OV-RASF: OPEN-VOCABULARY 3D INSTANCE SEGMENTATION WITH RELIABILITY-AWARE SEMANTIC FIELDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-Vocabulary 3D instance segmentation aims to localize objects and generate 3D instance masks from point clouds according to text queries. Existing methods struggle to balance efficiency and reliability: they either rely on foundation models for per-frame segmentation and classification, incurring high inference costs, or adopt bounding-box and multi-view label voting, which still suffers from in-box background interference, sparse semantic observations, and cross-view prediction inconsistency. The key issue is the lack of explicit modeling of instance-level semantic reliability. To this end, we propose OV-RASF, which, for the first time, introduces instance-level semantic reliability modeling into this task. The method projects and fuses multi-view vision-language features into 3D point space to construct an instance-queryable discrete 3D semantic field, and jointly quantifies reliability via semantic coverage and class response margin to adaptively adjust fusion weights. It leverages superpoint local consistency to constrain semantic propagation, completing masks and suppressing cross-instance leakage. Through class-residual gated fusion, it adaptively selects a more reliable prediction source for each class. The method requires no additional training and achieves more reliable 3D instance segmentation while exploiting the Open-Vocabulary recognition capability of 2D detectors. To validate the effectiveness of our method, we conduct experiments on two mainstream datasets (ScanNet200 and Replica). Results show that, compared with the current state-of-the-art methods, our method improves by 1.4% on ScanNet200 and 1.1% on Replica.The code will be made public later.

1 INTRODUCTION

In recent years, deep learning has accelerated progress in numerous cross-domain applications (Rombach et al., 2022; Xiang et al., 2025; Nguyen et al., 2025). Yet service robots and autonomous agents continue to struggle with localizing and retrieving targets in complex 3D scenes from open-ended queries. One pressing issue is how to tightly integrate natural language with 3D spatial understanding, thereby extending information retrieval from text or images into the 3D physical world. Open-vocabulary 3D instance segmentation (OV-3DIS) targets exactly this goal: from a 3D scene point cloud, multi-view RGB-D images, and a text query, the model is required to identify and output 3D masks of the relevant objects, independent of whether their categories have appeared during training. Such a capability links multimodal retrieval with spatial reasoning, supporting more natural, context-aware interaction with 3D environments. However, existing 3D instance segmentation approaches are usually trained on a predefined and limited category system, making generalization to novel objects difficult and severely limiting their deployment in real applications.

To break the closed-set limitation, most existing OV-3DIS methods (Takmaz et al., 2023; Nguyen et al., 2024) leverage pretrained 2D foundation models such as CLIP and SAM to extend recognition capability to unseen categories, or lift SAM's 2D segmentation results to 3D. Such methods have strong zero-shot generalization ability but incur high computational cost, requiring multiple forward passes through large 2D foundation models, often taking 5–10 minutes per scene, which is difficult to meet the demands of interactive or time-sensitive dynamic retrieval.To improve efficiency, OpenYOLO3D (Boudjoghra et al., 2025) uses a class-agnostic 3D instance segmentation model Mask3D (Schult et al., 2023) to generate candidates and combines it with a lightweight Open-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

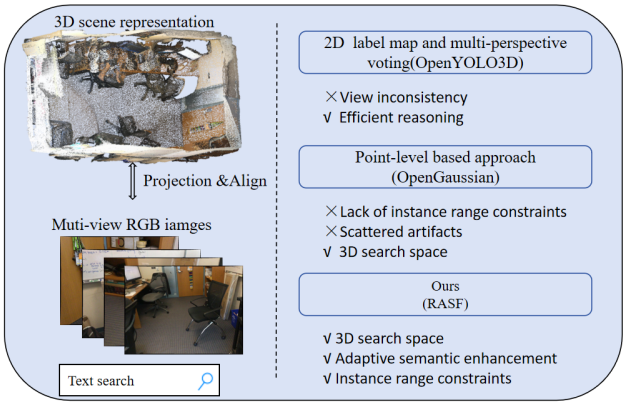

Figure 1: Comparison between 2D rendering methods (Boudjoghra et al., 2025), point-level methods (Wu et al., 2024), and the proposed reliability-aware instance-conditioned 3D semantic field. Previous methods either rely on 2D rendering supervision signals that lack cross-view consistency, or adopt point-level feature matching strategies, which often produce scattered or noisy activation maps. In contrast, our method builds a unified instance-level representation that preserves cross-view semantic coherence and supports precise retrieval of object-centric 3D instances through text queries.

