OVI-MAP: Open-Vocabulary Instance-Semantic Mapping

Zilong Deng[1,3], Federico Tombari[2,4], Marc Pollefeys[1,5], Johanna Wald[2,†], Daniel Barath[1,2,†]
[†]Equal contribution
[1]ETH Zurich, [2]Google, [3]University of Zurich, [4]TU Munich, [5]Microsoft

Abstract

Incremental open-vocabulary 3D instance-semantic mapping is essential for autonomous agents operating in complex everyday environments. However, it remains challenging due to the need for robust instance segmentation, real-time processing, and flexible open-set reasoning. Existing methods often rely on the closed-set assumption or dense per-pixel language fusion, which limits scalability and temporal consistency. We introduce OVI-MAP that decouples instance reconstruction from semantic inference. We propose to build a class-agnostic 3D instance map that is incrementally constructed from RGB-D input, while semantic features are extracted only from a small set of automatically selected views using Vision-Language Models. This design enables stable instance tracking and zero-shot semantic labeling throughout online exploration. Our system operates in real time and outperforms state-of-the-art open-vocabulary mapping baselines on standard benchmarks. Code available at https://ovi-map.github.io.

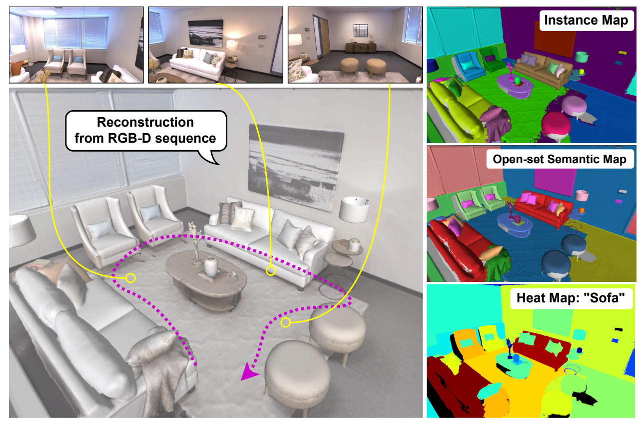


Figure 1. **Overview of OVI-MAP.** Given a streaming RGB-D sequence with camera poses, OVI-MAP incrementally reconstructs a volumetric 3D scene while maintaining a *class-agnostic* instance map. Semantic information is then assigned in a *zero-shot* manner using selectively chosen views, enabling open-set object recognition. Our method supports real-time, open-world scene reconstruction with instance-level semantic understanding.

1. Introduction

Semantic and instance 3D mapping are foundational capabilities for embodied perception, supporting downstream tasks like language-conditioned navigation, manipulation and facilitating queryable scene understanding for both AR/VR and robotics [1, 17, 19, 23, 47]. In indoor environments, voxel-based scene representations – most commonly Truncated Signed Distance Fields (TSDFs) – have become a standard choice due to their real-time fusion, robustness to drift, and ability to maintain dense and consistent geometry for planning and interaction [13, 34, 35]. Recent systems further couple volumetric reconstruction with semantics, yielding panoptic maps that are temporally consistent and easy to query [31, 49, 53, 58, 59].

However, existing pipelines are predominantly *closed-set*: they assume a fixed semantic ontology, learn class-dependent predictors, and store integer class labels per representation unit (*e.g.*, voxel, point). Extending these designs to *open-set* recognition is non-trivial for several reasons. First, open-set features extracted from Vision-Language Models (VLMs) are high-dimensional and continuous; naively storing them at voxel-level resolution leads to substantial computational and memory overhead.

Second, recent volumetric mapping systems [31, 53] rely on semantic labels to guide instance segmentation and association. Without semantics, object instance grouping becomes unstable and prone to fragmentation. Third, without consistent 3D instances, aggregating per-pixel open-set features over time is noisy due to occlusions, viewpoint changes, background noise and inconsistent 2D segmentations. While recent segmentation models, such as SAM [22], provide high-quality object proposals, running them per frame is computationally expensive and therefore unsuitable for real-time online mapping.

To address these limitations, we propose a pipeline that decouples instance mapping from open-set semantic recognition. Our approach first performs multi-view *class-agnostic* instance segmentation and incrementally lifts these instances into a global TSDF-based representation, produc-

Method	3D Recon.	Inst. Map	Zero-Shot Sem.	Online
Ours	TSDF	✔	✔	✔
OVO-SLAM [29]	PCL/3DGS	✔	✔	✔
OpenFusion [53]	TSDF	✘	✘	✔
ConceptFusion [19]	PCL	✘	✔	✔
OV-Octree-Graph [48]	PCL	✔	✔	✘
DCSEG [50]	3DGS	✔	✘	✘
OpenNeRF [10]	NeRF [32]	✘	✔	✘
Semantic Gauss [14]	3DGS	✘	✔	✘
OpenMask3D [46]	✘	✔	✔	✘
OpenScene [36]	✘	✘	✔	✘
Mask3D [43]	✘	✔	✘	✘

Table 1. State-of-the-art online and offline semantic/instance mapping methods. PCL stands for Point Cloud, and 3DGS denotes 3D Gaussian Splatting [20].

ing a consistent *instance map* across frames without requiring semantic labels. On top of this instance map, we introduce an *object-centric view selection strategy* that identifies informative viewpoints for semantic extraction. Semantic features are queried from VLMs only when a new viewpoint provides non-redundant coverage of the 3D object's surface, improving both efficiency and robustness while avoiding per-voxel storage of high-dimensional embeddings.
In summary, our contributions are:

- Class-agnostic online instance reconstruction pipeline that maintains a consistent 3D instance map independent of semantic categories.
- Object-centric incremental view selection mechanism that adaptively selects frames for feature extraction, reducing redundant VLM queries.
- An efficient, open-set semantic aggregation procedure that associates image regions with 3D instances and enables robust zero-shot semantic recognition.

We show on the ScanNet [7] and Replica [45] datasets that our proposed decoupled design leads to state-of-the-art instance mapping and open-set semantic segmentation while maintaining online processing rates.

