UNIGS-SLAM: MONOCULAR GAUSSIAN SLAM WITH ENHANCED GEOMETRY AND OPEN-VOCABULARY SEMANTICS

Anonymous authors
Paper under double-blind review

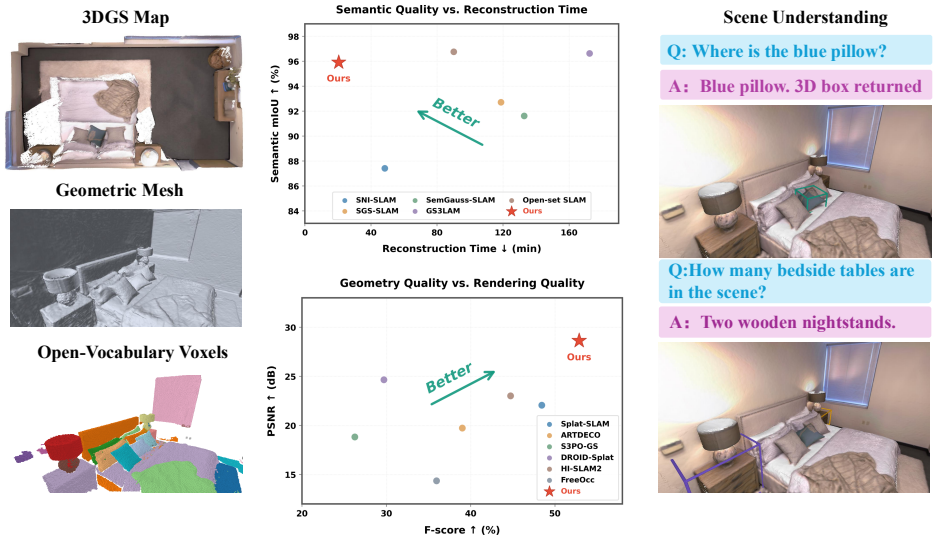

Figure 1: Our RGB-only framework reconstructs detailed geometry and open-vocabulary semantics (left), achieves strong quality and efficiency (middle), and supports 3D scene understanding (right).

ABSTRACT

Monocular semantic SLAM aims to reconstruct scene geometry and build semantic maps from RGB observations. Existing Gaussian-based methods often attach semantic labels or features to evolving Gaussian primitives, making semantic fusion difficult to maintain and increasing reconstruction overhead. Hybrid Gaussian–voxel representations provide more stable semantic support, but often rely on RGB-D input or known poses and do not revise historical semantic evidence after geometry updates. We present **UniGS-SLAM**, an RGB-only framework that decouples open-vocabulary semantics from Gaussian reconstruction while preserving bidirectional interaction. Foundation priors enhance monocular depth and pose estimation, while semantic observations are fused into an independent sparse voxel map with observation provenance, enabling historical contributions to be revised as reconstruction evolves. The fused semantics further guide Gaussian filtering and local surface refinement. Experiments on Replica, ScanNet, and KITTI-360 show leading tracking, rendering, and geometric reconstruction performance among representative RGB-only methods, together with strong open-set semantic mapping on Replica and ScanNet. We also achieves a 2.37× reconstruction speedup over the fastest compared semantic SLAM baseline.

1 INTRODUCTION

Recent advances in 3D Gaussian Splatting (3DGS) (Zhao et al., 2026a) have enabled efficient and high-fidelity SLAM systems with explicit scene representations and real-time rendering capabilities (Matsuki et al., 2024; Sandström et al., 2025). Beyond geometric reconstruction, embodied AI

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.

applications require semantic maps that associate open-vocabulary concepts with objects and their spatial locations, enabling tasks such as language-guided navigation, object search, and interaction (Yang et al., 2025a; Deng et al., 2025). Building such semantic-aware maps from a single RGB camera is particularly challenging, as semantic understanding relies on accurate geometry while the underlying scene representation continuously evolves during SLAM optimization (Yoo et al., 2025).

Early semantic Gaussian SLAM methods attach semantic labels or feature representations directly to Gaussian primitives, enabling unified reconstruction and semantic understanding (Li et al., 2024b; Zhu et al., 2025). More recent approaches incorporate foundation-model semantics to support open-vocabulary scene understanding through explicit Gaussian labels or semantic feature representations (Yang et al., 2025a; Yan et al., 2026). However, tightly coupling semantics with Gaussian primitives introduces two fundamental limitations. First, Gaussian primitives are continuously optimized, split, and pruned during reconstruction, making their spatial support unstable and complicating consistent semantic fusion for the same physical surfaces. Second, semantic attributes become part of the Gaussian optimization and rendering pipeline, introducing additional computational overhead and limiting the efficiency of online semantic reconstruction.

A natural solution is to decouple semantic representation from Gaussian primitives using more stable spatial structures. Hybrid Gaussian–voxel approaches maintain semantics in structured voxel spaces, providing stable units for multi-view semantic fusion (Yang et al., 2025b; Deng et al., 2025). However, existing systems typically rely on RGB-D observations and accurate camera poses, which limits their applicability to monocular settings. Moreover, incremental semantic fusion does not explicitly revise historical semantic evidence when backend SLAM optimization updates previous poses or geometry. As a result, semantic observations fused under outdated geometry may remain inconsistent even after the underlying reconstruction is corrected.

To address these challenges, we propose UniGS-SLAM, which decouples semantic representation from Gaussian optimization while preserving bidirectional interaction between semantics and geometry. Gaussian primitives remain dedicated to geometric reconstruction, whereas open-vocabulary semantics are fused in an independent sparse voxel map with observation-level provenance, allowing semantic evidence to be revised after geometric updates. Meanwhile, the fused semantic information is further used to refine Gaussian geometry through semantic-guided filtering and surface refinement.

Specifically, we first enhance monocular Gaussian SLAM with foundation-model geometry priors through geometric context selection, anchor alignment, confidence-aware depth fusion, and long-horizon pose guidance, providing reliable geometry for semantic lifting. We then maintain open-vocabulary semantics in a spatially decoupled sparse voxel representation, where observation-level contributions can be updated when SLAM corrects historical poses or depths. Finally, the semantic map feeds back into reconstruction by guiding floater removal and local surface refinement, allowing semantic understanding to improve Gaussian geometry.

Extensive experiments demonstrate that our RGB-only framework achieves strong performance in geometric reconstruction, rendering quality, and open-vocabulary semantic mapping. We further show that semantic decoupling enables efficient online reconstruction, while semantic-guided refinement improves the quality of the reconstructed Gaussian scene.

Our main contributions are summarized as follows:

Geometry-Enhanced Monocular Semantic Gaussian SLAM. We enhance monocular Gaussian SLAM with foundation-model geometry priors and confidence-aware fusion, providing reliable geometric support for reconstruction and semantic lifting.

Spatially Decoupled Open-Vocabulary Semantic Mapping. We maintain open-vocabulary semantics in an independent sparse voxel representation with observation-level provenance, enabling semantic evidence revision after geometric updates.

Semantic-Guided Gaussian Geometry Refinement. We use fused semantic evidence to guide floater filtering and local surface refinement, allowing semantic understanding to directly improve Gaussian reconstruction.

2 RELATED WORK

Gaussian Splatting SLAM. 3D Gaussian Splatting (Kerbl et al., 2023; Zhao et al., 2026b) enables efficient rendering and high-fidelity mapping in dense SLAM (Matsuki et al., 2024; Keetha et al., 2024; Yugay et al., 2023; Huang et al., 2024; Cheng et al., 2025a). Existing systems either track cameras through Gaussian rendering or combine separate geometric tracking with Gaussian mapping. DROID-Splat (Homeyer et al., 2024) extends DROID-SLAM (Teed & Deng, 2021) with Gaussian reconstruction, while HI-SLAM2 (Zhang et al., 2025) advances HI-SLAM (Zhang et al., 2024) with Gaussian mapping and adaptive depth-prior alignment. Splat-SLAM (Sandström et al., 2025) further maintains map consistency under global pose and depth updates. Rendering-based tracking can struggle under weak geometric constraints, while in RGB-only settings, monocular geometric ambiguity can lead to inaccurate pose and depth estimates that subsequently propagate into Gaussian mapping.

Recent advances use learned geometric priors to strengthen reconstruction (Cheng et al., 2025b; Yu et al., 2025). MASt3R-SLAM (Murai et al., 2025) demonstrates foundation-based monocular tracking, and ARTDECO (Li et al., 2025b) integrates foundation predictions into structured Gaussian reconstruction. S3PO-GS (Cheng et al., 2025c) addresses outdoor scale inconsistency through pointmap alignment, while KiloGS-SLAM (Yu et al., 2026) targets long-term tracking scalability. Despite this progress, integrating foundation priors requires reconciling their scale across prediction windows and rejecting unreliable geometry.

Our method integrates foundation priors into monocular Gaussian SLAM through cross-window alignment, confidence-aware depth fusion, and long-horizon pose guidance.

Semantic Gaussian Mapping and SLAM. Semantic SLAM has explored neural implicit representations in SNI-SLAM (Zhu et al., 2024) and multi-view feature consensus for open-vocabulary understanding in RADIO-ViPE (Nasser et al., 2026). For Gaussian-based semantic reconstruction, OpenGS-Fusion (Yang et al., 2025b) and OmniMap (Deng et al., 2025) combine Gaussian appearance modeling with volumetric semantic fusion, but require RGB-D input and supplied poses.

Semantic Gaussian SLAM brings together online localization, high-fidelity rendering, and scene understanding. Early methods attach semantic labels or features to Gaussians (Li et al., 2024b; Zhu et al., 2025; Li et al., 2024a), but are restricted to predefined categories. OpenGS-SLAM (Yang et al., 2025a), Open-Set Semantic Gaussian SLAM (Yan et al., 2026), and FeatureSLAM (Thirgood et al., 2026) incorporate foundation-model semantics to support open-set scene understanding, with Open-Set Semantic Gaussian SLAM further reducing storage through a shared expandable feature pool. However, these RGB-D systems tie semantics to evolving Gaussian primitives through labels, features, or indexing vectors; feature-based approaches also incur additional semantic optimization and rendering costs. This coupling complicates semantic maintenance, and cross-view agreement alone cannot correct historical evidence fused under outdated poses or depths.

