

OpenSplatGraph: From Dense Semantic Maps to Structured Scene Graphs for Open-Vocabulary Robot Perception

Binh Long Nguyen^{1,2} , Kien Nguyen¹ ,
Clinton Fookes¹ , and Peyman Moghadam^{1,2} 

¹ School of Electrical Engineering and Robotics, Queensland University of
Technology (QUT), Brisbane, QLD 4000, Australia {binhlong.nguyen,
k.nguyenthanh, c.fookes, peyman.moghadam}@qut.edu.au

² CSIRO Robotics, CSIRO, Brisbane, QLD 4069, Australia
{binhlong.nguyen, peyman.moghadam}@csiro.au

Abstract. Dense 3D mapping with semantic understanding is essential for robotic perception in complex environments. Recent 3D Gaussian Splatting-based mapping approaches enable high-fidelity geometry and efficient open-vocabulary perception, but typically represent semantics as unstructured feature fields that limit object-centric reasoning. In contrast, 3D scene graphs explicitly model objects and their relationships for structured reasoning, but are commonly constructed from sparse geometric representations that do not fully exploit dense semantic maps. In this work, we present OpenSplatGraph, a unified framework that constructs persistent 3D scene graphs directly from an online Gaussian-based open-vocabulary semantic map. The proposed framework augments the dense semantic map with a reliability-aware semantic field that maintains lightweight observation statistics for confidence-aware, query-conditioned object extraction. Extracted object instances are associated with persistent graph nodes, allowing object attributes and relationships to be incrementally updated across observations and queries. By tightly coupling dense semantic mapping with persistent object-centric representations, our framework supports both language-guided object grounding and structured relational reasoning while preserving the geometric fidelity of Gaussian-based mapping. Comprehensive evaluations on standard 3D scene understanding benchmarks and real-world robotic experiments demonstrate that OpenSplatGraph achieves competitive performance for online open-vocabulary perception and downstream robotic tasks. Project page: <https://csiro-robotics.github.io/OpenSplatGraph>.

Keywords: 3D Gaussian Splatting · 3D Scene Graph · Open-Vocabulary Scene Understanding · Robot Perception

1 Introduction

Dense 3D mapping with semantic understanding is fundamental to robotic perception, supporting navigation, manipulation, and interaction in complex environments [9, 11, 38]. Recent advances in vision-language foundation models [18,

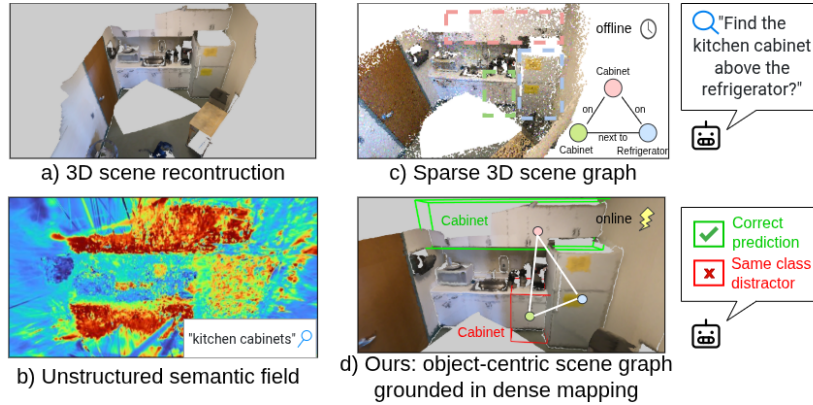


Fig. 1: Comparison of complementary representations for open-vocabulary 3D scene understanding. Dense semantic maps provide high-fidelity geometry and language-aligned features but represent semantics as unstructured feature fields. Scene graphs explicitly model objects and their relationships for structured reasoning, but typically rely on sparse geometric representations. Our framework bridges these paradigms by constructing a persistent 3D scene graph directly from an online dense semantic map.

24] have further enabled open-vocabulary 3D scene understanding that responds to natural language queries and generalizes to previously unseen objects.

Meanwhile, 3D Gaussian Splatting (3D-GS) [13] has established an efficient representation for high-fidelity 3D reconstruction with real-time rendering. On-line GS-based mapping systems [8, 12, 21] integrate Gaussian primitives with Simultaneous Localization and Mapping (SLAM) for geometrically consistent mapping. Building on this representation, open-vocabulary approaches [22, 23, 33, 34] incorporate vision-language features into dense Gaussian maps, enabling language-guided retrieval of scene regions. However, semantics are often represented as unstructured per-voxel or per-Gaussian feature fields rather than explicit object-level entities. Consequently, they provide limited support for object-centric reasoning and modeling inter-object relationships.

In parallel, 3D Scene Graphs (3DSGs) address this limitation by representing environments as graphs whose nodes correspond to objects and whose edges encode semantic or spatial relationships [1]. Recent advances have extended scene graphs to online incremental systems [10, 32], open-vocabulary reasoning [7, 16, 20], and large-scale robotic applications [5, 35]. Most open-vocabulary scene graph methods construct graphs from RGB-D observations by projecting 2D detections into point clouds or meshes. While these approaches provide compact and interpretable scene abstractions for reasoning, they typically rely on sparse geometric representations that cannot fully exploit the dense geometric and semantic information available in Gaussian-based maps.

As illustrated in Fig. 1, robotic systems benefit from these two complementary scene representations. Dense geometric-semantic maps provide high-fidelity 3D reconstruction for language-grounded perception and localization, whereas object-centric scene graphs explicitly model objects and their relationships for

structured reasoning and task planning. However, existing approaches typically emphasize one representation over the other, making it difficult to simultaneously achieve both high-fidelity mapping and structured scene understanding.

