GENANCHOR: OBSERVATION-ANCHORED PROGRESSIVE RECONSTRUCTION WITH GENERATIVE PRIORS FOR SPARSE-VIEW GAUSSIAN SPLATTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent sparse-view 3D reconstruction methods exploit generative priors to complete under-observed scene content. However, generated views are not necessarily grounded in the available input evidence and may therefore conflict with real observations, degrading the appearance and geometry of reliably reconstructed regions. To address this issue, we present **GenAnchor**, an observation-anchored progressive reconstruction framework that preserves reliably observed content while selectively applying generative supervision to uncertain and unobserved regions. Specifically, we first employ coverage-aware trajectory planning to explore inadequately reconstructed regions. Each planned view is then partitioned into preservation, refinement, and generation regions using a tri-state masking strategy, thereby adapting generative supervision to the reconstruction status of each region. Subsequently, we recover depth from the repaired RGB sequence using the planned camera poses and depth anchors from real observations, preserving observed geometry while guiding depth inference in generated regions. The resulting RGB-D views are integrated into the global scene representation, whose updated state guides the next round of trajectory planning and view repair, thereby forming a progressive reconstruction loop. Extensive experiments on multiple widely used public datasets demonstrate that our method consistently improves scene completeness while better preserving reliably observed regions, achieving state-of-the-art performance in both appearance fidelity and geometric accuracy.

1 INTRODUCTION

Recent advances in neural radiance fields (Mildenhall et al., 2020) and 3D Gaussian Splatting (Kerbl et al., 2023; Huang et al., 2024) have enabled high-quality novel-view synthesis from dense multi-view observations. Their performance, however, degrades substantially under sparse inputs, where limited view overlap provides insufficient photometric and geometric constraints for large portions of the scene. To alleviate this issue, prior methods have explored depth regularization and geometric priors (Zhu et al., 2024; Li et al., 2024; Guédon et al., 2025; Wu et al., 2025b), which improve reconstruction in observed regions but cannot reliably recover appearance and geometry in poorly observed or unseen areas. Consequently, a reconstruction may faithfully reproduce the input views while still exhibiting uncovered regions, distorted surfaces, or severe artifacts when rendered from novel viewpoints that lie outside reliable observational support.

To recover scene content beyond the observed views, recent methods introduce generative priors into 3D reconstruction (Wu et al., 2024; 2025c). Image and video diffusion models synthesize plausible novel observations where direct measurements are unavailable, while camera-controlled generation (Yu et al., 2024; Liu et al., 2024), geometric guidance (Zhong et al., 2025; Ni et al., 2026), and reference-view conditioning (Yin et al., 2025) further improve cross-view consistency and scene fidelity. These approaches typically form a progressive reconstruction–generation loop: the current scene is rendered, novel views are generated or repaired, and the resulting observations are incorporated as additional supervision to update the global 3D representation.

However, this reconstruction–generation loop introduces a fundamental failure mechanism. Generative models may modify regions that are already reliably constrained by real observations, intro-

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.

ducing unnecessary changes in color, texture, object boundaries, or local geometry even when no completion is required. Once such synthesized views are reused as reconstruction supervision, these deviations become competing targets against the original measurements and may accumulate across iterations, progressively contaminating previously faithful scene content. This reveals a fundamental mismatch between generation and reconstruction: generative models optimize for plausible synthesis, whereas reconstruction requires faithful preservation of real evidence.

In this paper, we introduce **GenAnchor**, an observation-anchored framework that uses generative priors for progressive scene reconstruction from sparse views. It extends the reconstruction where real evidence is insufficient while preserving content reliably constrained by the input images. Real observations determine which regions may be modified during generative repair and where generated views may provide reconstruction supervision.

To reconstruct complete scenes from sparse views while preserving content supported by real observations, we first plan continuous, coverage-aware trajectories from the input cameras. The selected views expose missing content while retaining reconstructed context for generation. We then construct a tri-state mask for each view from rendering support and correspondences geometrically verified against the real inputs, separating grounded, provisional, and uncovered regions. During repair, projected input colors anchor grounded regions throughout denoising and are restored after decoding. The provisional regions retain the current rendering as editable context. The uncovered regions are generated from reference views, neighboring context, and camera guidance without relying on rendered RGB. The same mask restricts generated-view supervision to provisional and uncovered regions, leaving grounded regions under real-input supervision. Observation-anchored geometric completion recovers depth from the repaired RGB sequence using the planned cameras and depth anchors verified against the real input views. This avoids assigning pre-repair depth to content changed or introduced during generation and provides paired RGB-D supervision for scene optimization. Accepted generated views accumulate across reconstruction rounds, and renderings of the updated scene guide the next camera trajectory.

Our main contributions are summarized as follows:

- We introduce *GenAnchor*, an observation-anchored framework that integrates generative priors into progressive sparse-view scene reconstruction. Its observation-verified tri-state partition preserves reliable input appearance while directing generative repair and generated-view supervision toward uncertain and missing regions.
- We establish a reconstruction loop linking coverage-aware camera exploration, tri-state masking for generative repair, and observation-anchored geometric completion. Renderings of the updated scene guide the next camera trajectory, allowing generated RGB-D views to target content that remains uncertain or missing.
- Extensive experiments on multiple public datasets show that our method better preserves observed content and improves scene completeness, outperforming strong baselines in appearance fidelity and geometric accuracy.

2 Related Work

2.1 Sparse-view 3D Reconstruction

Neural radiance fields and Gaussian Splatting achieve high-quality novel-view synthesis under dense observations, but become severely under-constrained when only a few input views are available. RegNeRF (Niemeyer et al., 2022) regularizes unobserved views to improve sparse-view radiance fields. Sparse-view 3DGS methods (Zhu et al., 2024; Li et al., 2024; Zhang et al., 2024; Guédon et al., 2025; Wu et al., 2025b) further introduce depth regularization, feature consistency, multi-view correspondences, and surface priors to stabilize reconstruction under limited observations. These methods substantially improve geometry and appearance near captured views, although weakly observed regions remain difficult to reconstruct reliably.

2.2 Generative Priors for Sparse-view Reconstruction

Generative models provide learned visual priors that supplement limited multi-view observations. Early approaches incorporate diffusion priors through 3D distillation or reconstruction regularization. SparseFusion (Zhou & Tulsiani, 2023) distills view-conditioned diffusion priors into 3D, while DiffusioNeRF (Wynn & Turmukhambetov, 2023) regularizes RGB-D patches with diffusion priors. ReconFusion (Wu et al., 2024) directly generates novel-view supervision for under-constrained regions. Subsequent methods employ image diffusion (Wang et al., 2024a; Kong et al., 2025; Wu et al., 2025a) or video diffusion (Yu et al., 2024; Liu et al., 2024; Tang et al., 2026) to synthesize or refine additional observations with improved visual quality and cross-view coherence.

Recent methods increasingly couple generation with the current reconstruction. Scene geometry, rendered context, reference views, and geometric memory are introduced to better constrain generated observations and improve their consistency with the captured scene (Zhong et al., 2025; Ni et al., 2026; Yin et al., 2025; Zhang et al., 2026). Other methods expand reconstruction coverage by generating observations from additional camera viewpoints and incorporating them into scene optimization (Wu et al., 2025c; Kim et al., 2025; de Lutio et al., 2026). Viewpoint selection based on geometric visibility, information gain, or current reconstruction coverage further improves where new observations are acquired (Zhong et al., 2025; Ni et al., 2026; Kim et al., 2025; Shen et al., 2026). Autoregressive generation and geometric memory also improve consistency over extended camera trajectories (de Lutio et al., 2026; Tang et al., 2026; Zhang et al., 2026), particularly as generated views move farther from the original observations.

Existing work has primarily focused on improving the fidelity, consistency, and coverage of generated observations. Building on these efforts, we further consider how generated observations can be incorporated as supervision while preserving reconstruction fidelity.

3 Method

We present **GenAnchor**, an observation-anchored framework for progressive sparse-view reconstruction. Given posed input images $\mathcal{O} = \{(I_i, K_i, P_i)\}_{i=1}^{M}$, we initialize a 2DGS scene following G4Splat (Ni et al., 2026) using geometry-aligned input depths D_i^{obs}. Here, K_i and P_i denote the camera intrinsics and world-to-camera transform. The input depths remain fixed for correspondence verification and depth anchoring, while the real images provide appearance references. As shown in Figure 1, each round explores novel views, repairs uncertain and missing regions, recovers depth for the repaired RGB, and updates the scene while preserving reliably observed content.

3.1 Coverage-Aware Camera Exploration

Progressive reconstruction requires novel views that reveal missing scene content while retaining enough reconstructed context for video repair. We therefore plan continuous trajectories from real input cameras. For a candidate view v, let a_v and $\tilde{D}_v$ be the opacity and depth rendered from the current reconstruction. We define rendering support as:

$$S_v(x) = \mathbf{1}[a_v(x) \geq \tau_a]\, \mathbf{1}\Big[0 < \tilde{D}_v(x) < \infty\Big].$$

Views with large uncovered areas may offer little context for repair. We therefore score a view by uncovered pixels within image-plane distance r of rendering support:

$$f(v) = \frac{1}{HW} \sum_x \big(1 - S_v(x)\big) \mathbf{1}[d(x, \{y : S_v(y) = 1\}) \leq r], \tag{1}$$

where $H \times W$ is the image resolution and d denotes image-plane distance to the nearest supported pixel. During trajectory planning, we search leftward and rightward camera motions subject to per-step displacement, collision avoidance, and minimum rendering-support constraints. These constraints maintain gradual camera motion and retain rendered context at every candidate view. Complete trajectories are ranked by the mean of their view scores. The candidate motion order and feasibility bounds are given in Appendix C.1. The selected camera sequence is used for video repair and geometry reconstruction. After each round of scene optimization, we recompute rendering coverage and plan the next trajectory based on the remaining uncovered regions.

