OAVP-SPLAT: OBJECT-AWARE VIEW PLANNING FOR ACTIVE RECONSTRUCTION WITH SEMANTIC 3D GAUSSIAN SPLATTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Active reconstruction aims to recover scene geometry and semantic information through autonomous exploration. However, semantic maps reconstructed with 3D Gaussian Splatting (3DGS) can exhibit inconsistent class predictions within individual objects and gaps across otherwise continuous surfaces in semantic renderings. We propose OAVP-Splat, a framework that uses object-level observation value to guide two-stage autonomous view planning and refines semantic maps through semantic contribution reweighting. We perform object-level Gaussian clustering and estimate the value of additional observations from semantic uncertainty, temporal disagreement, and observation history. In the coverage stage, the planner builds an initial map through frontier exploration. In the refinement stage, it balances progress toward the next frontier target with the value of revisiting mapped regions and uses geometric support to locally refine viewing directions. After mapping, we reweight semantic contributions using per-Gaussian class probabilities and support from spatially connected 3D components of the same predicted class to improve semantic rendering. Experiments on Replica and Matterport3D demonstrate that OAVP-Splat improves semantic segmentation accuracy while maintaining favorable geometric accuracy and reconstruction quality, combining autonomous exploration with high-quality semantic reconstruction.

Keywords: Active Reconstruction, 3D Gaussian Splatting, Autonomous View Planning, Object-Level Gaussian Clustering, Semantic Contribution Reweighting.

1 INTRODUCTION

Autonomous agents need scene models that capture spatial structure and semantic identities (Jatavallabhula et al., 2023; Gu et al., 2024). Active 3D reconstruction acquires these models by selecting observations under sensing and motion budgets (Bircher et al., 2016; Schmid et al., 2020). For semantic reconstruction, exploration must both expand coverage and strengthen evidence for objects that remain ambiguously or inconsistently labeled.

Neural radiance fields (Mildenhall et al., 2020) and dense neural mapping (Sucar et al., 2021; Zhu et al., 2022) support reconstruction from image sequences. The explicit representation and efficient differentiable rendering of 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) make it a practical foundation for active reconstruction. ActiveSplat (Li et al., 2025c) and ActiveGS (Jin et al., 2025) integrate Gaussian mapping with autonomous exploration; ActiveSGM (Chen et al., 2025b) adds rendered semantic entropy to view selection. However, geometric coverage does not ensure semantic consistency or continuity, as two recurring problems illustrate.

First, uneven observation support and changing predictions can fragment labels within one object. Current semantic entropy alone cannot determine observation value: a rarely seen object and one repeatedly viewed from similar directions may have comparable uncertainty. View selection must therefore account for observation history and directional coverage to identify missing evidence.

Second, semantic renderings can contain gaps across continuous surfaces. Labels combine overlapping Gaussian contributions whose strengths may not reflect the support for their predictions.

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.



Figure 1: Qualitative comparison of trajectories and selected viewpoints in room0 (left pair) and office4 (right pair). Within each pair, ActiveSGM is shown on the left and OAVP-Splat on the right. Cyan paths and yellow camera markers illustrate how OAVP-Splat combines smoother trajectories with targeted path planning and viewpoint selection for regions with weak semantic evidence, acquiring complementary observations of mapped objects.

Weakly supported fragments can disrupt surrounding semantic regions, motivating better aggregation alongside informative acquisition.

We propose OAVP-Splat, an active semantic reconstruction framework coupling object-aware view planning with semantic contribution reweighting. Object-level Gaussian groups maintain semantic states and observation histories to identify valuable complementary observations. A two-stage planner first establishes coverage, then balances exploration with targeted revisits; geometric support refines viewing poses. After mapping, per-Gaussian class probabilities and support from connected same-class regions regulate semantic contributions, reducing weakly supported fragments without additional training or changes to Gaussian geometry and opacity.

Our main contributions are as follows:

- **Object-level observation value.** We introduce ObjectVOI to score candidate views by combining semantic uncertainty, temporal disagreement, observation deficits, and view novelty to identify complementary object observations.
- **Semantic-guided two-stage view planning.** We first establish coverage, then balance coverage gain with object observation value, while geometry-guided pose refinement preserves semantic value.
- **Semantic contribution reweighting.** We combine per-Gaussian response strength with same-class 3D component support to reweight pixel-level contributions, improving semantic rendering without training or changing geometry or opacity.

2 Related Work

2.1 Semantic and Object-Centric Gaussian Representations

Dense neural mapping has developed hybrid encodings (Wang et al., 2023; Johari et al., 2023) and neural point-cloud representations (Sandström et al., 2023). Gaussian SLAM systems exploit explicit primitives for differentiable tracking and mapping (Keetha et al., 2024; Yan et al., 2024; Matsuki et al., 2024). Subsequent research has extended Gaussian primitives with language features and instance information to support scene understanding. LangSplat (Qin et al., 2024) learns language features for open-vocabulary queries, while Gaussian Grouping (Ye et al., 2024) uses identity encodings supervised by associated multi-view masks and spatial consistency regularization to enable segmentation and editing. SAGA (Cen et al., 2025) further introduces scale-gated affinity features for segmentation at multiple granularities. Feature distillation (Zhou et al., 2024; Shi et al., 2024) and point-level open-vocabulary understanding (Wu et al., 2024) further connect Gaussian representations with semantic features.

Alongside these developments, semantic mapping has progressed from integrating predictions into geometric maps, as in SemanticFusion (McCormac et al., 2017) and PanopticFusion (Narita et al., 2019), to jointly reconstructing geometry, appearance, and semantics with Gaussian representations. Semantic neural fields (Zhi et al., 2021; Zhu et al., 2024) and object-level neural maps (Kong et al.,

2023) support structured scene interpretation. SGS-SLAM (Li et al., 2024b), NEDS-SLAM (Ji et al., 2024), and SemGauss-SLAM (Zhu et al., 2025), and GS^{3}LAM (Li et al., 2024a) integrate semantics into incremental Gaussian mapping. These approaches establish mechanisms for organizing and integrating semantic observations. Their primary focus is reconstructing a meaningful representation from an incoming image sequence; selecting the observations that would most improve that representation is a complementary planning problem.

2.2 Active View Selection and 3D Reconstruction

Active reconstruction addresses this planning problem by selecting viewpoints and paths under sensing and motion constraints. Frontier-based exploration (Yamauchi, 1997) expands known space, with hierarchical planners improving exploration efficiency (Zhou et al., 2021a; Cao et al., 2021). Next-best-view methods evaluate coverage and information gain (Bircher et al., 2016; Schmid et al., 2020); neural reconstruction methods use uncertainty to guide acquisition (Pan et al., 2022; Ran et al., 2023; Feng et al., 2024). Gaussian-based systems exploit efficient rendering and explicit spatial structure for these decisions. ActiveGAMER (Chen et al., 2025a) uses rendering-based view evaluation and coarse-to-fine exploration. ActiveGS (Jin et al., 2025) combines Gaussian confidence with a voxel map to target under-reconstructed and unexplored regions, and ActiveSplat (Li et al., 2025c) uses sparse topology for viewpoint sampling and hierarchical planning. Other approaches extend the decision process through learned observation value, longer planning horizons, or object context. Learned exploration can couple map estimation with global and local policies (Chaplot et al., 2020a). For object reconstruction, PUN (Zhang et al., 2026) predicts uncertainty maps and aggregates previous predictions to suppress redundant views. NextBestPath (Li et al., 2025b) jointly predicts accumulated coverage gains and obstacle maps to plan toward longer-term goals. OccamView (Gao et al., 2026) maintains object memory and evaluates hidden-region coverage to guide complementary observations under frame budgets. These studies demonstrate that history, directional coverage, and object context already inform reconstruction planning. Semantic reconstruction adds a further consideration: whether additional observations can resolve ambiguity or disagreement in category predictions.

2.3 Active Semantic Mapping and Reconstruction

Active semantic mapping incorporates semantic information into observation planning. Semantic Curiosity (Chaplot et al., 2020c) rewards inconsistent semantic predictions for active visual learning, while goal-oriented semantic exploration (Chaplot et al., 2020b) uses semantic maps to guide object search. These approaches extend acquisition objectives beyond geometric coverage. Within Gaussian-based reconstruction, ActiveSGM (Chen et al., 2025b) provides a particularly relevant example. It maintains sparse semantic attributes and combines coverage estimated from rendered silhouettes, semantic entropy, and travel distance to evaluate candidate views. Its formulation allows geometric exploration and semantic uncertainty to jointly influence observation selection.

Current entropy, temporal disagreement, and viewing history capture different needs: two objects with similar entropy can require different next views. Our planner links persistent object memory to view valuation and seeks complementary evidence from object-level history, while contribution reweighting addresses the uneven semantic support that affects final rendering.