Our RGB-only framework fuses open-vocabulary semantics in an independent voxel map and revises historical contributions after pose or depth updates.

3 METHODOLOGY

As illustrated in Fig. 2, our framework takes a calibrated monocular RGB sequence as input and maintains a Gaussian map for geometry and appearance together with an independent sparse voxel map for open-vocabulary semantics. We first use foundation priors to improve monocular tracking and Gaussian reconstruction (Sec. 3.1). Based on this geometry, we fuse and revise semantic observations in the voxel map (Sec. 3.2). The resulting semantic map then provides feedback to refine Gaussian geometry (Sec. 3.3). Additional method and implementation details are provided in Appendix A.3.

3.1 GEOMETRY-ENHANCED MONOCULAR GAUSSIAN SLAM

Monocular SLAM methods (Teed & Deng, 2021) estimate camera poses and scene geometry from visual observations but remain subject to monocular geometric ambiguity. Recent geometry foundation models (Wang et al., 2025; Lin et al., 2025) provide complementary geometry over broader

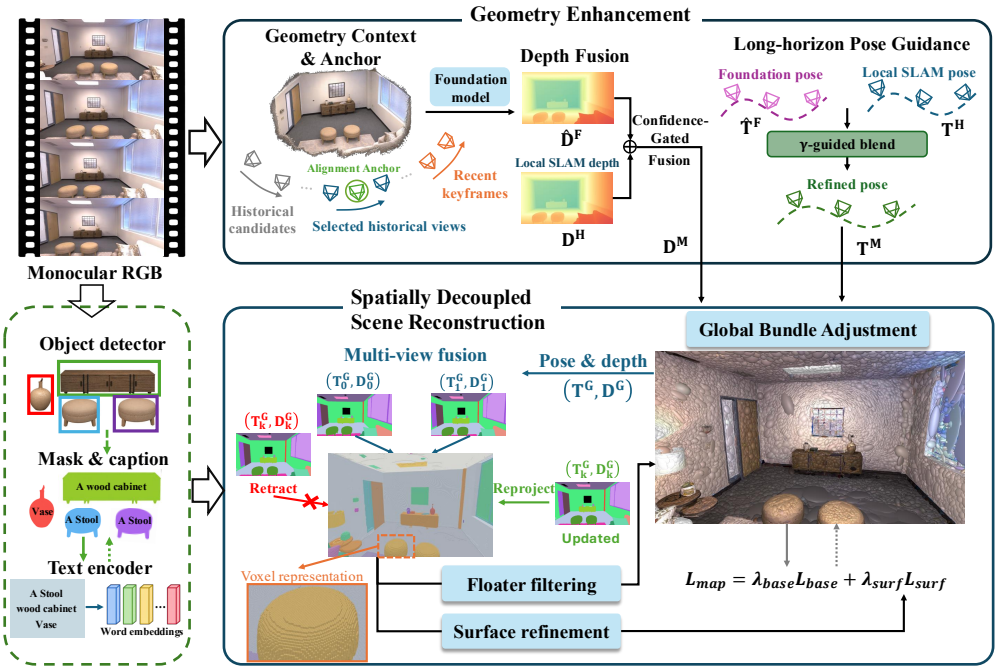

Figure 2: Overview of UniGS-SLAM. Foundation geometry priors enhance monocular reconstruction. An independent voxel map fuses semantic observations and revises historical evidence after geometry updates. Semantic feedback guides Gaussian floater filtering and surface refinement.

visual contexts, although their predictions may exhibit scale and pose inconsistencies across windows. We integrate these priors into monocular SLAM through geometric context selection, anchor alignment, confidence-aware depth fusion, and long-horizon pose guidance.

Geometric context and anchor alignment. We construct each prediction window from recent tracking keyframes and reliable historical views, increasing geometric diversity while preserving overlap. For candidate views i and j, we measure their geometric separation as

$$d(i,j) = \lambda_t \|\mathbf{t}_i^F - \mathbf{t}_j^F\|_2 + \frac{1}{\pi} \cos^{-1}\left(\frac{\mathrm{tr}((R_i^F)^\top R_j^F) - 1}{2}\right), \tag{1}$$

where $(R_i^F, \mathbf{t}_i^F)$ denotes a previously estimated foundation pose and λ_t balances translation and rotation. We favor confident historical views that complement the recent keyframes under this measure. For cross-window consistency, we select an alignment anchor using its historical translation and rotation residuals, and align the new prediction through the anchor and overlapping observations. This provides longer-range context within a bounded inference window.

Confidence-aware depth fusion. Let D_t^S denote local SLAM depth and $\widetilde{D}_t^F$ the foundation depth aligned to the current SLAM scale over reliable overlapping pixels. We determine the foundation contribution from prediction confidence and agreement with local geometry. For valid pixels in the blending regime, the weight is

$$\alpha_t(\mathbf{u}) = \bar{C}_t^F(\mathbf{u}) \exp\left(-\frac{|\log \widetilde{D}_t^F(\mathbf{u}) - \log D_t^S(\mathbf{u})|}{\tau_e}\right), \tag{2}$$

where $\bar{C}_t^F$ is the confidence rescaled to the blending interval and τ_e controls disagreement sensitivity. Unreliable or inconsistent predictions are rejected, while high-confidence consistent predictions receive full weight. We blend the aligned depths as

$$\bar{D}_t(\mathbf{u}) = (1 - \alpha_t(\mathbf{u}))D_t^S(\mathbf{u}) + \alpha_t(\mathbf{u})\widetilde{D}_t^F(\mathbf{u}). \tag{3}$$

A smooth residual scale correction, modulated by the same confidence gate, compensates for low-frequency alignment bias to obtain the mapping depth D_t^M. We initialize Gaussian primitives by

back-projecting D_t^M with the estimated camera poses, and optimize their geometry and appearance through differentiable rendering with photometric and geometric supervision (Matsuki et al., 2024).

Long-horizon pose guidance. The historical context also provides trajectory proposals beyond the local tracking window. We align these proposals to the SLAM coordinate system using a $\mathrm{Sim}(3)$ transformation and screen them for geometric consistency. Accepted proposals are blended with the current poses to initialize global visual bundle adjustment, with a coefficient γ controlling the guidance strength. Bundle adjustment then refines poses and depths under the visual correspondence constraints, allowing observation-driven optimization to correct residual prior errors. The updated geometry is propagated to Gaussian reconstruction.

3.2 Spatially Decoupled Open-Vocabulary Semantic Mapping

We maintain persistent open-vocabulary semantics in an independent sparse voxel map, retaining observation-level contributions to support revision as the reconstructed geometry evolves.

Observation-based voxel fusion. For each semantic keyframe, a region-language frontend extracts region masks, normalized language embeddings, and confidence scores (Deng et al., 2025). Using the current Gaussian-rendered depth and corresponding camera pose, we lift each region observation o into the voxel map. Its contribution to voxel v is denoted by $\Delta\mathbf{z}_{ov} = w_{ov}\mathbf{e}_o$, where $\mathbf{e}_o$ is the language embedding and w_{ov} is its fusion weight, which is zero outside the observation's voxel support. Each voxel accumulates these contributions, and its normalized feature can be compared with arbitrary text embeddings for open-vocabulary retrieval. We also associate observations across views using spatial overlap and language similarity to maintain entity identities. Alongside the fused statistics, we retain the original region observations and their individual voxel contributions.

Revision under geometry updates. Changes to camera poses or Gaussian geometry can invalidate the spatial support of previously fused observations. For each affected observation, we first retract its stored contribution, then reproject the cached region observation using the updated depth and pose and insert its new contribution:

$$\mathbf{z}_v = \sum_o \Delta\mathbf{z}_{ov}, \qquad \mathbf{z}'_v = \mathbf{z}_v - \Delta\mathbf{z}_{ov}^{\mathrm{old}} + \Delta\mathbf{z}_{ov}^{\mathrm{new}}. \tag{4}$$

The accumulated weights are updated analogously, and the revised semantic feature is obtained by normalizing $\mathbf{z}'_v$. The cached masks and language embeddings are reused, so geometry revision requires no semantic re-encoding; unaffected observations require no re-fusion.

3.3 Semantic-Guided Gaussian Geometry Refinement

Voxel semantics guide the selection of floater candidates and compatible surface neighborhoods for Gaussian refinement. Geometric and photometric evidence determines the corrections.

Verified floater filtering. Weak or missing semantic support identifies potential floaters in spatially sparse Gaussian neighborhoods, but is insufficient to justify removal. We verify these candidates using counter-evidence from multiple distinct views: either repeated free-space violations relative to reliable predicted mapping depths D_t^M from Sec. 3.1, or photometric evidence that reducing their opacity improves image reconstruction. Occluded Gaussians are not treated as free-space violations, and unobserved or semantically uncovered regions remain unknown. For verified candidates, we attenuate opacity and subsequently prune only those with sufficiently low opacity and renewed multi-view counterevidence. This filtering operates as map management rather than an additional semantic optimization loss.

Semantic-compatible surface refinement. For each Gaussian, we gather local support using spatial proximity, language-feature similarity, and compatibility of entity assignments. We recover surface samples from predicted mapping depths D_t^M across views, retaining samples that satisfy confidence, semantic compatibility, and geometric consistency checks. From the retained samples associated with the selected voxels, we estimate a local center $\bar{\mathbf{x}}_j$ and unit normal $\mathbf{n}_j$ by weighted PCA. Unlike purely spatial neighborhood selection, semantic and entity compatibility reduces the



Figure 3: RGB rendering comparisons on ScanNet and KITTI-360. Our method recovers finer scene details and achieves higher rendering quality.