2. Related Work

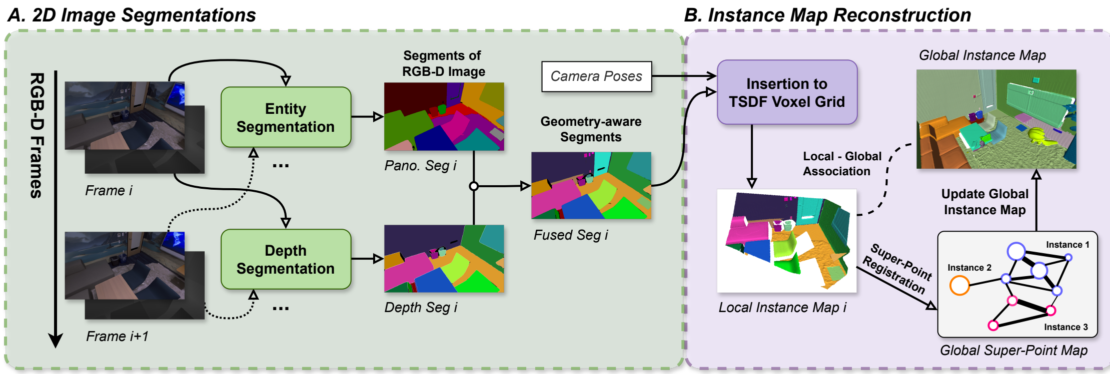
Open-Vocabulary Image Understanding has advanced rapidly with Vision-Language Models (VLMs), like CLIP [41] and SigLIP [60], that align visual and textual embeddings through large-scale contrastive pretraining. To enable pixel-level reasoning, many works produce dense or region-level semantic features. LSeg [24] predicts dense CLIP-aligned embeddings. OpenSeg [12] and OVSeg [28] refine segmentation using region proposals. SAM [22] and its derivatives [42, 51] offer category-agnostic masks, but produce over-segmented results. Works like Mask2Former [2, 25] provide semantic grounding only in closed-set, and X-Decoder [63, 64] provides a unified architecture for open-vocabulary segmentation. While these works show promising results on 2D images, they do not natively scale to 3D and there do not provide long-term, consistent segmentation across different views. We leverage these 2D models but focus on *incremental 3D* scenes and separate instance mapping from semantic recognition.

3D Semantic Understanding with VLMs. A growing class of methods lifts 2D open-vocabulary features into 3D by fusing multi-view semantic cues into volumetric, point-cloud, or neural field representations. OpenScene [36], PLA [9], and ConceptFusion [19] distill pixel-aligned VLM embeddings into point-level representations for open-vocabulary reasoning. Methods like CLIP-Fields [44], SemAbs [15], 3DSS-VLG [52], and PoVo [30] propagate CLIP-aligned features across views and align them with text embeddings. More recently, Gaussian splatting and NeRF approaches [10, 14, 21, 27, 37, 40, 58, 62] process 3D scenes by directly embedding semantics into 3D Gaussian primitives or neural implicit fields. However, these methods often require global scene optimization or dense 3D semantic fields, making them sub-optimal for online, incremental mapping. In contrast, our method avoids this limitation and computes semantics only *per-instance* via selective view aggregation, yielding scalable real-time mapping.

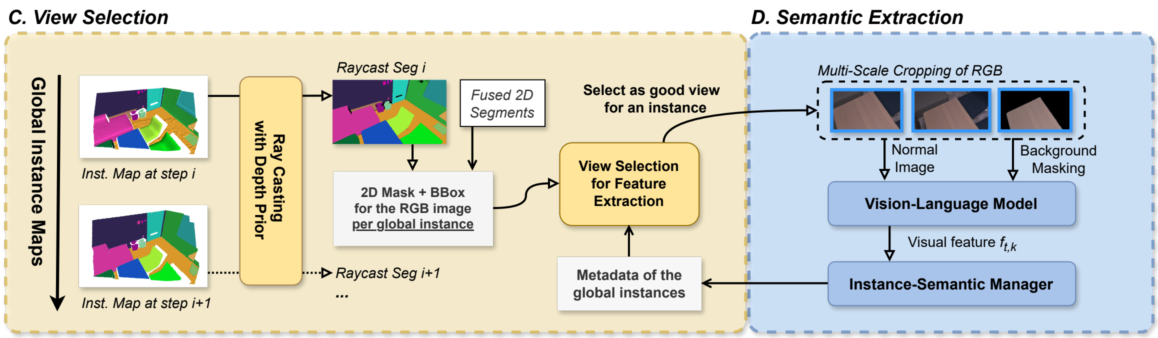
3D Instance Segmentation and Panoptic Mapping is commonly performed directly on point clouds using architectures such as PointNet++ [38], MinkowskiNet [4], and Mask3D [43]. These models, however, rely on large-scale annotated 3D datasets and are trained in closed-set settings. To reduce reliance on manual labels, recent works [18, 55–57] propagate 2D mask priors into 3D, but still require frequent segmentation updates and do not target online operation. Incremental panoptic mapping in indoor scenes has been explored using TSDF-based fusion [13, 26, 31, 53, 61], yet these systems couple instance association with closed-set semantic label prediction [2, 16, 64], making them incompatible with open-set recognition.

Open-Set Instance-Level Semantic Mapping. OpenMask3D [46], Search3D [47] extracts CLIP-like features for 3D instances from multi-view crops, demonstrating that *instance-level* semantic aggregation is more stable than pixel-level fusion [36]. However, these works assume instance masks are available and do not address online mapping. Works like O3D-SIM [33] and DCSEG [50] combine clustering-based instance segmentation and class-dependent semantic masks from 2D VLMs, achieving offline instance-semantic segmentation with pre-defined semantic labels. Recent, OVO-SLAM [29] and OpenGS-SLAM [54] combine SLAM with open-vocabulary instance queries, but they still rely on time-consuming and over-segmented SAM [22] outputs and do not reason about view selection efficiency. OpenVox [8] maintains the voxel-based instance map while its semantic codebook relies on Yolo-world [3] with pre-set label promptings.

In summary, existing open-set 3D scene understanding methods either (1) store dense semantic fields in 3D,



(a) Class-Agnostic Instance Map Reconstruction.



(b) Incremental Semantic Features Aggregation.

Figure 2. **Overview** of the proposed pipeline. **(a)** Class-agnostic instance map reconstruction: each RGB-D frame is segmented into entity proposals and refined with geometry-aware depth segmentation. The fused segments are lifted into 3D and incrementally integrated into a global TSDF-based instance map via super-point registration and spatial voting. **(b)** Incremental semantic feature aggregation: given the global instance map, each instance is re-projected into new frames via depth-guided ray casting. A view selection module identifies informative viewpoints based on object-centric coverage. Selected views are cropped at multiple scales and masked before being passed to a Vision-Language Model (VLM). The resulting features are aggregated per instance, producing stable open-set semantic embeddings.

which is computationally prohibitive for incremental mapping, or (2) rely on offline or closed-set instance segmentation pipelines. Our work differs in two key aspects: (i) we *reconstruct instances online in a class-agnostic manner*, avoiding the dependence on closed-set semantic labels for object grouping, and (ii) we introduce *object-centric view selection* to efficiently aggregate VLM features, avoiding dense or redundant semantic fusion.