3 Method

OAVP-Splat plans views with object-level observation value and reweights semantic contributions after mapping. Persistent object-level Gaussian groups store observation history and directional coverage, so ObjectVOI favors views that supply missing object evidence rather than relying on current entropy alone. A two-stage planner balances this value with scene coverage; geometric pose refinement and B-spline optimization realize selected views. Reweighting combines per-Gaussian response with connected-region support to improve semantic rendering (Figure 2).

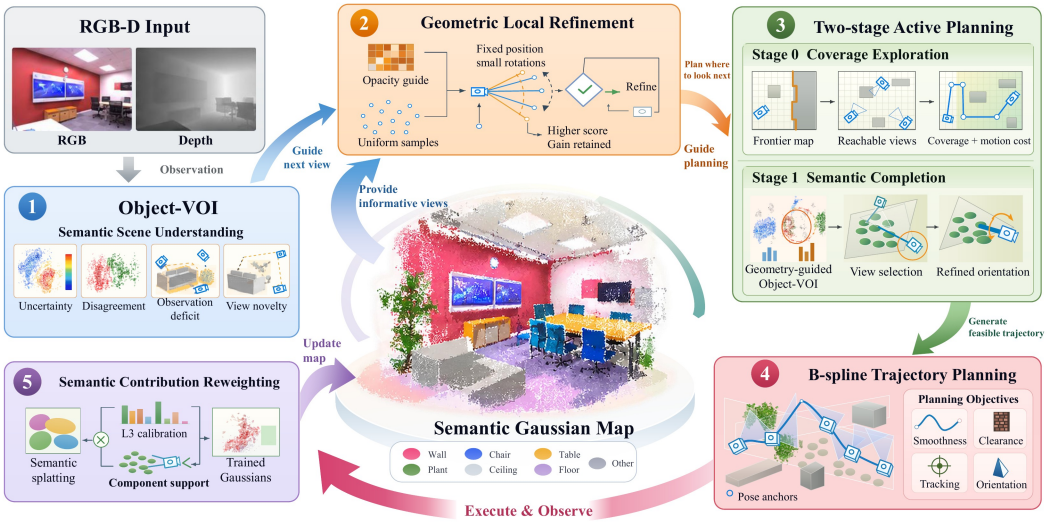

Figure 2: Overview of OAVP-Splat. ObjectVOI combines object semantic states and observation histories to score candidate views. Geometry-guided local refinement and two-stage active planning select acquisition poses, which a B-spline connects into a feasible trajectory. Acquired RGB-D observations update the map and the histories. After acquisition, semantic contribution reweighting adjusts Gaussian contributions using individual response strength and connected-component support.

3.1 Object-level observation value

Geometric coverage alone does not guarantee that the objects in the map have coherent semantic evidence. An object may remain ambiguous, be observed too infrequently, or be revisited from nearly identical directions. ObjectVOI turns these three failure modes into a value for acquiring a complementary view.

We associate reconstructed surface regions with predicted panoptic segments to form persistent object-level Gaussian groups. Each group stores its semantic state together with a history of supporting views. For every Gaussian, the observation count records depth-consistent support from acquired views, while near-duplicate camera poses are excluded from the history.

Identifying observation need. For object o and candidate pose v, we combine current ambiguity with temporal instability:

$$U_o(v) = (1-\beta)H_o(v) + \beta T_o, \tag{1}$$

where $H_o(v)$ is the opacity-weighted mean normalized entropy of the sampled Gaussians projected into v. The smoothed, normalized Jensen–Shannon divergence (Lin, 1991) T_o measures changes in the object's aggregated class distribution, and β balances the two signals. This temporal term keeps an object eligible for inspection when its label changes across map updates, even if its current distribution is confident.

We further measure the observation deficit $D_o(v)$. The deficit of each Gaussian decreases linearly to zero at a saturation count. We average the opacity-weighted deficits of the projected portion and of the complete sampled object with equal weights. Consequently, well-observed regions do not mask poor support elsewhere on the same object.

Valuing a complementary view. The novelty factor $\nu_o(v)$ combines exponentially decaying observation weights with the dissimilarity between the object-to-camera direction and the closest recorded direction bin. It therefore favors views that expose less observed portions from less familiar directions. The contribution of object o is

$$V_o(v) = \sqrt{f_o(v)}\,\nu_o(v)\left[U_o(v) + \lambda_D D_o(v) + 4\lambda_P f_o(v)\big(1-f_o(v)\big)\right], \tag{2}$$

where $f_o(v)$ is the fraction of sampled object opacity whose projected footprints intersect the image. The weights λ_D and λ_P control the observation deficit and the preference for partial projected support, respectively. Thus ObjectVOI separates observation need (uncertainty, temporal change, and deficit) from candidate-view contribution (projected support and directional novelty), rather than treating these signals as independent per-Gaussian scores.

The view value aggregates objects with sufficient projected support:

$$V(v) = \frac{\sum_{o \in \mathcal{O}_v} V_o(v)}{\sqrt{\max(1, |\mathcal{O}_v|)}}, \tag{3}$$

where $\mathcal{O}_v$ is the set of objects visible with sufficient support from v. Section 3.3 uses $V(v)$ to trade scene discovery against object inspection.

3.2 Geometry-guided local pose refinement

The planner produces a coarse endpoint from a finite candidate set. A small change around that endpoint can improve visibility, but an unconstrained change may destroy the semantic reason for visiting it. We therefore refine the parent pose v_0 locally while preserving its semantic value.

Among Gaussians that pass a depth-consistency check in the parent view, the Gaussian center with the largest positive geometric-support response provides an anchor. Let $\mathbf{p}_t$ and $\mathbf{p}_0$ denote the current and parent positions, and let $\hat{\mathbf{d}}$ be the unit direction from the former to the latter. We consider $\mathbf{p}_0$ and $\mathbf{p}_0 \pm \delta\hat{\mathbf{d}}$, with $\delta = 0.10\,\mathrm{m}$, and rotate each candidate toward the anchor by at most 5° from the parent orientation. When no valid anchor is available, the parent is retained; coincident positions allow rotation alone.

All candidates are evaluated against the same original pool $\mathcal{C}$. Let $\mathcal{F}(v_0)$ contain the parent and alternatives that improve the planning score while satisfying path and observation-budget constraints. We select

$$v^\star = \arg\max_{v \in \mathcal{F}(v_0)} J_{\mathcal{C}}(v) \quad \text{subject to} \quad S(v) \geq S(v_0) - \epsilon_S, \tag{4}$$

where $J_{\mathcal{C}}$ is the Stage 1 objective in Section 3.3. The semantic value $S(v)$ retains only $U_o(v)$ inside the brackets of Eq. (2) and then aggregates it as in Eq. (3); $\epsilon_S = 2 \times 10^{-6}$ accounts for numerical tolerance. An alternative must exceed both the recorded parent score and its recomputation under the fixed reference by 10^{-6}. An accepted refinement changes only the final route edge; otherwise, the parent pose and route are retained.

3.3 Two-stage active view planning

The acquisition schedule separates map bootstrapping from semantic evidence completion. Early views establish free-space coverage and the object histories needed for later revisits. Subsequent views can then balance new-space discovery against the value of inspecting already mapped objects.

Stage 0: establishing coverage. We sample camera positions and viewing directions in newly discovered free space. For a candidate pool $\mathcal{C}$ with positive total coverage gain, the Stage 0 score is

$$J_{\mathcal{C}}^{(0)}(v) = \omega_{\mathcal{C}}(v) \frac{I(v)}{\sum_{v' \in \mathcal{C}} I(v')} \widetilde{H}_{\mathcal{C}}(v). \tag{5}$$

Here, $I(v)$ counts pixels without mapped depth support, $\omega_{\mathcal{C}}(v)$ is one minus the softmax-normalized distance to the current camera, and $\widetilde{H}_{\mathcal{C}}(v)$ is the semantic-entropy factor evaluated over the same pool. The highest-scoring view balances coverage gain, travel distance, and semantic information. Newly discovered free space contributes new candidates, whereas low-gain or repeatedly visited candidates are retired.

From scene coverage to object inspection. An empty sampled pool does not imply complete coverage. We therefore augment it with views near the boundary of observed free space and select the highest- scoring feasible view with positive coverage gain using Eq. (5). This search is bounded so that budget remains for object inspection. When no admissible frontier view remains, or the frontier allowance or global exploration limit is reached, the method switches to Stage 1. The map and histories accumulated by Stage 0 initialize the object prior carried into that stage.

