R	12.12	0.201	0.756	12.29	0.223	11.93	0.195	**0.82**	**1.52**
Gen3R (Huang et al., 2026)	V	12.36	0.312	0.596	12.50	0.319	12.01	0.263	36.66	23.10
Gen3C (Ren et al., 2025)	V	14.04	0.376	0.546	14.48	0.391	12.78	0.308	1140.37	42.64
GLD (Jang et al., 2026)	J	12.11	0.374	0.589	12.33	0.395	11.84	0.290	40.20	15.90
GenRec (Çelen et al., 2026)	J	13.94	0.428	0.494	14.07	0.436	13.43	0.381	18.70	40.79
SEG (ours)	S	**15.08**	**0.461**	**0.472**	**15.23**	**0.467**	**14.71**	**0.424**	8.02	10.92

The backbone uses Plücker ray embeddings for camera conditioning and PRoPE (Li et al., 2025a) to incorporate relative camera geometry into attention. Different from joint multi-view setting (Gao et al., 2024), we use causal attention in both the encoder and decoder. The current view attends to all source views and earlier targets through KV reference memory $\mathcal{M}_{k-1}$ to capture the consistent geometric relationships during successive query, and historical KV memory is reused across denoising steps and queries.

Once the prediction $\hat{z}_k^{\text{tgt}}$ is complete, we perform a clean reference pass through the same diffusion network at $t = 0$ to complete the self-expanding process:

$$\Delta\mathcal{M}_k = \mathcal{C}_\theta\left(\hat{z}_k^{\text{tgt}}, P_k^{\text{tgt}}; \mathcal{M}_{k-1}\right), \tag{6}$$

where $\mathcal{C}_\theta$ denotes the network. The resulting $\Delta\mathcal{M}_k$ is appended to $\mathcal{M}_{k-1}$ by Eq. (5). This converts the completed prediction into reusable reference entries while keeping earlier latents and memory fixed, and still allow new queries to attend to all previous and generated contents.

Geometric Adapter and RGB Readout. After the self-expanding process, the Geometric Adapter $\mathcal{P}_\phi$ jointly propagates the archived Level 0 latents to the deeper feature levels of Depth Anything 3 (Lin et al., 2025). It follows the DA3 encoder architecture after Level 0 and is trained to accommodate differences between generated latents and encoder-extracted features. An RGB readout head comprising transformer blocks reconstructs each image from the concatenated multi-level features. These readout features are not written back to the latent archive or KV memory.

3.4 Causal-prefix Supervision Training

As shown in Figure 2(b), the shared reference pass builds KV memory from source latents and earlier clean targets defined in Eq. (3). Each noisy target is supervised using only this history; its clean counterpart becomes a reference for subsequent queries.

For M training queries, we perturb each clean target latent with Gaussian noise and train the DiT with flow matching:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\left[\frac{1}{M}\sum_{k=1}^{M}\left\|v_\theta\left(z_{k,t_k}^{\text{tgt}}, t_k, P_k^{\text{tgt}}; \mathcal{M}_{k-1}^{\star}\right) - \left(\epsilon_k - z_k^{\text{tgt},\star}\right)\right\|_w^2\right], \tag{7}$$

The geometric encoder remains frozen, while the remaining modules are jointly trained with RGB supervision. Appendix B.2 details noise construction, token weighting, and the full objective.