In this paper, we propose OpenSplatGraph, a framework that bridges these two paradigms by grounding 3D scene graphs directly in an online open-vocabulary 3D-GS semantic map. Rather than treating the dense semantic field and the scene graph as alternative representations, we maintain both throughout operation, allowing each to serve a complementary role. The dense semantic map preserves fine-grained geometric and semantic information for open-vocabulary perception, while the scene graph incrementally accumulates persistent object-centric knowledge for structured reasoning. To support object construction, we introduce a reliability-aware semantic field that augments the dense semantic map with lightweight observation statistics, enabling confidence-aware ranking of query-conditioned object instances. Extracted object instances are assigned persistent identities and incrementally incorporated into a 3D scene graph, allowing object attributes and relationships to be consistently accumulated across multiple queries. By jointly maintaining these two representations, OpenSplatGraph preserves the geometric fidelity of dense mapping while progressively constructing an interpretable object-centric representation for language grounding, relational reasoning, and downstream robotic tasks.

Our contributions are summarized as follows:

- We introduce OpenSplatGraph, a unified framework that constructs persistent 3D scene graphs directly from online open-vocabulary Gaussian semantic maps, enabling high-fidelity dense mapping and structured scene understanding within a single system.
- We propose a reliability-aware semantic field that leverages lightweight per-voxel observation statistics for robust query-conditioned object extraction under partial observations and semantic ambiguity.
- We design a scene graph construction strategy that associates extracted object instances with persistent identities and incrementally accumulates object-centric knowledge across queries for structured relational reasoning.
- We validate the framework on standard 3D scene understanding benchmarks and real-world robotic experiments, demonstrating competitive performance for online open-vocabulary perception and downstream robotic tasks.

2 Related Work

2.1 Open-Vocabulary 3D Scene Understanding

Vision-language foundation models [18, 24] have enabled open-vocabulary 3D scene understanding by aligning visual representations with natural language. Early methods, such as PointCLIP [37] and ConceptFusion [11], leverage language features for point cloud understanding, while LERF [14] extends language-guided retrieval to Neural Radiance Field (NeRF) representations. More recently, 3D-GS [13] has emerged as an efficient representation for open-vocabulary 3D

perception and robotic applications [22, 40]. Representative methods, including LangSplat [23], LEGaussian [26], OpenGaussian [33], and InstanceGaussian [17], associate language features with Gaussian primitives for open-vocabulary querying. However, these methods primarily represent semantics as unstructured feature fields, limiting their support for object-centric and relational reasoning.

2.2 3D Gaussian-Based Scene Mapping

3D-GS has been integrated into SLAM systems for efficient online dense mapping and high-fidelity rendering. Methods such as SplatTAM [12], MonoGS [21], and GS-ICP SLAM [8] jointly optimize Gaussian representations and camera poses for real-time reconstruction. Subsequent works incorporate semantic information through supervised labels [19, 39] or open-vocabulary features, as demonstrated by OpenGS-Fusion [34]. These approaches provide geometrically accurate and semantically rich maps for robotic perception. However, they primarily focus on online reconstruction and semantic fusion, without explicitly maintaining persistent object-centric representations for structured reasoning and task planning.

2.3 3D Scene Graphs

3DSGs [1] represent objects as nodes and their relationships as edges, providing interpretable scene abstractions for planning and human-robot interaction [25, 31]. Recent advances [3] highlight rapid progress from online incremental graph construction [10, 32] to open-vocabulary scene understanding through ConceptGraphs [7], Open3DSG [16], and Clio [20], and large-scale robotic applications [5, 35]. These methods typically construct graphs from point clouds or meshes, providing compact object-centric representations for downstream reasoning. Despite these advances, many open-vocabulary approaches infer object relationships from 2D observations using Vision-Language Models (VLM), resulting in graph representations that are weakly grounded in dense 3D geometry.

Recent studies [6, 30] have explored grounding scene graphs directly in Gaussian representations. Our work follows this emerging direction but adopts a different architectural perspective. Rather than treating the dense maps and the scene graph as alternative world models, we explicitly maintain both throughout operation. The dense semantic field serves as the geometric and semantic foundation for reliable object extraction, while the scene graph incrementally accumulates object-centric knowledge for structured and relational reasoning.

3 Methodology

We propose OpenSplatGraph, a unified framework that jointly maintains a dense open-vocabulary semantic field and a persistent 3D scene graph throughout online operation. The dense semantic field serves as a geometrically consistent foundation for query-conditioned object extraction, while the scene graph incrementally accumulates object-centric knowledge to support structured reasoning across multiple queries. Rather than maintaining a fixed object inventory,

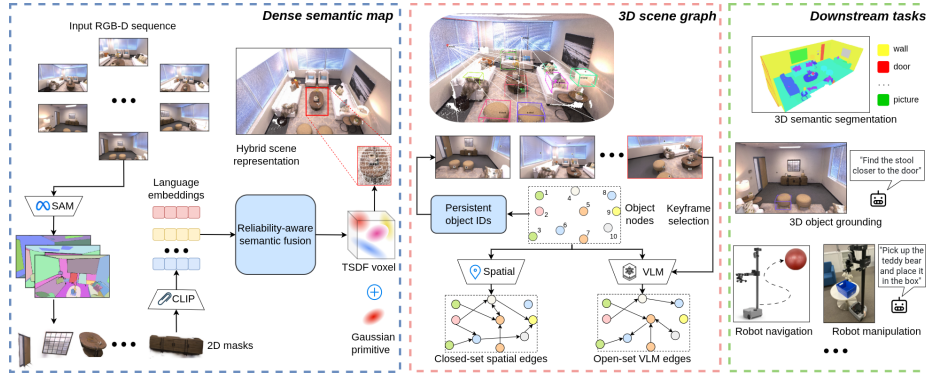


Fig. 2: Overview of OpenSplatGraph. Starting from a multi-view RGB-D sequence, the framework incrementally constructs an online open-vocabulary dense semantic map together with a reliability-aware semantic field for robust query-conditioned object extraction. Extracted object instances are then associated across queries and incrementally accumulated into a persistent 3D scene graph. By jointly maintaining the dense semantic field and the scene graph, OpenSplatGraph enables structured relational reasoning while preserving high-fidelity geometric and semantic information for downstream robotic tasks.

