OQ-ATOMS: DEFERRED SEMANTIC COMMITMENT FOR OPEN-VOCABULARY 3D SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary 3D segmentation is fundamental to indoor embodied agents, whose semantic maps are built incrementally as RGB-D observations stream in. However, existing systems follow a forward-prediction arc that collapses multi-view semantic alternatives before the query is known, and learned refiners that attempt to repair the result over-commit and degrade reliable predictions: current pipelines rarely ask *when* a semantic update should be allowed to overwrite the map. To address this limitation, we present OQ-ATOMS (Open-vocabulary Query-addressable Atoms), a three-stage framework that recasts semantic map updating as a query-time commitment decision. Our key idea is to group dense voxel evidence into compact atoms that preserve per-view observation provenance until query time, so that class scores are read out against the original multi-view support rather than an averaged posterior. A conservative committer then accepts an update only when four label-free evidence gates pass, keeping no-update as the default action. With fixed atoms and no per-scene tuning, OQ-ATOMS reaches 41.63 mIoU on Replica and 40.13 on ScanNet; this accuracy comes from provenance-weighted readout, while the committer is evaluated separately as a safety rule. Atom construction costs two orders of magnitude less than semantic scoring. We further introduce leave-one-scene-out (LOSO) calibration for label-free semantic committers, showing that same-set gains can be threshold artifacts and that deployment rules must be judged by held-out abstention behaviour.

1 INTRODUCTION

Powered by vision-language and segmentation foundation models Radford et al. (2021); Li et al. (2022); Ghiasi et al. (2022); Kirillov et al. (2023), open-vocabulary 3D scene understanding has advanced rapidly: an indoor embodied agent can now build its semantic map incrementally, lifting 2D vision-language evidence into 3D and fusing it into voxel-level class scores as RGB-D observations stream in Peng et al. (2023); Jatavallabhula et al. (2023); Engelmann et al. (2024); Kerr et al. (2023); Qin et al. (2024); Lee et al. (2025); Huang et al. (2023); Shafiullah et al. (2023), and then answering free-form language queries about the scene it inhabits Savva et al. (2019). Yet this streaming loop hides a decision that classical closed-vocabulary semantic mapping never had to face at this scale McCormac et al. (2017); Rosinol et al. (2020): every semantic update overwrites an existing map state, and a wrong overwrite propagates to all downstream queries the agent answers. Once evidence is stored, an update should be accepted only when its evidence is stronger than the no-update alternative, yet current systems rarely model *when* such an update should be committed. We refer to this missing decision as ***query-time semantic commitment***, and argue that it is a design axis of its own, separate from how semantic evidence is produced.

Existing open-vocabulary 3D systems can be categorized by where they place semantic evidence: feature-lifting maps and neural fields Peng et al. (2023); Jatavallabhula et al. (2023); Engelmann et al. (2024); Jiang et al. (2024); Zhu et al. (2024); Lee et al. (2025), radiance and Gaussian language fields Kerr et al. (2023); Qin et al. (2024); Zhu et al. (2025); Lu et al. (2025); Piekenbrinck et al. (2025), object-centric graphs and masks Gu et al. (2024); Werby et al. (2024); Takmaz et al. (2023); Nguyen et al. (2024); Takmaz et al. (2025), and training-free 2D groupings lifted to 3D Hajimiri et al. (2025); Shi et al. (2025); Bai et al. (2025); Barsellotti et al. (2025); Xiao et al. (2025); Alama et al. (2025). Despite this diversity, most pipelines follow the same ***forward-prediction arc*** (Fig. 1): multi-view semantic alternatives are collapsed *before* the query is known, so the commitment decision

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.

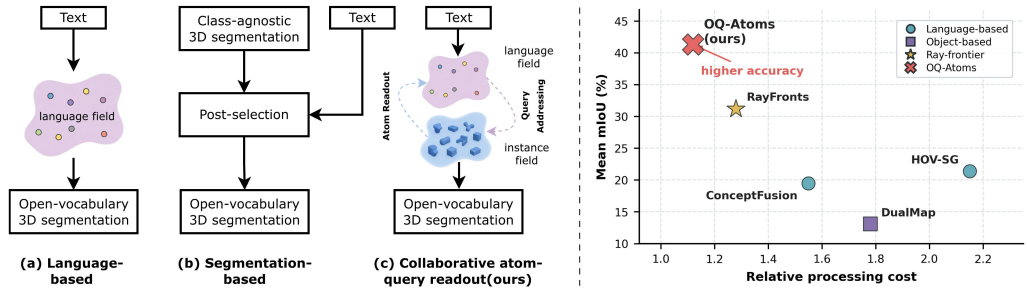

Figure 1: **Comparing open-vocabulary 3D segmentation paradigms.** (a) Language-based and (b) segmentation-based pipelines commit semantic evidence before the final query-time decision. (c) OQ-ATOMS instead keeps compact atom evidence query-addressable and defers commitment to provenance-weighted readout plus evidence gates. On matched indoor benchmarks, OQ-ATOMS outperforms existing methods in mIoU while targeting a lower relative query-time processing cost.

is made implicitly, early, and irreversibly. Learned refiners that try to repair the result afterwards miss the distinction between a useful edit and a safe one; trained to fit labels, they learn when an edit can reduce loss, not when it is safe to commit.