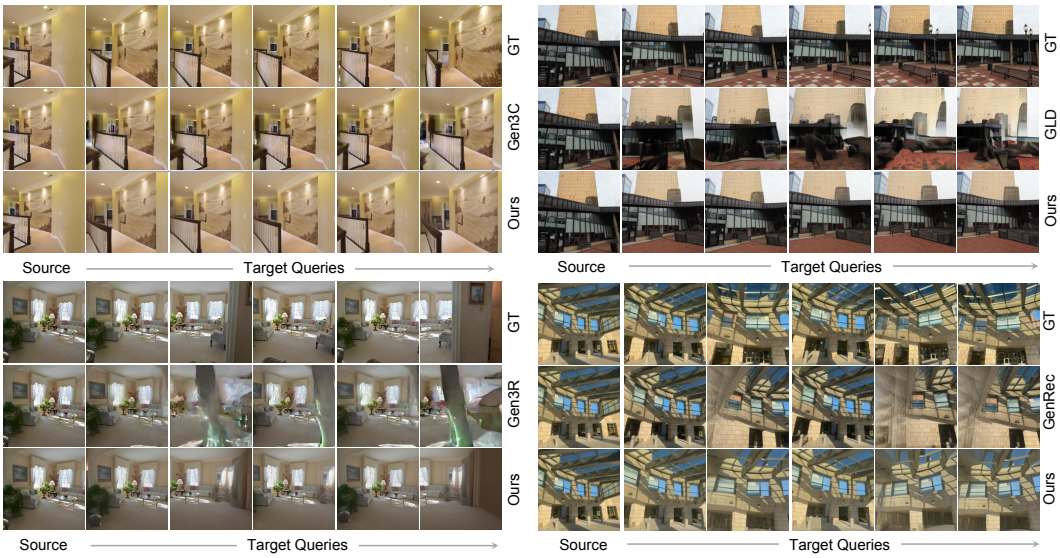

Figure 4: **Qualitative comparisons under successive view synthesis**. Examples from (a) RE10K and (b) DL3DV show better visualization results of our method compared to baselines, including both indoor and outdoor scenes.

Table 4: **Camera pose accuracy on RE10K and DL3DV.** Best results are **bold**.

Method	RE10K (Zhou et al., 2018)			DL3DV (Ling et al., 2024)		
	ATE ↓	RPE_t ↓	RPE_r ↓	ATE ↓	RPE_t ↓	RPE_r ↓
Gen3R (Huang et al., 2026)	0.595	0.455	2.074	0.587	0.694	3.032
Gen3C (Ren et al., 2025)	0.245	0.207	1.690	0.147	0.225	1.759
GLD (Jang et al., 2026)	0.218	0.313	3.043	0.576	0.660	2.755
GenRec (Çelen et al., 2026)	0.223	0.276	2.059	0.349	0.423	1.950
SEG (ours)	**0.157**	**0.204**	**1.687**	**0.133**	**0.185**	**1.322**

4 Experiments

4.1 Experimental setup

Datasets and Implementation. We train SEG on RealEstate10K (RE10K) (Zhou et al., 2018), DL3DV (Ling et al., 2024), and ScanNet++ (Yeshwanth et al., 2023) at 504×504 resolution. We follow the same evaluation procedures as previous protocols (Çelen et al., 2026) with 100 evaluation samples from RE10K and 100 from DL3DV. More implementation details about network architecture, training, evaluation settings along with hyper parameters are provided in Appendix B.

Baselines and Metrics. We compare against LVSM (Jin et al., 2025), GLD (Jang et al., 2026), GenRec (Çelen et al., 2026), Gen3R (Huang et al., 2026), and Gen3C (Ren et al., 2025) using their original checkpoints and same evaluation protocols for fair comparison. We measure image quality with PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018), and pose consistency with absolute trajectory error (ATE) and rotational/translational relative pose errors (RPE_r, RPE_t).

4.2 Quantitative Results

2D Image Quality. Table 3 shows our successive view synthesis 2D performance on two datasets with 1 source view and 7 target views. We also separate the observed and unobserved regions (Çelen et al., 2026) to better understand the model's performance. SEG achieves the best PSNR, SSIM, and LPIPS on both datasets compared with reconstruction based (R), joint multi-view based (J), and video generation based (V) methods, with more than 1dB gains in DL3DV and 0.8dB in RE10K. It

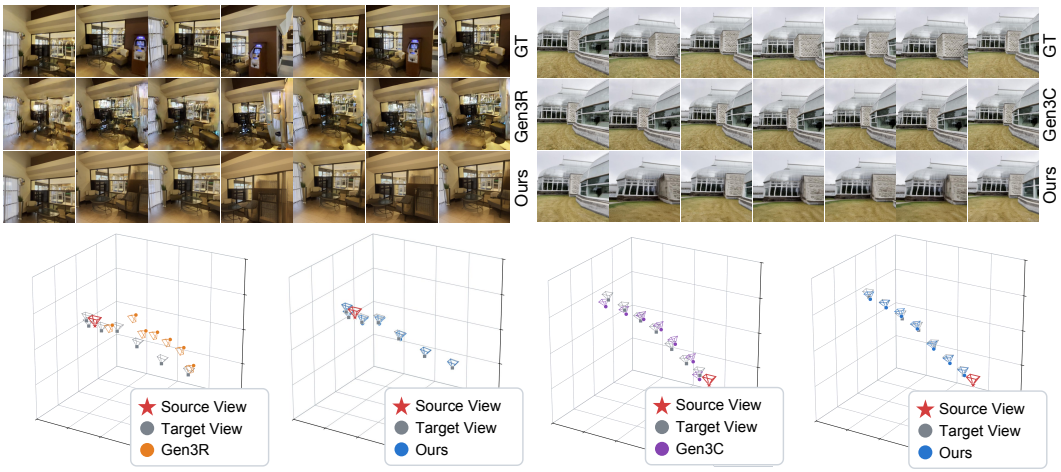

Figure 5: **Qualitative results of pose consistency**. SEG shows more consistent estimated viewpoints with ground truth, compared with other baselines (see more in Appendix C).