Table 1: Tracking and rendering comparisons on Replica and ScanNet. The best, second-best, and third-best results among RGB-only methods are highlighted, respectively.

Method	Input	Replica				ScanNet			
		ATE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	ATE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Open-set SLAM	RGB-D	0.230	37.95	0.989	0.053	7.54	–	–	–
GS3LAM	RGB-D	0.370	36.26	0.989	0.052	11.24	22.86	0.868	0.222
SemGauss-SLAM	RGB-D	0.380	34.83	0.978	0.069	9.46	23.75	0.821	0.241
Splat-SLAM	RGB	0.446	34.15	0.945	0.074	7.75	22.07	0.792	0.230
ARTDECO	RGB	1.717	30.33	0.910	0.171	7.63	19.73	0.784	0.357
S3PO-GS	RGB	2.291	30.05	0.895	0.203	35.10	18.82	0.772	0.571
DROID-Splat	RGB	0.408	34.56	0.944	0.073	23.82	24.65	0.800	0.295
HI-SLAM2	RGB	0.251	37.72	0.964	0.034	7.21	23.02	0.818	0.269
FreeOcc	RGB	0.941	19.03	0.693	0.280	7.44	14.36	0.450	0.563
Hier-SLAM++	RGB	0.310	35.68	0.981	0.066	11.59	–	–	–
Ours	RGB	0.247	38.65	0.975	0.026	7.15	28.65	0.859	0.239

mixing of nearby but distinct object surfaces. These cues determine the sample grouping; the surface position and orientation are estimated from geometry. We retain only neighborhoods with sufficient multi-view support, geometric consistency, and local planarity.

Let $\boldsymbol{\mu}_j$ and Σ_j denote the center and covariance of Gaussian j. We first encourage its center to approach the estimated surface through a robust point-to-plane constraint:

$$\mathcal{L}_P = \left\langle \rho\left(\frac{\mathbf{n}_j^\top(\boldsymbol{\mu}_j - \bar{\mathbf{x}}_j)}{h}\right)\right\rangle_{w_j}, \tag{5}$$

where h is the voxel size, ρ is a smooth-L_1 penalty, and $\langle\cdot\rangle_{w_j}$ denotes averaging weighted by surface reliability. We further suppress excessive Gaussian extent along the surface normal:

$$\mathcal{L}_H = \left\langle \log\left(1 + \left[\frac{\mathbf{n}_j^\top \Sigma_j \mathbf{n}_j}{h^2} - \tau_H^2\right]_+\right)\right\rangle_{w_j}, \tag{6}$$

where $[\cdot]_+ = \max(\cdot, 0)$ and τ_H specifies the permitted normal-direction thickness relative to the voxel size. The position term adjusts Gaussian centers, while the thickness term acts on their scales and orientations without directly penalizing tangential extent.

Optimization objective. We define the base reconstruction and surface losses as

$$\mathcal{L}_{\text{base}} = \lambda_{\text{rgb}}\mathcal{L}_{\text{rgb}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}}, \qquad \mathcal{L}_{\text{surf}} = \mathcal{L}_P + \lambda_H\mathcal{L}_H. \tag{7}$$

The combined mapping objective is

$$\mathcal{L}_{\text{map}} = \mathcal{L}_{\text{base}} + \lambda_{\text{surf}}\mathcal{L}_{\text{surf}}. \tag{8}$$

Surface constraints apply only to reliable neighborhoods, with targets held fixed during the corresponding optimization.

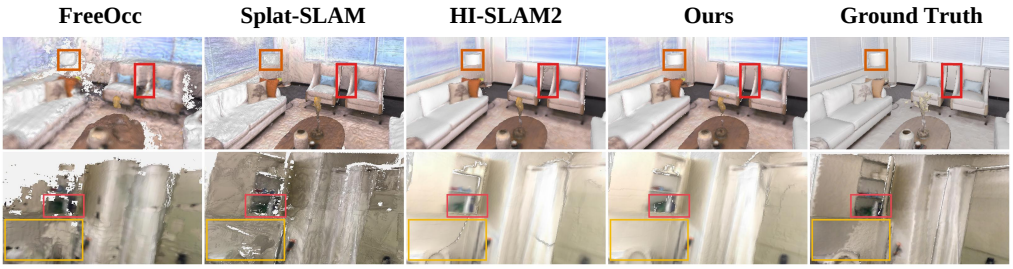

Figure 4: Mesh comparisons on Replica and ScanNet, showing cleaner surfaces and finer details.

Table 2: Geometric reconstruction on Replica and ScanNet.

Method	Replica			ScanNet		
	Acc.↓	L1↓	F-score↑	Acc.↓	L1↓	F-score↑
ARTDECO	12.15	12.565	57.43	63.46	14.467	39.01
DROID-Splat	7.07	5.690	66.59	39.08	19.400	29.70
Splat-SLAM	2.89	2.468	82.73	16.40	14.663	48.42
FreeOcc	4.25	12.820	63.37	15.23	19.716	35.94
S3PO-GS	173.06	11.523	20.83	22.70	42.427	26.25
HI-SLAM2	2.01	1.528	86.64	19.68	10.467	44.73
Ours	1.48	1.170	89.84	9.18	9.388	52.86

Table 3: Tracking and rendering on KITTI-360.

Method	ATE↓	PSNR↑	SSIM↑	LPIPS↓
Splat-SLAM	1.515	23.40	0.729	0.367
ARTDECO	11.534	22.17	0.722	0.283
S3PO-GS	12.162	20.65	0.668	0.407
DROID-Splat	2.075	23.73	0.752	0.324
FreeOcc	3.823	9.36	0.326	0.533
HI-SLAM2	1.668	23.42	0.754	0.339
Ours	1.366	25.39	0.800	0.261

4 EXPERIMENT

4.1 EXPERIMENT SETUP

Datasets. Experiments are conducted on eight scenes from Replica (Straub et al., 2019), six scenes from ScanNet (Dai et al., 2017), and four static segments from KITTI-360 (Liao et al., 2022). Dataset and scene-selection details are provided in Appendix A.4.1.

Baselines. We compare with representative visual SLAM methods (Sandström et al., 2025; Li et al., 2025b; Cheng et al., 2025c; Homeyer et al., 2024; Zhang et al., 2025) and semantic mapping systems (Zhu et al., 2024; Li et al., 2024b; Zhu et al., 2025; Li et al., 2025a; Yang et al., 2025a; Li et al., 2024a; Yan et al., 2026; Jiang et al., 2026; Nasser et al., 2026).

Implementation Details. All experiments are run on a server equipped with a single NVIDIA RTX A6000 GPU. Our system uses Depth Anything 3 (Lin et al., 2025) as geometric foundation model.

4.2 TRACKING AND APPEARANCE MAPPING RESULTS

We evaluate camera tracking, appearance reconstruction, and geometric reconstruction. Our RGB-only system achieves strong tracking accuracy and rendering quality in both indoor and outdoor scenes, together with substantial improvements in indoor geometric reconstruction.

Metrics. We evaluate RGB rendering using PSNR (dB), SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018). Camera tracking is measured by the root mean squared absolute trajectory error (ATE RMSE). Geometric reconstruction is assessed using accuracy (Acc.), L1 depth error (L1), and F-score (%).

Rendering and Camera Pose Estimation. Table 1 shows that our method achieves the highest PSNR among all compared methods on both indoor datasets, exceeding the strongest RGB-only baseline by 4.00 dB on ScanNet. On Replica, it surpasses the RGB-D baseline Open-set SLAM (Yan et al., 2026) by 0.70 dB and achieves the lowest LPIPS. It also achieves the lowest ATE among RGB-only methods on Replica and tracking accuracy comparable to HI-SLAM2 on ScanNet. Beyond indoor scenes, our method leads across all four metrics on the selected KITTI-360 segments (Table 3, ATE in meters), reducing ATE by 9.8% and improving PSNR by 1.66 dB over the best baseline for each metric.

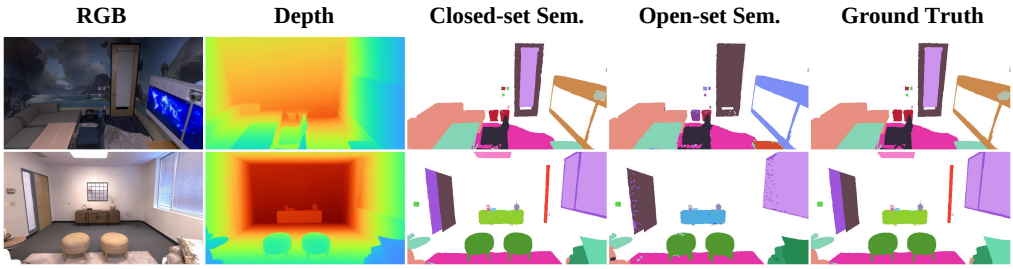

Figure 5: Semantic maps on Replica, showing coherent regions consistent with ground truth.

Table 4: Closed-set 2D semantic segmentation on Replica.

Method	Input	R0	R1	R2	O0	Avg.
SNI-SLAM	RGB-D	88.42	87.43	86.16	87.63	87.41
SGS-SLAM	RGB-D	92.95	92.91	92.10	92.90	92.72
SemGauss-SLAM	RGB-D	91.63	92.72	87.51	94.60	91.62
OpenGS-SLAM	RGB-D	93.24	94.11	92.79	93.22	93.34
GS3LAM	RGB-D	96.83	96.68	96.40	96.61	96.63
Open-set SLAM	RGB-D	96.93	96.82	96.67	96.62	96.76
FreeOcc	RGB	68.63	75.91	60.31	67.78	68.16
Hier-SLAM++	RGB	91.21	90.62	89.11	90.45	90.35
Ours	RGB	95.39	96.85	96.26	95.19	95.92

Table 5: Open-set 3D semantic segmentation.