3. Proposed Method

Given a streaming RGB-D sequence $\{(I_t, D_t), \mathbf{T}_t\}_{t=1}^{\infty}$, where $\mathbf{T}_t \in \mathrm{SE}(3)$ denotes the camera pose, I_t and D_t are the image and depth in the t-th frame, respectively, our goal is to incrementally construct: (i) a *class-agnostic 3D instance map*, and (ii) a *zero-shot semantic embedding* for each instance through selective multi-view feature extraction. The key idea is to decouple instance formation from semantic inference: instances are reconstructed purely from geometric and region-consistent evidence, and semantics are assigned only when *informative views* are observed. An overview of the pipeline is shown in Fig. 2.

3.1. Class-Agnostic Instance Map Reconstruction

We maintain the scene in a TSDF voxel grid $\mathcal{V}$ [6], following [13], where each voxel v stores the values for TSDF:

$$(v_{\text{tsdf}},\ v_{\text{weight}},\ v_{\ell}),$$

with $v_\ell \in \mathbb{N}$ denoting the instance identity, or 0 if unassigned. Instance identities are represented via a dynamic set of 3D *super-points* $\mathcal{S} = \{S_1, \ldots, S_K\}$, where each S_k corresponds to one globally consistent 3D instance and K denotes the current number of instances.

2D Entity Segmentation and Geometric Refinement. For each incoming RGB-D frame, we first compute a set of class-agnostic entity proposals as follows:

$$\{M_{t,j}\}_{j=1}^{N_t} = \mathrm{EntitySeg}(I_t),$$

where $M_{t,j}$ is a binary mask identifying a coherent image region likely corresponding to one physical object. Following [39], the segmentation emphasizes object-level boundaries while ignoring category semantics.

To improve segmentation around contacts and clutter, we compute a geometric segmentation G_t, in the t-th RGB-D frame, based on depth discontinuities [11] and use it to improve the obtained 2D masks as follows:

$$\hat{M}_{t,j} = \text{MaskFusion}(M_{t,j}, G_t),$$

where MaskFusion further split mask G_t into smaller pieces according to $\{M_{t,j}\}_{j=1}^{N_t}$ to avoid under-segmentation. This reduces over-merge errors where adjacent objects share texture or color, by using geometric discontinuities indicating boundaries not visible in appearance space.

3D Lifting. Each refined segment $\hat{M}_{t,j}$ is lifted into a 3D point cloud $P_{t,j}$ in the global coordinate frame using camera pose $\mathbf{T}_t$ and depth projection as follows:

$$P_{t,j} = \{\mathbf{x} = \mathbf{T}_t \cdot (D_t(\mathbf{u})K^{-1} \begin{bmatrix} \mathbf{u} \\ 1 \end{bmatrix}) : \mathbf{u} \in \hat{M}_{t,j}\},$$

where K denotes the camera intrinsics, $\mathbf{u}$ is a 2D point from the region defined by mask $\hat{M}_{t,j}$, $D_t(\mathbf{u})$ is the depth observed at $\mathbf{u}$, and $\mathbf{x}$ is the 3D point. This produces a partial surface patch corresponding to a single candidate object.

Instance Association via Spatial Voting. To determine if $P_{t,j}$ corresponds to an existing instance, we examine which instance label (*i.e.*, index of a 3D instance) appear the most frequently in voxels covering the points in $P_{t,j}$.

Let $V(\mathbf{x})$ denote the voxel contains $\mathbf{x}$, we define the voting of the super-point S_k with label k as:

$$\Omega_{j,k} = |\{\mathbf{x} \in P_{t,j} : V(\mathbf{x})_\ell = k\}|,$$

where $V(\mathbf{x})_\ell$ is the instance label of voxel $V(\mathbf{x})$. The assignment decision is the following:

$$k^* = \arg\max_k \Omega_{j,k}, \quad \text{if } \Omega_{j,k^*} > \theta_{\text{assoc}}, \text{ assign } P_{t,j} \mapsto S_{k^*};$$

where θ_{assoc} is a fixed overlapping threshold, otherwise a super-point S_{new} is created with a new label.

This association depends entirely on *spatial consistency*, not semantic similarity. Thus, unlike TSDF-based panoptic fusion pipelines [31, 61], we do not require a closed label set or hierarchical class priors.

Incremental TSDF Fusion and Label Stabilization. Once an instance with label k^* is selected for $P_{t,j}$, its surface points are integrated into the TSDF grid using standard weighted fusion [34]. Instance identity is reinforced via per-voxel label-support accumulation as follows:

$$O_v(k^*) \leftarrow O_v(k^*) + 1 \quad \forall v = V(\mathbf{x}), \mathbf{x} \in P_{t,j}.$$

After processing frame t, each voxel updates its label as:

$$v_\ell = \arg\max_k O_v(k) \quad \forall v \in \mathcal{V}.$$

This forms a temporally stable, class-agnostic 3D instance map $\mathcal{M}$ that progressively improves as the scene is observed from more viewpoints. Moreover, over-segmentation and multi-view fusion are handled by merging spatially close super-points, introduced in the supplementary materials.

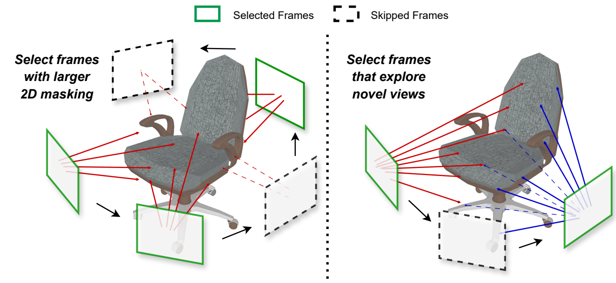


Figure 3. **Comparison of view selection strategies.** The left illustration shows the pixel-counting strategy [46], which prioritizes frames with larger object masking area, often leading to redundant front-facing views. The right illustration depicts our proposed *object-centric view coverage* method, which maintains a spherical map of explored viewing directions and selects frames that provide novel perspectives of the object. This yields a more diverse and informative set of viewpoints for semantic feature extraction.

3.2. View-Adaptive Semantic Feature Aggregation

Once object geometry is stable, we compute open-set descriptors. Rather than aggregating pixel-level VLM features across all frames – leading to noise, redundancy, and memory inefficiency – we extract semantic features *only* when a view provides novel evidence about object appearance.

Object-Centric View Coverage Representation. For each instance S_k, we maintain its *view coverage* on the unit sphere. Let $\text{Cov}_k \in \{0,1\}^{180\times240}$ denote the spherical coordinate map over bins, where if a bin contains 1, the instance has been observed from that direction.

We initialize S_k a centroid $\mathbf{c}_k$ of its 3D bounding box when it gets an instance mask with area larger than a fixed threshold. For every visible point $\mathbf{x} \in P_{t,k}$ at frame t, we compute its viewing direction:

$$\mathbf{d}_{t,\mathbf{x}} = \frac{\mathbf{x} - \mathbf{c}_k}{\|\mathbf{x} - \mathbf{c}_k\|},$$

and convert $\mathbf{d}_{t,\mathbf{x}}$ into spherical coordinates (θ, ϕ) to compute the current view coverage in the map Cov_k.