Table 5: **2-source view synthesis on RE10K and DL3DV**. Best in **bold**.

	Method	PSNR↑	SSIM↑	LPIPS↓
RE10K	GLD	16.11	0.586	0.354
	GenRec	16.01	0.575	0.378
	SEG (ours)	**17.32**	**0.630**	**0.335**
DL3DV	GLD	16.31	0.514	**0.383**
	GenRec	14.66	0.457	0.463
	SEG (ours)	**16.83**	**0.523**	0.390

Table 6: **Long-horizon with 15 target views on DL3DV**. Time (s), memory (GB). Best in **bold**.

Method	PSNR↑	SSIM↑	LPIPS↓	Time↓	Memory↓
GLD	13.08	0.404	0.583	346.00	46.58
Gen3R	11.61	0.302	0.619	77.04	41.19
Gen3C	12.92	0.379	0.586	1231.00	40.90
GenRec	12.34	0.385	**0.566**	29.40	72.83
SEG (ours)	**13.15**	**0.440**	0.567	**22.03**	**25.97**

is worth noting that the improvements are particularly significant in unobserved regions with around 1.3dB in DL3DV, supporting our geometric state reuse design for scene completion beyond source observations. Meanwhile, our method achieves better efficiency performance against all generation methods, indicating the effectiveness of our KV memory design.

3D Pose Consistency. Besides image quality, we also provide quantitative results of 3D pose consistency to evaluate the geometric ability of each method. Specifically, we use the pretrained VGGT (Wang et al., 2025a) to predict the per-view pose and compare it with the ground-truth target pose. The results in Table 4 demonstrate that our SEG achieves the lowest ATE and RPE on both datasets, indicating that our method is able to follow the target viewpoints more accurately, even these target viewpoints are unordered. This is consistent with the goal of using geometric latents to relate each query and retained context.

More Source views and Target queries. Besides the main evaluation with 1 source view and 7 target views, we also evaluate our method with 2 source views in Table 5, as well as 15 target views Table 6. For the 2-source evaluation, we use the same 7 target views as the main evaluation, and evaluate our model without extra fine-tuning. It is worth noting that our method still achieves the general best performance on both datasets against competitive multi-view baselines, indicating that our method is able to leverage more source observations with the help of self-expanding geometric state for better scene understanding. For long-horizon synthesis with 1 source view and 15 target views, SEG also keeps the leadership, which verifies that our method can be extended to longer sequences. Meanwhile, SEG uses the lowest runtime and GPU memory among generative methods.

4.3 Qualitative Results

Figure 4 shows the 2D view synthesis results in both RE10K and DL3DV, compared with generative baseline methods. The results demonstrate better preservation and consistency of hallway and living-room layouts. For example, GenRec shows clear artifacts in several viewpoints, while SEG

Table 7: **Ablation study on RE10K and DL3DV.** Best results are **bold**.

Exp. ID	Ablation Variants	RE10K (Zhou et al., 2018)			DL3DV (Ling et al., 2024)		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
I	w/ VGGT (Wang et al., 2025a)	14.17	0.523	0.471	14.59	0.448	0.501
II	w/o SEG State	13.43	0.477	0.495	13.70	0.441	0.569
III	w/o Geometric Adapter	14.46	0.523	0.463	14.87	0.437	0.484
IV	w/o causal-prefix supervision	14.30	0.525	0.477	14.75	0.452	0.497
V	**SEG (ours)**	**14.68**	**0.542**	**0.459**	**15.08**	**0.461**	**0.472**

Table 8: **Ablation of Ordered Query** on DL3DV. $\Delta\%$ represents the percentage improvement or degradation. Best scores are **bold**.