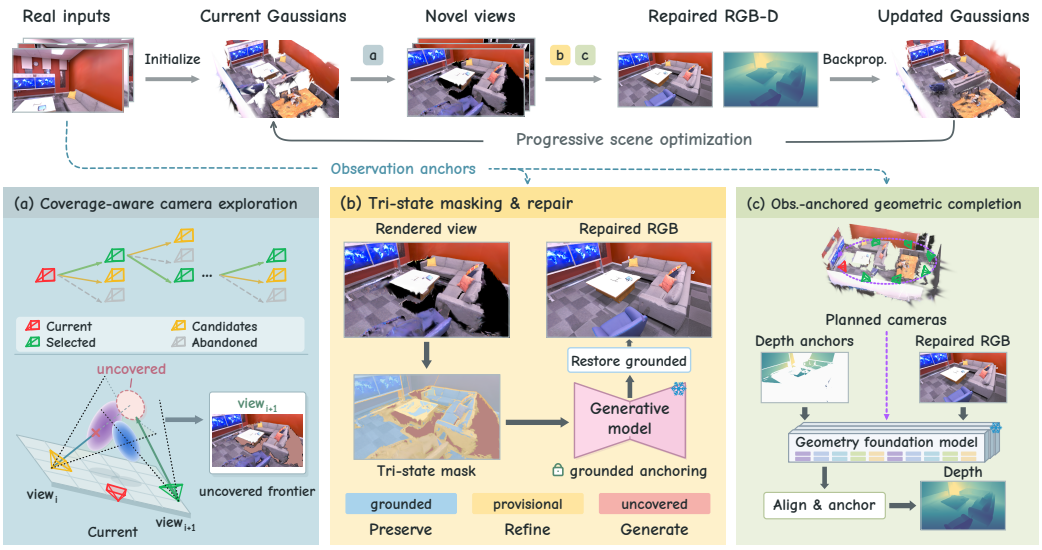

Figure 1: **Overview of GenAnchor.** **(a)** Coverage-aware exploration selects views with missing content and useful context. **(b)** Tri-state repair preserves observed content, refines uncertain renderings, and generates missing content. **(c)** Real-input depth anchors guide depth recovery from repaired RGB. The resulting RGB-D views update the reconstruction for the next round.

3.2 Tri-State Masking for Generative Repair

For a selected novel view, rendering support does not guarantee that the rendered content is reliable. Sparsely observed surfaces may exhibit blur, color shifts, or misaligned boundaries. We therefore verify rendered content against the real input views before determining which regions to preserve, refine, or generate. Specifically, we recover 3D surface points from the rendered depth $\tilde{D}_v$, project them into the real input views, and test depth agreement and local geometric consistency. Let V_v denote pixels with accepted correspondences, and let I_v^{obs} denote the corresponding colors from the real input images projected into target view v. The resulting tri-state masks are:

$$M_v^{\text{G}} = S_v V_v, \qquad M_v^{\text{P}} = S_v(1 - V_v), \qquad M_v^{\text{U}} = 1 - S_v. \tag{2}$$

The GROUNDED state contains rendered regions reliably associated with the real inputs. The PROVISIONAL state contains rendered but weakly constrained regions, while the UNCOVERED state has no valid rendering support and requires generation. Surface identities retain their first assigned labels across rounds, and correspondence verification always uses the fixed real inputs. This prevents generated content from being treated as trusted real evidence.

We use frozen WorldStereo (Zhang et al., 2026) as the camera-controlled video diffusion backbone, following its reference-image and planned-camera conditioning scheme. The grounded regions receive projected input colors, the provisional regions retain the current rendering $\tilde{I}_v$ as editable context, and the uncovered regions are generated from real reference views, neighboring context, and camera conditioning without relying on rendered RGB. These conditions give the diffusion input and editable region:

$$I_v^{\text{cond}} = M_v^{\text{G}} \odot I_v^{\text{obs}} + M_v^{\text{P}} \odot \tilde{I}_v, \qquad E_v = M_v^{\text{P}} + M_v^{\text{U}}. \tag{3}$$

To suppress generative drift in grounded regions, we conservatively downsample M_v^{G} and replace masked latents with reference latents at the same noise level at each denoising step. Let I_v^{diff} denote the decoded diffusion output. We then restore the verified input colors to obtain the repaired view:

$$\hat{I}_v = M_v^{\text{G}} \odot I_v^{\text{obs}} + E_v \odot I_v^{\text{diff}}. \tag{4}$$

The resulting $\hat{I}_v$ preserves verified input appearance while refining weakly constrained renderings and generating missing content. Figure 2 shows the combined effect from the same initial rendering. G4Splat (Ni et al., 2026) completes the missing background but leaves the foreground seating

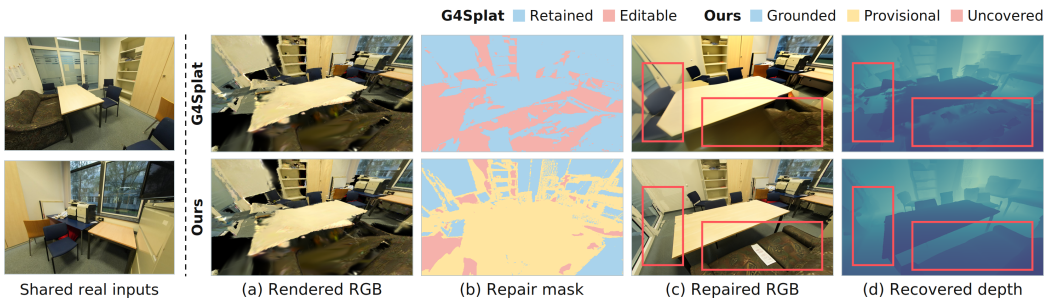

Figure 2: **Visualization of intermediate results.** Two nearby views from the five shared real inputs are shown left of the dashed line. On the right, G4Splat (top) and our method (bottom) use the same rendered input. **(a)** Rendered RGB. **(b)** Binary and tri-state repair masks. **(c)** Repaired RGB. **(d)** Recovered depth. Our method refines uncertain content while preserving observed appearance, yielding clearer foreground and under-table structure.

and under-table structure blurred, whereas our method completes the background, clarifies these structures, and retains appearance supported by the real inputs.

3.3 Observation-Anchored Geometric Completion

Generative repair may shift object boundaries, changing which surface a pixel depicts, or introduce content with no valid pre-repair depth. The rendered depth $\tilde{D}_v$ may therefore no longer describe the repaired RGB view $\hat{I}_v$. We recover depth from the repaired RGB sequence, using geometry verified against the real input views to anchor the result.

We adopt the frozen WorldMirror 2.0 model (Liu et al., 2025; Team HY-World et al., 2026) as our geometry foundation model. It receives the repaired RGB sequence, planned camera poses, and verified depth constraints from the real inputs. Appearance verification and depth anchoring require different levels of evidence. One reliable input-view correspondence can define a GROUNDED pixel, whereas a depth anchor requires consistent support from at least two sufficiently separated input cameras. Let B_v denote the accepted anchor region, D_v^{obs} the depth projected from the real inputs into view v, and D_v^{geo} the depth produced by the geometric foundation model. We estimate a scale s_v from the anchors and obtain the final depth:

$$\hat{D}_v = B_v \odot D_v^{\text{obs}} + (1 - B_v) \odot s_v D_v^{\text{geo}}. \tag{5}$$

The anchors preserve verified geometry and calibrate the model output, while the latter provides depth for refined and newly generated content. The resulting pair $(\hat{I}_v, \hat{D}_v)$ supplies RGB-D supervision at the same planned camera pose for progressive reconstruction.

3.4 Progressive Scene Reconstruction

Progressive reconstruction integrates repaired RGB-D views into a shared scene while retaining real observations as its reference. At round k, let $\mathcal{U}_k$ contain the accumulated generated views and masks, excluding each clip's real starting frame. In each round, we optimize the scene using the real inputs and accumulated generated views. Generated RGB supervises $E_v = M_v^{\text{P}} + M_v^{\text{U}}$, while depth additionally requires confidence filtering and new seeds are admitted only from reliable UNCOVERED regions. These distinct roles couple refinement with expansion: PROVISIONAL corrects existing weakly constrained content, whereas UNCOVERED adds geometric support for subsequent views. GROUNDED receives no direct synthetic target, limiting conflicts with observed content as generated supervision accumulates. Joint scene optimization then reconciles the admitted views in a common representation that supplies context for the next exploration round.

We separate real-observation preservation and generated-view completion in the expected objective:

$$\mathcal{L}_k(G, t) = (1 - p)\, \mathbb{E}_{i \sim \mathcal{O}}[\ell_{\text{real}}(G, i, t)] + p\, \mathbb{E}_{v \sim \mathcal{U}_k}[\ell_{\text{gen}}(G, v, t)]. \tag{6}$$

Here, G denotes scene parameters, t the within-round iteration, and p the fixed probability of sampling generated views. The real loss ℓ_{real} combines full-image RGB supervision, input-geometry priors, and inherited regularization. We define the generated-view loss as

$$\ell_{\text{gen}} = \lambda_{\text{rgb}}\ell_{\text{rgb}} + \lambda_d(t)\ell_d, \tag{7}$$

where ℓ_{rgb} combines L1 and SSIM losses against repaired RGB, and ℓ_d constrains rendered surface depth to repaired depth. The weights λ_{rgb} and $\lambda_d(t)$ control appearance and scheduled depth supervision. We restrict RGB supervision to E_v and admit SSIM windows only when fully contained in this region. Depth supervision additionally requires valid estimates that pass confidence filtering. We normalize the depth term by editable area to weight geometric constraints by reliable support, and apply a logarithmic penalty to attenuate large residuals.

We keep repaired RGB-D targets and masks fixed during scene optimization. Real and generated views are sampled with probabilities $1 - p$ and p, with shuffled traversal within each group. We retain earlier generated views in $\mathcal{U}_k$ to jointly constrain previously explored and newly completed regions. Fixed group probabilities maintain the frequency of real-view supervision as generated views accumulate. Appendix C.4 specifies the loss terms, weights, and schedules.

4 Experiments

We evaluate sparse-view scene reconstruction on synthetic and real datasets. Comparisons with existing methods assess overall appearance and geometry, while ablations and supplementary analyses examine observed-region fidelity, scene completeness, and the contribution of each component.

4.1 Settings

Datasets. Our main comparison uses eight Replica (Straub et al., 2019), four ScanNet++ (Yeshwanth et al., 2023), and four DL3DV (Ling et al., 2024) scenes. These comparisons use the same five inputs and 95 test views per scene across methods. Appendix A.6 additionally evaluates four Mip-NeRF 360 (Barron et al., 2022) scenes with 10 inputs and 90 test views per scene.

Metrics. We evaluate reconstruction using six metrics. PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) (VGG) assess rendering quality, while Chamfer distance (CD), F-Score, and normal consistency (NC) assess surface geometry. Appendix C.6 provides the evaluation details.

Baselines. Table 1 includes 2DGS (Huang et al., 2024), FSGS (Zhu et al., 2024), GenFusion (Wu et al., 2025c), Difix3D+ (Wu et al., 2025a), ArtiFixer (de Lutio et al., 2026), ExploreGS (Kim et al., 2025), GuidedVD (Zhong et al., 2025), and G4Splat (Ni et al., 2026). We also implement a 2DGS baseline augmented with the pretrained WorldStereo prior (Zhang et al., 2026). Reported scores are from local runs on the same split as detailed in Appendix C.6.

Implementation Details. We train for 7,000 iterations in the initial stage and in each of five subsequent progressive rounds. Each round produces a 20-frame clip with one real starting view and 19 novel views, yielding 95 generated views in total. WorldStereo repairs the RGB views, and WorldMirror 2.0 recovers their corresponding depths. In each round, we optimize the scene using the real inputs and accumulated generated views, and report results after the final round. Additional sampling and loss settings are provided in Appendix C.5.