Method	mAcc	fAcc	mIoU	fIoU
Replica				
FreeOcc	0.29	0.50	0.19	0.44
RADIO-ViPE	0.36	0.63	0.27	0.56
Ours	**0.42**	**0.73**	**0.38**	**0.69**
ScanNet				
FreeOcc	0.22	0.21	0.16	0.17
RADIO-ViPE	0.50	0.52	0.38	0.45
Ours	**0.61**	**0.66**	**0.50**	**0.56**

Geometric Reconstruction.

As shown in Table 2, our method achieves the best results across all three geometric metrics on both Replica and ScanNet. On Replica, it reduces Acc. and L1 depth error by 26.4% and 23.4%, respectively, relative to HI-SLAM2. On ScanNet, it reduces Acc. by 39.7% relative to FreeOcc and improves F-score by 4.44 percentage points over Splat-SLAM, the strongest competing RGB-only baselines on the respective metrics.

Qualitative Comparisons of Rendering and Geometry. Figure 3 shows that our method better preserves appearance details and object boundaries on ScanNet and KITTI-360. Figure 4 further demonstrates more regular object shapes and continuous surfaces on Replica and ScanNet, with fewer holes and spurious protrusions than the baselines.

4.3 CLOSED-SET AND OPEN-SET SEMANTIC MAPPING RESULTS

Metrics. We measure closed-set 2D segmentation using mean Intersection over Union (mIoU), reported in %. For open-set 3D object semantics, we report mean class accuracy (mAcc), frequency-weighted class accuracy (fAcc), mIoU, and frequency-weighted IoU (fIoU), following the RADIO-ViPE semantic segmentation protocol. Higher values indicate better semantic performance. Table 5 uses a 0–1 scale, with the best results in bold. Additional details of the semantic evaluation protocol are provided in Appendix A.4.2.

Closed-Set Semantic Mapping. We evaluate on four Replica scenes with ground-truth semantic supervision following Open-set SLAM (Yan et al., 2026), while our tracking and reconstruction remain RGB-only. Table 4 shows that our method outperforms the compared RGB-only baselines, improving mean mIoU over Hier-SLAM++ by 5.57 percentage points. Despite requiring no sensor depth, it also surpasses several RGB-D methods and approaches the strongest RGB-D results: the gaps to GS3LAM and Open-set SLAM are only 0.71 and 0.84 percentage points, respectively.

Open-Set Semantic Mapping. We follow the RADIO-ViPE (Nasser et al., 2026) protocol on Replica and ScanNet, comparing our RGB-only system with FreeOcc and RADIO-ViPE. Table 5 evaluates the semantic quality of the reconstructed 3D maps. Our method outperforms both FreeOcc and RADIO-ViPE across all four metrics on both datasets, improving mIoU over the stronger baseline, RADIO-ViPE, by 11 percentage points on Replica and 12 percentage points on ScanNet.

Qualitative Semantic Results. Figure 5 shows our closed-set and open-set semantic mapping results on Replica. Both recover distinct object regions and preserve their spatial structure.

4.4 ABLATION STUDY

We ablate three key components: geometry enhancement, historical semantic revision, and semantic-guided refinement. We also compare running time against semantic SLAM baselines.

Geometry Enhancement. Table 6 adds geometry enhancement components to the DA3 baseline one at a time on Replica office0. Context selection and anchor alignment improve pose and depth accuracy, while confidence-aware depth fusion further improves reconstruction quality. Long-horizon pose guidance provides an additional reduction in trajectory error. Each component improves all three metrics, demonstrating their complementary contributions to tracking and reconstruction.

Table 6: Geometry enhancement ablation.

Method	PSNR [dB] ↑	L1 depth [cm] ↓	ATE RMSE [cm] ↓
Foundation Prior (DA3)	33.68	1.964	2.010
+ Context & Anchor Alignment	34.79	1.310	0.713
+ Depth Fusion	40.34	1.083	0.623
+ Long-Horizon Guidance	**42.61**	**1.023**	**0.207**

Historical Semantic Revision. Table 7 compares three revision strategies averaged over all ScanNet scenes. Reprojection without retraction yields only modest gains, because the old semantic contributions remain in the map after the updated observations are inserted. Our retract-and-re-fuse strategy (Eq. 4) achieves the best performance, improving mAcc and mIoU by 4.95 and 4.43 percentage points over no revision, respectively. This demonstrates the importance of explicitly removing outdated semantic evidence when the underlying geometry is updated.

Table 7: Semantic revision ablation.

Method	mAcc↑	mIoU↑
w/o Historical Revision	55.85	45.43
Reprojection w/o Retraction	56.90	46.25
Ours	**60.80**	**49.86**

Semantic-Guided Geometry Refinement. Table 8 evaluates semantic-guided geometry refinement on ScanNet scene0207. Verified floater filtering reduces surface error and improves F-score, while adding semantic-compatible surface refinement yields further gains. These results show that the decoupled semantic map provides effective feedback for improving Gaussian geometry.

Table 8: Semantic-guided refinement.

Method	Acc. [cm] ↓	F-score [%] ↑
w/o both	13.76	48.86
+ Floater Filtering	10.08	50.28
+ Surface Refinement	**9.79**	**50.97**

Computational Efficiency. Table 9 reports reconstruction time on Replica. Our method takes 20.40 minutes, achieving a $2.37\times$ speedup over the fastest compared baseline, SNI-SLAM. This efficiency is consistent with our decoupled design, which fuses language features in voxels without optimizing or rendering them as Gaussian attributes. Detailed timing protocols and runtime calculations are provided in Appendix A.4.2.

Table 9: Reconstruction time.

Method	**Time (min) ↓**
SNI-SLAM	48.32
SGS-SLAM	118.70
SemGauss-SLAM	132.79
GS3LAM	172.40
Open-set SLAM	90.10
Ours	**20.40**

5 CONCLUSION

This work shows that geometry and semantics in monocular semantic SLAM can be decoupled in representation while remaining connected in optimization. UniGS-SLAM maintains geometry in a Gaussian map and open-vocabulary semantics in an independent voxel map, enabling semantic evidence to follow geometric updates without entering Gaussian optimization and rendering. The semantic map further feeds back to improve Gaussian reconstruction through targeted filtering and surface refinement. Our results demonstrate that this design supports accurate reconstruction, strong semantic mapping, and efficient RGB-only SLAM. More broadly, separating geometric and semantic state while preserving controlled interaction offers a useful design principle for semantic SLAM with evolving scene representations.

AI USE STATEMENT

We used large language models to aid manuscript writing and polish the text, including checking spelling and grammar. The authors take full responsibility for the final content, including all text, claims, and conclusions.

REFERENCES

Omar Alama, Avigyan Bhattacharya, Haoyang He, Seungchan Kim, Yuheng Qiu, Wenshan Wang, Cherie Ho, Nikhil Keetha, and Sebastian Scherer. RayFronts: Open-set semantic ray frontiers for online scene understanding and exploration. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5930–5937, 2025.

Chong Cheng, Yu Hu, Sicheng Yu, Beizhen Zhao, Zijian Wang, and Hao Wang. RegGS: Unposed sparse views gaussian splatting with 3DGS registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8100–8109, 2025a. URL https://arxiv.org/abs/2507.08136.

Chong Cheng, Zijian Wang, Sicheng Yu, Yu Hu, Nanjie Yao, and Hao Wang. Unposed 3DGS reconstruction with probabilistic procrustes mapping. *arXiv preprint arXiv:2507.18541*, 2025b. URL https://arxiv.org/abs/2507.18541.

Chong Cheng, Sicheng Yu, Zijian Wang, Yifan Zhou, and Hao Wang. Outdoor monocular slam with global scale-consistent 3d gaussian pointmaps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26035–26044, 2025c.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Yinan Deng, Yufeng Yue, Jianyu Dou, Jingyu Zhao, Jiahui Wang, Yujie Tang, Yi Yang, and Mengyin Fu. OmniMap: A general mapping framework integrating optics, geometry, and semantics. *IEEE Transactions on Robotics*, 2025. Accepted.

Christian Homeyer, Leon Begiristain, and Christoph Schnörr. DROID-Splat: Combining end-to-end slam with 3d gaussian splatting, 2024.

Huajian Huang, Longwei Li, Hui Cheng, and Sai-Kit Yeung. Photo-SLAM: Real-time simultaneous localization and photorealistic mapping for monocular, stereo, and RGB-D cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21584–21593, 2024.

Zeyu Jiang, Changqing Zhou, Xingxing Zuo, and Changhao Chen. FreeOcc: Training-free embodied open-vocabulary occupancy prediction, 2026.

Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. SplaTAM: Splat, track & map 3d gaussians for dense RGB-D slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21357–21366, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. doi: 10.1145/3592433.

Boying Li, Vuong Chi Hao, Peter J. Stuckey, Ian Reid, and Hamid Rezatofighi. Hier-SLAM++: Neuro-symbolic semantic slam with a hierarchically categorical gaussian splatting, 2025a.

Guanghao Li, Kerui Ren, Linning Xu, Zhewen Zheng, Changjian Jiang, Xin Gao, Bo Dai, Jian Pu, Mulin Yu, and Jiangmiao Pang. ARTDECO: Towards efficient and high-fidelity on-the-fly 3d reconstruction with structured scene representation, 2025b.

Linfei Li, Lin Zhang, Zhong Wang, and Ying Shen. GS3LAM: Gaussian semantic splatting slam. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM)*, pp. 3019–3027, 2024a. doi: 10.1145/3664647.3680739.

Mingrui Li, Shuhong Liu, Heng Zhou, Guohao Zhu, Na Cheng, Tianchen Deng, and Hongyu Wang. SGS-SLAM: Semantic gaussian splatting for neural dense slam. In *European Conference on Computer Vision (ECCV)*, pp. 163–179, 2024b. doi: 10.1007/978-3-031-72751-1_10.

Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022.

Haotong Lin, Sili Chen, Jun Hao 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.

Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Gaussian splatting slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18039–18048, 2024.

Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16695–16705, 2025.