Method	PSNR↑			SSIM↑			LPIPS↓		
	Ordered	Unordered	$\Delta\%$	Ordered	Unordered	$\Delta\%$	Ordered	Unordered	$\Delta\%$
Gen3R (Huang et al., 2026)	13.04	12.36	-5.21	0.350	0.312	-10.86	0.557	0.596	-7.00
Gen3C (Ren et al., 2025)	14.68	14.04	-4.36	0.409	0.376	-8.07	0.514	0.546	-6.16
SEG (ours)	**15.17**	**15.08**	-0.59	**0.464**	**0.461**	-0.65	**0.478**	**0.472**	+1.26

maintains a more reasonable appearance. In Figure 5, we visualize the estimated target pose of SEG compared with the baselines. Generally, we show better consistency with the ground truth target poses than the other methods. Meanwhile, we also provide the corresponding rendered target views, which also shows better visual performance. More qualitative results are provided in Appendix C.

4.4 Ablation Studies

Model components. We evaluate all the components of SEG in Table 7. The ablation variants are: (I) replacing DA3 geometric encoder Level 0 with VGGT (Wang et al., 2025a), (II) changing the SEG state as a fixed one, (III) replacing the Geometric Adapter with MLPs and upsamplings, and (IV) removing causal-prefix supervision. The qualitative results on both datasets show that each component contributes to the final performance, especially the usage of self-expanding state due to its ability to reuse historical geometric context for better scene understanding. Without causal-prefix supervision, performance also drops despite retaining causal attention.

Query order. We further evaluate the effect of query order on DL3DV (Table 8), compared to the video generation based methods. Switching from ordered to unordered queries degrades all three metrics for both video baselines, with a large margin, whereas SEG changes little and even slightly improves in LPIPS. The results demonstrate that our method is robust to unordered queries, which is consistent with the choice of geometric state for incremental and unordered view synthesis.

5 Conclusion

We presented SEG, an efficient and effective pipeline for successive novel view synthesis which requires incremental and unordered view queries. Our key insight is to turn context-dependent geometric latent into self-expanding geometric state with consistent representation under expansion and causal target construction under successive queries. We construct the SEG state with paired stable geometric latents and KV reference memory with causal-prefix supervision, and design a causal geometric diffusion model for state synthesis and expansion. Through extensive experiments on RealEstate10K and DL3DV, we show that SEG achieves better performance in both 2D view synthesis quality and 3D viewpoint accuracy efficiently, and ablation studies demonstrate the insight and effectiveness of each component in our method.

Limitations and future work. Although SEG shows promising results in successive view synthesis, it still has some limitations with respect to maintaining super-long-term consistency limited by the geometric latent backbone, which will be a promising future direction for further exploring successive view synthesis.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and for guidance on figure layout and presentation, to improve the readability of this paper. We have not used generative AI tools for any tasks with required disclosure, including generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. The remaining required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, sensitive data, or privacy concerns. All datasets used are publicly available, and we foresee no harmful applications or conflicts of interest arising from this research.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of our work. Architecture, training, and evaluation details are provided in the main text and appendix. All datasets used in our experiments are publicly available. The code and checkpoint of our model will be revealed in the article upon acceptance.

REFERENCES

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6045–6056, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Cao_MVGenMaster_Scaling_Multi-View_Generation_from_Any_Image_via_3D_Priors_CVPR_2025_paper.html.

Ata Çelen, Jaewoo Jung, Federico Tombari, Marc Pollefeys, Sunghwan Hong, Michael Niemeyer, and Daniel Barath. GenRec: Knowing where to reconstruct and where to generate. *arXiv preprint arXiv:2608.17832*, 2026. URL https://arxiv.org/abs/2608.17832.

Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3D-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4217–4229, 2023. URL https://arxiv.org/abs/2304.02602.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelSplat: 3D gaussian splats from image pairs for scalable generalizable 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19457–19467, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Charatan_pixelSplat_3D_Gaussian_Splats_from_Image_Pairs_for_Scalable_Generalizable_CVPR_2024_paper.html.

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 (ECCV)*, 2024. URL https://arxiv.org/abs/2403.14627.

Xin Fei, Wenzhao Zheng, Yueqi Duan, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Jiwen Lu. PixelGaussian: Generalizable 3D gaussian reconstruction from arbitrary views. *arXiv preprint arXiv:2410.18979*, 2024. URL https://arxiv.org/abs/2410.18979.

Ruiqi Gao, Aleksander Hołyński, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T. Barron, and Ben Poole. CAT3D: Create anything in 3D with multi-view diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 75468–75494, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/89e4433fec4b99f1d859db57af1e0a0f-Abstract-Conference.html.

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

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 (CVPR)*, 2026. URL https://arxiv.org/abs/2601.04090.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 167283–167308, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html.

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. In *European Conference on Computer Vision (ECCV)*, 2026. URL https://arxiv.org/abs/2603.22275.