4.2 Results

Novel-view synthesis quality. Our method combines stronger novel-view synthesis with better preservation of observed appearance. It achieves the best or second-best results on every rendering metric in Table 1. Figure 3 shows that other baselines render recognizable content but often blur objects already visible in the input views. By contrast, our method retains the shelving and dining furniture in Replica and the bed and chair in ScanNet++, while completing content beyond the input-view coverage. The outdoor examples show similar improvements around the garden table and foreground bin. A region-wise comparison in the appendix further supports this result, with the best rendering quality in both observed and unobserved areas (Table A2). Moreover, our method maintains its lead across input-view counts, indicating robustness to input sparsity (Table A4).

Table 1: **Quantitative comparison from five input views.** Our method improves rendering fidelity and surface reconstruction across the evaluated datasets. Colors mark first, second, and third.

Dataset	Method	Rendering			Surface geometry		
		PSNR ↑	SSIM ↑	LPIPS ↓	CD ↓	F-Score ↑	NC ↑
Replica	2DGS	14.97	0.652	0.417	0.0396	0.240	0.587
	FSGS	19.32	0.766	0.335	0.0193	0.434	0.579
	GenFusion	13.73	0.635	0.402	0.0345	0.298	0.538
	Difix3D+	17.30	0.712	0.364	0.0338	0.271	0.557
	ArtiFixer	17.71	0.760	0.355	0.0302	0.300	0.563
	ExploreGS	19.87	0.772	0.338	0.0163	0.495	0.632
	WorldStereo	14.34	0.638	0.463	0.0389	0.227	0.584
	GuidedVD	22.30	0.800	0.275	0.0199	0.362	0.710
	G4Splat	23.12	0.826	0.218	0.0104	0.734	0.816
	Ours	25.46	0.856	0.186	0.0057	0.864	0.876
ScanNet++	2DGS	12.02	0.614	0.494	0.0884	0.072	0.538
	FSGS	15.79	0.788	0.364	0.0231	0.310	0.577
	GenFusion	8.80	0.447	0.486	0.0949	0.088	0.522
	Difix3D+	14.64	0.722	0.422	0.0443	0.183	0.538
	ArtiFixer	15.08	0.771	0.422	0.0258	0.308	0.579
	ExploreGS	17.11	0.797	0.361	0.0260	0.312	0.565
	WorldStereo	12.49	0.663	0.477	0.0726	0.104	0.549
	GuidedVD	21.09	0.840	0.287	0.0211	0.278	0.719
	G4Splat	20.89	0.829	0.283	0.0103	0.724	0.800
	Ours	22.15	0.830	0.282	0.0072	0.793	0.830
DL3DV	2DGS	15.23	0.560	0.430	0.0370	0.276	0.585
	FSGS	17.05	0.635	0.388	0.0227	0.371	0.584
	GenFusion	16.00	0.638	0.436	0.0360	0.286	0.578
	Difix3D+	16.35	0.597	0.366	0.0292	0.365	0.573
	ArtiFixer	18.20	0.678	0.385	0.0347	0.278	0.569
	ExploreGS	17.88	0.651	0.390	0.0213	0.430	0.593
	WorldStereo	15.86	0.591	0.431	0.0296	0.356	0.585
	GuidedVD	19.08	0.692	0.323	0.0194	0.508	0.627
	G4Splat	17.89	0.685	0.335	0.0242	0.589	0.710
	Ours	19.03	0.699	0.313	0.0143	0.660	0.727

Geometry reconstruction quality. Beyond its rendering gains, our method delivers stronger surface reconstruction across the evaluated datasets, achieving the best results on all surface geometry metrics in Table 1. The paired RGB and mesh views in Figure 3 reveal the corresponding structural differences. In these examples, Difix3D+ and GuidedVD recover recognizable scene content but produce fragmented or distorted surfaces. G4Splat improves surface coherence, while our method more clearly resolves the furniture and round tabletop. In the square, it recovers a better-defined foreground bin and a more continuous ground surface. These meshes more closely match the object shapes in the paired RGB views. Additional results in the appendix show the lowest depth errors in both observed and unobserved regions (Table A3). Further mesh examples in Figure A5 and the overall depth comparison on Replica and ScanNet++ in Table A1 support the same conclusion.

4.3 Ablation Studies

We conduct cumulative ablations on Replica to assess generative priors (GP), tri-state observation constraints (TC), progressive camera exploration (PE), and observation-anchored geometric completion (OC). All variants use the same real input views and evaluation split. Table 2 summarizes the results. Detailed settings are provided in Appendix A.7.

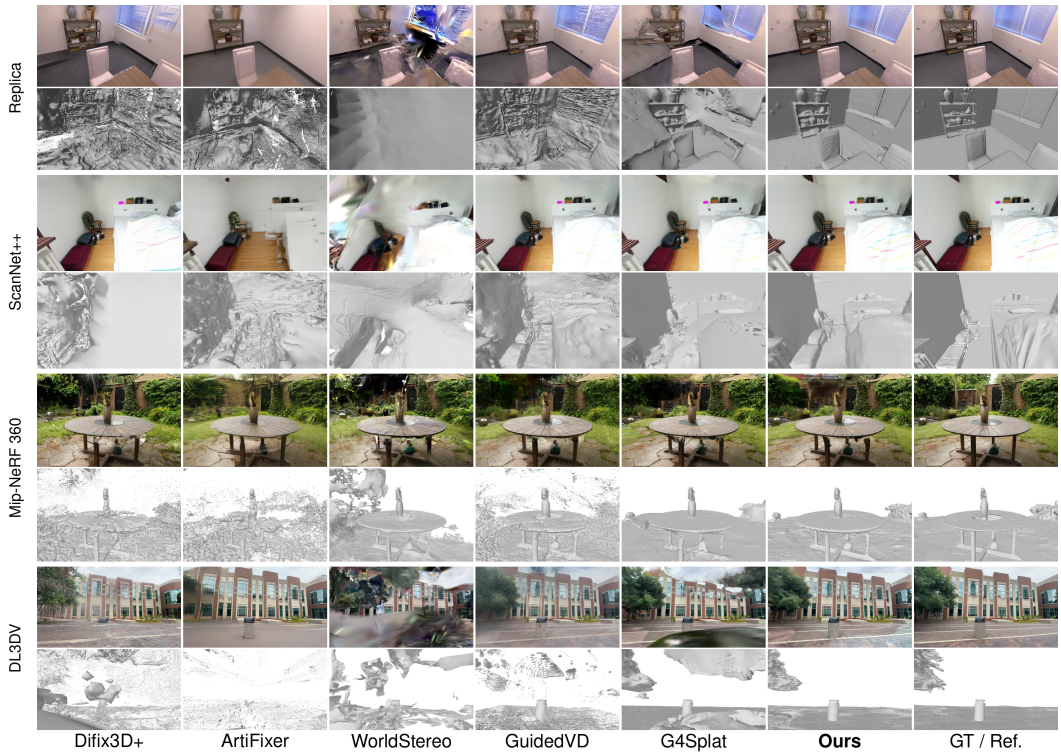

Figure 3: **Qualitative comparison.** Our method achieves better rendering quality and geometry reconstruction across indoor and outdoor scenes. RGB renderings (top) and reconstructed meshes (bottom) are shown alongside the corresponding references (rightmost column).

Table 2: **Cumulative component ablation on Replica.** GP: generative priors, TC: tri-state observation constraints, PE: progressive camera exploration, OC: observation-anchored geometric completion. Rows progressively enable the components. The last row is our complete method.

Components				Rendering			Depth		Surface geometry		
GP	TC	PE	OC	PSNR ↑	SSIM ↑	LPIPS ↓	AbsRel ↓	RMSE ↓	CD ↓	F-Score ↑	NC ↑
×	×	×	×	18.92	0.783	0.207	0.080	0.409	0.0074	0.834	0.859
✓	×	×	×	23.04	0.830	0.194	0.037	0.172	0.0070	0.823	0.844
✓	✓	×	×	24.51	0.847	0.186	0.033	0.154	0.0068	0.836	0.856
✓	✓	✓	×	25.25	0.853	0.185	0.029	0.127	0.0065	0.832	0.855
✓	✓	✓	✓	25.46	0.856	0.186	0.026	0.114	0.0057	0.864	0.876

1. **Generative priors.** Compared with reconstruction from real inputs alone, adding GP improves rendering fidelity and depth accuracy. Generated views introduce appearance constraints beyond the sparse input viewpoints and provide additional visual evidence for scene optimization. However, F-Score and NC decrease slightly despite the substantial rendering gains. This distinction suggests that generated RGB is effective for appearance completion but does not, by itself, provide sufficiently reliable supervision for surface reconstruction. These results motivate controlling where generated views contribute to training and supplementing their appearance information with observation-anchored geometric constraints for uncertain and uncovered regions.
2. **Tri-state observation constraints.** Adding TC to generative priors improves every reported metric. The grounded, provisional, and uncovered states separate appearance supported by real observations from uncertain renderings and missing content. This allows repair to preserve verified input appearance, refine weakly constrained structures, and generate content where rendering support is absent. The same distinction also prevents generated-view losses from directly super-

vising grounded regions during scene optimization. Figure 2(a–c) illustrates the resulting preservation and refinement in the complete repair pipeline, including clearer foreground seating and under-table structure. The gains over GP alone show the benefit of applying observation-based constraints throughout generation and reconstruction.

3. **Progressive camera exploration.** Enabling PE further improves rendering and depth accuracy relative to fixed trajectories. By planning from the updated reconstruction, PE directs subsequent views toward remaining uncovered regions as reconstructed content becomes available as context for repair. Figure A3 illustrates this progression through reduced coverage gaps around furniture and room boundaries. CD also improves, while F-Score and NC remain close to their values without PE. These results indicate that progressive exploration extends supervision into missing regions and improves novel-view reconstruction while maintaining comparable surface quality.
4. **Observation-anchored geometric completion.** Enabling OC produces the strongest depth and surface reconstruction in the ablation while largely preserving rendering fidelity. Depth recovered from repaired RGB provides explicit geometric supervision for refined and generated content. Anchors verified against real input views align the recovered depth with observed geometry, while reliable seeds in UNCOVERED regions initialize surfaces lacking geometric support. The gains in depth and surface metrics show that OC combines anchored depth supervision with surface initialization to incorporate generated content as more accurate and complete geometry.

4.4 FURTHER ANALYSIS

Reconstruction beyond the input views. Figure A2 evaluates reconstruction on reference surfaces initially unseen from the real inputs. Our method recovers more of these surfaces on Replica and ScanNet++, demonstrating that its geometric gains extend to regions lacking direct input coverage. Figure A4 locates this recovery in broader areas of the floor and room boundaries around furniture in Replica and the bed in ScanNet++. These recovered surfaces improve the completeness of the surrounding scene structure. Figure A3 traces the complementary expansion of camera coverage over successive rounds. Camera coverage identifies surfaces reached by the explored views, while final surface recovery evaluates their reconstruction in the optimized geometry. Together, expanded camera coverage and broader final recovery show that our method more fully explores regions beyond the original observations and extends the reconstruction into them.