Vocabulary 2D detector YOLO-World for classification, improving inference speed to about 22 seconds per scene while maintaining strong recognition performance; Box-Guided (Nguyen et al., 2026) uses 2D bounding boxes to guide the discovery of novel object masks in 3D point clouds, improving generalization to unseen categories. However, such methods rely entirely on pretrained 3D segmenters to produce candidates, often missing rare-category instances that are underrepresented in existing 3D datasets; meanwhile, multi-view voting is susceptible to occlusion and view inconsistency, leading to unstable class discrimination; the semantic quality within instances is not differentiated, making it easy for features from sparse observations or boundary contamination to participate in class decisions.

Therefore, this paper focuses not on simply stacking more 2D semantics, but on how to judge, at the instance level, whether semantic evidence is trustworthy, where it should be propagated, and when to adopt enhanced results. To this end, we propose a reliability-aware semantic field framework for Open-Vocabulary 3D instance segmentation, OV-RASF. This framework is organized around candidate instances, transforming continuous vision-language features aligned with the point cloud into assessable instance semantics, and using local geometric relations and class-level performance priors to respectively constrain mask refinement and prediction selection, thereby forming a progressive enhancement pipeline of "instance semantic construction—local structure optimization—class-level decision". The entire framework requires no retraining of the backbone network and can be integrated as an inference-stage enhancement mechanism into existing efficient Open-Vocabulary 3D segmentation pipelines. The main contributions of this paper are as follows:

1. We construct an instance-queryable 3D semantic field by fusing multi-view vision-language features, and adaptively adjust fusion weights via semantic coverage and class margin for reliable instance-level enhancement.

2. We propose a superpoint-constrained mask refinement method that limits semantic propagation via local 3D topology, reducing mask leakage between adjacent objects while preserving instance boundaries.

3. We propose an adaptive residual gated fusion module that fuses the base and enhancement branches according to class-wise performance differences, alleviating class imbalance caused by a unified post-processing strategy.

2 Related Work

Open-Vocabulary 2D Scene Understanding. Open-Vocabulary 2D scene understanding aims to break the limitation of predefined category sets, enabling models to simultaneously recognize base categories appearing during training and unseen novel categories. In the direction of Open-Vocabulary object detection (OVOD), various methods (Zhong et al., 2022; Pham et al., 2024; Zang et al., 2022; Wang et al., 2023; Kaul et al., 2023) have been proposed, and Open-Vocabulary image and instance segmentation has also been explored through image-level supervision (Ghiasi et al., 2022) and mask-free training (VS et al., 2023). Thanks to large-scale image-text contrastive learning, CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) have established a visual-language shared embedding space with good transferability, providing a unified semantic foundation for Open-Vocabulary recognition. On this basis, researchers have extended Open-Vocabulary capability in directions such as object detection, semantic segmentation, and instance segmentation: Open-Vocabulary detection methods localize objects and output bounding boxes according to text queries (Cheng et al., 2024; Yao et al., 2023; Zhou et al., 2022; Liu et al., 2024b), among which YOLO-World (Cheng et al., 2024) introduces language prompts into a real-time detection framework, maintaining high efficiency while recognizing unseen categories; Open-Vocabulary semantic and instance segmentation methods use vision-language features to generate pixel-level class responses and combine general segmentation models such as SAM (Kirillov et al., 2023) to obtain object masks (Liang et al., 2023; Liu et al., 2024a). These 2D foundation models have strong zero-shot generalization ability and thus become important semantic sources for Open-Vocabulary 3D scene understanding. However, directly transferring 2D predictions to 3D space is still affected by occlusion, viewpoint changes, and cross-frame prediction inconsistency, requiring cross-view association and semantic fusion combined with camera geometry and 3D structure.