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. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9676c5283df26cabca412ca66b164a7d-Abstract-Conference.html.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Xin Kong, Daniel Watson, Yannick Strümpler, Michael Niemeyer, and Federico Tombari. CausNVS: Autoregressive multi-view diffusion for flexible 3D novel view synthesis. *arXiv preprint arXiv:2509.06579*, 2025. URL https://arxiv.org/abs/2509.06579.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In *Advances in Neural Information Processing Systems*, 2025a. URL https://arxiv.org/abs/2507.10496.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. VMem: Consistent interactive video scene generation with surfel-indexed view memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25690–25699, 2025b. URL https://arxiv.org/abs/2506.18903.

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. URL https://arxiv.org/abs/2511.10647.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169, 2024. URL https://arxiv.org/abs/2312.16256.

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. In *European Conference on Computer Vision (ECCV)*, pp. 405–421, 2020. URL https://link.springer.com/chapter/10.1007/978-3-030-58452-8_24.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ren_GEN3C_3D-Informed_World-Consistent_Video_Generation_with_Precise_Camera_Control_CVPR_2025_paper.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Xiangyu Sun, Shijie Wang, Fengyi Zhang, Lin Liu, Caiyan Jia, Ziying Song, Zi Huang, and Yadan Luo. VGGT-World: Transforming VGGT into an autoregressive geometry world model. *arXiv preprint arXiv:2603.12655*, 2026. URL https://arxiv.org/abs/2603.12655.

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 (CVPR)*, 2026. URL https://arxiv.org/abs/2603.20176.

Wenbin Teng, Gonglin Chen, Haiwei Chen, and Yajie Zhao. ARSS: Taming decoder-only autoregressive visual generation for view synthesis from single view. *arXiv preprint arXiv:2509.23008*, 2025. URL https://arxiv.org/abs/2509.23008.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VGGT_Visual_Geometry_Grounded_Transformer_CVPR_2025_paper.html.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Continuous_3D_Perception_Model_with_Persistent_State_CVPR_2025_paper.html.

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. URL https://ece.uwaterloo.ca/~z70wang/research/ssim/.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. DepthSplat: Connecting gaussian splatting and depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://haofeixu.github.io/depthsplat/.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22, 2023. URL https://arxiv.org/abs/2308.11417.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4578–4587, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yu_pixelNeRF_Neural_Radiance_Fields_From_One_or_Few_Images_CVPR_2021_paper.html.

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. URL https://arxiv.org/abs/2409.02048.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. URL https://arxiv.org/abs/1801.03924.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3c4141c12660ad3625eb4ae845e0a6f9-Abstract-Conference.html.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12405–12414, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhou_Stable_Virtual_Camera_Generative_View_Synthesis_with_Diffusion_Models_ICCV_2025_paper.html.

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

Junhao Zhuang, Shiyi Zhang, Yuxuan Bian, Yaowei Li, Yawen Luo, Yijun Liu, Weiyang Jin, Songchun Zhang, Xianglong He, Xuying Zhang, Haoran Li, Haoyang Huang, Zeyue Xue, and Nan Duan. Self gradient forcing: Native long video extrapolation. *arXiv preprint arXiv:2607.20368*, 2026. URL https://arxiv.org/abs/2607.20368.

A Preliminary Details of Geometric Encoders

A.1 Depth Anything 3 Encoder

Depth Anything 3 (Lin et al., 2025) builds on a single Vision Transformer initialized from DINOv2 (Oquab et al., 2023). It combines within-view self-attention, which processes each image independently, with cross-view self-attention over tokens from all input images. The encoder contains 12 Transformer blocks with a hidden dimension of 768. Its first four blocks use within-view attention, and the remaining eight form four pairs of within-view and cross-view blocks. The original Depth Anything 3 model decodes these features into depth and ray maps through a dual-DPT head. We extract Level 0 to Level 3 after blocks 6, 8, 10, and 12, respectively.

A.2 VGGT Encoder

VGGT (Wang et al., 2025a) first extracts per-image patch tokens with a DINOv2 ViT-L encoder, then processes them with an aggregator containing 24 pairs of frame-wise and global self-attention blocks. Each block has a hidden dimension of 1024. Following its dense prediction heads, our VGGT ablation study selects block pairs 5, 12, 18, and 24 as Level 0 to Level 3. Table 9 summarizes the correspondence. Both encoders allow features to depend on the jointly encoded view set, motivating the stability analysis and causal-prefix targets used in SEG.