Performance under different input conditions. Table A4 compares different numbers of real inputs using the same test views. Each setting follows the prescribed protocol, with one reconstruction round per real input and the corresponding accumulation of generated supervision. Our method achieves stronger rendering and surface reconstruction than the compared baselines across these conditions. Its geometric advantage is particularly pronounced under the sparsest inputs, where real observations provide limited constraints on scene structure. With additional real inputs, image fidelity and surface accuracy improve together. More input views also expand opportunities for appearance verification and depth anchoring across a wider range of scene regions. The real inputs remain the references for verifying generated content throughout reconstruction. Since the number of rounds also increases with the input count, these results characterize the complete framework under each input condition. Its persistent advantage shows that the method remains effective as the balance between direct observations and generated supervision changes.

5 CONCLUSION

We present **GenAnchor**, an observation-anchored framework for progressive sparse-view scene reconstruction. Coverage-aware exploration exposes missing content, while tri-state observation constraints guide generative repair to preserve observed appearance, refine uncertain renderings, and complete uncovered regions. Depth recovered from repaired RGB is aligned with verified input-depth anchors, providing consistent RGB-D supervision for subsequent scene optimization. Experiments on Replica, ScanNet++, DL3DV, and Mip-NeRF 360, demonstrate stronger rendering and surface reconstruction than competitive baselines. Regional analyses further show improvements in initially unobserved areas while maintaining fidelity to observed content. Together, these results highlight the value of anchoring generative reconstruction to persistent real observations.

REFERENCES

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5470–5479, June 2022.

Riccardo de Lutio, Tobias Fischer, Yen-Yu Chang, Yuxuan Zhang, Jay Zhangjie Wu, Xuanchi Ren, Tianchang Shen, Katarina Tothova, Zan Gojcic, and Haithem Turki. ArtiFixer: Enhancing and extending 3D reconstruction with auto-regressive diffusion models. In *SIGGRAPH*, 2026.

Antoine Guédon, Tomoki Ichikawa, Kohei Yamashita, and Ko Nishino. MAtCha gaussians: Atlas of charts for high-quality geometry and photorealism from sparse views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2D gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH Conference Papers*, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Minsu Kim, Subin Jeon, In Cho, Mijin Yoo, and Seon Joo Kim. ExploreGS: Explorable 3D scene reconstruction with virtual camera samplings and diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27042–27051, 2025.

Hanyang Kong, Xingyi Yang, and Xinchao Wang. Generative sparse-view gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26745–26755, 2025.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3D with MASt3R, 2024.

Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. DNGaussian: Optimizing sparse-view 3D gaussian radiance fields with global-local depth normalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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*, pp. 22160–22169, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ling_DL3DV-10K_A_Large-Scale_Scene_Dataset_for_Deep_Learning-based_3D_Vision_CVPR_2024_paper.html.

Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. ReconX: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024.

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. doi: 10.48550/arXiv.2510.10726. URL https://arxiv.org/abs/2510.10726v1.

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*, 2020.

Junfeng Ni, Yixin Chen, Zhifei Yang, Yu Liu, Ruijie Lu, Song-Chun Zhu, and Siyuan Huang. G4Splat: Geometry-guided gaussian splatting with generative prior. In *International Conference on Learning Representations*, 2026.

Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. RegNeRF: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5480–5490, 2022. URL https://m-niemeyer.github.io/regnerf/.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Yiyang Shen, Yin Yang, Kun Zhou, and Tianjia Shao. Sparse-to-complete: From sparse image captures to complete 3d scenes. *arXiv preprint arXiv:2605.05664*, 2026.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Jimin Tang, Wenyuan Zhang, Junsheng Zhou, Zian Huang, Kanle Shi, Shenkun Xu, Yu-Shen Liu, and Zhizhong Han. VidSplat: Gaussian splatting reconstruction with geometry-guided video diffusion priors. In *SIGGRAPH 2026 Conference Papers*, 2026. doi: 10.1145/3799902.3811138.

Team HY-World, Chenjie Cao, Xuhui Zuo, Zhenwei Wang, Yisu Zhang, Junta Wu, Zhenyang Liu, Yuning Gong, Yang Liu, Bo Yuan, Chao Zhang, Coopers Li, Dongyuan Guo, Fan Yang, Haiyu Zhang, Hang Cao, Jianchen Zhu, Jiaxin Lin, Jie Xiao, Jihong Zhang, Junlin Yu, Lei Wang, Lifu Wang, Lilin Wang, Linus, Minghui Chen, Peng He, Penghao Zhao, Qi Chen, Rui Chen, Rui Shao, Sicong Liu, Wangchen Qin, Xiaochuan Niu, Xiang Yuan, Yi Sun, Yifei Tang, Yifu Sun, Yihang Lian, Yonghao Tan, Yuhong Liu, Yuyang Yin, Zhiyuan Min, Tengfei Wang, and Chunchao Guo. HY-World 2.0: A multi-modal world model for reconstructing, generating, and simulating 3D worlds. *arXiv preprint arXiv:2604.14268*, 2026. doi: 10.48550/arXiv.2604.14268. URL https://arxiv.org/abs/2604.14268.

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

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024b.

Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. DIFIX3D+: Improving 3D reconstructions with single-step diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26024–26035, 2025a.

Jiang Wu, Rui Li, Yu Zhu, Rong Guo, Jinqiu Sun, and Yanning Zhang. Sparse2DGS: Geometry-prioritized gaussian splatting for surface reconstruction from sparse views. *arXiv preprint arXiv:2504.20378*, 2025b.

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Hołyński. ReconFusion: 3D reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21551–21561, 2024.

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

Jamie Wynn and Daniyar Turmukhambetov. DiffusioNeRF: Regularizing neural radiance fields with denoising diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4180–4189, 2023.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Xingyilang Yin, Qi Zhang, Jiahao Chang, Ying Feng, Qingnan Fan, Xi Yang, Chi-Man Pun, Huaqi Zhang, and Xiaodong Cun. GSFixer: Improving 3D gaussian splatting with reference-guided video diffusion priors. *arXiv preprint arXiv:2508.09667*, 2025.

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

Jiawei Zhang, Jiahe Li, Xiaohan Yu, Lei Huang, Lin Gu, Jin Zheng, and Xiao Bai. CoR-GS: Sparse-view 3D gaussian splatting via co-regularization. In *European Conference on Computer Vision*, 2024.

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*, 2018.

Yisu Zhang, Chenjie Cao, Tengfei Wang, Xuhui Zuo, Junta Wu, Jianke Zhu, and Chunchao Guo. WorldStereo: Bridging camera-guided video generation and scene reconstruction via 3D geometric memories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40327–40339, 2026.

Yingji Zhong, Zhihao Li, Dave Zhenyu Chen, Lanqing Hong, and Dan Xu. Taming video diffusion prior with scene-grounding guidance for 3D gaussian splatting from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6133–6143, 2025.

Zhizhuo Zhou and Shubham Tulsiani. SparseFusion: Distilling view-conditioned diffusion for 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12588–12597, 2023.

Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. FSGS: Real-time few-shot view synthesis using gaussian splatting. In *European Conference on Computer Vision*, 2024.

APPENDIX

This appendix supplements the method and experimental protocol of the main paper. Section A organizes the supplementary experimental comparisons. Section B reviews the scene representation and initialization. Section C specifies camera exploration, tri-state repair, depth reconstruction, optimization, and evaluation. Section D discusses the limitations of the framework, and Section E contains the AI use statement.

A ADDITIONAL EXPERIMENTAL RESULTS

We supplement the main comparison with rendered-depth evaluation, regional fidelity, progressive coverage, and surface recovery on eight Replica and four ScanNet++ scenes. We also examine different input counts on Replica and ten-input reconstruction on indoor and outdoor Mip-NeRF 360 scenes.

Evaluation conventions. All comparisons use the metrics introduced in Section 4.1. CD averages the two directional mean nearest-neighbor Euclidean distances without squaring. NC averages absolute dot products of unit nearest-neighbor face normals in both directions. The diagonal D is computed from the axis-aligned bounding box of fixed reference samples within the evaluation extent, without padding. It is shared by all methods and ablations, independently of their predictions. CD is normalized by D before averaging, including the input-count comparison. Tables use the compact labels CD, F-Score, and NC. F-Score uses a distance threshold of $0.01D$. F-Score and NC are reported on a $[0, 1]$ scale. We additionally report AbsRel and RMSE for depth. RMSE is in metres on Replica and ScanNet++, and normalized by D on DL3DV and Mip-NeRF 360. Reference geometry serves only for evaluation and does not provide reconstruction or planning inputs.

Image scores are averaged over valid test views and then equally over scenes. Surface scores are averaged equally over scenes. Colors follow the ranking convention of Table 1. Rankings use displayed precision, except in Table A5, where the unrounded means determine the ordering. Displayed ties otherwise share a rank. Appendix C.6 details reference scope, depth evaluation, and surface extraction.

Regional and coverage measures. Let S be the fixed reference surface samples, V_0 the subset visible from the five real inputs under calibrated projection and depth-based occlusion checks, and $U = S \setminus V_0$. Missing reference geometry is excluded. For samples P from the final predicted mesh, unobserved-surface recovery is

$$R_U(\tau) = \frac{1}{|U|} \sum_{x \in U} \mathbf{1}\left[\min_{p \in P} \|x - p\|_2 < \tau\right]. \tag{A1}$$

We use $\tau = 0.01D$ and plot recovery over $\tau/D \in [0.001, 0.02]$. The meshes, samples, and evaluation extent match Table 1. In surface plots, "All" evaluates the entire set S and "Unseen" evaluates U. Camera coverage measures the fraction of S visible from the real inputs and generated cameras available at each stage. All fractions use the full reference sample set, including samples omitted by visualization cutaways, and are averaged equally over scenes.

A.1 RECONSTRUCTION ACROSS DATASETS

Table 1 compares complete-scene reconstruction on Replica, ScanNet++, and DL3DV. The following experiments examine how these reconstructions preserve observed content, complete unseen regions, and recover scene surfaces.

Rendered-depth accuracy. Table A1 shows that our method improves depth accuracy on both Replica and ScanNet++. Anchoring depth recovery to verified observations helps maintain spatial structure as repaired views extend supervision beyond the sparse inputs.

Table A1: **Rendered-depth accuracy from five input views.** Our method improves depth reconstruction on Replica and ScanNet++. Methods and test views match Table 1.