An over-commitment diagnostic makes the resulting failure mode concrete. On a representative ScanNet Dai et al. (2017) scene, a learned atom-query hypothesis improves mIoU by 2.65 points; on one Replica Straub et al. (2019) office scene, the *same* decoder reaches a mean gate of 0.94, commits almost everywhere, and degrades all four metrics. The hypothesis itself is useful; what the pipeline lacks is an evidence-based commitment rule that decides when to act on it. Deciding whether to act or to abstain is a classical question: it dates back to Chow's reject rule Chow (1970), underlies selective prediction El-Yaniv & Wiener (2010); Geifman & El-Yaniv (2017; 2019); Hendrycks & Gimpel (2017) and conformal abstention Vovk et al. (2005), and connects to uncertainty estimation Gal & Ghahramani (2016) and to the causal view of an update as an intervention whose value must be assessed before commitment Pearl (2009); Oberst & Sontag (2019). Open-vocabulary 3D mapping, however, rarely makes this abstention decision explicit at the map-update step.

In this work, we propose OQ-ATOMS, whose core contribution is a query-addressable evidence state for open-vocabulary 3D maps. No-update is preserved unless an edit passes label-free evidence gates. The framework operates in three stages (Fig. 2), each serving this single commitment problem. The first stage aggregates multi-view vision-language evidence into dense voxel scores. The second stage groups voxels into compact query-addressable atoms that keep per-view provenance, so class scores are read out against the original multi-view support rather than an averaged posterior. The third stage treats CAQD as a diagnostic proposal model and commits updates only when sparsity, confidence-gain, class-balance, and base-uncertainty gates pass. Section 4.5 stress-tests the commitment rule under leave-one-scene-out calibration, the held-out protocol we advocate before deployment.

Our contributions follow one spine: preserve query-addressable evidence, then commit only when an update is safe.

- We identify ***query-time semantic commitment*** as an unmodeled decision in streaming open-vocabulary 3D maps.
- We introduce query-addressable atoms and provenance-weighted readout, a compact, fixed representation that improves the dense evidence backbone without labels or per-scene tuning.
- We separate proposal quality from commitment safety: CAQD is used as an in-scene diagnostic of the representation ceiling, whereas the four-gate committer is evaluated as a label-free abstention rule.
- We establish LOSO calibration for semantic committers and show that same-set threshold selection can create apparent zero-regression behaviour; the accuracy gain comes from atom readout, while the committer exposes tail risk.

We will release our code upon acceptance.

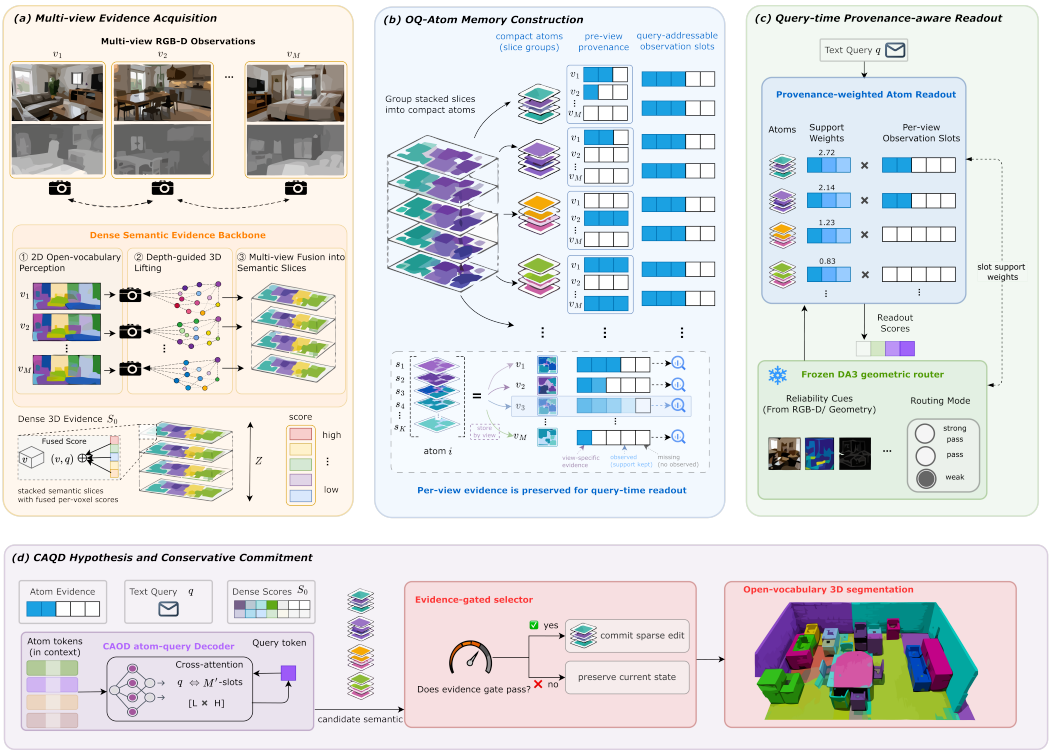

Figure 2: **Overview of OQ-ATOMS.** (1) Multi-view vision-language evidence is aggregated into dense voxel scores (Sect. 3.1). (2) Voxels are grouped into compact query-addressable atoms whose observation slots preserve per-view provenance (Sect. 3.2). (3) At query time, atom memory exposes provenance and view support to the provenance-weighted readout. (4) CAQD proposes atom-query hypotheses, and the selector commits them only when four label-free evidence gates pass; otherwise the backbone prediction is preserved (Sects. 3.3 and 3.4).