Table 9: **Feature-level correspondence for the geometric encoders.**

Method	Level 0	Level 1	Level 2	Level 3
Depth Anything 3 (Lin et al., 2025)	6	8	10	12
VGGT (Wang et al., 2025a)	5	12	18	24

B Implementation Details

B.1 Network Architecture Details

Geometric State Encoder. We use Depth Anything 3 encoder (Lin et al., 2025) described in Appendix A.1, evaluated through block 6 to obtain Level 0. For a 504×504 image, the output contains 36×36 spatial tokens and one non-spatial CLS token. Both spatial and CLS features are retained and generated as our geometric latent after normalization.

Geometric Causal Diffusion. We adapt the diffusion backbone similar with previous works (Gao et al., 2024; Jang et al., 2026), which retains a condition encoder of width 768 with 28 blocks and a velocity decoder of width 2048 with 6 blocks. Each block uses a causal attention with 16 attention heads with RMS and SwiGLU feed-forward layers. We use Gaussian Fourier timestep embeddings to modulate the condition encoder. Camera conditioning combines six Plücker-ray channels with a reference/query indicator. Source views attend jointly, while each subsequent query attends to the sources, committed views, and its own tokens. Each block caches spatial keys and values after PRoPE (Li et al., 2025a). A fixed session coordinate frame and scale keep the cached features consistent across queries. After denoising, a clean reference pass through the same network at $t = 0$ produces the new KV reference memory.

Geometric Adapter. The trainable adapter $\mathcal{P}_\phi$ follows the Depth Anything 3 architecture after Level 0 with 6 transformer blocks. It processes the denormalized archive jointly to produce multi-level inter-view and intra-view features. This adapts the generated latents for RGB readout while leaving the retained geometric state unchanged.

RGB Readout. The RGB readout module mainly contains 28 Transformer blocks, each block contains with a multi-head attention, LayerNorm, and GeLU feed-forward layer. We use a sinusoidal positional embeddings and a learned decoder CLS token to embed input features. Finally, a linear head predicts $14 \times 14 \times 3$ values per spatial token to reconstruct the image.

B.2 Training Details

Training Hyperparameters and Data Settings. Table 10 summarizes the training hyperparameters of our SEG. We train our model with samples from official training sets of DL3DV (Ling et al., 2024), RE10K (Zhou et al., 2018), and ScanNet++ (Yeshwanth et al., 2023) with sampling probabilities 0.48, 0.28, and 0.24, respectively. We use a training and evaluation resolution of 504×504. We use 1 source view and 7 target views per training sample. The source view is randomly selected from the available images, and the target views are randomly sampled from the remaining images. Each target view is assigned a random noise level $t \in [0, 1]$ and a Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ for flow-matching supervision. The camera dropout probability is set to 0.05 per scene, which randomly drops one source view during training to improve robustness.

Loss Functions. We jointly train the Geometric Causal DiT, Geometric Adapter, and RGB Readout using flow-matching and RGB reconstruction losses, while keeping the Geometric State Encoder $\mathcal{E}$ frozen. For each training query, we independently sample Gaussian noise $\epsilon_k \sim \mathcal{N}(0, \mathbf{I})$ and a shifted-uniform noise level $t_k \in [0, 1]$, and form the noisy latent $z^{\text{tgt}}_{k,t_k} = (1 - t_k) z^{\text{tgt},\star}_k + t_k \epsilon_k$. In Eq. (7), $\mathcal{M}^{\star}_{k-1}$ is the KV memory constructed from source latents and earlier clean targets through the shared reference pass, and the expectation is over training samples, noise, and timesteps. The weighted feature error in Eq. (7) treats spatial and CLS predictions as separate reductions. Let $e_{k,r}$ be the channel vector of velocity residuals for token r, where $r = 0$ denotes CLS and $r = 1, \ldots, n-1$ denotes spatial tokens. With $d_z = 1536$, the loss for one view is

$$\ell_k = \frac{1}{(n-1)d_z} \sum_{r=1}^{n-1} \|e_{k,r}\|_2^2 + \lambda_{\text{cls}} \frac{1}{d_z} \|e_{k,0}\|_2^2, \qquad \lambda_{\text{cls}} = 1. \tag{8}$$

The flow-matching objective averages ℓ_k across scenes and supervised horizons. Thus, the CLS contribution is not diluted by averaging it together with 1296 spatial tokens. The seed and reference-mode backbone passes remain differentiable, and training conditions on encoded ground-truth histories.