	Replica		ScanNet++	
Method	AbsRel ↓	RMSE ↓	AbsRel ↓	RMSE ↓
2DGS	0.368	1.586	0.482	0.921
FSGS	0.198	0.589	0.339	0.595
GenFusion	0.315	0.984	0.593	0.953
Difix3D+	0.365	1.084	0.575	1.015
ArtiFixer	0.315	1.166	0.338	0.644
ExploreGS	0.136	0.399	0.353	0.536
WorldStereo	0.347	1.166	0.484	0.896
GuidedVD	0.143	0.408	0.172	0.327
G4Splat	0.086	0.370	0.074	0.163
Ours	0.026	0.114	0.044	0.101

A.2 Observed-Region Fidelity and Unobserved-Region Completion

Evaluation. We compare Difix3D+, ArtiFixer, WorldStereo, GuidedVD, G4Splat, and our method using the shared visibility partition defined above. Regional scores average nonempty test views before averaging scenes. SSIM uses regional pixel centers and LPIPS uses region-weighted feature maps, retaining the original full-image context at boundaries.

Appearance and depth. Tables A2 and A3 show that the benefits of our method extend to both observed and unobserved regions. Preserving verified input colors helps retain appearance in supported regions, while selective repair improves completion beyond their coverage. The depth comparisons further support the role of observation anchors in constraining generated content. These results show that scene completion can improve alongside the reconstruction of already observed structures.

Table A2: **Appearance in observed and unobserved regions.** Comparison of five-input reconstructions using the same visibility partition.

	Observed			Unobserved		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Replica						
Difix3D+	18.30	0.758	0.339	15.45	0.610	0.455
ArtiFixer	19.25	0.792	0.337	16.44	0.676	0.435
WorldStereo	14.68	0.652	0.457	13.28	0.525	0.533
GuidedVD	23.83	0.826	0.255	17.89	0.632	0.464
G4Splat	24.96	0.848	0.196	18.01	0.649	0.413
Ours	27.48	0.874	0.163	19.57	0.682	0.392
ScanNet++						
Difix3D+	15.25	0.772	0.377	12.69	0.589	0.523
ArtiFixer	17.06	0.823	0.356	13.84	0.642	0.496
WorldStereo	12.63	0.681	0.463	11.86	0.526	0.564
GuidedVD	22.42	0.866	0.258	16.76	0.676	0.462
G4Splat	22.43	0.857	0.255	16.14	0.647	0.457
Ours	23.39	0.852	0.258	17.92	0.680	0.435

Figure A1 illustrates these effects across additional test views. The Replica examples show ceiling completion and better preservation of the window and dining-room furniture. The ScanNet++ examples extend the comparison to the floor around a bed and the chairs and floor of a meeting room. Our method reduces foreground artifacts in both RGB and depth while retaining the furniture layout. These examples illustrate observation preservation and completion, while the tables report all evaluated scenes.

Table A3: **Depth in observed and unobserved regions.** Reconstructions and visibility partitions match Table A2.

	Observed		Unobserved	
Method	AbsRel ↓	RMSE ↓	AbsRel ↓	RMSE ↓
Replica				
Difix3D+	0.358	1.091	0.396	1.018
ArtiFixer	0.302	1.148	0.362	0.961
WorldStereo	0.342	1.175	0.380	1.010
GuidedVD	0.136	0.391	0.201	0.526
G4Splat	0.087	0.367	0.111	0.381
Ours	0.022	0.096	0.057	0.207
ScanNet++				
Difix3D+	0.571	1.006	0.571	0.963
ArtiFixer	0.338	0.631	0.400	0.731
WorldStereo	0.483	0.888	0.457	0.786
GuidedVD	0.162	0.303	0.237	0.426
G4Splat	0.067	0.146	0.142	0.272
Ours	0.042	0.090	0.079	0.175

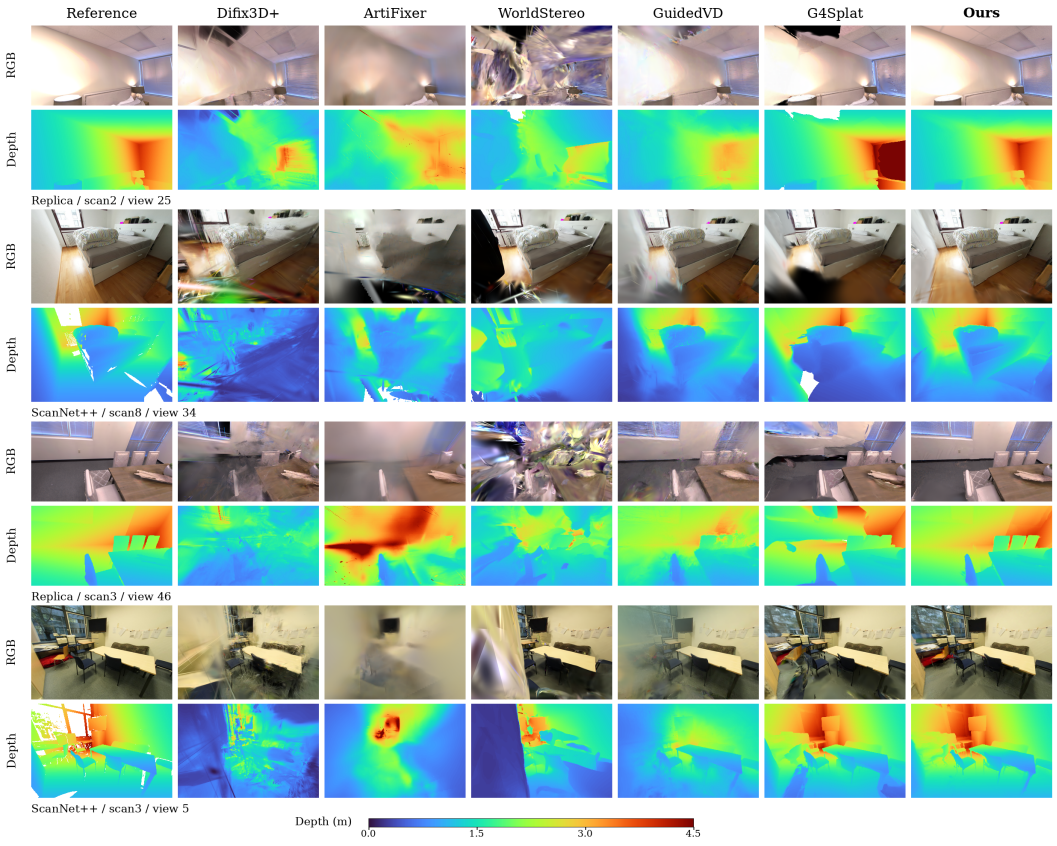

Figure A1: **Regional reconstruction from five input views.** Comparisons on Replica and ScanNet++ with RGB above depth in each scene group. Our method preserves observed structure while completing missing surfaces. All methods share the camera and depth color scale.

A.3 Progressive Coverage and Unobserved-Surface Recovery

Recovery beyond the input views. Figure A2 compares recovery of reference surfaces that are not visible from the real inputs. Our method leads all compared baselines on Replica and ScanNet++ across the tested distance thresholds. At tighter thresholds, recovered samples must lie closer to reconstructed geometry, so the higher recovery supports accurate placement of newly completed surfaces. The sustained advantage as the threshold increases also indicates broader completion. These curves therefore link the improvement to both the extent and spatial accuracy of recovered unseen surfaces across the evaluated scenes.

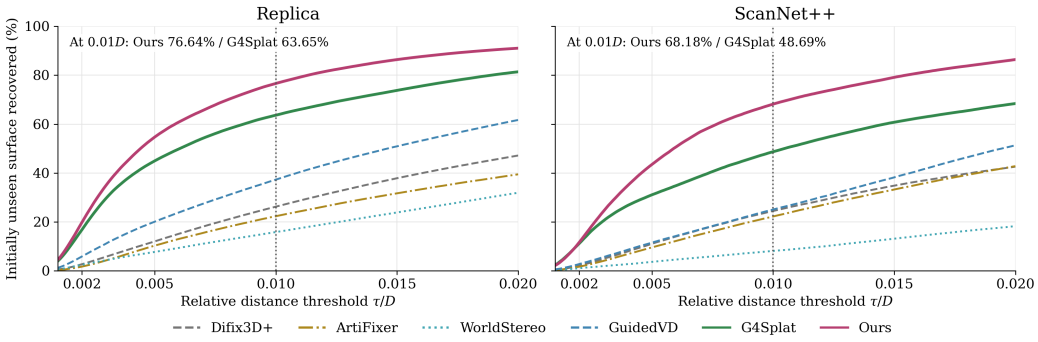

Figure A2: **Final recovery of initially unobserved surface.** Comparison on Replica (left) and ScanNet++ (right). Our method recovers more reference surface over a broad range of recovery tolerances.

Figure A3 shows the expansion of camera coverage through progressive completion, following the scene-view presentation of Shen et al. (2026). Later rounds reach gaps around the bed, floor, and room boundaries that are missed by the initial observations. This behavior illustrates how replanning from the updated reconstruction directs new views toward the remaining uncovered regions.

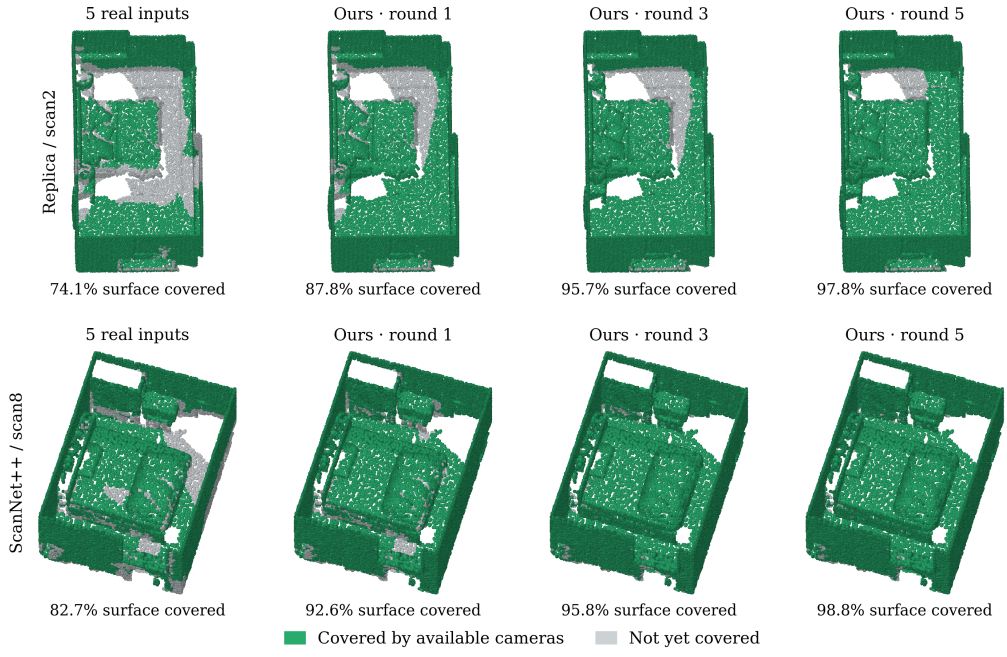

Figure A3: **Scene coverage expands over progressive completion.** Replica scan2 (top) and ScanNet++ scan8 (bottom). Columns show initialization and rounds 1, 3, and 5. Green indicates covered reference surface and gray indicates the remaining gaps. Each row uses a fixed bird's-eye cutaway.