2 RELATED WORK

Open-vocabulary semantic maps lift evidence from vision-language encoders, language-driven segmenters, or promptable masks into queryable 3D memories Radford et al. (2021); Li et al. (2022); Ghiasi et al. (2022); Kirillov et al. (2023); Huang et al. (2023); Shafiullah et al. (2023). Their representations include feature-lifting maps and neural fields Peng et al. (2023); Jatavallabhula et al. (2023); Engelmann et al. (2024), Gaussian fields and splatting Kerr et al. (2023); Qin et al. (2024); Lu et al. (2025), and object-centric graphs or masks Gu et al. (2024); Werby et al. (2024); Takmaz et al. (2023); Nguyen et al. (2024). Unlike voxel grouping and superpoints Landrieu & Simonovsky (2018), which typically commit labels during storage, OQ-ATOMS retains per-view provenance for query-time readout. Our committer is related to selective prediction and uncertainty estimation El-Yaniv & Wiener (2010); Geifman & El-Yaniv (2017; 2019); Gal & Ghahramani (2016), but evaluates an *edit* against the existing map, with no-update as the default.

3 METHOD

3.1 PROBLEM SETUP

Let a reconstructed 3D scene be represented by voxels $v \in \mathcal{V}$ and open-vocabulary queries $q \in \mathcal{Q}$. The first stage of our framework is a dense evidence backbone: each RGB-D view yields a per-view open-vocabulary score, and these are fused into a dense per-voxel score $S_0(v, q)$ by confidence-weighted multi-view aggregation, kept fully dense rather than reduced to object proposals. S_0 is the evidence state that all subsequent stages address, and we measure our contribution as the gain over S_0 rather than its absolute score.

A standard refiner learns $f : S_0 \mapsto S$, coupling hypothesis generation with the decision to overwrite the map. OQ-ATOMS decouples them through its remaining stages: query-addressable atoms that preserve multi-view provenance, provenance-weighted readout for the submitted prediction, a diagnostic atom-query probe, and a selector that commits only hypotheses supported by label-free evidence gates.

3.2 QUERY-ADDRESSABLE OQ-ATOMS

Each voxel v is assigned up to K *observation slots*, where each slot records the response of one past RGB-D observation to the voxel: the backbone-score vector under that view, its confidence, and a view-pose index. Observation slots are the "per-view provenance" that the representation preserves; they are stored per scene and do not change across query-time readouts.

OQ-ATOMS groups voxels into compact atoms $a \in \mathcal{A}$ via boundary-aware region growing. A voxel is merged into the nearest atom when its distance is below radius r, its backbone label agrees with the atom plurality, and the atom remains below size $N_{\max}$. We use r=0.05 m and $N_{\max}$=64, producing atoms at roughly half the voxel count. Each voxel has assignment $A(v)$, and each atom aggregates member slots into an evidence state

$$h_a = \mathrm{Agg}\left(\{\, \psi(v, \mathcal{O}, S_0) \mid A(v) = a \,\}\right), \tag{1}$$

where $\psi(v, \mathcal{O}, S_0)$ concatenates the backbone score with the mean and variance of the K=8 observation slots assigned to voxel v. Agg is a slot-weighted mean, so geometrically supported voxels contribute more than sparsely observed background voxels. Atom maps and provenance slots are stored once per scene and reused for all queries with fixed construction parameters.

3.3 DIAGNOSTIC ATOM-QUERY DECODING

Given atom evidence h_a, query embedding e_q, and the backbone score $S_0(a, q)$ pooled to the atom, CAQD predicts an atom-query semantic hypothesis for diagnostic analysis:

$$\Delta(a, q),\ \sigma(a, q) = \mathrm{CAQD}(h_a, e_q, S_0(a, q)). \tag{2}$$

The residual hypothesis $\Delta(a, q)$ describes how the atom's semantic belief would change for query q, while $\sigma(a, q)$ is an internal support signal. We use this module to probe how much supervised signal the atom state exposes; the fixed atom-readout path supplies the result in Table 1. A direct residual implementation would form

$$\widetilde{S}(a, q) = S_0(a, q) + \alpha\, \sigma(a, q) \cdot \Delta(a, q), \quad \alpha = 0.2, \tag{3}$$

with residual scale α. The Replica diagnostic shows that this direct form is insufficient: a gate trained as occupancy or objectness can become large on nearly all atoms and overwrite reliable semantic evidence. We therefore treat $\widetilde{S}$ as a hypothesis rather than the final score.

3.4 EVIDENCE-GATED COMMITMENT

The final module evaluates each candidate hypothesis as a counterfactual intervention. The selector does not simply choose the highest-scoring edit; it requires the edit to clear evidence gates before overwriting the current state.