OpenSplatGraph extracts objects on demand from language queries by grouping spatially and semantically consistent regions in the semantic field. Extracted instances are associated with persistent graph nodes and enriched with semantic and relational information as the map evolves. In this way, OpenSplatGraph preserves the fidelity of dense semantic mapping while progressively constructing an interpretable scene representation for downstream robotic tasks. An overview of the framework is shown in Fig. 2.

3.1 Reliability-Aware Semantic Field

The dense semantic field forms the perception layer of OpenSplatGraph, providing a geometrically consistent representation for query-conditioned object extraction. Built upon an online open-vocabulary Gaussian semantic map, the field maintains fused semantic features together with lightweight reliability statistics that characterize the quality of semantic observations. These statistics regulate semantic fusion and assess the reliability of extracted object candidates, improving robustness under partial visibility, occlusions, and semantic ambiguity.

Semantic Representation. OpenSplatGraph adopts a hybrid semantic mapping framework [34], combining voxel-based Truncated Signed Distance Function (TSDF) geometry with 3D Gaussian primitives. The TSDF supports incremental geometric and semantic fusion, while the Gaussians capture fine scene details and enable high-fidelity rendering. Given posed RGB-D observations, we extract 2D instance masks using SAM [15] and associate them with language-aligned CLIP embeddings [24]. These observations are incrementally integrated into a semantic field that remains consistent across viewpoints and over time.

Formally, the hybrid map is represented by a sparse voxel set $\mathcal{V} = \{v_j\}_{j=1}^V$, where each voxel v is associated with a set of Gaussian primitives $\mathcal{G}_v$. Let d denote the semantic embedding dimension (e.g., $d = 512$ for CLIP). Each voxel stores a fused semantic embedding $F_v \in \mathbb{R}^d$, an observation count $c(v)$, an accumulated semantic confidence $w(v)$, and a 3D coordinate $X_v \in \mathbb{R}^3$ in the global reference frame. For each observed voxel, an incoming semantic embedding f_v with confidence ω_v is integrated as:

$$F_v \leftarrow \frac{c(v)F_v + \omega_v f_v}{c(v) + \omega_v}. \quad (1)$$

Here, $c(v)$ controls the contribution of the stored embedding based on the number of prior observations, while ω_v weights the incoming embedding according to its observation confidence. This mixed weighting scheme reduces the influence of low-confidence observations on the accumulated embedding. After fusion, we update $c(v) \leftarrow c(v) + 1$ and $w(v) \leftarrow w(v) + \omega_v$, where the accumulated confidence $w(v)$ is used for subsequent instance scoring. Given a language query q , we obtain its normalized text embedding $e(q) \in \mathbb{R}^d$ and compute the voxel-level similarity $s_q(v) = \cos(e(q), F_v)$. This results in a dense semantic foundation that supports query-conditioned object extraction and subsequent scene graph construction.

Reliability-Aware Semantic Modeling. Semantic fusion over long RGB-D sequences naturally produces observations of varying reliability due to occlusions, partial visibility, and semantic ambiguity. To capture this uncertainty, OpenSplatGraph augments the semantic field with a per-voxel residual semantic disagreement $r(v)$. Together with the observation count $c(v)$ and accumulated confidence $w(v)$, it forms the per-voxel statistics:

$$\mathcal{S} = \{(c(v), w(v), r(v)) \mid v \in \mathcal{V}\}. \quad (2)$$

For each incoming observation, $r(v)$ is updated using an exponential moving average of the normalized cosine disagreement between the incoming and previously fused semantic embeddings, computed before semantic fusion. The residual is initialized to zero, with a fixed decay factor used throughout the sequence.

Given a candidate object region represented by a voxel subset $\mathcal{V}_k \subset \mathcal{V}$, we define an instance confidence score $C(\mathcal{V}_k) \in [0, 1]$ that favors well-observed and semantically consistent regions:

$$C(\mathcal{V}_k) = \lambda \cdot \tanh\left(\frac{1}{|\mathcal{V}_k|} \sum_{v \in \mathcal{V}_k} \frac{w(v)}{\kappa}\right) + (1 - \lambda) \cdot \text{sat}_{[0,1]}\left(1 - \frac{1}{|\mathcal{V}_k|} \sum_{v \in \mathcal{V}_k} r(v)\right), \quad (3)$$

where $\lambda \in (0, 1)$ controls the trade-off between observation strength and semantic stability, $\kappa > 0$ is a normalization constant, and $\text{sat}_{[0,1]}(x) = \min(1, \max(0, x))$. The confidence score is used solely for ranking and filtering candidate object instances and requires no supervised calibration.

Algorithm 1 Query-conditioned object extraction from voxel similarity scores.

Require: Query q ; voxel embeddings F ; voxel coordinates X ; voxel stats S ; threshold δ ; clustering params $(r, k_{\min})$; top- K

Ensure: Query-conditioned instances $\{O_k\}$ with voxel set $\mathcal{V}_k$, centroid μ_k , AABB, score S_k , confidence C_k

```

1:  $e \leftarrow \text{EncodeText}(q)$   $\triangleright$  normalized text embedding
2: for  $v = 1$  to  $V$  do
3:    $s(v) \leftarrow \cos(e, F_v)$ 
4: end for
5:  $\mathcal{M} \leftarrow \{v \mid s(v) > \delta\}$ 
6: if  $\mathcal{M} = \emptyset$  then
7:   return  $\emptyset$ 
8: end if
9:  $\mathcal{C} \leftarrow \text{DBSCAN}(\{X_v\}_{v \in \mathcal{M}}; r, k_{\min})$   $\triangleright$  a set of clusters
10: for all cluster  $\mathcal{V}_k \in \mathcal{C}$  do
11:    $\mu_k \leftarrow \text{mean}\{X_v \mid v \in \mathcal{V}_k\}$ 
12:    $\text{AABB}_k \leftarrow (\min_{v \in \mathcal{V}_k} X_v, \max_{v \in \mathcal{V}_k} X_v)$ 
13:    $S_k \leftarrow \text{mean}\{s(v) \mid v \in \mathcal{V}_k\}$ 
14:    $C_k \leftarrow \text{InstanceConfidence}(\mathcal{V}_k; S)$   $\triangleright$  computed via Eq. (3)
15:    $O_k \leftarrow (\mathcal{V}_k, \mu_k, \text{AABB}_k, S_k, C_k)$ 
16: end for
17:  $\{O_k\} \leftarrow \text{RankAndSelect}(\{O_k\}, K)$ 
18: return  $\{O_k\}$ 