Surface recovery across methods. Figure A4 locates the recovery gains in Replica scan2 and ScanNet++ scan8. Compared with the baselines, our method recovers broader, more continuous regions of the floor and room boundaries around the furniture in Replica and the bed in ScanNet++. Its higher recovery on both the full reference surface and the initially unseen subset shows that completing missing regions improves the overall scene structure. Figure A3 tracks which surfaces become visible to the accumulated cameras, while Figure A4 measures whether those surfaces are recovered by the final geometry. Together with Figure A2, these comparisons demonstrate that the expanded camera coverage is accompanied by more complete reconstruction of previously unseen surfaces. For display, the upper portion of the reference surface is cut away and the samples are spatially subsampled identically across methods.

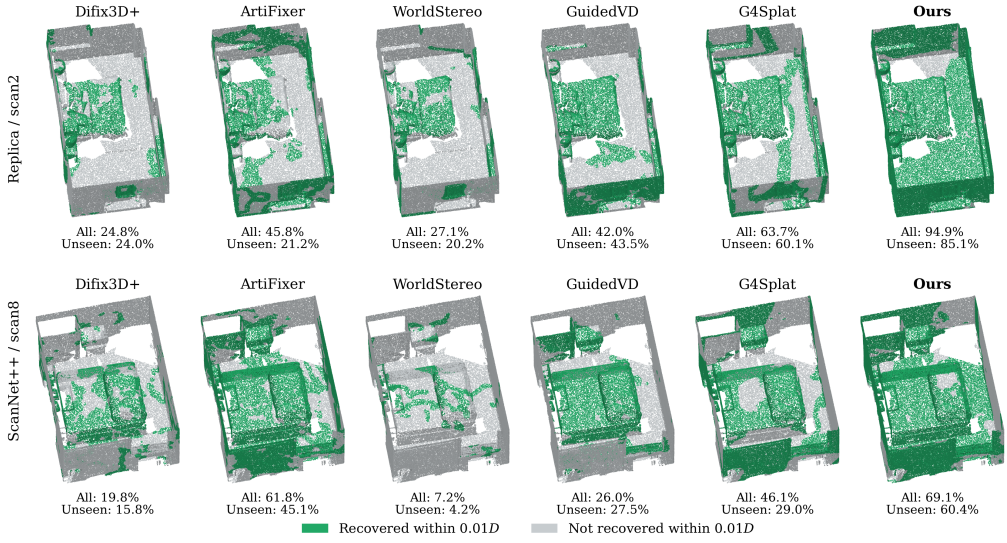

Figure A4: **Final surface recovery across reconstruction methods.** Replica scan2 (top) and ScanNet++ scan8 (bottom). Green and gray denote recovered and unrecovered reference surface, respectively. All columns share the reference samples and projection.

A.4 Additional Qualitative Comparisons

Figure A5 extends the comparison to additional indoor and outdoor scenes. Our method preserves window and furniture boundaries in the indoor examples and recovers coherent object surfaces around the bicycle and kitchen toy excavator. The agreement between the RGB and mesh comparisons supports improved reconstruction across these different scene structures.

A.5 Performance under Different Numbers of Input Views

Following G4Splat (Ni et al., 2026), we compare nested sets of 5, 10, and 15 real inputs on all eight Replica scenes (Table A4). Rendering uses the same 85 test views across input counts. The five- and ten-input scores are recomputed on this common subset, so they differ from evaluations over the complete test sets. Our method uses one progressive round per real input, with the same initialization, per-round optimization, sampling probabilities, and loss weights. Surface extraction follows each method's fixed protocol.

Our method improves appearance and surface reconstruction as more real observations become available, maintaining an advantage over the compared baselines across the input settings. Additional observations provide broader support for verification and depth anchoring, while progressive completion addresses the remaining gaps. The agreement between rendering and surface improvements supports the use of the framework at different levels of input sparsity.

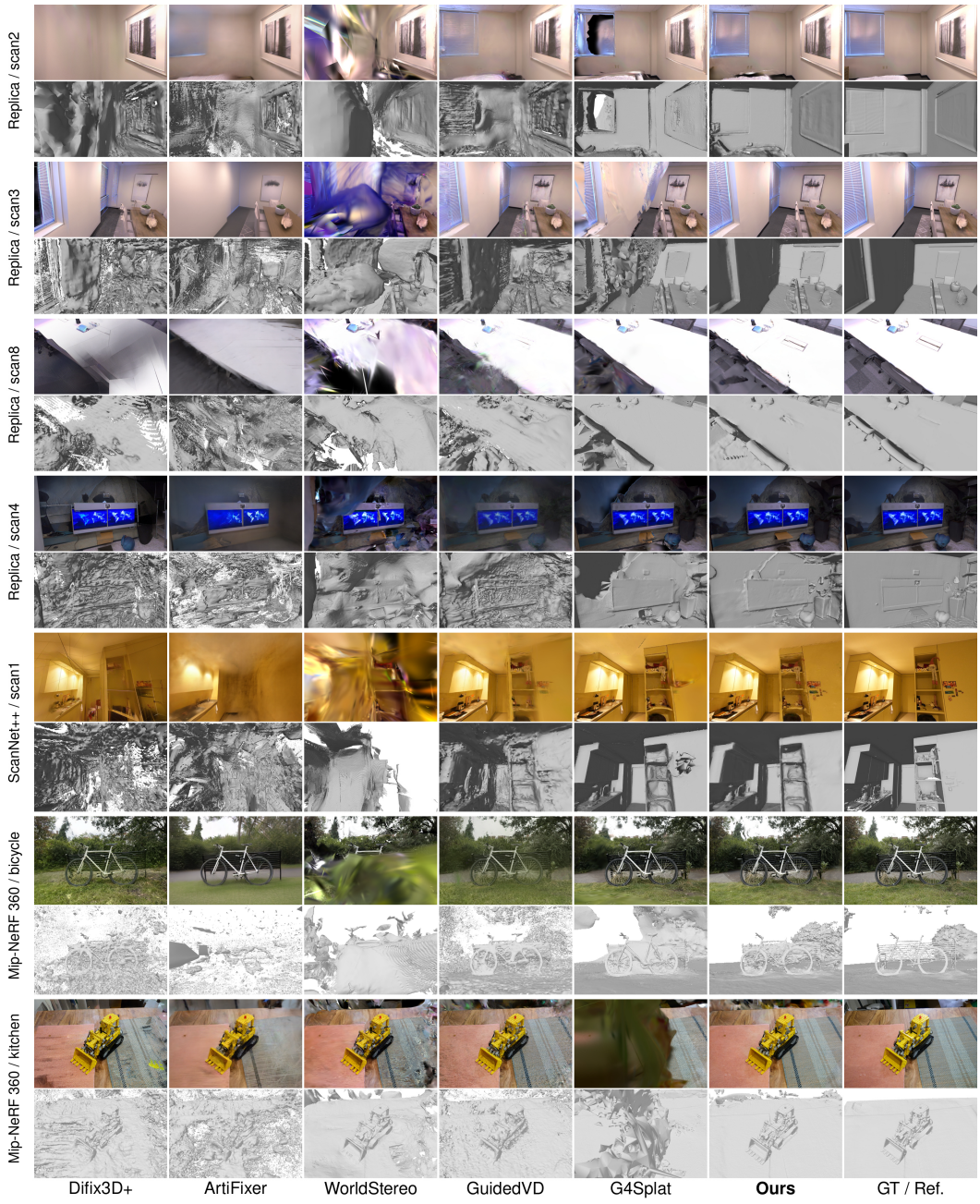

Figure A5: **Additional qualitative comparisons.** RGB outputs (top) and reconstructed meshes (bottom) for indoor and outdoor scenes, with shared camera viewpoints and neutral mesh shading.

A.6 Ten-Input Reconstruction on Mip-NeRF 360

Table A5 compares bicycle and garden as outdoor scenes and kitchen and room as indoor scenes. Every method uses the same ten real inputs and 90 test views at the stored image resolution. Our method runs ten progressive rounds with a common depth scale calibrated against verified real anchors within each generated clip. Sampling probabilities, loss weights, and the optimization schedule remain fixed.

Our method improves rendering and surface reconstruction across this collection of indoor and outdoor scenes. These results extend the comparison to more varied scene structure, where completion

Table A4: **Reconstruction with different numbers of input views on Replica.** Our method improves appearance and surface reconstruction across the input settings.

Method	Reconstruction (CD ↓ / NC ↑)			Rendering (PSNR ↑ / LPIPS ↓)		
	5 views	10 views	15 views	5 views	10 views	15 views
WorldStereo	0.0389 / 0.584	0.0231 / 0.681	0.0180 / 0.744	14.45 / 0.461	19.09 / 0.337	22.24 / 0.275
GuidedVD	0.0199 / 0.710	0.0125 / 0.793	0.0109 / 0.821	22.38 / 0.273	27.08 / 0.203	29.55 / 0.174
G4Splat	0.0104 / 0.816	0.0043 / 0.909	0.0035 / 0.920	23.22 / 0.215	29.25 / 0.152	32.06 / 0.122
Ours	0.0057 / 0.876	0.0037 / 0.913	0.0032 / 0.927	25.54 / 0.184	31.02 / 0.107	33.48 / 0.082

Table A5: **Ten-input reconstruction on Mip-NeRF 360.** Comparison across bicycle, garden, kitchen, and room.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	AbsRel ↓	RMSE ↓	CD ↓	F-Score ↑	NC ↑
WorldStereo	15.22	0.426	0.535	0.221	0.095	0.032	0.408	0.610
GuidedVD	17.77	0.530	0.426	0.166	0.058	0.014	0.644	0.610
G4Splat	17.31	0.522	0.476	0.201	0.084	0.019	0.616	0.690
Ours	19.02	0.569	0.367	0.144	0.051	0.013	0.670	0.713

must remain consistent with sparse observations despite broader changes in visibility. Calibrating generated depth against real anchors supports its integration into a common scene representation throughout the progressive loop.

A.7 Ablation Protocol

Table 2 combines rendering, depth, and surface metrics for five cumulative configurations on all eight Replica scenes. Every configuration uses the same five real inputs with real-input RGB and geometry supervision. We train for 7,000 iterations in the initial stage and in each of five subsequent progressive rounds. The first row uses only real inputs. GP adds generative priors on fixed trajectories, sampled uniformly within the initial model's safe, feasible region. Binary masks preserve rendered content and restrict generation and supervision to uncovered pixels. TC retains those trajectories and adds the tri-state partition, retention of each surface identity's first assigned label, and protection of verified input colors. PE then replans from the latest reconstruction using uncovered-frontier scores. The first four configurations use no generated depth supervision or generated geometry seeds. OC adds the complete anchored depth-recovery, geometry-supervision, and seed-injection pipeline.