Action space. Alongside the default no-update action, the selector considers a sparse top-k decoder edit and a residual class correction in uncertain regions. The sparse decoder edit gated by the four conditions below is the principled core, calibrated out-of-sample in Section 4.5; composite edits over regions where the decoder and residual hypotheses disagree are used only for diagnostic analysis.

Evidence gates. Each candidate intervention y is evaluated through four unsupervised gates, computed entirely from score-distribution statistics of the backbone and the hypotheses—no ground-truth labels:

1. **Minimum intervention.** The edit must change a small fraction of voxels: a counterfactual edit that rewrites most of the map is not a targeted repair but a new prediction. The sparsity gate requires $\mathrm{frac}(y) \le \tau_{\text{sparse}} = 0.04$.

2. **Positive confidence shift.** Edited voxels must show a net increase in prediction confidence and margin, so the gate requires both $\Delta\text{conf}(y) > 0$ and $\Delta\text{margin}(y) > 0$ to hold strictly.

3. **Class-balanced impact.** No single class may dominate the edit, so the gate requires $\max_c \text{share}_c(y) \leq \tau_{\text{conc}} = 0.6$.

4. **Targeted uncertainty.** The edited region must be genuinely uncertain under the evidence backbone. The gate requires $u_{\text{base}}(y) \geq \tau_{\text{uncert}} = 0.3$, which bounds backbone confidence in the edited region by 0.7.

Selection rule. Let $\mathcal{Y}$ be the set of candidate actions and $\phi(y)$ the score-distribution proxy vector for action y, distinct from the per-voxel feature ψ of Eq. (1). Each evidence gate $g_k : \phi(y) \mapsto \{0, 1\}$ implements one of the four conditions above—sparsity, confidence gain, class balance, and base uncertainty—applied to the statistics in $\phi(y)$. The selector applies a cascade of feasibility filters:

$$\mathcal{Y}_{\text{feasible}} = \{\, y \in \mathcal{Y} \mid g_k(\phi(y)) = 1 \;\; \forall k \in \{1, 2, 3, 4\} \,\}. \tag{4}$$

If $\mathcal{Y}_{\text{feasible}}$ is non-empty, the final action is $y^* = \arg\max_{y \in \mathcal{Y}_{\text{feasible}}} \text{score}(\phi(y))$, where $\text{score}(\phi(y)) = \Delta\text{conf}(y) + \Delta\text{margin}(y) - \lambda_{\text{conc}} \cdot \text{conc}(y)$ with $\lambda_{\text{conc}} = 0.5$; otherwise y^* is the no-update action. The thresholds are fixed across both Replica and ScanNet, so a single selector operates on both datasets without dataset-specific tuning. The principled four-gate core is calibrated out-of-sample in Section 4.5; we distinguish this protocol from same-set threshold selection, which can yield optimistic diagnostics.

Conservative by design. The cascade structure encodes a deliberate asymmetry: it is harder to commit an edit than to preserve the current state. This reflects the conservative-commitment principle that an edit requires positive evidence, while no-update remains the default null hypothesis. Voxel scores are obtained by applying the selected action and transporting atom scores back through the atom assignment.

3.5 Implementation Details

CAQD is implemented as a two-layer MLP: the first layer projects the concatenation of h_a, e_q, and $S_0(a, q)$ to 256 dimensions with a ReLU activation; the second produces $\Delta(a, q) \in \mathbb{R}^C$ and a sigmoidal $\sigma(a, q) \in [0, 1]$. The model is trained per scene for 3 epochs with a cross-entropy loss on atom-level observation targets, using batches of 512 atom-query pairs. CAQD serves as an in-scene diagnostic model, trained and evaluated on the same scene; the trainable-refinement rows in Table 2 measure the representation ceiling within this setting.

Scope of training. The submitted path uses the dense backbone, fixed atoms, and provenance-weighted readout. The trainable CAQD rows are reported separately as in-scene diagnostics.

Atom construction uses a fixed random seed, and over 3 random seeds the Replica 8-scene mIoU varies by at most 0.001, confirming stable grouping boundaries. The resulting atom maps are compact and inexpensive to reuse: in representative ScanNet and Replica diagnostics, atom construction takes about 0.02 s, two orders of magnitude below the several-second semantic scoring stage, so the representation adds essentially no overhead to scene-level inference.

4 Experiments

4.1 Experimental Setup

We evaluate the streaming indoor mapping scenario on five ScanNet Dai et al. (2017) scenes and eight Replica Straub et al. (2019) scenes, the standard testbeds for indoor embodied semantic mapping, under a fixed dense-voxel evaluation protocol. All rows in Table 1, baselines and our method alike, use the same voxelization and unified 3D-projection evaluation. The dense evidence backbone (atoms and provenance disabled; first row of each block in Table 2) reaches 39.55 mIoU on Replica and 39.64 on ScanNet; adding query-addressable atoms and provenance-weighted readout contributes +2.08 mIoU on Replica and +0.49 on ScanNet, the isolated representation gain of our method.