Stage 1: balancing discovery and inspection. Coverage alone can undervalue a revisit to an object whose semantic state is still unresolved. We therefore add ObjectVOI as an independent term:

$$J_{\mathcal{C}}(v) = \mathcal{N}_{\mathcal{C}}\left[\omega_{\mathcal{C}} I e^{H_{\text{img}}}\right](v) + \lambda_V \mathcal{N}_{\mathcal{C}}[\omega_{\mathcal{C}} V](v). \tag{6}$$

Here, $H_{\text{img}}(v)$ is the mean entropy of the rendered semantic image, $V(v)$ is the object-level value in Eq. (3), and $\omega_{\mathcal{C}}(v)$ is the same distance weight used in Stage 0. The operator $\mathcal{N}_{\mathcal{C}}$ normalizes nonnegative scores by their sum over the pool, and λ_V controls the contribution of object inspection.

The first term rewards discovery, while the second turns accumulated object history into targeted revisits. Their additive form allows a view with little new-space coverage to compete when it provides valuable object evidence, while separate normalization keeps λ_V independent of the raw scales of the two terms.

Executing and updating the decision. The highest-scoring view becomes v_0 for local pose refinement (Section 3.2). Its refined endpoint and route then define the target for trajectory optimization (Section 3.4). Acquired RGB-D frames update both the map and the object histories before the next decision. Stage 1 continues until the candidate pool or observation budget is exhausted, after which the map is refined and semantic contribution reweighting is applied.

3.4 B-SPLINE TRAJECTORY OPTIMIZATION

View planning specifies where evidence should be acquired; trajectory optimization determines how the sensor reaches those endpoints. We connect the planned observation positions with a cubic B-spline, as used in smooth local planning (Zhou et al., 2021b), and optimize the intervening path while keeping acquisition poses and timing fixed. The objective is

$$J_{\text{traj}} = \lambda_s J_{\text{smooth}} + \lambda_t J_{\text{track}} + \lambda_c J_{\text{clearance}} + \lambda_o J_{\text{orientation}}, \tag{7}$$

where the λ coefficients balance four normalized costs. J_{smooth} penalizes acceleration and jerk to avoid abrupt motion; J_{track} limits deviation from the reference path; $J_{\text{clearance}}$ discourages proximity to obstacles using the map's distance field; and $J_{\text{orientation}}$ measures deviation from the reference rotations. Interpolating these rotations preserves the intended viewing directions while translation is optimized.

3.5 SEMANTIC CONTRIBUTION REWEIGHTING

After acquisition, the map is fixed but the final semantic label at a pixel still depends on how overlapping Gaussian predictions are combined. We adjust their contributions using two sources of evidence: the response strength of each Gaussian and the support of the spatial region to which it belongs.

Individual response strength. For Gaussian i, softmax with temperature τ_s converts stored semantic coefficients b_{ic} into class probabilities p_{ic}. Because this normalization gives every Gaussian unit probability mass, we separately retain its positive response amplitude:

$$a_i = \text{clip}_{[a_{\min}, a_{\max}]}\left(\frac{\sum_c [b_{ic}]_+}{b_{\text{ref}}}\right), \tag{8}$$

where $[\cdot]_+$ retains positive coefficients, b_{ref} is the map-wide median of the per-Gaussian positive sums with a positive lower bound, and $[a_{\min}, a_{\max}]$ bounds the amplitude. Thus, p_{ic} expresses class preference and a_i controls the strength of that Gaussian in rendering. Both quantities come from its stored class support; coefficients for other classes contribute zero to the amplitude.

Regional support. To account for support from the surrounding semantic region, we group Gaussians with the same predicted global class into connected components on a voxel grid. These components are computed from the completed map and are distinct from the persistent planning objects. For component k, we combine view support, size, and class confidence:

$$z_k = \frac{m_k}{m_k + \lambda_m}\left(\frac{\log(1+n_k)}{\ell_{\text{ref}}}\right)^{\eta} \bar{p}_k, \tag{9}$$

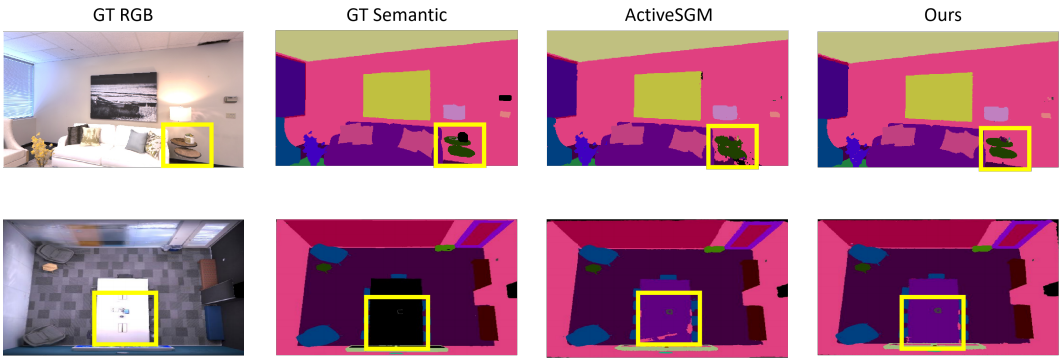

Figure 3: Semantic reconstruction on room0 (top) and office4 (bottom). Columns show RGB, semantic ground truth, ActiveSGM, and OAVP-Splat. The highlighted regions exhibit more coherent labels and fewer gaps in OAVP-Splat's semantic renderings.

where m_k counts views in which at least one member passes a depth-consistency visibility check, n_k is the number of member Gaussians, and $\bar{p}_k$ is their mean probability of the shared class. The reference ℓ_{ref} is the median log-transformed component size; λ_m and η control the influence of view count and size. View support is computed from the completed map's rendered depth at a fixed probe set of 36 camera poses, namely the same poses used for RGB/depth evaluation. This is a deterministic post-hoc visibility probe: it reads no RGB, semantic labels, or ground-truth depth, and it cannot affect view selection, map optimization, or trajectory generation.

Normalizing positive scores by their Gaussian-count-weighted median and clipping yields a shared weight κ_k for each component. Components with zero support retain unit weight.

Applying semantic contribution reweighting. The final label at pixel u is

$$\widehat{y}(u) = \arg\max_c \sum_i w_i(u)\, a_i\, \kappa_{k(i)}\, p_{ic}, \tag{10}$$

where $w_i(u)$ is the alpha-compositing weight and $k(i)$ identifies the component containing Gaussian i. The two positive factors preserve each Gaussian's class ranking while adjusting its influence on the pixel-level decision. Gaussian geometry and opacity remain unchanged.

4 Experiments

We evaluate whether OAVP-Splat improves semantic reconstruction while preserving geometry and appearance, and examine how its components contribute to these gains.

4.1 Experimental Setup

Datasets and comparisons. We evaluate on eight Replica (Straub et al., 2019) scenes and four Matterport3D (Chang et al., 2017) scenes in Habitat (Savva et al., 2019), with exploration budgets of 2,000 and 5,000 steps per scene, respectively. Mapping uses RGB-D observations, known camera poses, and semantic predictions from OneFormer (Jain et al., 2023). ActiveSGM (Chen et al., 2025b) is the primary baseline; SGS-SLAM (Li et al., 2024b), TGS-SLAM (Sun et al., 2026), and Hier-SLAM (Li et al., 2025a) provide semantic SLAM references on four Replica scenes. We report mIoU and F1 for semantic quality, with mAP in the component analysis. Accuracy, completeness, and the 5 cm completeness ratio assess geometry; PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and Depth L1 assess rendering. RGB and depth are evaluated on 36 novel views per scene, and scene averages are equally weighted.

4.2 Semantic Reconstruction Results

Table 1: Semantic reconstruction results (%). Top: per-scene Replica comparison, with better scores bold. Bottom: means over four Replica scenes (left) and four Matterport3D scenes (right).

(a) Replica: eight scenes

	mIoU ↑		F1 ↑	
Scene	ActiveSGM	Ours	ActiveSGM	Ours
room0	82.76	**86.15**	86.84	**90.46**
room1	**89.33**	89.06	**93.28**	92.98
room2	81.35	**85.74**	85.07	**89.82**
office0	81.10	**82.52**	85.36	**86.72**
office1	80.18	**81.72**	83.72	**85.25**
office2	86.81	**89.48**	91.07	**93.38**
office3	**89.98**	89.22	**93.65**	92.99
office4	81.35	**81.66**	83.52	**83.98**
Mean	84.11	**85.69**	87.82	**89.45**

(b) Replica: four scenes

Method	mIoU ↑	F1 ↑
SGS-SLAM	83.06	88.84
TGS-SLAM	84.34	88.97
Hier-SLAM	83.82	89.53
Ours	**85.87**	**90.00**

(c) Matterport3D: four scenes

Method	mIoU ↑	F1 ↑	PSNR ↑
ActiveSGM	66.73	72.53	**26.42**
Ours	**67.66**	**73.46**	26.06

(b): room0–room2 and office0; SLAM uses GT supervision and GT-present averaging, Ours uses predictions and union averaging. (a,c) use union; (c) uses 36 novel views, with PSNR in dB. Protocols and scene IDs: Appendices A.1, D, and E.