Open-Vocabulary 3D Instance Segmentation. Open-Vocabulary 3D instance segmentation typically adopts a two-stage pipeline: first generate class-agnostic 3D instance candidates, then use a vision-language model to match candidate instances with text categories. In the candidate generation stage, one class of methods directly uses pretrained 3D instance segmenters to generate point cloud masks (Jiang et al., 2020; Vu et al., 2022; Jain et al., 2024; **?**). OpenMask3D (Takmaz et al., 2023) uses a class-agnostic 3D segmentation network to generate candidates, then combines SAM and CLIP to aggregate candidate region features from multi-view RGB-D images. Another class of methods generates instance masks from 2D images and lifts them to 3D space. OVIR-3D (Lu et al., 2023) generates 3D candidates by fusing 2D instance masks across views; Open3DIS (Nguyen et al., 2024) combines 2D and 3D candidates and adopts hierarchical clustering to obtain a more complete instance set. Methods based on 2D information have good generalization ability for small objects and long-tail categories, but usually require repeatedly running SAM and CLIP on many image frames, bringing high computational overhead.

To improve inference efficiency, OpenYOLO3D (Boudjoghra et al., 2025) uses an Open-Vocabulary 2D detector to replace computationally intensive 2D segmentation and candidate-region CLIP encoding. This method constructs a low-granularity label map from 2D detection boxes, then completes instance classification through 3D projection, visibility filtering, and multi-view label aggregation. Box-Guided (Nguyen et al., 2026) further uses 2D detection boxes to guide RGB-D-based 3D candidate generation, enhancing retrieval of rare categories while avoiding the use of 2D segmentation models. Such methods provide an efficient path for Open-Vocabulary 3D instance segmentation, but discrete labels based on detection boxes easily contain background and adjacent objects, and may be affected by occlusion and cross-view prediction inconsistency. This paper, while retaining the efficient 2D detection path, introduces continuous 3D vision-language features aligned with the point cloud to provide complementary semantic information for instance classification and mask optimization.

3D Vision-Language Representation. The basic idea of 3D vision-language representation is to project and fuse visual features extracted by 2D foundation models into 3D space, thereby supporting text-driven point cloud understanding. PLA (Ding et al., 2023) aligns language with 3D scenes for Open-Vocabulary understanding, and 3D concept learning (Hong et al., 2023) further studies multi-view concept reasoning. OpenScene (Peng et al., 2023) lifts 2D Open-Vocabulary semantic features to 3D point space, so that each valid 3D point has a vision-language representation matching text embeddings. ConceptGraphs (Gu et al., 2024) builds a scene graph on the basis of Open-Vocabulary object representations, providing support for downstream tasks such as ob-

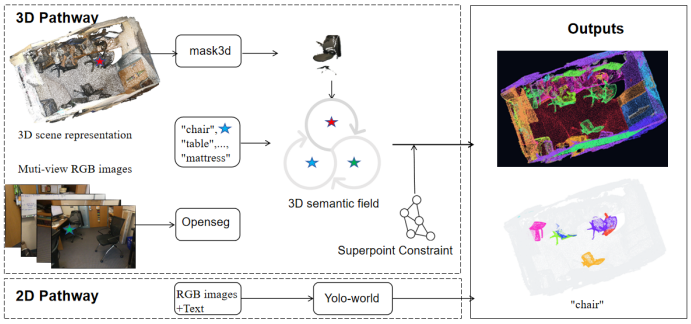

Figure 2: Architecture of the reliability-aware semantic field model for Open-Vocabulary 3D instance segmentation.

ject localization, navigation, and manipulation. Open3DIS (Nguyen et al., 2024) further shows that, compared with constructing a single feature only for a complete mask, per-point 3D vision-language features can preserve finer-grained spatial semantics.