Table 1: **Open-vocabulary 3D semantic segmentation on Replica Straub et al. (2019) and ScanNet Dai et al. (2017).** We report mIoU, frequency-weighted mIoU (fMIoU), and pixel accuracy (pAcc) under one fixed dense-voxel full-scene protocol; mAcc is omitted because it is unavailable for most baselines. Baseline numbers are obtained under our existing protocol with a unified 3D projection evaluation; see Sect. 4.1. The best results are in **bold**.

	Replica			ScanNet		
Method	**mIoU↑**	**fMIoU↑**	**pAcc↑**	**mIoU↑**	**fMIoU↑**	**pAcc↑**
ConceptFusion Jatavallabhula et al. (2023)	0.2038	0.3575	0.4158	0.1857	0.2306	0.2877
ConceptGraphs Gu et al. (2024)	0.1172	0.2135	0.2828	0.2083	0.2361	0.3580
HOV-SG Werby et al. (2024)	0.1929	0.3064	0.3517	0.2348	0.2892	0.3852
NACLIP-3D Hajimiri et al. (2025)	0.1530	0.1698	0.2623	0.2232	0.2432	0.3346
Trident-3D Shi et al. (2025)	0.2063	0.3853	0.5031	0.2480	0.2777	0.3843
SC-CLIP Bai et al. (2025)	0.2198	0.4431	0.5569	0.2859	0.4089	0.5114
Talk2DINO Barsellotti et al. (2025)	0.2394	0.3686	0.4847	0.2913	0.3447	0.4665
TextRegion Xiao et al. (2025)	0.2835	0.5090	0.6004	0.3583	**0.4609**	0.5508
RayFronts Alama et al. (2025)	0.3128	0.5328	0.6302	0.3105	0.4230	0.5169
Ours	**0.4163**	**0.6514**	**0.7079**	**0.4013**	0.4577	**0.5579**

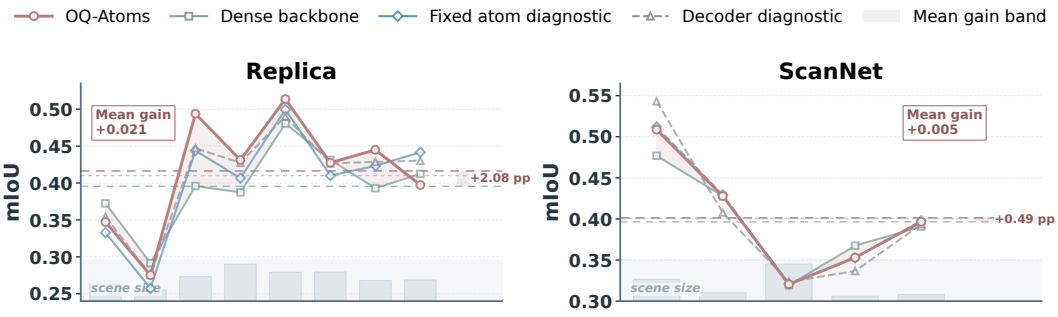

Figure 3: **Readout profiling with a matched dense backbone.** The upper traces compare dense backbone scores, atom-readout variants, and decoder diagnostics for each evaluated scene; dashed reference lines and the shaded band mark the mean readout gain. The pale bars embedded at the bottom of each panel encode evaluated scene size, so paired gains can be read together with scene scale without introducing a separate metric axis.

4.2 Main Results

Table 1 reports the submitted atom-readout path. OQ-Atoms obtains the best mIoU on both datasets: 41.63 on Replica, 10.35 points above RayFronts, and 40.13 on ScanNet, 4.30 points above TextRegion; it ranks best on every Replica metric and on ScanNet mIoU and pixel accuracy. The margin comes from preserved provenance, not from a trained refiner, per-scene threshold selection, or extra inference cost. The decoder and committer are evaluated separately through the held-out calibration in Section 4.5 as commitment-safety diagnostics.

Fig. 3 explains why OQ-Atoms is a representation and readout contribution rather than a refiner: paired scene traces show that Replica gains are larger, while ScanNet gains are smaller but consistent.

4.3 Representation Ablation

Table 2 isolates the representation path. The first row of each block is our dense evidence backbone without atoms or provenance readout. Fixed atoms and provenance-weighted readout explain the dependable gains. The CAQD rows estimate the supervised proposal ceiling inside each scene, whereas Table 1 uses the fixed readout path. We omit the ScanNet fixed-atom row because no matched configuration exists under the same protocol.

Table 2: **Representation ablation on Replica and ScanNet.** Framework stages are enabled incrementally over our dense evidence stage; CAQD is reported as an in-scene diagnostic rather than the main held-out result. The best and second best results are highlighted.

Dataset	Atoms	Prov.	CAQD	mIoU	fMIoU	mAcc	pAcc
Replica	×	×	×	0.3955	0.6138	0.5002	0.6799
Replica	✓	×	×	0.4019	0.6406	0.5265	0.7041
Replica	✓	✓	×	**0.4163**	**0.6514**	0.5174	**0.7079**
Replica	✓	✓	✓	0.4100	0.6483	**0.5346**	0.7027
ScanNet	×	×	×	0.3964	0.4505	**0.5338**	0.5497
ScanNet	✓	✓	×	**0.4013**	0.4577	0.5266	0.5579
ScanNet	✓	✓	✓	0.4009	**0.4591**	0.5271	**0.5595**

Prov. = provenance-weighted readout; CAQD is an in-scene diagnostic.