For RGB reconstruction, we apply pixel-wise mean squared error $\mathcal{L}_{\text{MSE}}$ and perceptual loss $\mathcal{L}_{\text{LPIPS}}$ (Zhang et al., 2018) between the images reconstructed through the Geometric Adapter and RGB Readout and their corresponding ground-truth RGB images. The combined RGB loss is weighted by 0.1 relative to flow matching, giving the joint objective

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{FM}} + 0.1 \left(\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{LPIPS}} \right). \tag{9}$$

Table 10: **Configuration** of SEG training.

Setting	Value
GPU	8 NVIDIA H20-3e
Learning rate	5×10^{-5}
Optimizer	AdamW
Scheduler	Constant schedule
Optimizer betas	$(0.9, 0.95)$
Global batch size	16
Weight decay	0
EMA decay	0.999
Gradient clipping	1.0
Camera dropout probability	0.05 per scene
Iterations	100000

B.3 Evaluation Details

Baseline inference. All baselines (Jin et al., 2025; Jang et al., 2026; Ren et al., 2025; Huang et al., 2026; Çelen et al., 2026) use their original pretrained checkpoints without additional fine-tuning. For joint multi-view synthesis methods, we implement the successive evaluation protocol by inference on the same source observations for each query. We use the original resolutions supported

by each checkpoint, and we crop the generated outputs to match the same evaluation resolution for fair comparison. We notice that the source code of GenRec (Çelen et al., 2026) has information leakage during inference, so we fix the problem for fair comparison.

C Additional Qualitative Results

Two-source view synthesis. Figure 6 shows that SEG maintains clear window frames and seating arrangements in the lounge, and sharper bed and room boundaries in the bedroom. GenRec introduces blur and distortions in the lounge, while GLD produces softer details in several bedroom views. These examples complement Table 5 and support initializing the geometric state from multiple source observations for successive view synthesis.

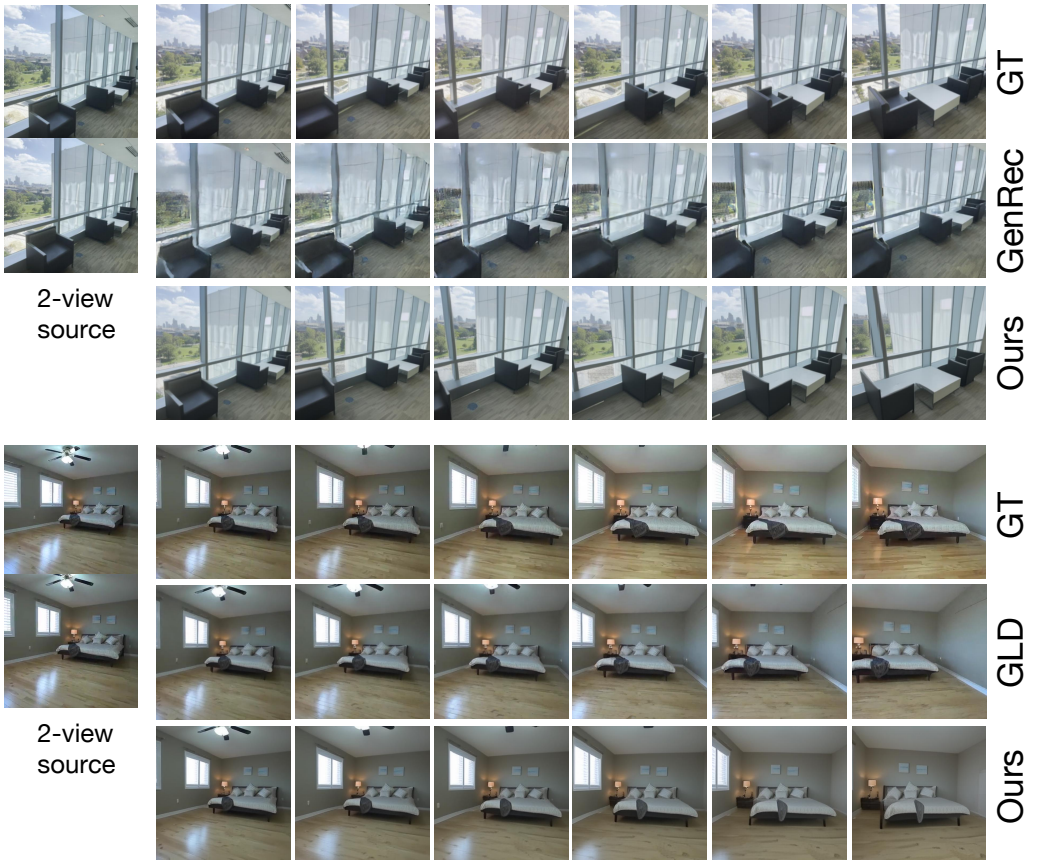

Figure 6: **Two-source view synthesis**: given the two source images on the left, SEG synthesizes successive target views and is compared with GenRec (top) and GLD (bottom).