All configurations with generation use WorldStereo, diffusion seed 1024, and five 20-frame clips, yielding 95 generated training views after excluding the real start frames. Generated views are sampled with probability 0.6. The generated and real RGB weights are 0.1 and 2.0. The generated geometry weight is zero without OC and 0.5 with OC. Each scene is evaluated on the same 95 test views. Surface evaluation uses the shared reference, extent, sampling, and TSDF settings in Appendix C.6. The real-only configuration fuses 100 original calibrated cameras. Configurations with generation also fuse their own 95 generated cameras. All configurations follow the shared evaluation conventions in Appendix A.

Full-model run selection. The aggregate scores of our method in Tables 1, 2, and A1 use the original accepted runs for seven Replica scenes. For scan5, all eight metrics come from the second of three complete reruns, selected for the highest test-view PSNR after all three test results were available. Each rerun repeated initialization and all five rounds with training seed 0 and diffusion seed 1024. This is a post hoc run selection, not an average over repeated trials. The final surface-recovery analyses use these same accepted reconstructions. Regional image comparisons, progressive camera coverage, and other qualitative analyses retain their originally recorded reconstructions and measurements.

B BACKGROUND

Our method uses a differentiable surface representation for reconstruction and fixed input geometry as the reference for observation verification.

B.1 2D GAUSSIAN SPLATTING

We represent the scene with oriented Gaussian surfels following 2DGS (Huang et al., 2024). Differentiable rasterization provides RGB, opacity, and depth for camera exploration and reconstruction. For camera v, let $w_j(x)$ be the front-to-back compositing weight of surfel j at pixel x, including its opacity and the transmittance of preceding surfels. With color $c_j(v)$ and camera-space depth $z_j(x)$, the rendered quantities are

$$\widetilde{I}_v(x) = \sum_j w_j(x) c_j(v), \qquad a_v(x) = \sum_j w_j(x), \qquad \widetilde{D}_v(x) = \frac{\sum_j w_j(x) z_j(x)}{a_v(x)}. \tag{A2}$$

We use a black rendering background. Camera planning uses this expected depth where opacity is nonzero. Valid support additionally requires finite opacity of at least 0.20 and finite positive depth. This support describes where the scene renders content, while verification against real inputs determines whether that content is reliable.

B.2 SCENE INITIALIZATION

Following G4Splat (Ni et al., 2026), we obtain input geometry through MASt3R correspondences (Leroy et al., 2024), MAtCha chart alignment (Guédon et al., 2025), and plane refinement. Backprojecting the refined depths initializes the scene and supplies geometric priors for optimization. The original images, cameras, and depths remain fixed: they define the reference set for GROUNDED correspondences and depth anchors. Generated views expand the training supervision but never become reference observations. This separation keeps verification tied to captured evidence throughout progressive reconstruction.

C IMPLEMENTATION DETAILS

This section complements Section 3 with the criteria that govern camera selection, observation verification, depth recovery, and scene optimization. The original images, cameras, and input depths form a fixed reference set throughout the progressive reconstruction.

C.1 COVERAGE-AWARE CAMERA EXPLORATION

As described in Section 3.1, we formulate camera exploration as a constrained trajectory search that maximizes the coverage of unreconstructed content while retaining sufficient rendered context. Each real input anchors a clip $\tau = (v_0, \ldots, v_{L-1})$ of length $L = 20$, with its measured camera retained as v_0. Starting from the first input, subsequent anchors are ordered by farthest-first traversal of their camera centers. The remaining $L - 1$ views are selected using the current 2DGS and the anchor intrinsics, rescaled to the 832×480 conditioning resolution.

Search domain and motion parameterization. Camera motion is expressed in a common vertical frame. For input coordinates without a consistent vertical alignment, a rigid transform aligns the mean image-up direction of the observed cameras with the planning vertical axis and centers the frame at their mean position. This transform preserves scene scale, and selected poses are mapped back to the reconstruction coordinates. Translational candidates lie within the axis-aligned envelope of the input camera centers, expanded by $(0.85, 0.85, 0.40)$ along the planning axes; their height is additionally restricted to the observed height range with a margin of 0.20.

For each anchor, we search both yaw directions. At each step, candidate motions comprise an orbit about the current pivot, an orbit with the pivot recentered along the preceding horizontal heading, a forward translation, and a rotation in place, in this order. The nominal orbit radius and forward displacement are 1.5 and 0.175, respectively, in reconstruction coordinates. Viewing elevation is

decoupled from the orbit position: pitch increments of $\{0^\circ, 4^\circ, -4^\circ\}$ are evaluated for each motion. Among the candidates of the first feasible motion primitive, the view maximizing the frontier score in Eq. 1 is selected, with unchanged pitch preferred in the case of equal scores.

Geometric feasibility and rendering support. Let c_l and R_l denote the center and rotation of camera v_l. We impose the following constraints on consecutive generated views:

$$\|c_l - c_{l-1}\|_2 \le 0.215, \qquad d_{\mathrm{SO}(3)}(R_l, R_{l-1}) \le \omega_l, \qquad \frac{1}{HW}\sum_x S_{v_l}(x) \ge 0.5,$$

where $d_{\mathrm{SO}(3)}$ is the relative rotation angle, $\omega_1 = 21^\circ$, and $\omega_l = 9.2^\circ$ for $l \ge 2$. The support mask S_v follows Appendix B.1. The rotation bound applies to the composed camera orientation, jointly accounting for yaw, pitch, and roll. Translational segments must remain in observed free space and maintain a clearance of 0.25 from the input surface. These conditions are evaluated along each segment, together with ray–mesh intersection tests against a TSDF surface fused from the fixed input depths. For frontier scoring, we use the image-space radius $r = \max\{1, 0.025\min(H, W)\}$ and assign zero score to views with no supported pixels or no uncovered pixels.

View elevation and anchor compatibility. The nominal pitch interval is $[-35^\circ, \theta_{\max}]$, with

$$\theta_{\max} = \min\left\{15^\circ, \arctan\left(\frac{H}{4f_y}\right)\right\},$$

where f_y is the vertical focal length at the conditioning resolution. This field-of-view-dependent upper bound limits upward tilt to preserve context in the lower image region. A measured anchor outside the nominal interval enters it directly when the relative-rotation constraint permits. If the standard search yields no feasible complete clip, anchor compatibility is handled by retaining the measured image-plane roll as needed and approaching the nearest pitch boundary monotonically in steps of at most 4°. The nominal interval is enforced after entry, and the final view must lie within it. Rendering support, collision avoidance, and relative-rotation bounds remain active throughout this transition.

Adaptive angular extent and trajectory selection. To accommodate restricted visibility while retaining complete clips, we evaluate yaw increments in $\Delta = \{8^\circ, 4^\circ, 2^\circ, 1^\circ\}$, in decreasing order. Let $\mathcal{T}_\delta$ denote the complete feasible trajectories returned by the bidirectional search at increment δ. We select the largest feasible angular extent and then maximize the mean frontier score of the generated views:

$$\begin{aligned} \delta^* &= \max\{\delta \in \Delta : \mathcal{T}_\delta \neq \varnothing\}, \\ \tau^* &= \arg\max_{\tau \in \mathcal{T}_{\delta^*}} \frac{1}{L-1}\sum_{l=1}^{L-1} f(v_l), \qquad L = 20. \end{aligned} \tag{A3}$$

Only complete trajectories with distinct poses are retained. The real starting frame is excluded from scoring, and mean translation does not affect feasibility or ranking. The selected cameras remain fixed during RGB repair and depth recovery. After the resulting RGB-D supervision updates the scene, exploration resumes from the next real anchor.

C.2 Tri-State Masking and Generative Repair

Intermediate comparison. For Figure 2, both methods receive identical rendered RGB, depth, and opacity. G4Splat uses See3D repair and native depth extraction before global plane refinement. Our method uses its saved paired RGB-D sequence. The comparison retains each method's repair pipeline.

Geometrically verified correspondences. For a supported target pixel, rendered depth proposes a 3D point on its camera ray. Projecting this point into real input i gives source pixel u_i and camera-space depth z_i. A correspondence must have positive depth, lie within the source image, and satisfy

$$\left|D_i^{\mathrm{obs}}(u_i) - z_i\right| \le \delta_a + \delta_r z_i, \qquad \delta_a = 0.03, \quad \delta_r = 0.025. \tag{A4}$$

We refine the proposed target depth twice along its ray using the source depth residual, checking validity before each correction. Nearly degenerate projections are rejected. Source confidence must be positive, and the depth range in a 3×3 source neighborhood must remain within the same absolute-plus-relative tolerance. Among valid sources, we choose the camera nearest to the target center and retain its paired color and corrected depth. A 5×5 target-depth stability test and a 3×3 source-consistency test reject discontinuities and inconsistent source assignments. Accepted correspondences define V_v and provide I_v^{obs} and D_v^{obs}.

Tri-state conditioning. The partition in Eq. 2 separates verified appearance, editable rendered context, and missing content. Surface identities retain their first assigned labels across rounds. The image masks are obtained at each target camera, and color anchoring still requires a valid correspondence in that view. Generated observations never enter the reference set used for verification. Equation 3 therefore supplies real colors in GROUNDED, rendered context in PROVISIONAL, and no rendered color in UNCOVERED. All real images remain available as diffusion references, ordered by support in the target frusta and camera proximity. At the real starting frame, we use the resized input RGB and mark the full image as GROUNDED. This appearance override does not create depth anchors.

Latent anchoring and color restoration. We encode a reference clip after extending the nearest GROUNDED color outside the protected region to avoid artificial black boundaries. This extension affects reference encoding only. A latent cell is protected by M_z only when its entire spatiotemporal pooling region is GROUNDED. Frames without such pixels have no protected cells. For proposed denoised latent $\bar{z}_{j+1}$, reference latent z^{ref}, and scheduler forward-noise operator $\mathcal{Q}$, we use

$$z_{j+1} = M_z \odot \mathcal{Q}_{s_{j+1}}(z^{\mathrm{ref}}, \epsilon) + (1 - M_z) \odot \bar{z}_{j+1}. \tag{A5}$$

The reference and generated branches share the initial noise ϵ. The last step uses the clean reference latent. The same conservative mask is used throughout the four denoising steps. After decoding, Eq. 4 restores every protected image pixel, including boundary pixels excluded by latent pooling. Consequently, the saved repaired RGB agrees exactly with its verified colors on GROUNDED. This property applies to the repaired images, while subsequent renderings also depend on scene optimization.