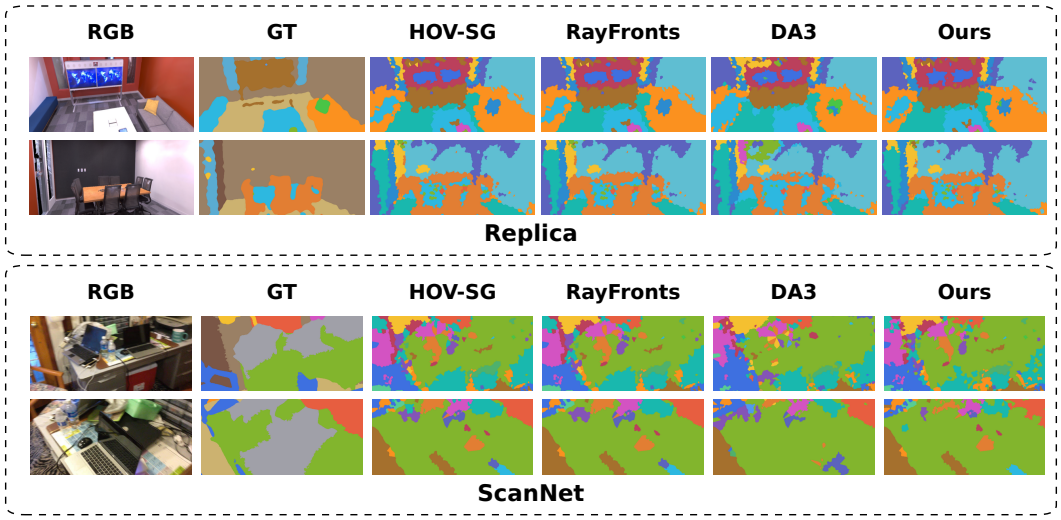

Figure 4: **Qualitative comparison on Replica and ScanNet.** Each row projects 3D predictions of one camera frame back to image pixels; columns compare RGB, ground truth, our dense evidence stage, the RayFronts baseline, DA3-conditioned residual routing, and the selected OQ-ATOMS output on the same view. OQ-ATOMS preserves reliable regions of the dense evidence stage and commits only targeted corrections in uncertain regions. Best viewed zoomed in.

4.4 QUALITATIVE EVIDENCE

Fig. 4 shows the same pattern in real views: reliable dense-evidence regions are preserved, while uncertain regions receive targeted corrections.

4.5 HELD-OUT CALIBRATION

We go beyond same-set evaluation with leave-one-scene-out (LOSO) calibration. We calibrate the *principled* core rule—commit a sparse decoder top-k edit only when four label-free gates pass, otherwise no-update—by sweeping gate thresholds on all *other* scenes, then freezing and evaluating on the held-out scene. This held-out protocol tests whether a label-free semantic committer transfers beyond its calibration scenes.

Fig. 5 delivers the methodological payoff. Same-set threshold selection can create apparent non-regression, yet LOSO shows that this does not follow from the gates alone. Under LOSO the gates mostly abstain, and mixed-scene transfer yields 2 of 5 strict improvements with a mean of $+0.0056$ mIoU. The result separates two roles: atom readout supplies the dependable accuracy gain, while LOSO measures how the committer transfers as a safety wrapper.

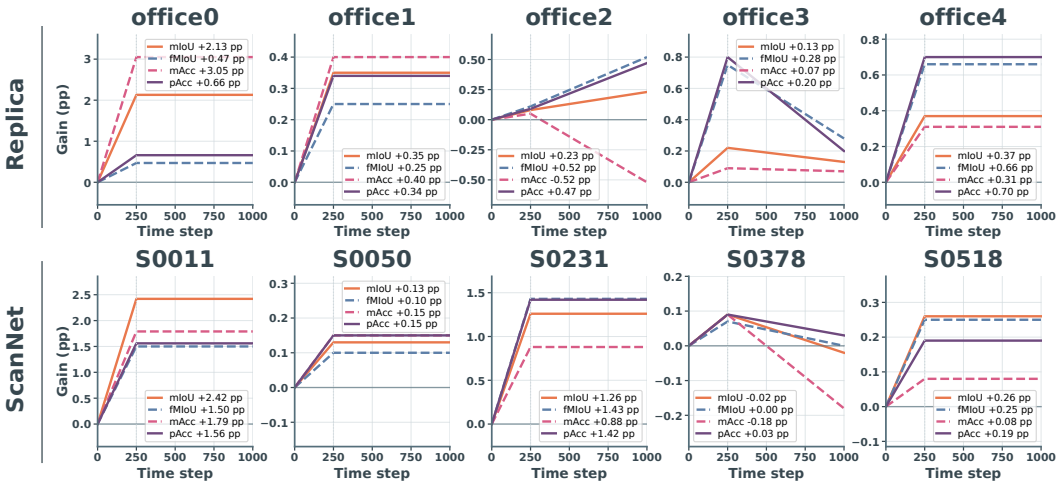

Figure 5: **Held-out calibration exposes same-set threshold artifacts.** Each panel shows four metric-gain curves for one scene. Gains are reported in percentage points, and each small panel uses a local y-scale so that small held-out changes remain visible. The dotted vertical marker denotes the same-set calibration point; the final value reports the LOSO-calibrated held-out gain.