Semantic quality on Replica (panel a). OAVP-Splat achieves 85.69% mIoU and 89.45% F1, compared with ActiveSGM's 84.11% and 87.82% (Table 1a). Both metrics improve in six of eight scenes; room0 and room2 gain 3.39 and 4.39 mIoU points. Overall, OAVP-Splat shows a consistent semantic advantage over ActiveSGM across the Replica benchmark.

Reference comparison on Replica (panel b). On the four Replica reference scenes, OAVP-Splat reaches 85.87% mIoU and 90.00% F1 under union averaging. The supervised GT-present references range from 83.06–84.34% mIoU and 88.84–89.53% F1, so this panel provides reference context rather than a controlled ablation because the class-averaging and supervision protocols differ. Overall, OAVP-Splat remains broadly competitive with these supervised references under the stated protocol differences.

Reconstruction on Matterport3D (panel c). Across the four scenes completed for both methods, OAVP-Splat reaches 67.66% mIoU and 73.46% F1, improving over ActiveSGM's 66.73% and 72.53% by 0.93 percentage points; PSNR is 26.06 versus 26.42 dB (Table 1c). Overall, OAVP-Splat improves semantic quality over ActiveSGM while maintaining comparable rendering quality on Matterport3D.

4.3 Component Analysis

Beyond frontier coverage. Table 2 separates the planning and aggregation controls on office0, office1, and room0. Relative to frontier exploration, Ours improves mIoU by 3.76, 3.67, and 1.55 points, and F1 by 3.92, 3.81, and 1.82 points, respectively. Removing ObjectVOI lowers mIoU/F1 to 82.31/86.36, 81.50/85.03, and 85.35/89.74; Ours is best in all three scenes.

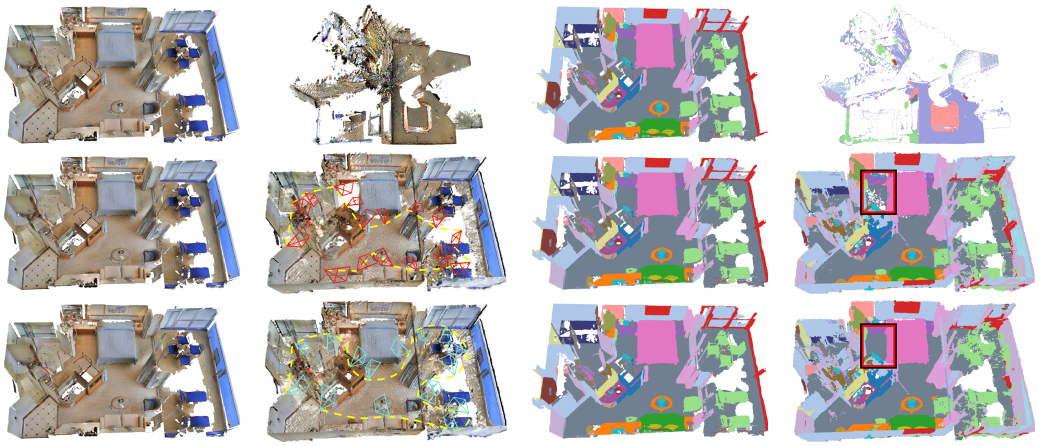

Figure 4: HxpKQynjfin reconstruction. Rows: GS^{3}LAM (passive path), ActiveSGM, and OAVP-Splat. Columns: RGB ground truth/prediction and semantic ground truth/prediction. Trajectory overlays compare motion; boxes highlight semantic regions.

Table 2: Component comparisons on office0, office1, and room0 (%). Best values are bold. The no-reweighting variant uses the same reconstructed map as Ours.

	office0			office1			room0		
Configuration	mIoU ↑	mAP ↑	F1 ↑	mIoU ↑	mAP ↑	F1 ↑	mIoU ↑	mAP ↑	F1 ↑
Frontier exploration	78.76	96.28	82.80	78.05	**95.31**	81.44	84.60	**97.10**	88.64
W/o ObjectVOI	82.31	95.71	86.36	81.50	94.88	85.03	85.35	96.32	89.74
W/o reweighting	80.34	**96.66**	84.33	79.18	95.04	82.57	85.46	97.08	89.74
Ours	**82.52**	96.40	**86.72**	**81.72**	94.33	**85.25**	**86.15**	96.78	**90.46**

Contribution of semantic reweighting. With the same reconstructed maps, reweighting adds 2.18/2.39 mIoU/F1 points on office0, 2.54/2.68 on office1, and 0.69/0.72 on room0 relative to W/o reweighting. Across eight Replica scenes, mean mIoU follows 84.60%→85.30%→85.69% as semantic contribution terms are enabled.

5 Conclusion

OAVP-Splat combines object-aware view planning with semantic contribution reweighting for active semantic reconstruction. Object histories guide coverage and targeted inspection, while per-Gaussian responses and same-class spatial support improve semantic aggregation without additional training. Replica results show improved semantic accuracy with comparable geometric fidelity; Matterport3D extends the evaluation to scanned indoor scenes. Fixed-map ablations establish reweighting's contribution, supporting the joint use of complementary observation acquisition and effective semantic aggregation.

AI use statement

AI tools assisted writing, language refinement, and literature discovery; the authors retain responsibility for all claims, interpretations, references, and final content.

REFERENCES

Andreas Bircher, Mina Kamel, Kostas Alexis, Helen Oleynikova, and Roland Siegwart. Receding Horizon "Next-Best-View" Planner for 3D Exploration. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1462–1468, 2016. doi: 10.1109/icra.2016.7487281. URL https://doi.org/10.1109/icra.2016.7487281.

Chao Cao, Hongbiao Zhu, Howie Choset, and Ji Zhang. TARE: A Hierarchical Framework for Efficiently Exploring Complex 3D Environments. In *Robotics: Science and Systems XVII*, 2021. doi: 10.15607/rss.2021.xvii.018. URL https://doi.org/10.15607/rss.2021.xvii.018.

Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment Any 3D Gaussians. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1971–1979, 2025. doi: 10.1609/aaai.v39i2.32193. URL https://doi.org/10.1609/aaai.v39i2.32193.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D Data in Indoor Environments. In *International Conference on 3D Vision (3DV)*, pp. 667–676, 2017. doi: 10.1109/3dv.2017.00081. URL https://doi.org/10.1109/3dv.2017.00081.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to Explore using Active Neural SLAM. In *International Conference on Learning Representations (ICLR)*, 2020a. URL https://arxiv.org/abs/2004.05155.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object Goal Navigation using Goal-Oriented Semantic Exploration. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4247–4258, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/2c75cf2681788adaca63aa95ae028b22-Abstract.html.

Devendra Singh Chaplot, Helen Jiang, Saurabh Gupta, and Abhinav Gupta. Semantic Curiosity for Active Visual Learning. In *European Conference on Computer Vision (ECCV)*, pp. 309–326, 2020c. doi: 10.1007/978-3-030-58539-6_19. URL https://doi.org/10.1007/978-3-030-58539-6_19.

Liyan Chen, Huangying Zhan, Kevin Chen, Xiangyu Xu, Qingan Yan, Changjiang Cai, and Yi Xu. ActiveGAMER: Active GAussian Mapping through Efficient Rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16486–16497, 2025a. doi: 10.1109/cvpr52734.2025.01537. URL https://doi.org/10.1109/cvpr52734.2025.01537.

Liyan Chen, Huangying Zhan, Hairong Yin, Yi Xu, and Philippos Mordohai. Understanding while Exploring: Semantics-driven Active Mapping. In *Advances in Neural Information Processing Systems*, volume 38, pp. 37293–37320, 2025b. doi: 10.52202/085713-1113. URL https://doi.org/10.52202/085713-1113.

Ziyue Feng, Huangying Zhan, Zheng Chen, Qingan Yan, Xiangyu Xu, Changjiang Cai, Bing Li, Qilun Zhu, and Yi Xu. NARUTO: Neural Active Reconstruction from Uncertain Target Observations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21572–21583, 2024. doi: 10.1109/cvpr52733.2024.02038. URL https://doi.org/10.1109/cvpr52733.2024.02038.