The novelty of the current view is measured by the ratio:

$$\eta_{t,k} = \frac{|\text{BinsNewOccupied}(P_{t,k})|}{|\text{BinsOccupied}(P_{t,k})|}.$$

If $\eta_{t,k} > \theta_{\text{novel}}$, the view reveals new geometric or appearance information and is selected for semantic extraction, where θ_{novel} is a predefined threshold. After a view is selected, all bins in Cov_k occupied by (θ, ϕ) are set to 1.

As illustrated in Fig. 3, this object-centric criterion differs fundamentally from the "visible pixel counting" heuristic [46]: while the latter repeatedly selects large, front-facing views that fail to capture full object shape, our coverage-based selection explicitly favors novel viewpoints that expand the explored surface area, yielding more diverse and informative observations.

Object-Level VLM Feature Extraction. For the selected frame of an instance, we extract the visual embeddings from two kinds of object crops: 1) a cropped image containing the object extent, and 2) a masked version where background pixels are removed:

$$\mathbf{f}_{t,k}^{(1)} = F_{\text{VLM}}(I_t, P_{t,k}), \quad \mathbf{f}_{t,k}^{(2)} = F_{\text{VLM}}(I_t \odot M_{t,k}).$$

We average these and update a running semantic feature f_k per instance using visibility-aware weighting:

$$\mathbf{f}_k \leftarrow \frac{w_{\text{sum}}}{w_k + w_{\text{sum}}}\mathbf{f}_k + \frac{w_k}{2(w_k + w_{\text{sum}})} \cdot (\mathbf{f}_{t,k}^{(1)} + \mathbf{f}_{t,k}^{(2)}),$$

where w_k denotes the number of visible object pixels in frame k and w_{sum} is the cumulative pixel count over all previous observations. This adaptive update produces a stable, view-invariant embedding that captures semantic consistency across frames, enabling robust zero-shot reasoning and language-conditioned retrieval.

Why Decoupling Works. The decoupled design separates temporally stable instance mapping from open-vocabulary semantic inference. Since instance mapping depends only on geometry and does not require consistent labeling, it remains stable and efficient in open-world environments. Semantic reasoning is then performed only when evidence is strong and informative, mitigating accumulated noise and reducing VLM calls significantly and controllably.

4. Experiments

Implementation Details. We adopt CropFormer [39] for 2D entity segmentation and the depth-based segmentor from [11] to obtain geometric boundaries. Our pipeline is implemented on top of Voxblox++ [13] using a TSDF [34] representation with voxel size of 0.1m, following [31] but without closed-set semantic dependencies. For semantic extraction, we use SigLIP [60] as the vision-language backbone to obtain zero-shot embeddings from RGB crops. For evaluation, text labels are encoded with the same model, and matched to instance features via cosine similarity. We run all methods on an Nvidia RTX3090 GPU and an Intel Core i7-12700K CPU.

Datasets. We evaluate our method on the *Replica* [45] and *ScanNet* [7] datasets, which provide dense RGB-D sequences with ground-truth camera poses, instance labels, and semantic annotations. For fair comparison, all methods are evaluated under identical input trajectories and reconstructed geometry. We run each sequence with 200 frames evenly sampled, as done in baselines [10, 14, 36]. For evaluation, we use the 51-class label set defined for Replica and the ScanNet200 label set for ScanNet.

4.1. Instance Segmentation Results

We first evaluate the quality of the reconstructed instance maps. Each reconstructed map is projected to the ground-truth mesh using a kNN-based nearest-neighbor mapping [5] with a fixed distance threshold, allowing per-vertex comparison with the annotated ground truth. We compare our approach with three state-of-the-art open-set methods: Mask3D [43], Segment3D [18], and OVO-SLAM [29]. Mask3D and Segment3D are offline volumetric networks predicting per-vertex instance masks, whereas OVO-SLAM and our method perform online, incremental mapping.

	Method	Online	mIoU	AP_{75}	AP_{50}	AP_{25}
Replica	Mask3D	✘	23.1	14.3	31.2	56.2
	Segment3D	✘	29.3	6.5	14.5	24.9
	OVO-SLAM	✔	**42.7**	11.1	23.6	32.8
	Ours	✔	36.3	**22.0**	**50.8**	**76.7**
ScanNet	Mask3D	✘	**47.6**	**16.9**	**36.1**	**47.8**
	Segment3D	✘	42.0	2.5	9.3	16.5
	OVO-SLAM	✔	39.8	2.0	7.4	14.4
	Ours	✔	41.2	9.8	24.0	37.4

Table 2. **Instance segmentation performance** on Replica [45] and ScanNet [7]. The *Online* column indicates whether the method performs incremental mapping during exploration. We achieve comparable or better accuracy to offline approaches while consistently outperforming prior online systems, particularly at high IoU thresholds (AP_{75}). Mask3D was trained on ScanNet.

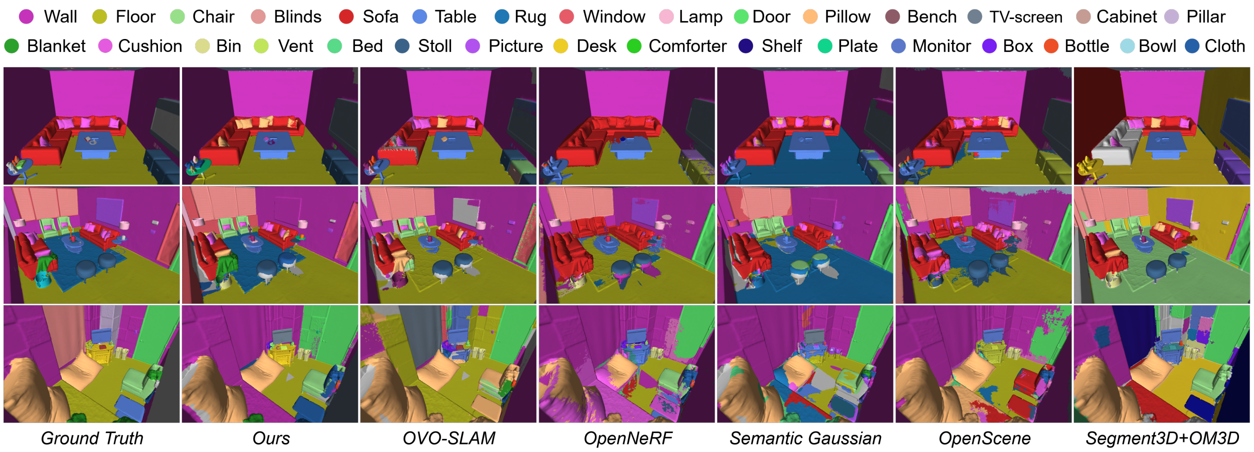
Metrics. We report mean IoU (mIoU) and instance-level average precision (AP) at IoU thresholds of 25%, 50%, and 75% following [18]. Best and second-best scores are highlighted. Offline methods are evaluated on the reconstructed meshes mapped to the ground truth ones.