Zaid Nasser, Mikhail Iumanov, Tianhao Li, Maxim Popov, Jaafar Mahmoud, and Sergey Kolyubin. RADIO-ViPE: Online tightly coupled multi-modal fusion for open-vocabulary semantic slam in dynamic environments, 2026.

Erik Sandström, Ganlin Zhang, Keisuke Tateno, Michael Oechsle, Michael Niemeyer, Youmin Zhang, Manthan Patel, Luc Van Gool, Martin Oswald, and Federico Tombari. Splat-SLAM: Globally optimized rgb-only slam with 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1686–1697, 2025.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Yuheng 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 A. Newcombe. The Replica dataset: A digital replica of indoor spaces. *CoRR*, 2019.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual slam for monocular, stereo, and rgb-d cameras. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 16558–16569, 2021.

Christopher Thirgood, Oscar Mendez, Erin Ling, Jon Storey, and Simon Hadfield. FeatureSLAM: Feature-enriched 3d gaussian splatting slam in real time, 2026.

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, June 2025.

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. doi: 10.1109/TIP.2003.819861.

Yucheng Yan, Chen Liang, Wenguan Wang, and Yi Yang. Open-set semantic gaussian splatting slam with expandable representation. In *International Conference on Learning Representations (ICLR)*, 2026.

Dianyi Yang, Yu Gao, Xihan Wang, Yufeng Yue, Yi Yang, and Mengyin Fu. OpenGS-SLAM: Open-set dense semantic slam with 3d gaussian splatting for object-level scene understanding, 2025a.

Dianyi Yang, Xihan Wang, Yu Gao, Shiyang Liu, Bohan Ren, Yufeng Yue, and Yi Yang. OpenGS-Fusion: Open-vocabulary dense mapping with hybrid 3d gaussian splatting for refined object-level understanding, 2025b.

Jisang Yoo, Gyeongjin Kang, Hyun kyu Ko, Hyeonwoo Yu, and Eunbyung Park. OpenMonoGS-SLAM: Monocular gaussian splatting slam with open-set semantics, 2025.

Sicheng Yu, Chong Cheng, Yifan Zhou, Xiaojun Yang, and Hao Wang. RGB-only gaussian splatting SLAM for unbounded outdoor scenes. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11068–11074, 2025. doi: 10.1109/ICRA55743.2025.11128286. URL https://arxiv.org/abs/2502.15633.

Sicheng Yu, Dongxu Shen, Beizhen Zhao, Guanzhi Ding, and Hao Wang. Robust and efficient monocular 3d gaussian slam for kilometer-scale outdoor scenes, 2026.

Vladimir Yugay, Yue Li, Theo Gevers, and Martin R. Oswald. Gaussian-SLAM: Photo-realistic dense slam with gaussian splatting, 2023.

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)*, pp. 586–595, 2018.

Wei Zhang, Tiecheng Sun, Sen Wang, Qing Cheng, and Norbert Haala. HI-SLAM: Monocular real-time dense mapping with hybrid implicit fields. *IEEE Robotics and Automation Letters*, 2024. doi: 10.1109/LRA.2023.3347131.

Wei Zhang, Qing Cheng, David Skuddis, Niclas Zeller, Daniel Cremers, and Norbert Haala. HI-SLAM2: Geometry-aware gaussian slam for fast monocular scene reconstruction. *IEEE Transactions on Robotics*, 2025. doi: 10.1109/TRO.2025.3626627.

Beizhen Zhao, Boyi Fu, Sicheng Yu, Zijian Wang, Kaiyong Zhao, Pengcheng Wu, Hao Wang, Steven Hoi, and Hui Xiong. 3D gaussian splatting rasterization: A survey. *Preprints*, 2026a. URL https://www.preprints.org/manuscript/202607.0776/v2. Preprint, version 2.

Beizhen Zhao, Sicheng Yu, Ziran Yin, Dongxu Shen, and Hao Wang. MMGS: 10× compressed 3DGS through optimal transport aggregation based on multi-view ranking. *arXiv preprint arXiv:2605.19304*, 2026b. URL https://arxiv.org/abs/2605.19304.

Siting Zhu, Guangming Wang, Hermann Blum, Jiuming Liu, Liang Song, Marc Pollefeys, and Hesheng Wang. SNI-SLAM: Semantic neural implicit slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21167–21177, 2024.

Siting Zhu, Renjie Qin, Guangming Wang, Jiuming Liu, and Hesheng Wang. SemGauss-SLAM: Dense semantic gaussian splatting slam. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 21174–21181, 2025.

A Appendix

A.1 APPLICATIONS

Our voxel semantic map provides a queryable 3D spatial representation that supports both semantic reasoning and metric 3D grounding. Figure 6 illustrates these complementary capabilities.

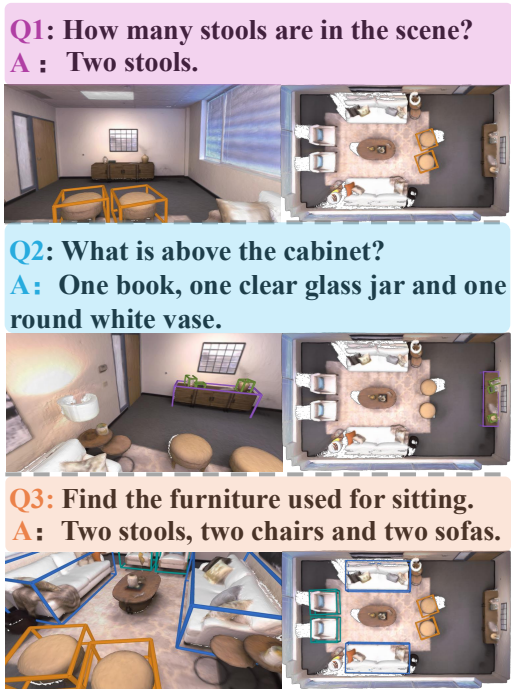

Figure 6: Scene understanding applications.

Open-Vocabulary Scene Question Answering. The examples cover object counting ("How many stools are in the scene?"), spatial relations ("What is above the cabinet?"), and functional descriptions ("Find the furniture used for sitting."). The responses identify two stools, the objects above the cabinet, and three categories of seating furniture, respectively. These examples illustrate how the reconstructed semantic map supports language-driven access to scene contents, including queries that span multiple object categories and express relationships between objects.

Language-Guided 3D Object Localization. Language features in the voxel map are associated with explicit 3D positions, allowing queried concepts to be grounded in the reconstructed geometry. The corresponding 3D boxes localize the retrieved objects in perspective (left) and top-down (right) views, linking each answer to its spatial support in a common scene coordinate system. Together, these examples demonstrate the utility of open-vocabulary mapping for scene understanding and spatial object retrieval beyond segmentation evaluation.

Implementation. Object queries are encoded using the same language encoder as the semantic map and matched against entity-level voxel embeddings by semantic similarity. For counting queries, we count the retrieved entity instances. Spatial queries are answered using the 3D locations and extents of the retrieved entities. For functional descriptions, a lightweight rule-based parser maps the query to a small set of candidate object concepts, which are then retrieved from the open-vocabulary voxel map. The spatial extent of the retrieved voxels is used to construct the 3D bounding boxes shown in both views.

A.2 LIMITATIONS

Our current system assumes a largely static scene and does not explicitly model independently moving objects. Extending the framework to dynamic environments remains an important direction for future work.

Our method also relies on foundation-model priors for geometry and semantics. Although confidence-aware fusion and consistency checks are introduced to reduce the influence of unreliable predictions, the overall performance is still bounded by the capability and generalization of the underlying foundation models.

A.3 ADDITIONAL METHOD DETAILS

A.3.1 GEOMETRY-ENHANCED MONOCULAR GAUSSIAN SLAM

This section expands the geometry enhancement in Section 3.1. We use camera-to-world poses $T_t = [R_t, \mathbf{t}_t]$, with superscripts S and F denoting local SLAM estimates and foundation predictions, respectively. All depth and pose alignment is relative to the current SLAM coordinate system.

Context and anchor selection. We maintain a bounded foundation context consisting of recent keyframes and geometrically complementary historical views, extending inference beyond the local tracking window. Historical views are selected according to confidence, geometric diversity under Eq. equation 1, and overlap with the current context. A reliable overlapping historical view serves as the alignment anchor, with reliability updated from cross-window pose consistency. Overlapping observations align successive predictions to a common coordinate system.

Depth alignment and gating. We align foundation depth to the current SLAM scale over reliable pixels $\Omega_t^{\cap}$ where both estimates are valid, using a confidence-weighted robust fit of their log-depth difference:

$$\ell_t^* = \arg\min_{\ell} \sum_{\mathbf{u}\in\Omega_t^{\cap}} C_t^F(\mathbf{u})\rho_\delta\big(\log D_t^S(\mathbf{u}) - \log D_t^F(\mathbf{u}) - \ell\big), \qquad \widetilde{D}_t^F = e^{\ell_t} D_t^F, \tag{9}$$

where ρ_δ is the Huber penalty and ℓ_t is the temporally smoothed scale estimate. For jointly valid pixels, define $e_t = |\log \widetilde{D}_t^F - \log D_t^S|$ and $\bar{C}_t^F = \mathrm{clip}((C_t^F - c_l)/(c_h - c_l), 0, 1)$, with confidence thresholds $c_l < c_h$. The complete gate is

$$\alpha_t(\mathbf{u}) = \begin{cases} 0, & \text{invalid prior, } C_t^F < c_l, \text{ or } e_t > \tau_e, \\ 1, & C_t^F \geq c_h \text{ and } e_t \leq \tau_e, \\ \bar{C}_t^F \exp(-e_t/\tau_e), & \text{otherwise.} \end{cases} \tag{10}$$

The gate first produces the blended depth $\bar{D}_t$ in Eq. equation 3. We represent s_t as a per-frame low-resolution scale field with bilinear parameterization in log space, fitted to reliable SLAM-to-blended-depth ratios and temporally smoothed. Applying $s_t^{\alpha_t} = \exp(\alpha_t \log s_t)$ modulates the correction in log-scale space and preserves valid SLAM depth wherever the prior is rejected. The final mapping depth is $D_t^M(\mathbf{u}) = s_t(\mathbf{u})^{\alpha_t(\mathbf{u})} \bar{D}_t(\mathbf{u})$. A valid aligned prior can supply missing local depth; pixels without either valid source are excluded. The final depth D_t^M initializes and supervises Gaussian mapping.