Hongbo Gao, Wei Zhang, Zeyu Ni, Dihao Zhu, Ruifeng Li, Yunke Wang, and Chang Xu. OccamView: Object-Conditioned View Selection for Frame-Budgeted Active 3D Gaussian Reconstruction. *arXiv preprint arXiv:2608.16499*, 2026. URL https://arxiv.org/abs/2608.16499.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning. In *IEEE International*

Conference on Robotics and Automation (ICRA), pp. 5021–5028, 2024. doi: 10.1109/icra57147.2024.10610243. URL https://doi.org/10.1109/icra57147.2024.10610243.

Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. OneFormer: One Transformer to Rule Universal Image Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2989–2998, 2023. doi: 10.1109/cvpr52729.2023.00292. URL https://doi.org/10.1109/cvpr52729.2023.00292.

Krishna Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Ganesh Iyer, Soroush Saryazdi, Tao Chen, Alaa Maalouf, Shuang Li, Nikhil Keetha, Ayush Tewari, Joshua Tenenbaum, Celso Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. ConceptFusion: Open-set multimodal 3D mapping. In *Robotics: Science and Systems XIX*, 2023. doi: 10.15607/rss.2023.xix.066. URL https://doi.org/10.15607/rss.2023.xix.066.

Yiming Ji, Yang Liu, Guanghu Xie, Boyu Ma, Zongwu Xie, and Hong Liu. NEDS-SLAM: A Neural Explicit Dense Semantic SLAM Framework Using 3D Gaussian Splatting. *IEEE Robotics and Automation Letters*, 9(10):8778–8785, 2024. doi: 10.1109/lra.2024.3451390. URL https://doi.org/10.1109/lra.2024.3451390.

Liren Jin, Xingguang Zhong, Yue Pan, Jens Behley, Cyrill Stachniss, and Marija Popović. ActiveGS: Active Scene Reconstruction Using Gaussian Splatting. *IEEE Robotics and Automation Letters*, 10(5):4866–4873, 2025. doi: 10.1109/lra.2025.3555149. URL https://doi.org/10.1109/lra.2025.3555149.

Mohammad Mahdi Johari, Camilla Carta, and François Fleuret. ESLAM: Efficient Dense SLAM System Based on Hybrid Representation of Signed Distance Fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17408–17419, 2023. doi: 10.1109/cvpr52729.2023.01670. URL https://doi.org/10.1109/cvpr52729.2023.01670.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21357–21366, 2024. doi: 10.1109/cvpr52733.2024.02018. URL https://doi.org/10.1109/cvpr52733.2024.02018.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433. URL https://doi.org/10.1145/3592433.

Xin Kong, Shikun Liu, Marwan Taher, and Andrew J. Davison. vMAP: Vectorised Object Mapping for Neural Field SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 952–961, 2023. doi: 10.1109/cvpr52729.2023.00098. URL https://doi.org/10.1109/cvpr52729.2023.00098.

Boying Li, Zhixi Cai, Yuan-Fang Li, Ian Reid, and Hamid Rezatofighi. Hier-SLAM: Scaling-Up Semantics in SLAM with a Hierarchically Categorical Gaussian Splatting. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9748–9754, 2025a. doi: 10.1109/icra55743.2025.11127775. URL https://doi.org/10.1109/icra55743.2025.11127775.

Linfei Li, Lin Zhang, Zhong Wang, and Ying Shen. GS^{3}LAM: Gaussian Semantic Splatting SLAM. In *ACM International Conference on Multimedia*, pp. 3019–3027, 2024a. doi: 10.1145/3664647.3680739. URL https://doi.org/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. URL https://doi.org/10.1007/978-3-031-72751-1_10.

Shiyao Li, Antoine Guédon, Clémentin Boittiaux, Shizhe Chen, and Vincent Lepetit. NextBestPath: Efficient 3D Mapping of Unseen Environments. In *International Conference on Learning Representations (ICLR)*, 2025b. URL https://openreview.net/forum?id=7WaRh4gCXp.

Yuetao Li, Zijia Kuang, Ting Li, Qun Hao, Zike Yan, Guyue Zhou, and Shaohui Zhang. ActiveSplat: High-Fidelity Scene Reconstruction Through Active Gaussian Splatting. *IEEE Robotics and Automation Letters*, 10(8):8099–8106, 2025c. doi: 10.1109/lra.2025.3580331. URL https://doi.org/10.1109/lra.2025.3580331.

J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115. URL https://doi.org/10.1109/18.61115.

Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Gaussian Splatting SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18039–18048, 2024. doi: 10.1109/cvpr52733.2024.01708. URL https://doi.org/10.1109/cvpr52733.2024.01708.

John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. SemanticFusion: Dense 3D semantic mapping with convolutional neural networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4628–4635, 2017. doi: 10.1109/icra.2017.7989538. URL https://doi.org/10.1109/icra.2017.7989538.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision (ECCV)*, pp. 405–421, 2020. doi: 10.1007/978-3-030-58452-8_24. URL https://doi.org/10.1007/978-3-030-58452-8_24.

Gaku Narita, Takashi Seno, Tomoya Ishikawa, and Yohsuke Kaji. PanopticFusion: Online Volumetric Semantic Mapping at the Level of Stuff and Things. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4205–4212, 2019. doi: 10.1109/iros40897.2019.8967890. URL https://doi.org/10.1109/iros40897.2019.8967890.

Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. ActiveNeRF: Learning Where to See with Uncertainty Estimation. In *European Conference on Computer Vision (ECCV)*, pp. 230–246, 2022. doi: 10.1007/978-3-031-19827-4_14. URL https://doi.org/10.1007/978-3-031-19827-4_14.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3D Language Gaussian Splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20051–20060, 2024. doi: 10.1109/cvpr52733.2024.01895. URL https://doi.org/10.1109/cvpr52733.2024.01895.

Yunlong Ran, Jing Zeng, Shibo He, Jiming Chen, Lincheng Li, Yingfeng Chen, Gimhee Lee, and Qi Ye. NeurAR: Neural Uncertainty for Autonomous 3D Reconstruction With Implicit Neural Representations. *IEEE Robotics and Automation Letters*, 8(2):1125–1132, 2023. doi: 10.1109/lra.2023.3235686. URL https://doi.org/10.1109/lra.2023.3235686.

Erik Sandström, Yue Li, Luc Van Gool, and Martin R. Oswald. Point-SLAM: Dense Neural Point Cloud-based SLAM. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18387–18398, 2023. doi: 10.1109/iccv51070.2023.01690. URL https://doi.org/10.1109/iccv51070.2023.01690.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9338–9346, 2019. doi: 10.1109/iccv.2019.00943. URL https://doi.org/10.1109/iccv.2019.00943.

Lukas Schmid, Michael Pantic, Raghav Khanna, Lionel Ott, Roland Siegwart, and Juan Nieto. An Efficient Sampling-Based Method for Online Informative Path Planning in Unknown Environments. *IEEE Robotics and Automation Letters*, 5(2):1500–1507, 2020. doi: 10.1109/lra.2020.2969191. URL https://doi.org/10.1109/lra.2020.2969191.

Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language Embedded 3D Gaussians for Open-Vocabulary Scene Understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5333–5343, 2024. doi: 10.1109/cvpr52733.2024.00510. URL https://doi.org/10.1109/cvpr52733.2024.00510.

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. URL https://arxiv.org/abs/1906.05797.

Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J. Davison. iMAP: Implicit Mapping and Positioning in Real-Time. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6209–6218, 2021. doi: 10.1109/iccv48922.2021.00617. URL https://doi.org/10.1109/iccv48922.2021.00617.

Guoxi Sun, Handong Shen, Xiaohao Liu, Xinchao Li, Lingyu Liang, Beibei Liu, and Shuangping Huang. TGS-SLAM: Tri-Plane Gaussian Splatting for Semantic SLAM. *IEEE Robotics and Automation Letters*, 11(7):8300–8307, 2026. doi: 10.1109/lra.2026.3692078. URL https://doi.org/10.1109/lra.2026.3692078.

Hengyi Wang, Jingwen Wang, and Lourdes Agapito. Co-SLAM: Joint Coordinate and Sparse Parametric Encodings for Neural Real-Time SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13293–13302, 2023. doi: 10.1109/cvpr52729.2023.01277. URL https://doi.org/10.1109/cvpr52729.2023.01277.

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. URL https://doi.org/10.1109/tip.2003.819861.

Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, and Jian Zhang. OpenGaussian: Towards Point-Level 3D Gaussian-based Open Vocabulary Understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 19114–19138, 2024. doi: 10.52202/079017-0604. URL https://doi.org/10.52202/079017-0604.

B. Yamauchi. A frontier-based approach for autonomous exploration. In *IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)*, pp. 146–151, 1997. doi: 10.1109/cira.1997.613851. URL https://doi.org/10.1109/cira.1997.613851.

Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. GS-SLAM: Dense Visual SLAM with 3D Gaussian Splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19595–19604, 2024. doi: 10.1109/cvpr52733.2024.01853. URL https://doi.org/10.1109/cvpr52733.2024.01853.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian Grouping: Segment and Edit Anything in 3D Scenes. In *European Conference on Computer Vision (ECCV)*, pp. 162–179, 2024. doi: 10.1007/978-3-031-73397-0_10. URL https://doi.org/10.1007/978-3-031-73397-0_10.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018. doi: 10.1109/cvpr.2018.00068. URL https://doi.org/10.1109/cvpr.2018.00068.

Zhengquan Zhang, Feng Xu, and Mengmi Zhang. Peering into the Unknown: Active View Selection with Neural Uncertainty Maps for 3D Reconstruction. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=e7gbgdw05A.

Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J. Davison. In-Place Scene Labelling and Understanding with Implicit Scene Representation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15818–15827, 2021. doi: 10.1109/iccv48922.2021.01554. URL https://doi.org/10.1109/iccv48922.2021.01554.

Boyu Zhou, Yichen Zhang, Xinyi Chen, and Shaojie Shen. FUEL: Fast UAV Exploration Using Incremental Frontier Structure and Hierarchical Planning. *IEEE Robotics and Automation Letters*, 6(2):779–786, 2021a. doi: 10.1109/lra.2021.3051563. URL https://doi.org/10.1109/lra.2021.3051563.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3DGS: Supercharging 3D Gaussian Splatting to Enable Distilled Feature Fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21676–21685, 2024. doi: 10.1109/cvpr52733.2024.02048. URL https://doi.org/10.1109/cvpr52733.2024.02048.

Xin Zhou, Zhepei Wang, Hongkai Ye, Chao Xu, and Fei Gao. EGO-Planner: An ESDF-Free Gradient-Based Local Planner for Quadrotors. *IEEE Robotics and Automation Letters*, 6(2):478–485, 2021b. doi: 10.1109/lra.2020.3047728. URL https://doi.org/10.1109/lra.2020.3047728.

Siting Zhu, Guangming Wang, Hermann Blum, Jiuming Liu, Liang Song, Marc Pollefeys, and Hesheng Wang. SNI-SLAM: Semantic Neural Implicit SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21167–21177, 2024. doi: 10.1109/cvpr52733.2024.02000. URL https://doi.org/10.1109/cvpr52733.2024.02000.

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. doi: 10.1109/iros60139.2025.11246462. URL https://doi.org/10.1109/iros60139.2025.11246462.

Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R. Oswald, and Marc Pollefeys. NICE-SLAM: Neural Implicit Scalable Encoding for SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12776–12786, 2022. doi: 10.1109/cvpr52688.2022.01245. URL https://doi.org/10.1109/cvpr52688.2022.01245.

APPENDIX ROADMAP

This appendix is organized around four questions. Appendix A specifies the implementation and evaluation protocol. Appendix B gives the complete Replica reconstruction results by scene. Appendix C isolates the contributions of planning, semantic reweighting, and parameter sensitivity. Appendix D reports the Matterport3D transfer results, and Appendix E documents the semantic SLAM reference protocol and scores. Tables retain the primary metrics used in the paper; repeated secondary metrics are omitted when they do not add evidence for a main claim.

A REPRODUCIBILITY AND EVALUATION PROTOCOL

Implementation settings. All experiments use Habitat RGB-D observations at 680×1200 pixels with vertical/horizontal fields of view of 60° /90°. The exploration voxel size is 5 cm. Replica uses eight scenes and a 2,000-step budget; Matterport3D uses four scenes and a 5,000-step budget. Candidate exhaustion can terminate acquisition early. Mapping uses simulator ground-truth poses. The dataset-specific fine-tuned OneFormer checkpoints are loaded before each run and are not updated during active mapping. The map keeps static Top-16 semantic entries with raw parameters and a sparse renderer. Experiments use NVIDIA RTX 4090 GPUs. The complete method combines object-level observation value, geometry-guided local pose refinement, frontier completion with B-spline trajectory processing, and semantic contribution reweighting at the individual and component levels. Stage 0 establishes the map and object-history prefix, which initializes Stage 1. Reweighting is applied after training and adds no optimization steps. We set the object-value weight to $\lambda_V = 1.5$, individual amplitude bounds to [0.25, 4], and component-factor bounds to [0.5, 2]. Local refinement and the semantic softmax temperature follow Section 3.

Planning implementation details. Object-value aggregation requires at least four sampled Gaussians whose projected footprints intersect the image. Local refinement fully evaluates at most three additional candidates. A replacement may add at most eight path frames and must fit the remaining terminal-capture budget. Candidate positions must be in known free space; the replacement edge is checked with a voxel supercover. The semantic tolerance is $\epsilon_S = 2 \times 10^{-6}$, and the alternative must exceed both the recorded parent score and its recomputation under the fixed reference by 10^{-6}. Frontier search evaluates at most twelve candidates per round and permits at most two tasks per stage.

A.1 EVALUATION PROTOCOL

Semantic scores are computed from rendered class predictions on valid pixels with nonzero ground-truth labels. The active-method and ablation results use union class averaging: each view averages IoU and F1 over nonzero classes present in its prediction or ground truth. For ActiveSGM and OAVP-Splat on Replica, this evaluation comprises the 37 indices 0, 4, 9, ..., 179. RGB and depth use the 36 novel-view indices 4, 9, ..., 179. Matterport3D uses these 36 indices for both semantic and RGB/depth evaluation. The two active methods share each metric's view set.

The SGS-SLAM, TGS-SLAM, and Hier-SLAM references use the same 36 novel views on room0, room1, room2, and office0. Their GT-present metrics average classwise IoU and F1 over nonzero classes appearing in each ground-truth view, then average over views and scenes. Predictions outside the ground-truth class set remain errors for the affected ground-truth classes but do not add separate zero-valued classes to the macro average. These references are reported in a separate panel in Table 1; their per-scene results and metric definitions are given in Appendix E.

Top-1, Top-3, and mAP assess the semantic predictions with the same valid-pixel mask. PSNR, SSIM, LPIPS, and Depth L1 are evaluated on the 36 RGB/depth views.

Component calibration uses a fixed probe set of 36 camera poses (the same poses used for RGB/depth evaluation, excluding frame 0) and the completed map's rendered depth. The probe poses are used only to count geometric visibility in the post-hoc readout; no RGB, semantic labels, or ground-truth depth are read. The same probe set is used for every method and ablation, and calibration cannot feed back into view selection, trajectory processing, or map optimization. Scene means give equal weight to each scene; semantic differences are in percentage points (pp).

3D geometry. Accuracy is the mean nearest-neighbor distance from sampled Gaussian centers to reference-mesh samples; completeness reverses this direction. C.R. is the percentage of reference samples within 5 cm of a predicted point. The common evaluator uses 100,000 predicted and 100,000 reference points, a 1 cm voxel size, and geometry sampling seed 0 (reference-mesh sampling uses seed 17). ActiveSGM and Ours use the same evaluator and sampling definition. These point-based 3D metrics are separate from rendered Depth L1.

Reconstruction consistency. Semantic, RGB/depth, and geometric metrics for each OAVP-Splat scene use the same complete reconstructed map. The Replica room1 run uses seed 42. Component variants are reported separately in the ablation results.

B Replica Results by Scene

This section expands the main Replica averages into scene-level semantic, rendering, and geometry results. The same ActiveSGM and OAVP-Splat configurations, view sets, and evaluators are used throughout.

The tables below report the same complete Ours configuration as the main text.