Results. As shown in Table 2, our method achieves similar or higher instance-level precision compared to offline networks, while maintaining online operation. In particular, it surpasses the recent OVO-SLAM across all thresholds, with a substantial margin in AP_{75}, demonstrating the robustness of our class-agnostic instance tracking and fusion. It is important to note that, while Mask3D performs the best on ScanNet, it was trained on that dataset.

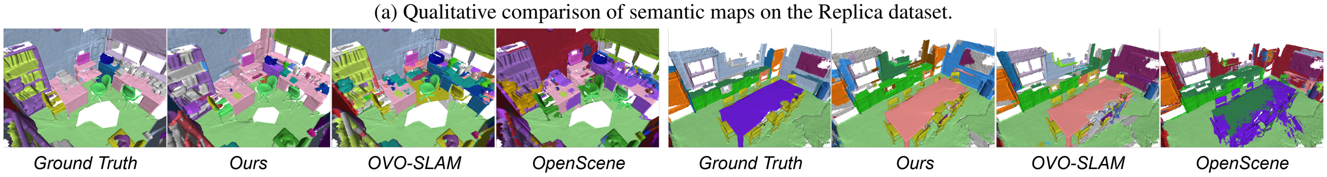
4.2. Semantic Segmentation Results

We next evaluate the open-vocabulary semantic performance. To further assess real-time capability, we benchmark both our OVI-MAP and the baseline OVO-SLAM [29] under a *30 FPS constraint*, where only every n-th frame is processed for semantic extraction to run in real time. On the Replica dataset, semantics are computed every 10th frame for our method and every 50th frame for OVO-SLAM, while on ScanNet, we process every 10th and 30th frame, respectively. Thanks to our lightweight and view-selective semantic extraction pipeline, OVI-MAP can perform semantic updates more frequently without exceeding real-time limits.

Baselines. We compare with offline and online open-vocabulary mapping systems. Among the offline methods, Mask3D+OpenMask3D (OM3D) and Segment3D+OM3D



(a) Qualitative comparison of semantic maps on the Replica dataset.



(b) Qualitative comparison of semantic maps on the ScanNet dataset.

Figure 4. **Open-vocabulary 3D semantic maps** aligned to the ground-truth label sets of the respective datasets. We compare our method with online [29] and offline [10, 14, 18, 36, 46] approaches on the ***Replica*** (a) and ***ScanNet*** (b) datasets. Colors correspond to the semantic classes defined in each dataset. Gray regions indicate unobserved areas. OVI-MAP produces spatially coherent and semantically accurate reconstructions, maintaining sharp instance boundaries and consistent semantics throughout incremental mapping. Minor discrepancies in color (e.g., *pillows* in (a) and the *table* in (b)) arise from closed-set label mapping, where the ground truth uses alternative class names such as "cushion" or "dining table".

	Method	Online	mIoU	mAcc	$\mathbf{AP_{25}}$	$\mathbf{AP_{50}}$	$\mathbf{AP_{all}}$
Replica	Mask3D+OM3D	✘	18.7	29.4	6.8	4.5	3.2
	Segment3D+OM3D	✘	17.3	30.9	19.9	15.2	9.0
	Semantic Gaussian	✘	16.1	22.1	–	–	–
	OpenScene	✘	19.8	33.9	–	–	–
	OpenNeRF	✘	19.2	30.9	–	–	–
	OpenFusion	✔	14.9	22.4	–	–	–
	OVO-SLAM	✔	24.9	**34.0**	28.1	17.5	**9.1**
	Ours	✔	26.5	32.2	**34.5**	**21.2**	8.5
	OVO-SLAM (30 fps)	✔	21.8	27.5	21.5	15.2	8.1
	Ours (30 fps)	✔	**27.0**	32.5	31.8	17.7	8.0
ScanNet	Mask3D+OM3D	✘	8.6	17.5	10.4	8.0	5.1
	Segment3D+OM3D	✘	4.5	13.5	5.0	3.9	2.3
	Semantic Gaussian	✘	10.7	17.8	–	–	–
	OpenScene	✘	10.3	17.3	–	–	–
	OpenNeRF	✘	Fail	Fail	–	–	–
	OpenFusion	✔	8.3	11.5	–	–	–
	OVO-SLAM	✔	14.6	**27.8**	19.4	12.6	5.5
	Ours	✔	**17.5**	27.6	**23.4**	**15.7**	**7.2**
	OVO-SLAM (30 fps)	✔	15.5	26.9	19.4	13.7	5.8
	Ours (30 fps)	✔	16.3	25.4	21.1	14.4	7.0

Table 3. Comparison of **open-vocabulary 3D semantic and instance segmentation** performance on Replica [45] and ScanNet [7]. OM3D denotes OpenMask3D [46]. The *30 FPS* rows correspond to experiments where only every n-th frame is processed to match real-time performance. Our method achieves the best open-vocabulary semantic-instance accuracy among online systems, and maintains performance comparable to offline approaches even under real-time constraints.

combine an instance segmentation backbone with zero-shot semantic labeling: Mask3D [43] and Segment3D [18] generate 3D instance masks from the mesh input, while OpenMask3D [46] assigns open-vocabulary semantic feature to each reconstructed instance with the RGB-D input. We also include Semantic Gaussian [14], OpenScene [36], OpenNeRF [10], and OpenFusion [53], which perform per-vertex open-vocabulary reasoning directly from scene representations without explicit instance segmentation. For online mapping, we compare to OVO-SLAM [29].

Metrics. We report mean IoU (mIoU) and accuracy (mAcc) following [10], as well as instance-level precision at different IoU thresholds (AP_{25}, AP_{50}, AP_{all}) as done in [46]. Since the ground-truth annotations in both datasets are provided as closed-set labels, open-vocabulary features are projected onto the label set defined by the datasets, following [10]. The semantic feature f_k of each reconstructed instance is compared against the SigLIP-encoded text embeddings of these class names, and the most similar label is assigned based on cosine similarity to enable consistent quantitative evaluation.

Results. As shown in Table 3, our method achieves the highest instance-level precision among the online systems and remains competitive with, or superior to, offline pipelines such as OpenScene and OpenNeRF. On both Replica and ScanNet, our approach improves the semantic-

Figure 5. **Instance highlighting from arbitrary text queries.** Given natural language prompts, our system retrieves and highlights corresponding 3D instances based on the learned vision-language embeddings. The examples demonstrate zero-shot grounding of both concrete ("*pillow*", "*toilet*") and abstract ("*where to sleep*", "*where is the music*") concepts in reconstructed scenes. Darker tones indicate higher cosine similarity between an object and the query.