Table 3: **Same-set selector gate ablation.** One label-free gate is removed per row; safe scenes have non-negative changes on all reported metrics.

Variant	Replica ΔmIoU	ScanNet ΔmIoU	Overall	Safe scenes
Full	+0.0005	+0.0070	+0.0030	8/13
w/o sparsity	-0.0008	+0.0104	+0.0035	4/13
w/o confidence	+0.0005	+0.0066	+0.0028	5/13
w/o balance	+0.0005	+0.0066	+0.0028	7/13
w/o uncertainty	+0.0005	+0.0068	+0.0029	8/13

4.6 Gate Robustness and Same-Set Diagnostics

We additionally isolate the role of each evidence gate using the fixed decoder top-k action pool. This is a same-set diagnostic rather than a held-out performance claim: thresholds are fixed on the evaluated scenes, whereas the transfer result is reported by LOSO above. Removing sparsity reduces the number of scenes with non-negative changes on all four metrics from 8/13 to 4/13 and admits a worst-case -0.0081 mIoU edit. Removing confidence gain yields 5/13 such scenes. Class balance and base uncertainty provide smaller safeguards. Thus sparsity and confidence gain are the load-bearing checks: they constrain tail risk even when removing a gate can slightly improve a mean same-set score. This diagnostic explains the gate structure; it does not replace held-out calibration as evidence of deployment safety.

Table 3 makes the trade-off explicit. The full selector is not the largest mean-gain variant, but it gives the strongest same-set safety rate. In particular, removing sparsity raises the ScanNet average while halving the safe-scene count. This is why we treat the gates as safety constraints rather than score maximizers, and reserve LOSO for judging whether that conservative behavior transfers beyond calibration scenes.

4.7 Training-Free Geometric Conditioning

We also test whether geometric reliability can condition the observation slots without rewriting semantic targets. Frozen Depth Anything 3 Lin et al. (2026) supplies only a per-view reliability summary; the reliable-view fraction selects one of three fixed slot-support strengths. Across paired scenes, this provenance-level conditioning improves all four metrics, with a mean gain of 1.52 mIoU points. The result is consistent with the central design principle: complementary signals should

determine which evidence supports readout, while semantic edits remain subject to conservative commitment.

4.8 Over-Commitment Diagnostic

CAQD can produce a strong atom-query hypothesis on supportive local evidence, yet the same gated decoder harms one Replica office scene: with a mean gate of 0.94 it commits almost everywhere and loses 0.36 mIoU and 2.37 fMIoU points. This failure motivates making no-update the default; the headline gain still comes from the representation.

Runtime. Atom construction is lightweight once the atom map is available: on the representative ScanNet scene, construction maps 80,212 voxels to 40,106 atoms in 0.019 s against 4.19 s of semantic scoring—a two-order-of-magnitude gap, so the representation gain does not come from extra inference cost.

5 Discussion

The experiments separate *where* semantic evidence is stored from *when* it is allowed to change the map. OQ-Atoms keeps observation slots attached to each atom and makes readout query-addressable, so no-update remains a defined fallback whenever evidence for an edit is weak.

For embodied mapping, an update is an intervention on persistent memory, not an isolated prediction. The limit is upstream coverage—OQ-Atoms can reject risky edits, but not recover classes absent from all views—so stronger evidence sources remain complementary to the atom state and fallback.

This boundary also clarifies how the representation should be used. Atoms complement stronger 2D evidence, better geometry, and scene-specific reasoning by keeping these signals separable until a query and a risk tolerance are known. This motivates evaluating the selector as a commit rule rather than as another mean-score optimizer. Reporting only same-scene gains can hide policies that improve the average by spending tail risk, whereas LOSO calibration asks whether a label-free rule still abstains when its thresholds are selected outside the test scene. For open-vocabulary maps that persist across robot actions, this distinction is operational: a reversible readout error is less costly than an irreversible map edit that every later query inherits.

6 Conclusion

This paper establishes query-time semantic commitment as a design axis for open-vocabulary 3D mapping. By keeping per-view provenance available until query time, the submitted atom-readout path improves the dense evidence stage by 2.08 mIoU on Replica, yields consistent gains on ScanNet, and reaches 41.63 and 40.13 mIoU with fixed construction parameters and no scene-specific training. The fixed-atom diagnostic attributes the gain to provenance-weighted readout rather than boundary alignment alone.

The broader lesson is that persistent semantic maps should preserve evidence until the query and update risk are both known. LOSO calibration exposes same-set threshold artifacts. Query-addressable atoms provide the storage substrate, readout turns provenance into class evidence, and the committer makes abstention an explicit alternative to irreversible semantic collapse.