```

Query-conditioned Object Extraction. Given a language query q , we extract a set of object instances $\mathcal{O}_q = \{O_1, \dots, O_K\}$ from the voxel-level similarity field $s_q(v)$. Each instance corresponds to a spatially coherent voxel cluster with associated semantic and geometric attributes. We first threshold the similarity field using $m_q(v) = \mathbb{I}[s_q(v) > \delta]$ to obtain the corresponding voxel coordinates $X_q = \{X_v \mid m_q(v) = 1\}$. We then cluster X_q in Euclidean space using DBSCAN with neighborhood radius r and minimum number of points $k_{\min}$. The radius is set proportional to the voxel size to ensure scale consistency across maps.

For each resulting voxel cluster $\mathcal{V}_k$, we compute its centroid $\mu_k = \frac{1}{|\mathcal{V}_k|} \sum_{v \in \mathcal{V}_k} X_v$, axis-aligned bounding box (AABB), instance score S_k as the mean voxel-level similarity, and reliability-based confidence $C(\mathcal{V}_k)$ from Eq. (3). Candidate instances are ranked by a weighted combination of S_k and $C(\mathcal{V}_k)$, prioritizing semantic similarity and using cluster size as a tie-breaker. The top-ranked candidates are retained for subsequent processing. The complete extraction procedure is summarized in Alg. 1. Importantly, object extraction is conditioned on the language query, whereas the underlying semantic field remains persistent and is incrementally updated during mapping.

3.2 Persistent 3D Scene Graph Construction

The persistent scene graph forms the reasoning layer of OpenSplatGraph, incrementally accumulating object-centric knowledge extracted from the reliability-aware semantic field. Rather than reconstructing a scene graph for every query,

Algorithm 2 Persistent object association via label-aware centroid matching.

Require: New instances $\{O_i\}$ with label ℓ_i and centroid μ_i ; previous instances $\{\hat{O}_j\}$ with identifier $\hat{id}_j$, label $\hat{\ell}_j$, centroid $\hat{\mu}_j$; distance threshold τ

Ensure: Assigned persistent IDs for all O_i

- 1: $\text{Prev}[\ell] \leftarrow \{\hat{O}_j \mid \hat{\ell}_j = \ell\}$ $\triangleright$ group previous instances by label
- 2: Mark all previous IDs as unused
- 3: **for all** new instance O_i **do**
- 4: Candidates $\leftarrow \{\hat{O}_j \in \text{Prev}[\ell_i] \mid \hat{id}_j \text{ unused, } \|\mu_i - \hat{\mu}_j\|_2 \leq \tau\}$
- 5: **if** Candidates $\neq \emptyset$ **then**
- 6: $j^* \leftarrow \arg \min_{\hat{O}_j \in \text{Candidates}} \|\mu_i - \hat{\mu}_j\|_2$
- 7: $O_i.\text{id} \leftarrow \hat{id}_{j^*}$
- 8: Mark $\hat{id}_{j^*}$ as used
- 9: **else**
- 10: $O_i.\text{id} \leftarrow \text{NewID}()$
- 11: **end if**
- 12: **end for**
- 13: **return** $\{O_i\}$

confidence-ranked object instances are associated with existing graph nodes whenever possible, enabling incremental updates to node attributes and inter-object relations. The resulting graph provides a compact and interpretable representation for structured reasoning while preserving the underlying dense semantic field for perception. We represent the scene as a graph $G = (\mathcal{N}, \mathcal{E})$, where nodes $\mathcal{N}$ define object instances and edges $\mathcal{E}$ encode inter-object relations.

Object Association. To maintain persistent object identities during mapping, OpenSplatGraph maintains a set of previously extracted object instances and associates newly extracted objects with this set before updating the scene graph.

Let $\hat{\mathcal{O}} = \{\hat{O}_j\}$ denote the set of previously extracted object instances, each associated with a persistent identifier (ID), semantic label, and centroid. For each newly extracted instance O_i with label ℓ_i and centroid μ_i , candidate matches are restricted to instances in $\hat{\mathcal{O}}$ with the same semantic label. Among these candidates, the nearest unmatched centroid within a distance threshold τ is selected. If a valid match is found, its persistent identifier is reused; otherwise, a new identifier is assigned. This one-to-one association preserves object identities across extraction steps and supports incremental graph updates. The complete association procedure is summarized in Alg. 2.

Nodes. Each graph node $N_i \in \mathcal{N}$ represents an associated object instance O_i and stores its persistent identifier, semantic label, centroid, axis-aligned bounding box (AABB), confidence score, and other geometric attributes. As new observations become available, node attributes are updated while preserving the persistent identifier. Optionally, descriptive captions can be attached by selecting informative object views and applying a lightweight vision-language captioning model [7]. Since caption generation is decoupled from graph construction, it can be performed asynchronously without affecting the online mapping pipeline.

Edges. Edges encode spatial or semantic relationships between graph nodes. We distinguish two complementary relation types: (i) online spatial relations inferred directly from object geometry, and (ii) optional open-set semantic relations inferred asynchronously using a VLM. This separation enables efficient low-latency access to spatial relations for robot planning while allowing richer semantic knowledge to be incorporated without interrupting online mapping.

1) Online Spatial Relations: Spatial relations are inferred online using simple geometric tests based on object centroids and AABBs. We consider relations such as *near* (centroid distance below a threshold), *on/under* (sufficient horizontal overlap with limited vertical separation), *inside* (bounding-box containment), and relative ordering relations derived from centroid offsets in the global coordinate frame. Each spatial relation is assigned a confidence score computed from normalized geometric measures.