	Method	mIoU	mAcc	AP_{25}	AP_{50}	AQ↓
Replica	Random 8 Views	23.8	30.4	31.4	18.6	50.3
	Pixel Counting	**26.5**	31.8	33.2	19.8	18.7
	View Coverage (Ours)	**26.5**	**32.2**	**34.5**	**21.2**	**8.6**
	GT Inst. + Pixel Cnt.	38.0	50.3	37.6	37.6	20.7
	GT Inst. + View Cov.	37.6	48.4	36.2	36.2	10.4
ScanNet	Random 8 Views	14.7	22.4	19.2	13.7	23.0
	Pixel Counting	**17.5**	**27.5**	**24.7**	**17.0**	15.8
	View Coverage (Ours)	**17.5**	27.2	23.4	15.7	**7.6**
	GT Inst. + Pixel Cnt.	34.8	48.9	39.8	39.8	23.0
	GT Inst. + View Cov.	30.7	44.8	34.7	34.7	12.3

Table 4. **View selection strategies** for semantic feature extraction: (1) random view selection, (2) larger visible-pixel counting similar to [46], and (3) our proposed *object-centric view coverage* method. *AQ* denotes the average number of VLM queries per instance. Our approach achieves comparable zero-shot instance-level semantic precision (AP_{25}/AP_{50}) to pixel-counting while requiring substantially fewer (47%) VLM queries. The *GT Inst. + Pixel Cnt.* and *GT Inst. + View Cov.* configurations provide an upper bound using ground-truth instance masks.

instance AP metrics by a large margin over OVO-SLAM, particularly at higher IoU thresholds (AP_{50} and AP_{25}), demonstrating that our view-selection-based semantic aggregation yields more reliable and consistent semantics. Under the *30 FPS* real-time constraint, OVI-MAP retains most of its original performance, achieving only marginal drops in $mIoU$ and AP scores, whereas OVO-SLAM exhibits a substantial degradation on Replica. Our results outperform offline methods even in the real-time regime.

Figure 4 presents qualitative results of the reconstructed semantic maps on Replica and ScanNet. Our method produces spatially coherent and semantically rich reconstructions, maintaining clear object boundaries and accurate label assignments across diverse viewpoints.

Open Query Searching. Figure 5 shows the map highlighting the most related objects in the scene with the given text queries. These results also demonstrate the accurate instance segmentation accuracy of our method.

4.3. Ablation Studies

We conduct ablation studies to analyze the effect of the proposed components, with a focus on the view-selection strategy for semantic feature aggregation and the trade-off between accuracy and computational efficiency.

View Selection Strategy. We evaluate the influence of different view selection strategies on semantic feature aggregation, focusing on the trade-off between recognition accuracy and computational efficiency. The goal is to minimize redundant queries to the vision-language model (VLM) while maintaining accurate open-vocabulary semantics.

We compare three strategies for selecting views per instance: (1) *Random* selection of eight frames per instance, serving as a naive baseline; (2) *Pixel-counting* selection, which prioritizes frames with the largest visible mask area [46]; and (3) our proposed *object-centric view coverage* selection, which measures how much of an object's surface has been observed and selects frames that contribute novel surface regions.

As shown in Table 4, random view selection, as expected, yields poor performance. Our object-centric view coverage strategy achieves comparable accuracy to pixel-counting, while using less than half the number of VLM queries (47.0% on average). This efficiency gain arises from explicitly modeling viewpoint novelty, which avoids redundant observations while preserving semantic consistency.

Incremental Performance. Fig. 6 shows how semantic accuracy evolves as more views are observed. Both our *view coverage* and *pixel-counting* strategies exhibit consistent improvement as additional frames are processed, demonstrating the benefits of multi-view semantic aggregation. Our view coverage approach achieves slightly better accuracy while requiring significantly fewer VLM queries per instance, as shown in the bottom-right plot. In contrast, post-hoc semantic extraction methods – such as OpenMask3D, Segment3D, and OpenScene – only converge af-

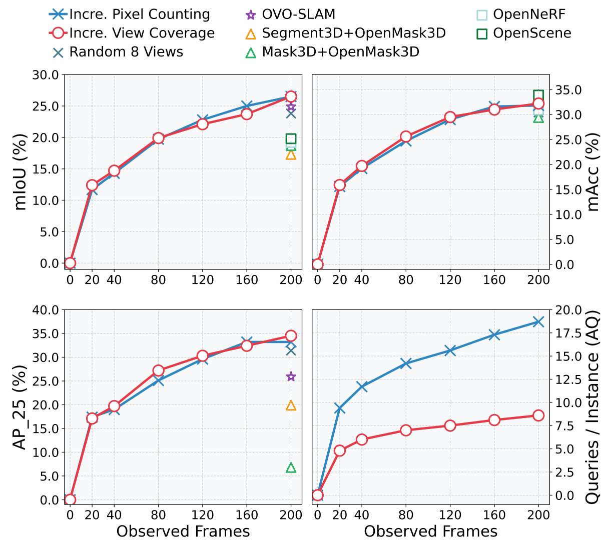


Figure 6. **Incremental performance** on Replica as the number of observed views increases. We compare our incremental semantic aggregation using the proposed *view coverage* strategy against pixel-counting and other baselines. Our method achieves comparable semantic accuracy (mIoU, mAcc, AP_{25}) while requiring significantly fewer VLM queries per instance, demonstrating efficient and scalable open-vocabulary mapping during online exploration.

ter the full sequence is processed, offering no incremental feedback during exploration. The results confirm that our coverage-based selection enables efficient, real-time semantic reasoning that scales with scene exploration, effectively bridging the gap between accuracy and online performance.

Influence of 2D Instance Segmentation. We further study how the quality of instance segmentation affects downstream semantic feature extraction. Table 5 compares different methods for 2D instance segmentation: (1) *SAM2*, which provides high-recall but noisy over-segmentations; (2) *CropFormer* [39], our chosen segmentor, producing compact and spatially coherent entities; and (3) *GT 2D instance masks*, serving as an oracle reference.

The results show that improving instance segmentation quality directly enhances semantic accuracy. CropFormer outperforms SAM2 in both instance- and semantic-level metrics, yielding a +5.7 improvement in AP_{50} and a +6.8 gain in semantic mIoU. This confirms that cleaner and more consistent instance boundaries lead to more discriminative VLM embeddings. The use of ground-truth instance masks further boosts all metrics, demonstrating the upper bound achievable with perfect segmentation.

Feature Fusion. We analyze how semantic features from multiple selected views should be fused to form stable instance-level embeddings. Table 6 compares several strategies: simple *averaging*, *weighted averaging* by the number of visible object pixels, and two *feature clustering* variants based on cosine similarity and L_1 distance.