C.3 Observation-Anchored Geometric Completion

Multiple-view depth anchors. Depth anchoring requires stronger evidence than appearance preservation. Let $n_v(x)$ count real cameras supporting the selected corrected depth, including its source camera. An additional camera must be separated from the source by at least 0.02 and pass the correspondence tests when initialized at that corrected depth. The anchor mask is

$$B_v(x) = S_v(x) V_v(x) \mathbf{1}[n_v(x) \geq 2]. \tag{A6}$$

This test is applied before the full-frame appearance override at the real starting camera. Thus, every accepted anchor has independent geometric support and $B_v \leq M_v^{\mathrm{G}}$.

Confidence and scale calibration. Frozen WorldMirror 2.0 (Liu et al., 2025; Team HY-World et al., 2026) predicts depth from the final repaired RGB, conditioned on the planned cameras and accepted anchors. We normalize its confidence using the 10th and 90th percentiles of finite values in the reconstruction window, clip the result to $[0, 1]$, and clamp the percentile span below by 10^{-6}. For view v, let $\mathcal{Q}_v$ contain the ratios of anchor depth to predicted depth at valid anchor pixels with normalized confidence at least 0.20. We retain finite positive predictions and ratios in $[0.25, 4]$, then calibrate scale as

$$s_v = \begin{cases} \mathrm{median}(\mathcal{Q}_v), & \mathcal{Q}_v \neq \emptyset, \\ 1, & \mathcal{Q}_v = \emptyset. \end{cases} \tag{A7}$$

Without valid overlap, the view retains the predicted scale. Equation 5 restores accepted anchor depths and assigns them confidence one. RGB, depth, masks, and intrinsics share the planned camera grid. When resolutions differ, anchor depths are resized by support-normalized interpolation, with at least 0.5 interpolated support required for calibration. The clip-level scale calibration used for DL3DV is specified in Appendix C.5.

C.4 PROGRESSIVE SCENE RECONSTRUCTION

Admitted supervision. Let A_v accept pixels marked valid by depth reconstruction whose depth is finite and positive and whose normalized confidence is finite and at least 0.20. With the editable mask $E_v = M_v^{\mathrm{P}} + M_v^{\mathrm{U}}$, the domains for generated RGB, generated depth, and new Gaussian seeds are

$$W_v^{\mathrm{rgb}} = E_v, \qquad W_v^{\mathrm{depth}} = E_v A_v, \qquad W_v^{\mathrm{seed}} = M_v^{\mathrm{U}} A_v. \tag{A8}$$

When resizing, an output pixel remains editable only if its entire area-resampling footprint is editable, up to numerical tolerance 10^{-6}. Its depth weight is multiplied by the area-resampled admission mask. This excludes direct synthetic targets from pixels overlapping GROUNDED. Reliable UNCOVERED pixels are backprojected to initialize Gaussian seeds with repaired colors and depth-derived normals, subject to spatial subsampling and the point limit.

Generated and real-view losses. Generated RGB uses the $0.8/0.2$ mixture of L1 and SSIM losses (Wang et al., 2004) over the editable region. An SSIM term is accepted only when its entire 11×11 window lies within that region. Both terms are normalized by the editable pixel and channel count, rather than by the number of accepted SSIM windows. An empty editable region contributes zero. For rendered surface depth $D_v(G)$, the generated depth residual is

$$\ell_d(G, v) = \frac{\sum_x E_v(x) A_v(x)\, \rho\Big(D_v(G)(x) - \widehat{D}_v(x)\Big)}{\max(\sum_x E_v(x), 1)}, \qquad \rho(e) = \tfrac{1}{2}\log(1 + |e|). \tag{A9}$$

The generated-view loss ℓ_{gen} combines masked RGB supervision with this depth term under the inherited geometric schedule. Its RGB and geometry multipliers are fixed at 0.1 and 0.5, respectively. Generated views receive no additional normal, curvature, depth-order, or representation regularization terms.

The real-view loss ℓ_{real} uses full-image RGB supervision with multiplier 2.0 and the same L1–SSIM mixture. It also retains the input-depth, normal, curvature, and depth-order priors of G4Splat (Ni et al., 2026) and MAtCha (Guédon et al., 2025), together with depth–normal consistency and surfel-anisotropy regularization. Input geometry is supervised where its confidence exceeds 0.5. These inherited terms use their original schedules and are applied on real-view updates. Geometric losses use the mean of expected and median rendered depth. Camera planning uses expected depth as described in Appendix B.1.

Accumulation and optimization. Let $\mathcal{R}_k$ contain the repaired novel RGB-D views and their masks from round k, excluding the real starting frame, and let $\mathcal{U}_k = \bigcup_{j=0}^{k} \mathcal{R}_j$. The expected objective is

$$\mathcal{L}_k(G, t) = 0.4\, \mathbb{E}_{i\sim\mathcal{O}}[\ell_{\mathrm{real}}(G, i, t)] + 0.6\, \mathbb{E}_{v\sim\mathcal{U}_k}[\ell_{\mathrm{gen}}(G, v, t)]. \tag{A10}$$

This instantiates Eq. 6 with $p = 0.6$. t is the iteration within the current reconstruction round. The fixed group probabilities preserve the frequency of real-image supervision as the generated set grows. Sampling within each group traverses shuffled views. In each round, we optimize the scene using the real inputs and accumulated generated views. The resulting scene supplies the trajectory and rendering conditions for the next round. Generated masks remain attached to their saved RGB-D views. Although these masks exclude direct synthetic supervision from GROUNDED, shared Gaussian parameters are jointly optimized, so later renderings are not subject to an exact appearance-invariance constraint.

C.5 IMPLEMENTATION SETTINGS

We use a fixed configuration throughout the main experiments. Frozen WorldStereo repairs 832×480 clips with four denoising steps, and frozen WorldMirror 2.0 reconstructs their geometry. Generated and real views are sampled with probabilities 0.6 and 0.4, independent of the accumulated view count. Their RGB loss multipliers are 0.1 and 2.0, respectively, and the generated geometry multiplier is 0.5. Scene optimization and diffusion use random seeds 0 and 1024. The training schedule is given in Section 4.1. Component-specific thresholds are described with their corresponding procedures above. For DL3DV, the reported complete five-round variant calibrates one common depth scale per generated clip against verified real-input depth anchors before fitting

the scene. All sampling probabilities, loss weights, and iteration counts retain the fixed settings above.

C.6 Evaluation Protocol

Views and reference geometry. For the main comparison, each scene contains 100 posed images, split into five fixed inputs and 95 test views. The input-count comparison in Appendix A.5 uses nested sets of 5, 10, and 15 inputs with a common set of 85 test views. The Mip-NeRF 360 comparison in Appendix A.6 uses ten inputs and the remaining 90 test views. Evaluation uses native image resolution, including 1200×680 for Replica and 1752×1168 for ScanNet++. DL3DV retains each scene's stored image dimensions. Our method's generated training views are distinct from the test views. RGB images from the test views and reference meshes are used only for evaluation. Exceptions in baseline use of camera poses are described below.

Depth evaluation. Reference camera-space Z depth is obtained by ray casting the reference meshes. Every pixel with finite positive reference depth is included, including pixels with zero predicted depth. We apply no opacity filtering, scale or shift fitting, or depth truncation. Errors are computed per image, averaged over the 95 test views, and then averaged equally over scenes. RGB metrics use the same view and scene averaging. This protocol applies to every evaluated method, including 2DGS.

Surface extraction and evaluation. We use each method's available native mesh extraction or its documented native-depth TSDF adapter. On Replica and ScanNet++, our method fuses final-model depths from 100 calibrated and 95 generated cameras, using a 2 cm TSDF voxel size, an 8 cm truncation distance, and a 10 m maximum camera depth without opacity filtering. G4Splat uses its native adaptive TSDF and marching tetrahedra extraction. We sample each mesh uniformly by surface area with 200,000 points (seed 0), then clip to the common calibrated-camera frusta and reference bounding-box extent, without fitted alignment. Scoring follows the definitions and units in Appendix A, using the same reference samples and evaluation extent for all methods.

DL3DV relative-scale evaluation. For DL3DV, we follow the relative-scale conventions in Appendix A. We use the same fixed spherical and camera-frustum extent for all methods, sample 200,000 points from each surface, and fit neither alignment nor scale to the prediction. Depth-based TSDF extraction uses voxel size $0.002D$ and truncation $0.008D$. Reference depths are the first ray intersections expressed as camera-space Z. All methods share the valid-reference mask and zero predictions count as errors. We retain raw COLMAP-unit RMSE and RMSE/D in the evaluation records and average each metric equally over the four scenes.

Baseline initialization and outputs. We initialize 2DGS, GenFusion, and Difix3D+ from sparse COLMAP points (Schönberger & Frahm, 2016) computed from the five inputs. GuidedVD uses DUSt3R (Wang et al., 2024b), and G4Splat uses its geometry-based initialization. WorldStereo denotes the WorldStereo-DMD + 2DGS baseline, with all metrics evaluated on direct renderings of the reconstructed scene. Its auxiliary views are constructed solely from the five input cameras. The reused Difix3D+ and ArtiFixer reconstructions used camera poses of the test views for augmentation, while GenFusion's coordinate normalization used all calibrated cameras. These uses of camera calibration do not provide RGB supervision from the test views. GenFusion depths are restored to the original metric scale before evaluation.

D Limitations

Dependence on estimated input geometry. Observation verification uses depths inferred from the sparse input images. Errors in these fixed priors can affect both the tri-state partition and the depth anchors. An incorrect accepted correspondence may preserve an inaccurate color or depth, whereas conservative verification can reduce the region that benefits from input anchoring. Projected input colors also provide a limited description of view-dependent appearance, particularly for reflective or transparent surfaces.

Scope of observation preservation. Color restoration guarantees agreement with the projected references on GROUNDED pixels of the repaired images. Excluding these pixels from generated-view losses removes direct synthetic supervision at those locations, but does not impose a hard constraint on the shared Gaussian parameters. Updates driven by other viewpoints can therefore change subsequent renderings of the same surfaces. The proposed supervision policy reduces conflicting targets without guaranteeing global invariance of observation-supported appearance or geometry.

Restricted exploration and completion ambiguity. The planner searches a finite set of motions within free space inferred from the real inputs and requires substantial rendered context. These constraints can limit access to heavily occluded or disconnected regions. The frontier score does not guarantee complete scene coverage. Content without observational support is inferred from pretrained priors and may admit multiple plausible completions. Moreover, when no reliable depth anchors overlap a repaired view, scale calibration retains the model's predicted scale, leaving that view without an observation-based correction.

E AI USE STATEMENT

Generative AI tools were used to assist with drafting and polishing the manuscript, including its wording, grammar, and organization, and with literature retrieval and the discovery of related work. We also used AI assistance for code development. The authors reviewed and revised AI-assisted work and retain responsibility for the research design, interpretation of experimental results, and scientific conclusions. We take full responsibility for the final text, claims, references, code, and other artifacts in this work.