2) Open-set Semantic Relations: To capture relations that cannot be reliably inferred from geometry alone, OpenSplatGraph optionally enriches the scene graph using a VLM. This process is independent of object extraction and graph construction, and is applied only to augment the graph with open-set semantic relations. Specifically, we render a set of keyframes following [8] from the reconstructed scene with persistent instance identifiers overlaid near visible objects, and query the VLM to predict directed relation triplets (i, r, j) . Predictions from individual keyframes are aggregated through multi-view voting. For each object pair, edge confidence is computed as the ratio of positive relation predictions to the number of keyframes in which both objects are co-visible. Relations with sufficient multi-view support are retained, reducing viewpoint-dependent ambiguity while improving the robustness of semantic relation prediction.

4 Experiments

We evaluate OpenSplatGraph from three complementary perspectives corresponding to the proposed dual-representation architecture. First, we assess the reliability-aware semantic field through open-vocabulary 3D semantic segmentation (Sec. 4.2). Next, we evaluate the persistent scene graph on 3D object grounding from relational natural language queries (Sec. 4.3). Finally, we demonstrate the practical utility of the complete system on a mobile robotic platform in real-world environments (Sec. 4.4). Ablation studies further examine the contributions of individual components (Sec. 4.5).

4.1 Implementation Details

We summarize the key implementation details of the proposed framework. For 2D feature extraction, SAM [15] is used on benchmark datasets, while MobileSAMv2 [36] is adopted for real-world robotic experiments to enable real-time

Table 1: Quantitative comparison of 3D semantic segmentation on ScanNet and Replica datasets. (Best and second-best results are highlighted in **bold** and underlined.)

Mode	Method	ScanNet					Replica				
		ScanNet20		NYUv2-40			ScanNet20		NYUv2-40		
		mIoU↑	mAcc↑	mIoU↑	mAcc↑	FPS↑	mIoU↑	mAcc↑	mIoU↑	mAcc↑	FPS↑
	LangSplat [23]	3.78	9.11	3.46	8.43	–	3.84	8.75	3.56	8.26	–
Offline	OpenGaussian [33]	30.08	45.59	27.53	42.17	–	25.63	36.24	23.45	31.59	–
	InstanceGaussian [17]	40.32	55.78	36.88	52.91	–	32.24	<u>45.99</u>	29.88	41.01	–
	ConceptFusion [11]	12.80	22.62	11.72	20.93	0.52	13.05	29.89	11.51	27.66	0.49
Online	OpenGS-Fusion [34]	27.09	39.62	24.69	35.95	1.22	<u>34.84</u>	45.15	<u>31.98</u>	<u>41.93</u>	1.02
	OpenSplatGraph (Ours)	<u>36.75</u>	<u>53.82</u>	<u>33.52</u>	<u>50.64</u>	<u>1.20</u>	35.37	47.67	32.87	42.85	1.02

operation. The resulting segmentation masks are then encoded using the OpenCLIP ViT-B/16 model to obtain language-aligned semantic embeddings. For query-conditioned object extraction, DBSCAN clustering is performed with radius $r = \beta \cdot \text{voxel_size}$, where $\text{voxel_size} = 0.05\text{m}$ and $\beta = 1.5$. We set $k_{\min} = 5$ and discard clusters containing fewer than $\text{min_voxels} = 10$ voxels. The semantic residual $r(v)$ is updated with decay factor $\alpha = 0.9$. Reliability-aware confidence estimation uses $\lambda = 0.6$ and $\kappa = 5.0$. Object association employs label-aware centroid matching with a distance threshold of $\tau = 0.35\text{m}$. When open-set semantic relations are enabled, GPT-4o-mini predicts relation triplets, which are retained with at least three votes across up to 30 keyframes.

4.2 3D Semantic Segmentation

Settings. (i) *Task:* We evaluate the quality of the proposed reliability-aware semantic field through open-vocabulary 3D semantic segmentation. Given a language query, object instances are extracted from the dense semantic field and compared against ground-truth semantic annotations. (ii) *Baselines:* We compare our approach against recent 3D-GS-based open-vocabulary methods, including LangSplat [23], OpenGaussian [33], and InstanceGaussian [17]. In addition, point cloud-based methods, including ConceptFusion [11] and OpenGS-Fusion [34], are compared to assess performance under online mapping settings. For all baselines, we use the official hyperparameters or tune them on validation scenes for fair comparison. (iii) *Metrics:* Experiments are conducted on ScanNet [4] and Replica [28] with ground-truth 3D semantic annotations. Performance is reported using $mIoU$ and $mAcc$. Following common practice, Replica’s 88 semantic categories are remapped to the ScanNet20 [4] and NYUv2-40 [27] label sets to ensure consistent evaluation.

Results. Table 1 reports quantitative results for open-vocabulary 3D semantic segmentation on ScanNet and Replica. Among the *offline* methods, InstanceGaussian achieves the best performance, benefiting from its bottom-up instance aggregation strategy [17]. Although OpenSplatGraph is designed for online semantic mapping rather than offline segmentation, it achieves competitive performance with InstanceGaussian while consistently outperforming other Gaussian-

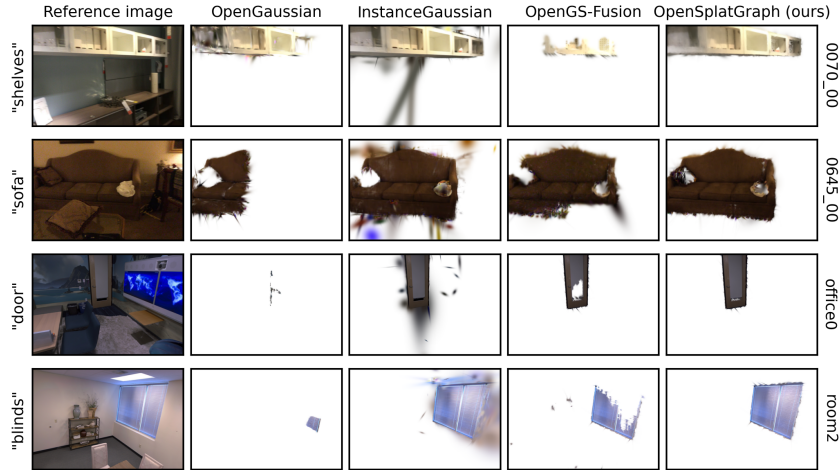


Fig. 3: Qualitative results of 3D semantic segmentation on ScanNet and Replica datasets.

based methods, including LangSplat and OpenGaussian. More importantly, under the *online* setting, OpenSplatGraph consistently surpasses ConceptFusion and OpenGS-Fusion on both mIoU and mAcc, demonstrating that the proposed reliability-aware semantic field produces more accurate and geometrically coherent object instances.