Long-horizon pose guidance. Historical context supplies trajectory proposals beyond the local tracking window. We transfer these proposals to the SLAM frame through a $\mathrm{Sim}(3)$ alignment over common frames $\mathcal{I}$:

$$(s_a, R_a, \mathbf{t}_a) = \arg\min_{s>0, R\in\mathrm{SO}(3), \mathbf{t}} \sum_{i\in\mathcal{I}} \|\mathbf{t}_i^S - (sR\mathbf{t}_i^F + \mathbf{t})\|_2^2. \tag{11}$$

The aligned poses are $\widetilde{\mathbf{t}}_i^F = s_a R_a \mathbf{t}_i^F + \mathbf{t}_a$ and $\widetilde{R}_i^F = R_a R_i^F$. Acceptance depends on trajectory coverage, continuity, translation and rotation agreement, and scale stability, using estimated geometry without ground-truth poses. Let $\mathcal{A} \in \{0, 1\}$ denote acceptance and $\gamma = \mathcal{A}\gamma_0$ the guidance coefficient. We initialize global visual bundle adjustment with

$$\begin{aligned} \mathbf{t}_i^0 &= (1-\gamma)\mathbf{t}_i^S + \gamma\widetilde{\mathbf{t}}_i^F, \\ R_i^0 &= R_i^S \,\mathrm{Exp}\Big[\gamma\,\mathrm{Log}\Big((R_i^S)^\top \widetilde{R}_i^F\Big)\Big]. \end{aligned} \tag{12}$$

Accepted proposals provide long-horizon guidance to the initialization; subsequent optimization refines poses and depths through visual correspondence constraints, allowing observation-driven correction of residual foundation-prediction errors. Rejected proposals retain the SLAM initialization.

Gaussian map optimization. We build on the base Gaussian mapper (Sandström et al., 2025) and initialize new primitives from the fused mapping depth D_t^M. Gaussian parameters are optimized using photometric, depth, and normal supervision, with standard densification and pruning. Following backend pose updates, primitives are re-anchored to their spawning keyframes and globally refined. Unless otherwise stated, Gaussian optimization follows the base mapper; our semantic-guided refinement is applied subsequently.

A.3.2 Spatially Decoupled Open-Vocabulary Semantic Mapping

We use Gaussian-rendered depth for semantic lifting so that region observations are attached to the current reconstructed surface. Foundation geometry improves lifting indirectly through the fused depth supervision of Gaussian geometry and the accepted pose guidance.

The semantic map retains observation provenance so that accumulated language evidence can follow changes in reconstructed geometry. Each region observation $o = (t, r)$ stores its original mask M_o, normalized language embedding $\mathbf{e}_o$, confidence c_o, and its current voxel contributions. The region-language frontend follows Deng et al. (2025), with c_o combining detector and mask confidence.

Projection. Processed region masks are lifted using a consistent snapshot of Gaussian-rendered depth, pose, and intrinsics. Valid pixels are back-projected as $\mathbf{x} = R_t(\widehat{D}_t K_t^{-1}\widetilde{\mathbf{u}}) + \mathbf{t}_t$ and assigned to voxel keys $\lfloor \mathbf{x}/h \rfloor$. The resulting support $\mathcal{Q}_o$ contains unique keys, so each region contributes at most once per voxel. Mask boundaries and invalid depths are excluded before projection.

Weighting. We use $w_{ov} = \max(c_o, \epsilon)\,\mathbf{1}\{v \in \mathcal{Q}_o\}$. and accumulate $\mathbf{z}_v = \sum_o w_{ov}\mathbf{e}_o$ and $W_v = \sum_o w_{ov}$. Open-vocabulary retrieval compares normalized $\mathbf{z}_v$ with text embeddings. We retain the unnormalized statistics and their contributing observation identifiers, allowing each observation's contribution to be retracted independently of other evidence in the voxel.

Entity association. For an observation o, we consider entities supported in its spatial neighborhood. With entity support $\mathcal{Q}_e$ and voxel neighborhood $\mathcal{N}(v)$, association combines spatial overlap and language similarity:

$$\begin{aligned} g(o, e) &= \frac{1}{|\mathcal{Q}_o|} \sum_{v \in \mathcal{Q}_o} \mathbf{1}\{\mathcal{N}(v) \cap \mathcal{Q}_e \neq \emptyset\}, \\ a(o, e) &= \lambda_g g(o, e) + \lambda_f\, \mathbf{e}_o^\top \mathbf{f}_e. \end{aligned} \tag{13}$$

where $\mathbf{f}_e$ is the entity's normalized aggregated embedding. The highest-scoring candidate is accepted above a threshold; otherwise, a new entity is created. Entity identities organize object-level support and constrain compatible surface neighborhoods; open-vocabulary retrieval operates directly on voxel embeddings.

Observation revision and update scheduling. For each integrated frame, we retain the original semantic packet, its observation identifiers and voxel supports, and the geometry snapshot used at fusion. Semantic revision is triggered for observations whose associated pose, depth, or calibration changes exceed predefined thresholds; revisions are processed asynchronously to preserve online operation.

Revision operates on the stored observations. We first retract each affected observation's old contribution, then reproject its cached mask with the updated geometry and insert the new contribution. For an observation o, the feature update follows Eq. equation 4, while its weight is updated as

$$W_v' = W_v - w_{ov}^{\text{old}} + w_{ov}^{\text{new}}. \tag{14}$$

Observation membership, confidence statistics, and entity votes are updated together, and voxels with no remaining support are removed. The observation identifier is preserved: reprojection relocates existing evidence without counting it as a new observation. Cached masks and embeddings are reused, and unaffected observations retain their existing contributions. Retraction from unnormalized accumulators preserves other observations' evidence even when their supports overlap.

This incremental mechanism maintains semantic evidence during ongoing tracking and mapping, including revisions of historical views, without re-encoding their semantics.

A.3.3 Semantic-Guided Gaussian Geometry Refinement

We detail the depth-reference selection and validity checks used for the geometric evidence in Section 3.3. Reference-depth back-projection constructs local surface targets for refinement; it does not insert new observations into the semantic voxel map.

Depth-reference selection. At the start of each refinement stage, we snapshot the available depth evidence and its camera geometry. When predicted mapping depth is available, the reference is $D_t^E = D_t^M$, using the aligned and gated depth from Appendix A.3.1. We retain its associated depth confidence and explicitly record its predicted origin. An unidentified depth array is not accepted as predicted evidence. If this source is unavailable, we use a fixed depth rendering of the current

Gaussian map and record it as a rendered reference. Such a reference can support local surface self-consistency, but cannot supply free-space counter-evidence for floater removal. It introduces no independent depth measurement. The reference depths, reliability masks, and derived surface targets remain fixed during the corresponding optimization stage.

Camera and pixel validity. We require depth, semantic observations, and camera parameters to share the same frame, camera pose, and field of view. Intrinsics are adjusted consistently under image resizing, and inconsistent camera/depth snapshots are excluded.

Depth-reference reliability. We retain finite positive depths with reliable confidence and suppress evidence near depth discontinuities and holes. Predicted references use their associated depth confidence, while rendered references additionally require sufficient opacity coverage. Depth agreement uses the tolerance $\delta_t(\mathbf{u}) = \max(\tau_d h, \tau_{\mathrm{rel}} D_t^E(\mathbf{u}))$, combining voxel-scale and depth-relative uncertainty.

Validity for surface estimation. We project each candidate semantic voxel into the reference views and sample the nearest image pixel. A view contributes a surface sample only if the projection lies inside the image, its camera depth agrees with D_t^E within δ_t, and both depth and semantic weights are positive. Semantic compatibility requires $\mathbf{f}_v^\top \mathbf{e}_o \geq \tau_{\mathrm{sem}}$, where $\mathbf{f}_v$ and $\mathbf{e}_o$ are the normalized voxel and region embeddings, respectively. For a voxel with an established entity identity, the region must additionally be associated with the same entity. We back-project the sampled reference depth to obtain a continuous surface position, using the voxel center only to select the image ray. Samples are weighted by depth reliability and semantic confidence. Their weighted spatial variance must remain below a voxel-scaled threshold, and sufficient distinct views must support them before they enter the local plane fit. The subsequent PCA rejects insufficient, nearly collinear, or nonplanar neighborhoods, as well as planes too far from the Gaussian being constrained.

Validity for free-space counter-evidence. Weak semantic support in sparse Gaussian neighborhoods proposes floater candidates; removal requires geometric or photometric counter-evidence. Only reliable predicted mapping depths provide free-space evidence. We conservatively test the projected spatial extent of each Gaussian against the reference depth rather than using its center alone. Let z_{tj} denote its center's camera depth and $\sigma_{tj}^z = \sqrt{\mathbf{a}_t^\top \Sigma_j \mathbf{a}_t}$ its extent standard deviation along the camera depth axis $\mathbf{a}_t$ in world coordinates. A free-space contradiction requires the finite extent to lie in front of the reliable reference surface:

$$z_{tj} + 2\sigma_{tj}^z < D_t^E(\mathbf{u}) - \delta_t(\mathbf{u}). \tag{15}$$

The contradiction must recur across geometrically distinct views. Gaussians behind the reference surface may be occluded and provide no free-space counter-evidence; invalid depth or missing semantic coverage also cannot establish free space.

Photometric verification. For photometric verification, we evaluate the gradient of the RGB reconstruction error with respect to each candidate's opacity logit over views in which it contributes to the rendering. Consistently positive gradients across sufficiently many distinct views, together with a positive aggregate gradient, provide evidence that reducing its opacity would decrease the reconstruction error. This is local directional evidence; pruning additionally requires sufficiently low opacity and renewed multi-view counter-evidence.