Reproducibility Statement

All experiments use the fixed dense-voxel protocol, scenes, and metrics specified in Section 4.1; the fixed atom-construction parameters and decoder configuration are given in Sections 3.2–3.4. Section 4.5 documents the held-out calibration protocol. We will provide anonymous supplementary materials containing implementation, evaluation scripts, scene splits, fixed seeds, and data-processing details to reproduce the reported results.

AI Use Statement

Generative AI tools were used to assist with restructuring and editing the manuscript for clarity and formatting. They were not used to generate experimental data, develop the method or theoretical claims, design or interpret experiments, implement the method, or analyze results. The authors reviewed and verified all AI-assisted text and take responsibility for the final content of this work.

References

Omar Alama, Avigyan Bhattacharya, Haoyang He, et al. RayFronts: Open-set semantic ray frontiers for online scene understanding and exploration. In *IROS*, 2025.

Sule Bai, Yong Liu, Yifei Han, Haoji Zhang, Yansong Tang, Jie Zhou, and Jiwen Lu. Self-calibrated CLIP for training-free open-vocabulary segmentation. *IEEE Transactions on Image Processing*, 34:8271–8284, 2025.

Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, et al. Talking to DINO: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *ICCV*, 2025.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theory*, 16:41–46, 1970.

Angela Dai, Angel X. Chang, Manolis Savva, et al. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *J. Mach. Learn. Res.*, 11:1605–1641, 2010.

Francis Engelmann, Fabian Manhardt, Michael Niemeyer, et al. OpenNeRF: Open set 3D neural scene segmentation with pixel-wise features and rendered novel views. In *ICLR*, 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *NeurIPS*, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *ICML*, 2019.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, et al. ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning. In *ICRA*, 2024.

Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *WACV*, 2025.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

Chen Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *ICRA*, 2023.

Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, et al. ConceptFusion: Open-set multimodal 3D mapping. In *RSS*, 2023.

Li Jiang, Shaoshuai Shi, and Bernt Schiele. Open-vocabulary 3D semantic segmentation with foundation models. In *CVPR*, 2024.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language embedded radiance fields. In *ICCV*, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. Segment Anything. In *ICCV*, 2023.

Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *CVPR*, 2018.

Junha Lee, Chunghyun Park, Jaesung Choe, et al. Mosaic3D: Foundation dataset and model for open-vocabulary 3D segmentation. In *CVPR*, 2025.

Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022.

Haotong Lin, Sili Chen, Jun Hao Liew, et al. Depth Anything 3: Recovering the visual space from any views. In *ICLR*, 2026.

Yiren Lu, Yunlai Zhou, Yiran Qiao, et al. Segment then Splat: Unified 3D open-vocabulary segmentation via Gaussian Splatting. In *NeurIPS*, 2025.

John McCormac, Ankur Handa, Andrew J. Davison, et al. SemanticFusion: Dense 3D semantic mapping with convolutional neural networks. In *ICRA*, 2017.

Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, et al. Open3DIS: Open-vocabulary 3D instance segmentation with 2D mask guidance. In *CVPR*, 2024.

Michael Oberst and David A. Sontag. Counterfactual off-policy evaluation with Gumbel-Max structural causal models. In *ICML*, 2019.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2 edition, 2009.

Songyou Peng, Kyle Genova, Chiyu Max Jiang, et al. OpenScene: 3D scene understanding with open vocabularies. In *CVPR*, 2023.

Jens Piekenbrinck, Christian Schmidt, Alexander Hermans, et al. OpenSplat3D: Open-vocabulary 3D instance segmentation using Gaussian Splatting. In *CVPR Workshops (OpenSUN3D)*, 2025.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3D language Gaussian Splatting. In *CVPR*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: An open-source library for real-time metric-semantic localization and mapping. In *ICRA*, 2020.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, et al. Habitat: A platform for embodied AI research. In *ICCV*, 2019.

Nur Muhammad (Mahi) Shafiullah, Chris Paxton, Lerrel Pinto, et al. CLIP-Fields: Weakly supervised semantic fields for robotic memory. In *RSS*, 2023.

Yuheng Shi, Minjing Dong, and Chang Xu. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. In *ICCV*, 2025.

Julian Straub, Thomas Whelan, Lingni Ma, et al. The Replica dataset: A digital replica of indoor spaces, 2019.

Ayca Takmaz, Elisabetta Fedele, Robert W. Sumner, et al. OpenMask3D: Open-vocabulary 3D instance segmentation. In *NeurIPS*, 2023.

Ayca Takmaz, Alexandros Delitzas, Robert W. Sumner, et al. Search3D: Hierarchical open-vocabulary 3D segmentation. *IEEE RA-L*, 10(3):2558–2565, 2025.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer US, 2005.

Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3D scene graphs for language-grounded robot navigation. In *RSS*, 2024.

Yao Xiao, Qiqian Fu, Heyi Tao, Yuqun Wu, Zhen Zhu, and Derek Hoiem. TextRegion: Text-aligned region tokens from frozen image-text models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Runsong Zhu, Ka-Hei Hui, Zhengzhe Liu, et al. COS3D: Collaborative open-vocabulary 3D segmentation. In *NeurIPS*, 2025.

Xiaoyu Zhu, Hao Zhou, Pengfei Xing, et al. Open-vocabulary 3D semantic segmentation with text-to-image diffusion models. In *ECCV*, 2024.