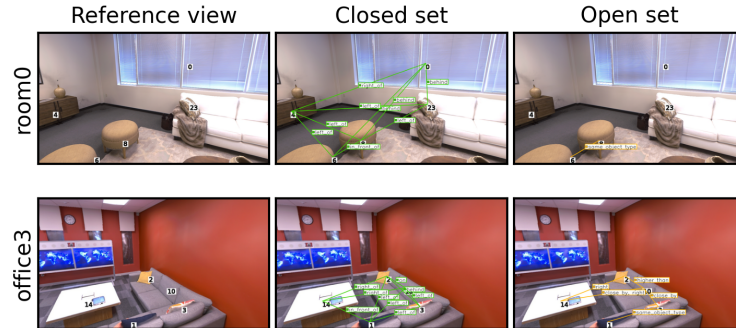
Figure 3 presents qualitative comparisons on both datasets. Compared with prior methods, OpenSplatGraph produces more spatially coherent object instances with better alignment to object boundaries, particularly for thin structures and partially observed objects such as “shelves”, “door”, and “blinds”. Modeling observation reliability effectively suppresses fragmented and noisy predictions. These qualitative observations are consistent with the quantitative results and demonstrate that the proposed reliability-aware semantic field provides a stronger foundation for persistent scene graph construction.

4.3 3D Object Grounding

Settings. (i) *Task:* We evaluate the persistent scene graph through 3D object grounding under complex natural language queries. Following the ConceptGraphs protocol [7], we consider three types of queries: *a) descriptive* (e.g., “a pillow on top of the sofa”), *b) affordance* (e.g., “something to tell the time”), and *c) negation* (e.g., “something to sit on other than a chair”). (ii) *Baselines:* We compare against ConceptGraphs [7], the only existing open-vocabulary 3D scene graph method with a publicly available grounding benchmark and evaluation protocol on Replica [28]. We adopt its query set and protocol for direct comparison, evaluating both CLIP retrieval based on query-embedding similarity and LLM retrieval based on structured node descriptions. (iii) *Metrics:* Performance is evaluated using the standard Recall@*k* ($R@k$) metric [29] following the ConceptGraphs protocol.

Table 2: Quantitative comparison of 3D object grounding on Replica across different query types and retrieval strategies. (ConceptGraphs results are reported from [7].)

Query Type	CLIP Retrieval						LLM Retrieval					
	ConceptGraphs			OpenSplatGraph			ConceptGraphs			OpenSplatGraph		
	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3
Descriptive	0.59	0.82	0.86	0.45	0.82	0.92	0.61	0.64	0.64	0.50	0.59	0.69
Affordance	0.43	0.57	0.63	0.54	0.61	0.79	0.57	0.63	0.66	0.61	0.71	0.75
Negation	0.26	0.60	0.71	0.51	0.64	0.82	0.80	0.89	0.97	0.80	0.89	0.94

**Fig. 4:** Qualitative visualization of open-vocabulary 3D scene graph construction on the Replica dataset. For visualization, asymmetric relations are displayed from nodes with lower instance IDs to those with higher IDs.

Results. Table 2 reports quantitative results for 3D object grounding on the Replica dataset across three query categories and two retrieval strategies. Overall, OpenSplatGraph achieves competitive grounding performance and consistently improves Recall@ k over ConceptGraphs in several retrieval settings. The largest improvements are observed for the *affordance* and *negation* queries, which require reasoning beyond direct semantic similarity. These gains demonstrate the benefit of maintaining a persistent object-centric scene graph, where explicit object identities and inter-object relations provide complementary information to the underlying dense semantic field during language grounding.

Figure 4 visualizes representative scene graphs predicted by OpenSplatGraph on two Replica scenes. For clarity, the 3D scene graphs are projected into 2D graph visualizations from a reference viewpoint. The predicted object relationships are spatially and semantically consistent with the underlying scene geometry. In addition, open-set relation prediction enables diverse semantic predicates, such as “same object type”, that cannot be predefined by fixed relation vocabularies, further enriching the scene graph for downstream reasoning. Additional direct scene-graph evaluation is provided in the supplementary material.

4.4 Real-world Robotic Experiments

We evaluate the practical utility of OpenSplatGraph in a real-world indoor environment using a Stretch 3 (Hello Robot) mobile manipulator equipped with

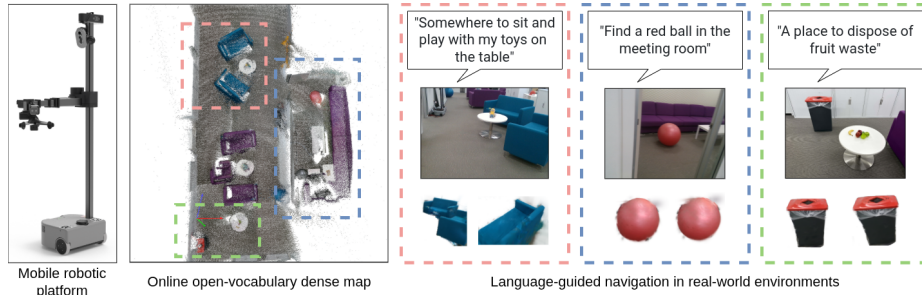


Fig. 5: Real-world robotic navigation experiments with OpenSplatGraph. The framework constructs an open-vocabulary semantic map online and supports language-guided navigation in indoor environments. For each query, we visualize the rendered target object from the reference view and from an additional arbitrary viewpoint.

an Intel RealSense D435i RGB-D camera and an arm with a gripper. During exploration, the robot incrementally scans the environment to acquire RGB-D observations while constructing an online Gaussian semantic map. Following OpenGS-Fusion [34], camera poses are estimated using ORB-SLAM3 [2]. The proposed reliability-aware semantic field and persistent scene graph are then built online from the fused semantic map to support language-guided robotic perception.