A.4 ADDITIONAL EXPERIMENTAL DETAILS

A.4.1 DATASET DETAILS

Replica. Replica (Straub et al., 2019) provides digital reconstructions of indoor spaces. We evaluate on eight scenes: `office0`, `office1`, `office2`, `office3`, `office4`, `room0`, `room1`, and `room2`, comprising five office scenes and three room scenes.

ScanNet. ScanNet (Dai et al., 2017) contains annotated real-world indoor scans. Following the scene selection of Open-set SLAM (Yan et al., 2026), we evaluate on six scenes: `scene0000_00`, `scene0059_00`, `scene0106_00`, `scene0169_00`, `scene0181_00`, and `scene0207_00`.

KITTI-360. KITTI-360 (Liao et al., 2022) is an outdoor urban-driving dataset with long trajectories spanning diverse street environments. Our system assumes a static scene and does not explicitly handle dynamic objects. We therefore select continuous segments without moving pedestrians or vehicles, yielding four evaluation scenes from three original driving sequences. Each scene contains 300 consecutive rectified RGB frames from camera `image_00`, with no intermediate frames skipped. Table 10 lists the original sequence identifiers and inclusive source-frame ranges; the two segments from sequence `0004` are evaluated separately.

Table 10: KITTI-360 segments with inclusive source-frame ranges.

Sequence	Frame range	Frames
`0002`	13419–13718	300
`0004`	1946–2245	300
`0004`	6192–6491	300
`0005`	2030–2329	300

A.4.2 Evaluation Details

Closed-set semantic evaluation. Following Open-set SLAM (Yan et al., 2026), we evaluate on four Replica scenes (room0, room1, room2, and office0) with ground-truth semantic supervision, while our tracking and reconstruction remain RGB-only. We report 2D segmentation mIoU in percent. For its closed-set evaluation, Open-set SLAM trains a classification head on rendered semantic features using cross-entropy supervision from ground-truth labels. We use the results reported by Open-set SLAM (Yan et al., 2026) for Open-set SLAM itself and for the compared baselines SGS-SLAM (Li et al., 2024b) and GS3LAM (Li et al., 2024a). We reproduce SemGauss-SLAM (Zhu et al., 2025) and FreeOcc (Jiang et al., 2026) using their publicly available code with ground-truth semantic supervision. For OpenGS-SLAM (Yang et al., 2025a), we use the *Ours (Prior)* results reported in its Table I, where all methods use ground-truth semantic labels as semantic priors for mapping.

Open-set semantic evaluation. To match our input setting, we restrict the quantitative comparison to systems operating with RGB-only input and no externally supplied camera poses. We consider OpenGS-SLAM (Yang et al., 2025a), Open-set SLAM (Yan et al., 2026), RADIO-ViPE (Nasser et al., 2026), and FreeOcc (Jiang et al., 2026) as related open-vocabulary semantic mapping systems. At the time of our experiments, complete SLAM implementations and evaluation code for OpenGS-SLAM and Open-set SLAM were not publicly available, preventing reproducible comparisons under a common protocol. We therefore compare against RADIO-ViPE and FreeOcc using their publicly available implementations in RGB-only configurations, with camera poses estimated by each system. We adopt the 3D semantic segmentation protocol used by RADIO-ViPE, which follows RayFronts (Alama et al., 2025), and apply it to Replica and ScanNet. Following the without-background setting in RADIO-ViPE, we exclude background categories, including structural classes such as walls and floors, from evaluation. We report mean class accuracy (mAcc), overall accuracy (fAcc), mean IoU (mIoU), and frequency-weighted IoU (fIoU), all on a 0–1 scale.

Time Efficiency evaluation. We evaluate the end-to-end runtime of our method and four publicly available baselines: SNI-SLAM (Zhu et al., 2024), SGS-SLAM (Li et al., 2024b), SemGauss-SLAM (Zhu et al., 2025), and GS3LAM (Li et al., 2024a). All five methods are measured on a single NVIDIA RTX A6000 GPU under the same experimental setup and with a common end-to-end timing scope. Timing includes data processing, semantic segmentation, tracking and mapping, and the final refinement stages used to produce the evaluated reconstruction. Each baseline retains the iteration counts and parameter settings specified in its paper and released implementation.

Open-set SLAM (Yan et al., 2026) is reported separately as a literature-derived estimate because its complete SLAM implementation was not publicly available at the time of our experiments. Its approximately 90-minute runtime is obtained from $T = N/(60\,\mathrm{FPS})$, using $N = 2000$ frames and the reported 0.37 FPS for the LoopSplat-based variant on Replica room0. The original measurement uses a single NVIDIA RTX 3090 GPU.

A.4.3 Detailed Reconstruction Results

We provide per-scene rendering quality (PSNR, SSIM, and LPIPS) and geometric reconstruction error (L1 depth) on Replica and ScanNet, as well as per-scene rendering quality on KITTI-360.

Table 11: Progressive ablation of geometry enhancement and semantic-guided refinement, averaged over all eight Replica scenes.

Method	PSNR [dB] ↑	SSIM ↑	LPIPS ↓	L1 Depth [cm] ↓
w/o both	36.79	0.958	0.036	1.564
+ Geometry Enhancement	38.37	0.974	0.030	1.260
+ Semantic-Guided Refinement	**38.65**	**0.975**	**0.026**	**1.170**

Table 12: Ablation of semantic guidance in Gaussian geometry refinement on ScanNet scene0207. The geometry-only variant retains the same geometric refinement while removing semantic guidance.

Refinement Variant	**Acc. [cm] ↓**	**F-score [%] ↑**
No Refinement	13.76	48.86
Geometry-Only Refinement	12.07	49.14
Semantic-Guided Refinement (ours)	**9.79**	**50.97**

Tables 13 and 14 report the Replica results, while Tables 15 and 16 report the ScanNet results. KITTI-360 rendering results are presented in Table 17.

We further provide additional qualitative results. Figure 7 compares reconstructed meshes with those of baseline methods and the ground truth. Figure 8 visualizes our closed-set Gaussian semantics and open-set voxel semantics alongside RGB images, depth maps, and ground-truth semantic annotations. Figure 9 presents RGB rendering comparisons on ScanNet and KITTI-360, with enlarged regions highlighting local details.

A.4.4 More Ablation Studies

Overall Component Ablation. We evaluate the contributions of geometry enhancement and semantic-guided geometry refinement on all eight Replica scenes. Our SLAM backbone follows the architecture of DROID-SLAM (Teed & Deng, 2021) and related Gaussian SLAM systems (Homeyer et al., 2024; Zhang et al., 2025). Table 11 reports scene-averaged results as we progressively add the two components to this backbone. The first row disables both components, and the last row includes both.

Adding geometry enhancement increases PSNR from 36.79 to 38.37 dB ($+1.58$ dB) and SSIM from 0.958 to 0.974, while reducing LPIPS from 0.036 to 0.030. It also reduces L1 depth error from 1.564 to 1.260 cm, a 19.4% reduction. Adding semantic-guided refinement further improves PSNR to 38.65 dB ($+0.28$ dB) and SSIM to 0.975, and lowers LPIPS to 0.026. L1 depth error decreases further to 1.170 cm, a 7.1% reduction relative to geometry enhancement alone. These results show that geometry enhancement improves both appearance and depth reconstruction, while semantic-guided refinement provides additional gains across all four metrics.

Effect of Semantic Guidance. To isolate the contribution of semantic guidance, we compare our full refinement with a geometry-only variant that retains the same geometric refinement procedure but removes semantic cues for candidate and neighborhood selection. As shown in Table 12, geometry-only refinement yields a modest improvement over no refinement, whereas semantic guidance further reduces Acc. from 12.07 to 9.79 cm and improves F-score from 49.14% to 50.97%. This indicates that the gain is not merely due to additional geometric refinement; semantic cues improve the selection of reliable floater candidates and compatible surface neighborhoods.

A.5 Legal and Ethical Considerations

A.5.1 Asset License

Replica and ScanNet are available under the Replica Dataset Research Terms and ScanNet Terms of Use, respectively. KITTI-360 is released under CC BY-NC-SA 3.0. For pretrained models, Depth

Anything 3 (DA3NESTED-GIANT-LARGE) uses CC BY-NC 4.0, YOLO-World uses GPL-3.0, and Tokenize Anything and Sentence-BERT (all-MiniLM-L6-v2) use Apache-2.0.

Table 13: Rendering quality on Replica.

Method	Metric	ro-0	ro-1	ro-2	of-0	of-1	of-2	of-3	of-4	Avg.
Splat-SLAM	PSNR ↑	30.75	32.69	32.63	39.00	36.15	33.69	33.52	34.73	34.15
	SSIM ↑	0.914	0.930	0.944	0.973	0.964	0.930	0.943	0.959	0.945
	LPIPS ↓	0.091	0.076	0.070	0.046	0.066	0.113	0.066	0.066	0.074
ARTDECO	PSNR ↑	29.28	28.34	30.44	35.63	34.38	27.88	28.29	28.40	30.33
	SSIM ↑	0.883	0.875	0.920	0.951	0.936	0.915	0.901	0.902	0.910
	LPIPS ↓	0.172	0.182	0.153	0.137	0.189	0.188	0.173	0.174	0.171
S3PO-GS	PSNR ↑	27.64	27.03	28.41	34.05	33.13	28.97	30.84	30.36	30.05
	SSIM ↑	0.852	0.829	0.882	0.931	0.935	0.902	0.914	0.918	0.895
	LPIPS ↓	0.196	0.266	0.220	0.171	0.195	0.224	0.163	0.191	0.203
DROID-Splat	PSNR ↑	30.68	34.36	32.46	37.80	38.93	33.34	33.79	35.08	34.56
	SSIM ↑	0.885	0.950	0.937	0.963	0.974	0.939	0.955	0.949	0.944
	LPIPS ↓	0.130	0.056	0.089	0.057	0.045	0.080	0.061	0.067	0.073
FreeOcc	PSNR ↑	15.89	19.20	17.99	24.48	24.34	15.91	16.43	18.00	19.03
	SSIM ↑	0.594	0.677	0.679	0.747	0.785	0.668	0.654	0.738	0.693
	LPIPS ↓	0.369	0.274	0.312	0.193	0.206	0.306	0.304	0.276	0.280
HI-SLAM2	PSNR ↑	34.46	36.45	37.13	41.42	42.66	35.82	35.87	37.98	37.72
	SSIM ↑	0.951	0.955	0.966	0.967	0.981	0.955	0.964	0.971	0.964
	LPIPS ↓	0.041	0.039	0.033	0.021	0.019	0.041	0.035	0.041	0.034
Ours	PSNR ↑	35.24	37.48	38.29	42.55	43.53	36.74	36.52	38.83	38.65
	SSIM ↑	0.963	0.972	0.976	0.987	0.986	0.972	0.972	0.974	0.975
	LPIPS ↓	0.034	0.029	0.025	0.014	0.015	0.030	0.027	0.032	0.026

Table 14: Depth reconstruction error on Replica (L1 depth ↓ [cm]).