Method	Instance Segmentation			Semantic Segmentation			
	mIoU	AP_{75}	AP_{50}	mIoU	mAcc	AP_{25}	AP_{50}
SAM2	36.6	14.2	27.8	20.2	26.4	26.2	18.6
Cropformer (Ours)	36.3	22.0	50.8	**26.9**	33.2	**36.4**	22.0
GT 2D Inst. Seg.	**53.9**	**38.8**	**65.7**	26.6	**33.8**	36.1	**27.3**
GT Instances Map	100	100	100	38.0	50.3	37.6	37.6

Table 5. **Influence of instance segmentation:** instance map built with 2D instance masks from (1) SAM2, (2) CropFormer, and (3) ground-truth 2D instance masks. All methods use the same view-selection strategy for semantic fusion. CropFormer yields the highest zero-shot semantic precision, while the *GT Instances* row represents an upper bound with perfect instance masks.

Method	mIoU	mAcc	AP_{25}	AP_{50}	AP_{all}
Averaging	26.5	31.8	33.2	19.8	8.3
+ Weighted by Pixel Count	**26.9**	**33.2**	**36.4**	**22.0**	**8.6**
Cluster - Max. Cosine Sim.	25.1	31.9	33.1	19.7	8.4
Cluster - Min. L_1 Distance	24.8	31.2	32.8	19.5	8.3

Table 6. **Semantic feature fusion by** (1) averaging, (2) weighted averaging by normalized pixel counts (Ours), (3) selecting the feature with maximum average cosine similarity to others, and (4) selecting the feature with minimum average L_1 distance to others.

Averaging provides a strong baseline, confirming that multi-view aggregation mitigates noise in individual predictions. Weighting by visible pixel count yields the best results across all metrics, as it naturally emphasizes clearer and better-framed object observations. Clustering-based methods offer no additional benefit and sometimes degrade performance due to overemphasis on redundant features.

These results indicate that a simple yet visibility-aware fusion is sufficient for robust zero-shot semantic reasoning, avoiding the complexity of feature clustering.

5. Conclusions

Our work OVI-MAP demonstrates that open-vocabulary semantic mapping can be performed incrementally and efficiently without sacrificing accuracy. By combining class-agnostic instance reconstruction with language-guided feature aggregation, we show that reliable semantic reasoning is possible under online constraints. The proposed object-centric view coverage strategy is central to this capability, enabling informed view selection that maintains semantic consistency while substantially reducing VLM queries. Together, these components bridge the gap between offline open-set understanding and real-time 3D perception.

Limitations. Our method still depends on the quality of 2D segmentation [39], which limits performance on small or visually complex objects. Semantic embeddings extracted from masked RGB crops are also affected by segmentation errors and background bias. Finally, current vision-language models provide only weak alignment between visual and textual spaces, causing ambiguous label assignments. Future research will explore tighter vision-language coupling and more adaptive feature fusion to enhance robustness in cluttered, real-world environments.

Acknowledgments

This work was supported by an ETH Zurich Career Seed Award, the ETH Foundation Project 2025-FS-352, and the SNSF Advanced Grant 216260.

References

[1] Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S Ryoo, Austin Stone, and Daniel Kappler. Open-vocabulary queryable scene representations for real world planning. In *IEEE Int. Conf. on Robotics and Automation*, pages 11509–11522, 2023. 1

[2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1290–1299, 2022. 2

[3] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16901–16911, 2024. 2

[4] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3075–3084, 2019. 2

[5] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1): 21–27, 1967. 5

[6] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Conference on Computer Graphics and Interactive Techniques*, pages 303–312, 1996. 3

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5828–5839, 2017. 2, 5, 6

[8] Yinan Deng, Bicheng Yao, Yihang Tang, Tianxing Zhou, Yi Yang, and Yufeng Yue. Openvox: Real-time instance-level open-vocabulary probabilistic voxel representation. In *IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, pages 1305–1311. IEEE, 2025. 2

[9] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7010–7019, 2023. 2

[10] Francis Engelmann, Fabian Manhardt, Michael Niemeyer, Keisuke Tateno, Marc Pollefeys, and Federico Tombari. Opennerf: open set 3d neural scene segmentation with pixel-wise features and rendered novel views. In *Int. Conf. Learn. Represent.*, 2024. 2, 5, 6

[11] Fadri Furrer, Tonci Novkovic, Marius Fehr, Abel Gawel, Margarita Grinvald, Torsten Sattler, Roland Siegwart, and Juan Nieto. Incremental object database: Building 3d models from multiple partial observations. In *IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, pages 6835–6842. IEEE, 2018. 4, 5

[12] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Eur. Conf. Comput. Vis.*, pages 540–557. Springer, 2022. 2

[13] Margarita Grinvald, Fadri Furrer, Tonci Novkovic, Jen Jen Chung, Cesar Cadena, Roland Siegwart, and Juan Nieto. Volumetric instance-aware semantic mapping and 3d object discovery. *IEEE Robotics and Automation Letters*, 4(3): 3037–3044, 2019. 1, 2, 3, 5

[14] Jun Guo, Xiaojian Ma, Yue Fan, Huaping Liu, and Qing Li. Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting. *arXiv preprint arXiv:2403.15624*, 2024. 2, 5, 6

[15] Huy Ha and Shuran Song. Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models. In *Conf. Robot. Learn.*, pages 643–653, 2022. 2

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Int. Conf. Comput. Vis.*, pages 2961–2969, 2017. 2

[17] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *IEEE Int. Conf. on Robotics and Automation*, pages 10608–10615, 2023. 1

[18] Rui Huang, Songyou Peng, Ayca Takmaz, Federico Tombari, Marc Pollefeys, Shiji Song, Gao Huang, and Francis Engelmann. Segment3d: Learning fine-grained class-agnostic 3d segmentation without manual labels. In *Eur. Conf. Comput. Vis.*, pages 278–295. Springer, 2024. 2, 5, 6

[19] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, et al. Conceptfusion: Open-set multimodal 3d mapping. In *Robotics: Science and Systems*, 2023. 1, 2

[20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[21] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Int. Conf. Comput. Vis.*, pages 19729–19739, 2023. 2

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Int. Conf. Comput. Vis.*, pages 4015–4026, 2023. 1, 2

[23] Sebastián Barbas Laina, Simon Boche, Sotiris Papatheodorou, Simon Schaefer, Jaehyung Jung, and Stefan Leutenegger. Findanything: Open-vocabulary and object-centric mapping for robot exploration in any environment. *arXiv preprint arXiv:2504.08603*, 2025. 1

[24] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *Int. Conf. Learn. Represent.*, 2022. 2

[25] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, Lei Zhang, and Jianfeng Gao. Segment and recognize anything at any granularity. In *Eur. Conf. Comput. Vis.*, pages 467–484. Springer, 2024. 2