To demonstrate the effectiveness of the proposed perception framework for downstream robotics, we evaluate two representative tasks, as illustrated in Fig. 2 and Fig. 5. In a *navigation* task, the robot receives abstract natural language queries that may include semantic, relational, or functional descriptions. Using the persistent scene graph, the queried object is grounded in the reconstructed environment, after which the robot plans and executes navigation toward the target location. In a *manipulation* task, the robot performs language-guided pick-and-place by grounding an object query (e.g., “teddy bear”) within the persistent scene graph, localizing the corresponding object in the dense semantic map, and executing grasping and placement into a designated container. These experiments illustrate that OpenSplatGraph provides a unified structured and open-vocabulary perception backbone that supports closed-loop robotic navigation and manipulation in real-world environments. A demonstration video of the complete real-world experiments is included in the supplementary material.

4.5 Ablation Study

Table 3 presents a progressive ablation of the proposed framework. Starting from a baseline model, introducing reliability-aware semantic modeling significantly improves segmentation performance, demonstrating that modeling observation quality leads to more accurate and coherent object instances. Adding persistent object identities does not affect segmentation metrics but improves object grounding by enabling consistent object association across multiple queries. In this setting, object grounding relies solely on semantic similarity without exploiting relational information. Finally, incorporating scene graph reasoning

Table 3: Progressive ablation of OpenSplatGraph components.

Configuration	mIoU↑	mAcc↑	R@1↑
Baseline	30.97	42.39	–
+ Reliability-aware semantic modeling	36.06	50.74	–
+ Persistent object association	36.06	50.74	0.36
+ Scene graph reasoning (Full)	36.06	50.74	0.50

yields the largest gain in relational grounding, highlighting the benefit of explicit inter-object modeling. Overall, the full framework achieves the best performance across all metrics, validating the effectiveness of the proposed design.

4.6 Limitations

While OpenSplatGraph enables persistent scene graph construction from an online open-vocabulary semantic map, several limitations remain. First, object extraction is query-conditioned, and the framework does not maintain a complete query-independent object inventory. Instead, object instances are accumulated incrementally as new queries are issued. Second, object association relies on lightweight semantic and geometric cues, which may become less reliable in highly cluttered scenes or under severe viewpoint changes. Finally, open-set semantic relations are inferred using an external VLM, introducing additional computational overhead and dependence on the model’s reasoning capability. Addressing these limitations through more robust object association and query-independent proposal mechanisms is a promising direction for future work.

5 Conclusion

We presented OpenSplatGraph, a unified framework that jointly maintains a reliability-aware open-vocabulary semantic field and a persistent 3D scene graph for online robotic perception. The proposed semantic field enables robust query-conditioned object extraction from dense Gaussian-based maps, while the persistent scene graph incrementally accumulates object-centric knowledge to support structured relational reasoning. Extensive experiments on standard benchmarks demonstrate competitive performance in open-vocabulary semantic segmentation and 3D object grounding, and real-world robotic experiments validate the effectiveness of the proposed framework for language-guided navigation and manipulation. Future work will address the discussed limitations and explore integration with multi-step task planning in more complex environments.

Acknowledgements This work was supported in part by the Australian Research Council Discovery Project under Grant DP250103634, and in part by the Commonwealth Scientific and Industrial Research Organisation (CSIRO). The authors acknowledge continued support from the CSIRO’s Embodied AI Cluster.

References

1. Armeni, I., He, Z.Y., Gwak, J., Zamir, A.R., Fischer, M., Malik, J., Savarese, S.: 3d scene graph: A structure for unified semantics, 3d space, and camera. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5664–5673 (2019)
2. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D.: Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam. *IEEE transactions on robotics* **37**(6), 1874–1890 (2021)
3. Catalano, I., Zumaya, C.C., Placed, J.A., Civera, J., Bessa, W.M., Peña-Queralta, J.: 3d scene graphs in robotics: A unified representation bridging geometry, semantics, and action. *Authorea Preprints* (2025)
4. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
5. Deng, Y., Wang, J., Zhao, J., Tian, X., Chen, G., Yang, Y., Yue, Y.: Opengraph: Open-vocabulary hierarchical 3d graph representation in large-scale outdoor environments. *IEEE Robotics and Automation Letters* **9**(10), 8402–8409 (2024)
6. Ge, L., Zhu, X., Yang, Z., Li, X.: Dynamicgsg: Dynamic 3d gaussian scene graphs for environment adaptation. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2232–2239. IEEE (2025)
7. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., et al.: Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5021–5028. IEEE (2024)
8. Ha, S., Yeon, J., Yu, H.: RGBD gs-icp SLAM. In: European Conference on Computer Vision. pp. 180–197 (2024)
9. Huang, C., Mees, O., Zeng, A., Burgard, W.: Visual language maps for robot navigation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 10608–10615. IEEE (2023)
10. Hughes, N., Chang, Y., Carlone, L.: Hydra: A real-time spatial perception system for 3d scene graph construction and optimization. *arXiv preprint arXiv:2201.13360* (2022)
11. Jatavallabhula, K., Kuwajerwala, A., Gu, Q., Omama, M., Iyer, G., Saryazdi, S., Chen, T., Maalouf, A., Li, S., Keetha, N., et al.: Conceptfusion: Open-set multi-modal 3d mapping (2023)
12. Keetha, N., Karhade, J., Jatavallabhula, K.M., Yang, G., Scherer, S., Ramanan, D., Luiten, J.: Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21357–21366 (2024)
13. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **42**(4), 139–1 (2023)
14. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19729–19739 (2023)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)