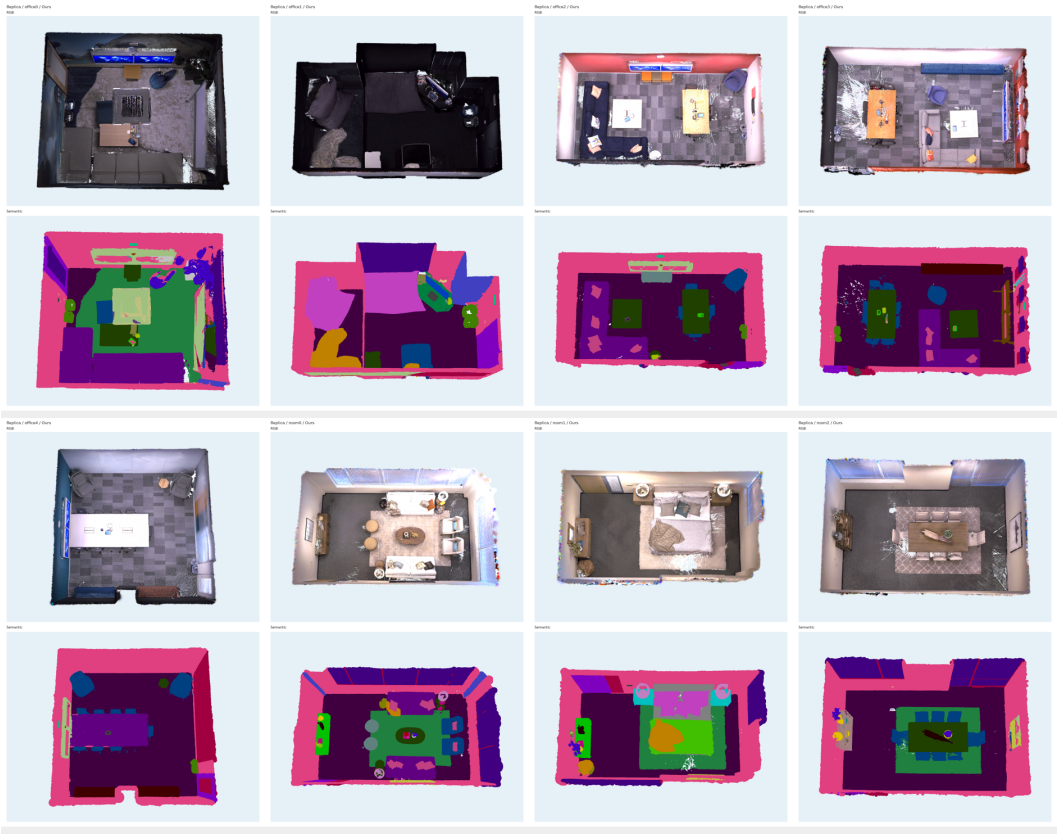

Figure 5: Qualitative OAVP-Splat reconstructions on Replica. For each displayed scene, the upper view is the RGB reconstruction and the paired lower view is the semantic reconstruction.

Figure 5 gives a scene-level overview of the reconstructed appearance and semantic map across the Replica rooms and offices.

Table S1 verifies that the semantic improvements are not obtained by sacrificing the reconstructed geometry: Ours has nearly identical accuracy and slightly better completeness and coverage.

Table S1: Replica 3D reconstruction over eight scenes.

Method	Acc. (cm) ↓	Comp. (cm) ↓	C.R. (%) ↑
ActiveSGM	2.4439	2.9879	88.723
Ours	2.4510	2.9483	88.950

Table S2: Replica rendering over eight scenes, using 36 views per scene.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	D-L1 (cm) ↓
ActiveSGM	31.65	0.9672	0.1191	1.31
Ours	31.97	0.9685	0.1203	1.15

Table S2 gives the aggregate RGB/depth rendering context for the semantic results; Table S4 breaks the same evaluation down by scene.

Table S3: Replica semantic segmentation by scene (%).

	mIoU ↑		F1 ↑	
Scene	ActiveSGM	Ours	ActiveSGM	Ours
room0	82.76	86.15	86.84	90.46
room1	89.33	89.06	93.28	92.98
room2	81.35	85.74	85.07	89.82
office0	81.10	82.52	85.36	86.72
office1	80.18	81.72	83.72	85.25
office2	86.81	89.48	91.07	93.38
office3	89.98	89.22	93.65	92.99
office4	81.35	81.66	83.52	83.98
Mean	84.11	85.69	87.82	89.45

Table S3 exposes scene variability behind the mean semantic scores. The complete method improves the mean while some individual scenes remain comparable or slightly lower, so the table is included to make that distribution explicit.

Table S4: Replica rendering by scene on 36 views. PSNR is in dB; Depth L1 is in cm.

	PSNR ↑		SSIM ↑		LPIPS ↓		D-L1 ↓	
Scene	SGM	Ours	SGM	Ours	SGM	Ours	SGM	Ours
room0	28.03	28.46	0.9416	0.9430	0.1822	0.1872	1.54	1.59
room1	30.42	30.58	0.9597	0.9601	0.1369	0.1359	0.89	0.86
room2	30.99	30.76	0.9657	0.9590	0.1047	0.1263	1.34	1.53
office0	35.69	35.52	0.9798	0.9774	0.0833	0.0965	1.27	1.42
office1	35.89	36.30	0.9730	0.9749	0.1242	0.1257	0.91	0.66
office2	28.69	31.23	0.9617	0.9782	0.1351	0.1052	2.27	1.17
office3	31.36	30.22	0.9790	0.9736	0.0946	0.1070	1.17	1.09
office4	32.11	32.73	0.9770	0.9818	0.0921	0.0785	1.11	0.90
Mean	31.65	31.97	0.9672	0.9685	0.1191	0.1203	1.31	1.15

Table S4 checks whether the aggregate rendering differences are driven by a single scene; the mean values match the main-text rendering comparison.

The largest completeness improvements over ActiveSGM occur on office3 (0.1824 cm) and office2 (0.0873 cm). These distances measure reference surface coverage by the reconstructed Gaussian map, complementing the semantic scores in Table S3.

Table S5: Replica 3D geometry by scene. Accuracy (A) and completeness (C) are in cm; C.R. is in %.

	ActiveSGM			Ours		
Scene	A ↓	C ↓	C.R. ↑	A ↓	C ↓	C.R. ↑
room0	2.738	3.522	82.62	2.772	3.549	81.94
room1	2.284	2.546	94.30	2.291	2.541	94.31
room2	2.414	2.636	94.23	2.385	2.620	94.00
office0	2.189	2.332	95.97	2.162	2.308	95.90
office1	1.627	1.806	97.19	1.679	1.813	97.70
office2	2.623	3.622	83.92	2.653	3.534	84.51
office3	2.924	3.958	78.00	2.919	3.776	79.21
office4	2.754	3.481	83.55	2.748	3.446	84.02
Mean	2.444	2.988	88.72	2.451	2.948	88.95

C Controlled Component Evidence

This section keeps comparisons that identify where the reported gains come from: frontier exploration versus the complete planner, and fixed-map semantic contribution reweighting.

C.1 Ablation Controls

Frontier control uses the SGM score with frontier fallback in both stages, disabling object value, local pose refinement, B-spline processing, and semantic contribution reweighting. Frontier tasks are considered only when the native candidate pool is exhausted, with at most two tasks per stage. The full method enables all components. The no-reweighting control reuses the full model's training weights with both reweighting factors disabled and performs no additional optimization. The two-stage workflow is not an independently toggled module.

Frontier and fixed-map reweighting comparisons cover office0–office3. The fixed-map comparisons isolate semantic contribution reweighting by keeping observations, geometry, and evaluation views unchanged. RGB/depth evaluation uses the same 36 views in all configurations.

Table S6 extends the main-text office0/office1 comparison to office0–office3. The full pipeline improves mIoU-G over the frontier control by 3.77, 3.67, and 3.04 pp on office0, office1, and office3. On office2 the frontier control reaches 89.86%, compared with 89.48% for Ours. Across the four scenes, Ours reaches 85.74% mIoU-G and 89.58% F1-G, versus 83.21% and 87.01% for frontier control. Adding the full reweighting to the no-reweighting control adds 1.21 pp mIoU-G on average across these four fixed maps. It changes how semantic evidence is aggregated without altering RGB/depth predictions.

Table S6: Four-scene planning and aggregation controls. Semantic scores are percentages; all rows share the same evaluation protocol.

	Frontier control		No reweighting		OAVP-Splat	
Scene	mIoU-G	F1-G	mIoU-G	F1-G	mIoU-G	F1-G
office0	78.76	82.80	80.34	84.33	82.52	86.72
office1	78.05	81.44	79.18	82.57	81.72	85.25
office2	89.86	93.97	89.19	93.01	89.48	93.38
office3	86.17	89.83	89.41	93.13	89.22	92.99
Mean	83.21	87.01	84.53	88.26	85.74	89.58

Table S6 reports the four-scene semantic controls; the fixed-map table below isolates reweighting with observations and geometry held constant.

C.2 Fixed-Map Semantic Contribution Reweighting

These ablations use all eight final Replica maps with three semantic aggregation settings for each selected checkpoint. No reweighting removes both reweighting factors; per-Gaussian response applies the individual transformation; the full OAVP-Splat setting additionally uses component support. All three share the same map and evaluation views. The eight-scene means are computed on this same set.

Table S7: Fixed-map semantic aggregation ablation over the same eight Replica maps. All rows use identical maps, observations, and evaluation views; only semantic contribution reweighting differs. mIoU and F1 are percentages.