Additional pose comparisons. Figure 7 complements Figure 5 with comparisons against GenRec and GLD.

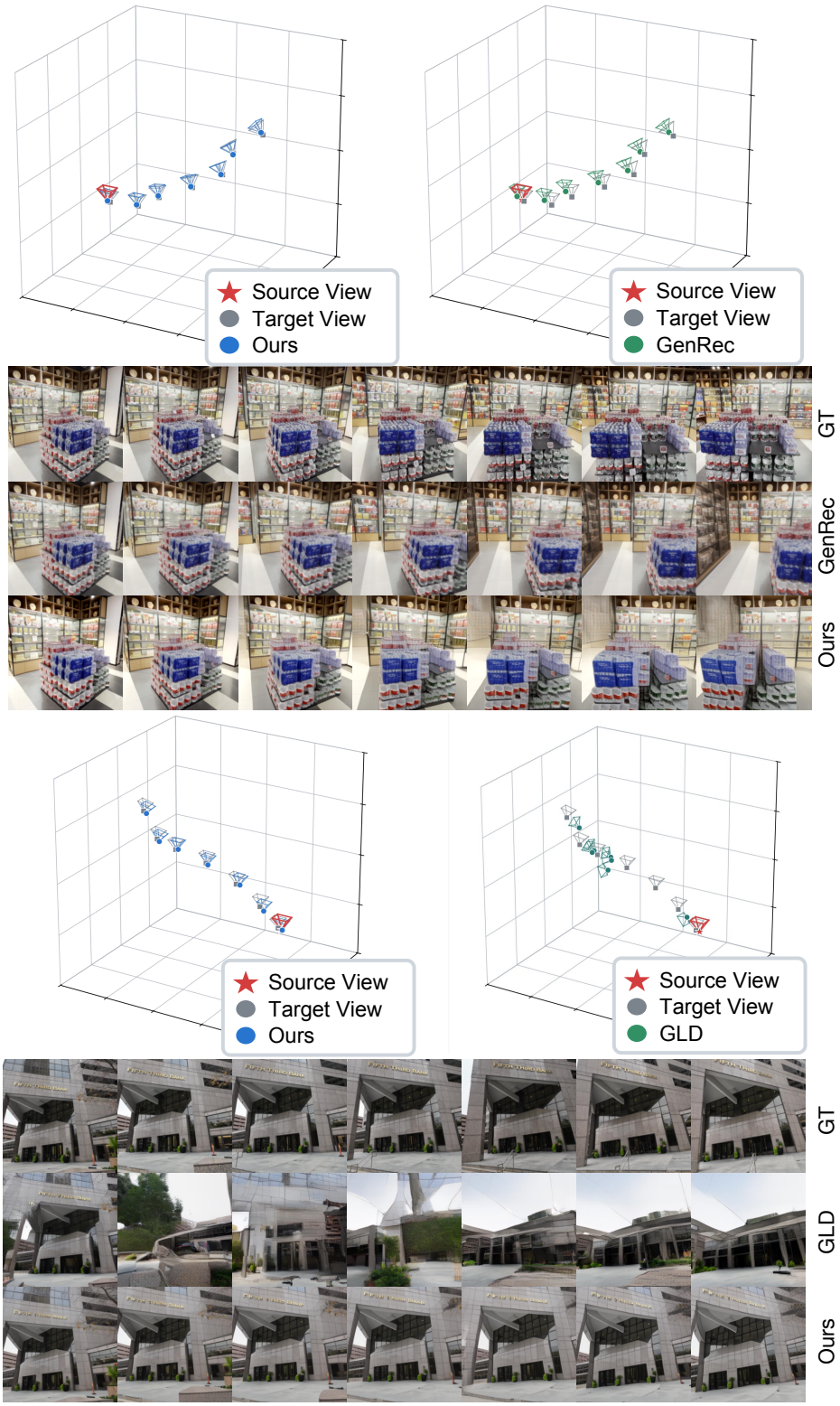

Figure 7: **Additional pose comparisons**: estimated camera poses and generated views of SEG are compared with GenRec (top) and GLD (bottom).