16. Koch, S., Vaskevicius, N., Colosi, M., Hermosilla, P., Ropinski, T.: Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14183–14193 (2024)
17. Li, H., Wu, Y., Meng, J., Gao, Q., Zhang, Z., Wang, R., Zhang, J.: Instance-gaussian: Appearance-semantic joint gaussian representation for 3d instance-level perception. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 14078–14088 (2025)
18. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: *International conference on machine learning*. pp. 12888–12900. PMLR (2022)
19. Li, M., Liu, S., Zhou, H., Zhu, G., Cheng, N., Deng, T., Wang, H.: Sgs-slam: Semantic gaussian splatting for neural dense slam. In: *European Conference on Computer Vision*. pp. 163–179. Springer (2024)
20. Maggio, D., Chang, Y., Hughes, N., Trang, M., Griffith, D., Dougherty, C., Cristofalo, E., Schmid, L., Carlone, L.: Clio: Real-time task-driven open-set 3d scene graphs. *IEEE Robotics and Automation Letters* **9**(10), 8921–8928 (2024)
21. Matsuki, H., Murai, R., Kelly, P.H., Davison, A.J.: Gaussian splatting slam. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 18039–18048 (2024)
22. Nguyen, B.L., Nguyen, K., Sridharan, S., Fookes, C., Moghadam, P.: Ilov3splat: Instance-level open-vocabulary 3d scene understanding in gaussian splatting. In: *International Conference on Pattern Recognition*. pp. 251–266. Springer (2026)
23. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 20051–20060 (2024)
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PmLR (2021)
25. Rana, K., Haviland, J., Garg, S., Abou-Chakra, J., Reid, I., Suenderhauf, N.: SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Robot Task Planning. In: *7th Annual Conference on Robot Learning*
26. Shi, J.C., Wang, M., Duan, H.B., Guan, S.H.: Language embedded 3d gaussians for open-vocabulary scene understanding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5333–5343 (2024)
27. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgb-d images. In: *European conference on computer vision*. pp. 746–760. Springer (2012)
28. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., et al.: The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797* (2019)
29. Wald, J., Dhano, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3961–3970 (2020)
30. Wang, X., Yang, D., Gao, Y., Yue, Y., Yang, Y., Fu, M.: Gaussiangraph: 3d gaussian-based scene graph generation for open-world scene understanding. In: *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 4091–4098. IEEE (2025)

31. Werby, A., Huang, C., Büchner, M., Valada, A., Burgard, W.: Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In: First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024 (2024)
32. Wu, S.C., Wald, J., Tateno, K., Navab, N., Tombari, F.: Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7515–7525 (2021)
33. Wu, Y., Meng, J., Li, H., Wu, C., Shi, Y., Cheng, X., Zhao, C., Feng, H., Ding, E., Wang, J., et al.: Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. *Advances in Neural Information Processing Systems* **37**, 19114–19138 (2024)
34. Yang, D., Wang, X., Gao, Y., Liu, S., Ren, B., Yue, Y., Yang, Y.: Opengs-fusion: Open-vocabulary dense mapping with hybrid 3d gaussian splatting for refined object-level understanding. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 21135–21142. IEEE (2025)
35. Yin, H., Xu, X., Wu, Z., Zhou, J., Lu, J.: Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. *Advances in neural information processing systems* **37**, 5285–5307 (2024)
36. Zhang, C., Han, D., Zheng, S., Choi, J., Kim, T.H., Hong, C.S.: Mobilesamv2: Faster segment anything to everything. *arXiv preprint arXiv:2312.09579* (2023)
37. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8552–8562 (2022)
38. Zhou, G., Hong, Y., Wu, Q.: Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7641–7649 (2024)
39. Zhu, S., Qin, R., Wang, G., Liu, J., Wang, H.: Semgauss-slam: Dense semantic gaussian splatting slam. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 21174–21181. IEEE (2025)
40. Zieliński, M., Hall, D., Belter, D., Moghadam, P.: Neo: Nerf it once, edit it many times for continuous object manipulation. *IEEE Robotics and Automation Letters* (2026)

OpenSplatGraph: From Dense Semantic Maps to Structured Scene Graphs for Open-Vocabulary Robot Perception

Supplementary Material

A Additional 3D Semantic Segmentation Results

We provide additional comparisons with GaussianGraph [30] and DynamicGSG [6] for open-vocabulary 3D semantic segmentation on ScanNet and Replica. Table 4 reports the results alongside OpenSplatGraph. Compared with GaussianGraph, OpenSplatGraph achieves higher mIoU on both datasets. DynamicGSG uses a different class and evaluation mapping; therefore, its Replica results are included for reference rather than direct comparison.

Table 4: Additional open-vocabulary 3D semantic segmentation results on ScanNet and Replica. [†]DynamicGSG uses a different evaluation mapping.

Method	ScanNet		Replica	
	mIoU $\uparrow$	mAcc $\uparrow$	mIoU $\uparrow$	mAcc $\uparrow$
GaussianGraph [30]	31.09	48.91	31.18	49.14
DynamicGSG [6]	–	–	31.06 [†]	54.04 [†]
OpenSplatGraph (Ours)	36.75	53.82	35.37	47.67

B Direct Scene-Graph Evaluation

To directly evaluate graph correctness and identity persistence, we compare OpenSplatGraph with ConceptGraphs [7]. Following ConceptGraphs, we report node precision for object instances, edge precision for predicted inter-object relations, and duplicate-node rate to assess identity consistency across sequential observations.

Table 5: Direct scene-graph evaluation against ConceptGraphs.

Method	Node Prec. $\uparrow$	Edge Prec. $\uparrow$	Dup. Rate (%) $\downarrow$
ConceptGraphs [7]	0.71	0.88	6.77
OpenSplatGraph (Ours)	0.73	0.91	5.94

As shown in Table 5, OpenSplatGraph achieves higher node and edge precision while reducing the duplicate-node rate compared to ConceptGraphs. These results complement the object grounding evaluation by directly assessing graph correctness and identity consistency.