[26] Wei Li, Junhua Gu, Benwen Chen, and Jungong Han. Incremental instance-oriented 3d semantic mapping via rgb-d

cameras for unknown indoor scene. *Discrete Dynamics in Nature and Society*, 2020(1):2528954, 2020. 2

[27] Yue Li, Qi Ma, Runyi Yang, Huapeng Li, Mengjiao Ma, Bin Ren, Nikola Popovic, Nicu Sebe, Ender Konukoglu, Theo Gevers, et al. Scenesplat: Gaussian splatting-based scene understanding with vision-language pretraining. In *Int. Conf. Comput. Vis.*, 2025. 2

[28] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7061–7070, 2023. 2

[29] Tomas Berriel Martins, Martin R Oswald, and Javier Civera. Ovo-slam: Open-vocabulary online simultaneous localization and mapping. *CoRR*, 2024. 2, 5, 6

[30] Guofeng Mei, Luigi Riz, Yiming Wang, and Fabio Poiesi. Vocabulary-free 3d instance segmentation with vision-language assistant. In *Int. Conf. on 3D Vis.*, pages 1197–1210. IEEE, 2025. 2

[31] Yang Miao, Iro Armeni, Marc Pollefeys, and Daniel Barath. Volumetric semantically consistent 3d panoptic mapping. In *IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, pages 12924–12931. IEEE, 2024. 1, 2, 4, 5

[32] 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. *Commun. ACM*, 65(1):99–106, 2021. 2

[33] Laksh Nanwani, Kumaraditya Gupta, Aditya Mathur, Swayam Agrawal, AH Abdul Hafez, and K Madhava Krishna. Open-set 3d semantic instance maps for vision language navigation–o3d-sim. *Advanced Robotics*, 38(19-20): 1378–1391, 2024. 2

[34] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136. IEEE, 2011. 1, 4, 5

[35] Helen Oleynikova, Zachary Taylor, Marius Fehr, Roland Siegwart, and Juan Nieto. Voxblox: Incremental 3d euclidean signed distance fields for on-board mav planning. In *IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, pages 1366–1373. IEEE, 2017. 1

[36] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 815–824, 2023. 2, 5, 6

[37] Jens Piekenbrinck, Christian Schmidt, Alexander Hermans, Narunas Vaskevicius, Timm Linder, and Bastian Leibe. Opensplat3d: Open-vocabulary 3d instance segmentation using gaussian splatting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5246–5255, 2025. 2

[38] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inform. Process. Syst.*, 30, 2017. 2

[39] Lu Qi, Jason Kuen, Weidong Guo, Tiancheng Shen, Jiuxiang Gu, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High-quality entity segmentation. *arXiv preprint arXiv:2211.05776*, 2022. 3, 5, 8

[40] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 20051–20060, 2024. 2

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, pages 8748–8763. PmLR, 2021. 2

[42] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[43] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *IEEE Int. Conf. on Robotics and Automation*, pages 8216–8223, 2023. 2, 5, 6

[44] Nur Muhammad Mahi Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. *arXiv preprint arXiv:2210.05663*, 2022. 2

[45] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 5, 6

[46] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. In *Adv. Neural Inform. Process. Syst.*, 2023. 2, 4, 6, 7

[47] Ayca Takmaz, Alexandros Delitzas, Robert W Sumner, Francis Engelmann, Johanna Wald, and Federico Tombari. Search3d: Hierarchical open-vocabulary 3d segmentation. *IEEE Robotics and Automation Letters*, 2025. 1, 2

[48] Zhigang Wang, Yifei Su, Chenhui Li, Dong Wang, Yan Huang, Xuelong Li, and Bin Zhao. Open-vocabulary octree-graph for 3d scene understanding. In *Int. Conf. Comput. Vis.*, pages 7037–7047, 2025. 2

[49] Silvan Weder, Hermann Blum, Francis Engelmann, and Marc Pollefeys. Labelmaker: automatic semantic label generation from rgb-d trajectories. In *Int. Conf. on 3D Vis.*, pages 334–343. IEEE, 2024. 1

[50] Luis Wiedmann, Luca Wiehe, and David Rozenberszki. Dcseg: Decoupled 3d open-set segmentation using gaussian splatting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5217–5226, 2025. 2

[51] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang,

Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16111–16121, 2024. 2

[52] Xiaoxu Xu, Yitian Yuan, Jinlong Li, Qiudan Zhang, Zequn Jie, Lin Ma, Hao Tang, Nicu Sebe, and Xu Wang. 3d weakly supervised semantic segmentation with 2d vision-language guidance. In *Eur. Conf. Comput. Vis.*, pages 87–104. Springer, 2024. 2

[53] Kashu Yamazaki, Taisei Hanyu, Khoa Vo, Thang Pham, Minh Tran, Gianfranco Doretto, Anh Nguyen, and Ngan Le. Open-fusion: Real-time open-vocabulary 3d mapping and queryable scene representation. In *IEEE Int. Conf. on Robotics and Automation*, pages 9411–9417. IEEE, 2024. 1, 2, 6

[54] Dianyi Yang, Yu Gao, Xihan Wang, Yufeng Yue, Yi Yang, and Mengyin Fu. Opengs-slam: Open-set dense semantic slam with 3d gaussian splatting for object-level scene understanding. *arXiv preprint arXiv:2503.01646*, 2025. 2

[55] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023. 2

[56] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *Eur. Conf. Comput. Vis.*, pages 162–179. Springer, 2024.

[57] Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3292–3302, 2024. 2

[58] Xuan Yu, Yili Liu, Chenrui Han, Sitong Mao, Shunbo Zhou, Rong Xiong, Yiyi Liao, and Yue Wang. Panopticrecon: Leverage open-vocabulary instance segmentation for zero-shot panoptic reconstruction. In *IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, pages 12947–12954. IEEE, 2024. 1, 2

[59] Xuan Yu, Yuxuan Xie, Yili Liu, Haojian Lu, Rong Xiong, Yiyi Liao, and Yue Wang. Leverage cross-attention for end-to-end open-vocabulary panoptic reconstruction. *arXiv preprint arXiv:2501.01119*, 2025. 1

[60] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Int. Conf. Comput. Vis.*, pages 11975–11986, 2023. 2, 5

[61] Jianhao Zheng, Daniel Barath, Marc Pollefeys, and Iro Armeni. Map-adapt: Real-time quality-adaptive semantic 3d maps. In *Eur. Conf. Comput. Vis.*, pages 220–237. Springer, 2024. 2, 4

[62] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Int. Conf. Comput. Vis.*, pages 15838–15847, 2021. 2

[63] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15116–15127, 2023. 2

[64] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Adv. Neural Inform. Process. Syst.*, 36:19769–19782, 2023. 2