Method	ro-0	ro-1	ro-2	of-0	of-1	of-2	of-3	of-4	Avg.
Splat-SLAM	2.93	2.17	2.19	2.50	1.79	2.83	2.76	2.57	2.468
ARTDECO	7.11	6.70	6.04	4.37	4.42	10.62	8.83	52.43	12.565
S3PO-GS	9.84	12.02	7.94	5.62	20.92	10.99	12.66	12.21	11.523
DROID-Splat	4.58	2.75	3.78	2.06	2.01	10.82	15.60	3.93	5.690
FreeOcc	16.90	8.24	14.22	7.90	6.37	13.05	19.33	16.56	12.820
HI-SLAM2	1.93	1.17	1.47	1.32	0.86	1.54	2.19	1.75	1.528
Ours	1.35	0.98	1.16	1.00	0.67	1.19	1.76	1.26	1.170

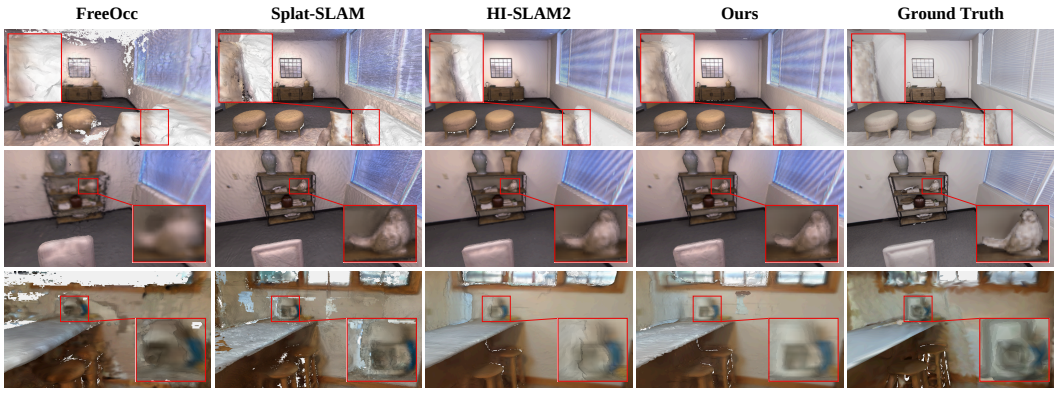

Figure 7: Additional qualitative comparisons of reconstructed meshes against baseline methods and the ground truth.

Table 15: Rendering quality on ScanNet.

Method	Metric	0000	0059	0106	0169	0181	0207	Avg.
Splat-SLAM	PSNR ↑	23.29	21.30	22.49	24.66	17.93	22.72	22.07
	SSIM ↑	0.800	0.793	0.803	0.801	0.762	0.794	0.792
	LPIPS ↓	0.220	0.216	0.248	0.218	0.249	0.230	0.230
ARTDECO	PSNR ↑	19.37	19.99	19.36	20.48	19.02	20.16	19.73
	SSIM ↑	0.732	0.794	0.812	0.806	0.806	0.752	0.784
	LPIPS ↓	0.363	0.343	0.341	0.324	0.397	0.372	0.357
S3PO-GS	PSNR ↑	16.44	17.22	18.47	18.16	21.44	21.16	18.82
	SSIM ↑	0.716	0.707	0.769	0.782	0.846	0.812	0.772
	LPIPS ↓	0.689	0.566	0.530	0.575	0.505	0.563	0.571
DROID-Splat	PSNR ↑	24.09	22.82	24.82	25.01	24.37	26.79	24.65
	SSIM ↑	0.754	0.786	0.834	0.804	0.828	0.792	0.800
	LPIPS ↓	0.321	0.289	0.259	0.300	0.331	0.267	0.295
FreeOcc	PSNR ↑	15.90	12.65	11.55	15.86	14.72	15.46	14.36
	SSIM ↑	0.459	0.365	0.359	0.517	0.511	0.466	0.45
	LPIPS ↓	0.528	0.560	0.651	0.523	0.564	0.553	0.563
HI-SLAM2	PSNR ↑	23.35	21.64	22.66	28.67	19.65	22.17	23.02
	SSIM ↑	0.808	0.808	0.844	0.849	0.828	0.772	0.818
	LPIPS ↓	0.271	0.265	0.250	0.225	0.273	0.327	0.269
Ours	PSNR ↑	28.69	26.65	27.37	30.22	29.81	29.16	28.65
	SSIM ↑	0.835	0.863	0.884	0.862	0.894	0.814	0.859
	LPIPS ↓	0.254	0.228	0.214	0.208	0.230	0.302	0.239

Table 16: Depth reconstruction error on ScanNet (L1 depth ↓ [cm]).

Method	0000	0059	0106	0169	0181	0207	Avg.
Splat-SLAM	8.89	12.57	24.53	12.36	16.83	12.80	14.663
ARTDECO	8.45	15.67	18.72	15.33	12.71	15.92	14.467
S3PO-GS	43.78	20.05	22.70	83.01	50.43	34.59	42.427
DROID-Splat	17.59	17.33	19.00	31.13	12.65	18.70	19.400
FreeOcc	12.66	26.35	31.87	19.87	11.23	16.31	19.716
HI-SLAM2	5.24	10.16	14.92	13.83	6.90	11.75	10.467
Ours	5.27	9.31	14.85	11.03	6.44	9.43	9.388

Table 17: Rendering quality on KITTI-360.

Method	Metric	scene01	scene02	scene03	scene04	Avg.
ARTDECO	PSNR ↑	21.43	22.81	23.03	21.43	22.17
	SSIM ↑	0.731	0.705	0.720	0.731	0.722
	LPIPS ↓	0.285	0.284	0.276	0.285	0.283
DROID-Splat	PSNR ↑	22.69	25.71	23.30	23.22	23.73
	SSIM ↑	0.759	0.768	0.720	0.760	0.752
	LPIPS ↓	0.305	0.301	0.362	0.327	0.324
FreeOcc	PSNR ↑	10.22	8.71	9.96	8.54	9.36
	SSIM ↑	0.370	0.286	0.296	0.353	0.326
	LPIPS ↓	0.471	0.562	0.541	0.559	0.533
HI-SLAM2	PSNR ↑	22.78	25.56	24.35	20.97	23.42
	SSIM ↑	0.771	0.778	0.749	0.719	0.754
	LPIPS ↓	0.309	0.299	0.334	0.412	0.339
Ours	PSNR ↑	24.70	27.64	25.39	23.84	25.39
	SSIM ↑	0.820	0.824	0.777	0.781	0.800
	LPIPS ↓	0.226	0.219	0.285	0.313	0.261

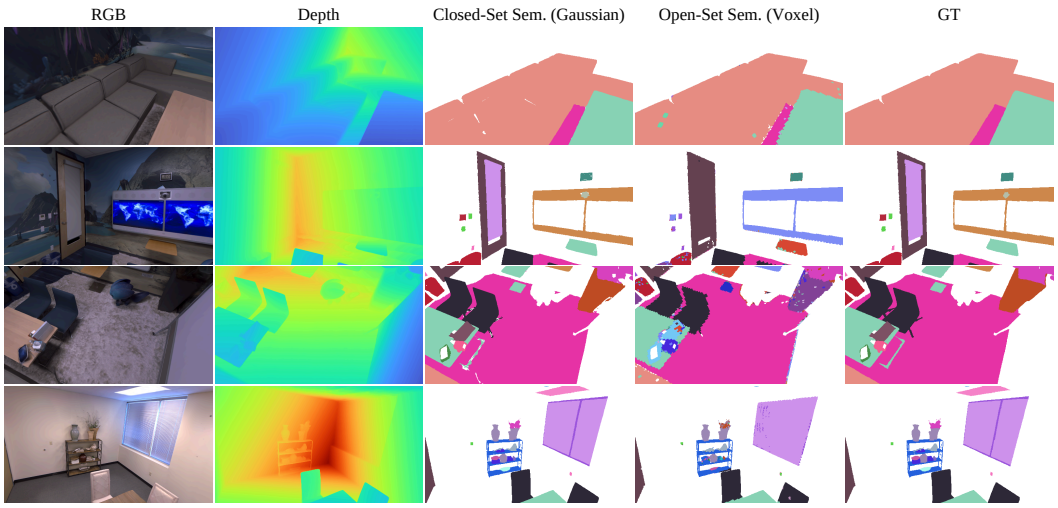

Figure 8: Additional semantic reconstruction results. From left to right: RGB images, depth maps, closed-set Gaussian semantics, open-set voxel semantics, and ground-truth semantic annotations.



Figure 9: Additional RGB rendering comparisons on ScanNet and KITTI-360. Enlarged regions highlight local appearance details.