Configuration	mIoU ↑	F1 ↑
No reweighting	84.60	88.28
Per-Gaussian response	85.30	88.99
Per-Gaussian + component support (OAVP-Splat)	85.69	89.45

On these fixed maps, per-Gaussian response weighting and component support improve mean mIoU by 0.70 and 0.39 pp, respectively. The maps and acquired observations remain fixed, attributing these changes to semantic aggregation.

C.3 Parameter Sensitivity

We evaluate parameter sensitivity on room0 and office0 with one seed and the union class-averaging protocol over the same 36 novel views. Readout sweeps recompute the semantic factors on a frozen checkpoint without optimization; planning sweeps reuse each scene's Stage 0 snapshot, rerun Stage 1 and trajectory processing, and then perform the final readout. Thus the first table measures aggregation sensitivity, whereas the second measures planning sensitivity.

Table S8: Fixed-map semantic readout sensitivity. Entries are mIoU-G percentages over 36 views; the default values are $\tau_s = 0.35$, $\lambda_m = 3$, and $\eta = 0.25$.

Parameter	Value	room0	office0
τ_s	0.20	85.877	81.749
τ_s (default)	0.35	85.966	82.221
τ_s	0.50	85.951	82.632
λ_m	1	85.978	82.240
λ_m (default)	3	85.966	82.221
λ_m	6	85.953	82.189
η	0.125	86.033	82.308
η (default)	0.25	85.966	82.221
η	0.5	85.756	81.925

Across the tested readout ranges, room0 varies by 0.089, 0.025, and 0.277 mIoU points for τ_s, λ_m, and η, respectively; office0 varies by 0.883, 0.051, and 0.383 points. The default is therefore within a high-performing region for both scenes, while τ_s and η show the larger scene-dependent changes. Default RGB/depth values are 28.456 dB PSNR, 0.943 SSIM, 0.187 LPIPS, and 1.590 cm Depth L1 on room0, and 35.525 dB, 0.977, 0.097, and 1.419 cm on office0; these metrics were not recomputed for parameter variants.

The planning sweep shows a consistent penalty when λ_V is increased to 3, especially on office0. The preferred β is scene dependent: room0 benefits from 0.6, whereas office0 benefits most from 0.15. These are observations within the tested values, not claims about unmeasured ranges; missing or failed runs are not treated as zero.

Table S9: Stage 1 planning parameter sensitivity. Values are final mIoU-G percentages; Δ is relative to the default setting, and frames count the Stage 1 trajectory.

Scene	Parameter	Value	mIoU-G (Δ pp)	Stage 1 frames
room0	λ_V	0.75	85.626 (-0.339)	325
room0	λ_V	3	84.863 (-1.103)	390
room0	β	0.15	86.093 (+0.127)	325
room0	β	0.6	86.903 (+0.938)	359
office0	λ_V	0.75	80.594 (-1.626)	212
office0	λ_V	3	80.213 (-2.007)	179
office0	β	0.15	84.623 (+2.402)	235
office0	β	0.6	81.948 (-0.273)	235

D MATTERPORT3D TRANSFER RESULTS

Table S10 reports the matched ActiveSGM and OAVP-Splat results on four Matterport3D scenes. Semantic and RGB/depth metrics use 36 novel views per scene. All metrics describe the same reconstructed map for each scene, and means give equal weight to the four scenes.

Table S10: Matched ActiveSGM and OAVP-Splat results on four completed Matterport3D scenes. mIoU-G and F1-G are percentages; PSNR is in dB.

Method	Scene	mIoU-G ↑	F1-G ↑	PSNR ↑
ActiveSGM	GdvgFV5R1Z5	70.87	76.86	26.42
Ours	GdvgFV5R1Z5	71.53	77.71	25.64
ActiveSGM	gZ6f7yhEvPG	71.14	76.96	26.21
Ours	gZ6f7yhEvPG	70.10	75.80	25.82
ActiveSGM	HxpKQynjfin	57.20	63.62	26.52
Ours	HxpKQynjfin	59.50	65.86	26.51
ActiveSGM	pLe4wQe7qrG	67.72	72.68	26.06
Ours	pLe4wQe7qrG	69.52	74.47	26.27
ActiveSGM	Four-scene mean	66.73	72.53	26.42
Ours	Four-scene mean	67.66	73.46	26.06

Table S10 is a transfer check on four Matterport3D scenes under the same 36-view protocol. Because both methods are now complete on all four scenes, the HxpKQynjfin qualitative example in Figure 4 is accompanied by a matched quantitative comparison here.

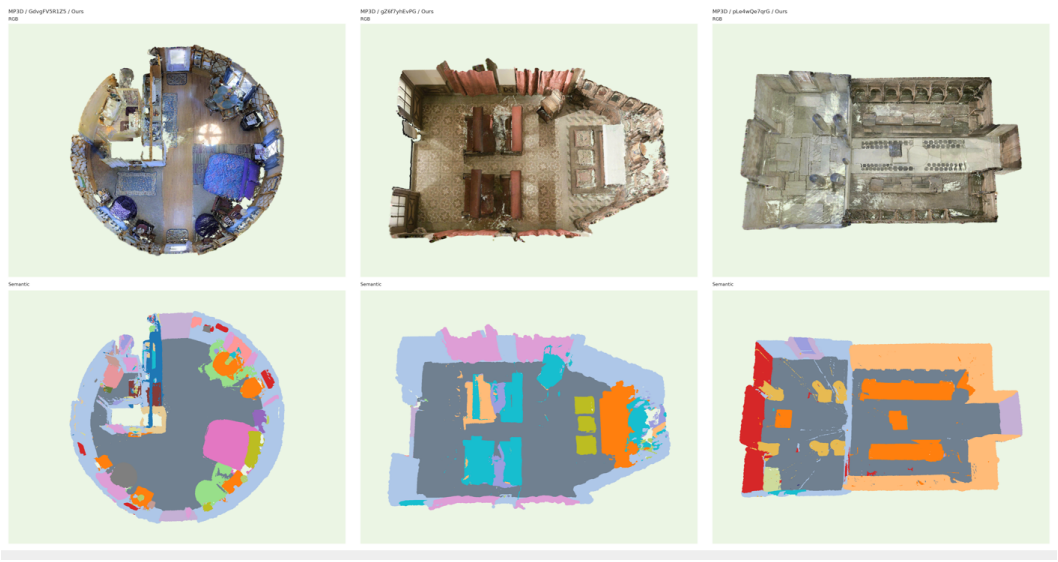

Figure 6: Qualitative OAVP-Splat reconstructions on Matterport3D. The upper row shows RGB reconstructions and the lower row shows the corresponding semantic reconstructions for the displayed scenes.

Figure 6 provides a qualitative view of the method's 3D and semantic reconstruction behavior on the four Matterport3D scenes.

E Semantic SLAM Reference Protocol

We evaluate SGS-SLAM (Li et al., 2024b), TGS-SLAM (Sun et al., 2026), and Hier-SLAM (Li et al., 2025a) on room0, room1, room2, and office0 using 36 novel views per scene. These reference configurations use ground-truth semantic supervision during mapping. Table S11 reports their GT-present mIoU and F1. Each scene score is the mean of 36 frame-level macro scores, and the overall result weights the four scenes equally. These supervised, protocol-matched scores provide reference context for the active-method results rather than an apples-to-apples causal ablation of OAVP-Splat.

Class averaging. For each view, let TP_c, FP_c, and FN_c denote the classwise pixel counts after excluding invalid and zero-labeled ground-truth pixels. We compute $\mathrm{IoU}_c = \mathrm{TP}_c/(\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c)$ and $\mathrm{F1}_c = 2\mathrm{TP}_c/(2\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c)$ independently. GT-present averages these scores over classes occurring in that view's ground truth. Union averaging, used for the active-method comparisons, also includes classes occurring only in the prediction. Both retain the same pixel-level prediction errors; their distinction is the set of classes included in the macro average.

Table S11: Semantic SLAM reference results on four Replica scenes, with 36 novel views per scene. Scores are frame-mean GT-present percentages; Mean weights scenes equally.

Method	Scene	mIoU ↑	F1 ↑
SGS-SLAM	room0	87.16	91.79
SGS-SLAM	room1	87.11	91.94
SGS-SLAM	room2	79.48	86.29
SGS-SLAM	office0	78.49	85.35
SGS-SLAM	Mean	83.06	88.84
TGS-SLAM	room0	90.56	94.51
TGS-SLAM	room1	90.26	94.22
TGS-SLAM	room2	74.98	79.95
TGS-SLAM	office0	81.54	87.19
TGS-SLAM	Mean	84.34	88.97
Hier-SLAM	room0	87.06	92.12
Hier-SLAM	room1	88.48	93.00
Hier-SLAM	room2	80.88	87.60
Hier-SLAM	office0	78.87	85.41
Hier-SLAM	Mean	83.82	89.53