However, existing 3D vision-language representations usually take point-level features as the main carrier. When directly averaging or indiscriminately using these features, instance boundaries and differences in effective semantic coverage across instances are not explicitly considered. Occlusion, insufficient projection viewpoints, and sparse point clouds may cause some instances to have only a few points with valid semantic features, while similar categories may also produce close responses in the vision-language space. To address this issue, this paper queries the 3D semantic field conditioned on candidate instance masks, aggregates valid point features inside the instance into an instance-level representation, and jointly evaluates its sufficiency and separability via semantic coverage and class margin. This mechanism does not redefine the point-level 3D semantic field, but enables its output to be adaptively used according to the semantic quality of a specific instance.

3 METHOD

As shown in Figure 2, our method takes as input a 3D point cloud with N points, multi-view RGB-D images and their camera parameters, Open-Vocabulary text categories, and class-agnostic 3D instance candidates. The goal is to predict a mask, class label, and confidence for each candidate instance in the scene. The overall pipeline consists of three interconnected stages: first, multi-view vision-language features are lifted and fused into 3D space aligned with the point cloud (Landrieu & Simonovsky, 2018; Geist et al., 2025), instance-level semantic representations are constructed conditioned on candidate masks, and the enhancement strength of 3D semantics on original class evidence is controlled according to semantic coverage and class margin; subsequently, local superpoint topology (Edelsbrunner & Harer, 2008; Hui et al., 2021; Robert et al., 2023) is used to constrain the propagation range of vision-language semantics, refining missing regions and boundaries of candidate masks; finally, class-aware residual gating is performed according to reliability differences of the base branch and the enhancement branch on different classes, selecting a more appropriate prediction result. Through the above process, the proposed framework progressively optimizes 3D instance prediction from three levels—instance semantics, mask structure, and class decision—while retaining an efficient Open-Vocabulary inference path. Similar to recent studies (Peng et al., 2023; Takmaz et al., 2023; Nguyen et al., 2024), we assume that the poses and camera parameters of the input 3D scene are known.

3.1 RELIABILITY-AWARE INSTANCE-CONDITIONED 3D SEMANTIC FIELD

First, camera parameters are used to project multi-view image features into 3D point space. For a 3D point x_j, aggregate its vision-language features across valid views to obtain a point-level 3D semantic representation:

$$f_j = \mathrm{Norm}\left(\frac{\sum_v a_{vj}\Phi_v(u_{vj})}{\sum_v a_{vj} + \epsilon}\right), \tag{1}$$

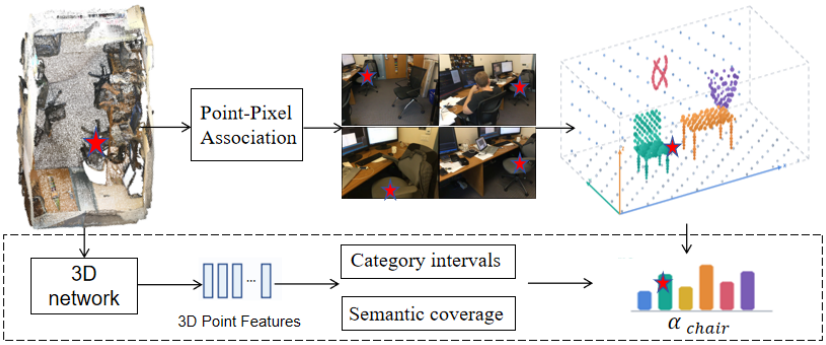

Figure 3: Implementation architecture of the 3D semantic field.

where $\Phi_v(u_{vj})$ denotes the 2D feature at the position corresponding to point x_j in the v-th view, and a_{vj} indicates whether the projection is valid. The features of all points constitute a 3D vision-language semantic field aligned with the scene point cloud. For points without valid features, we set them to zero vectors to avoid introducing spurious semantics into unobserved regions.

Then, conditioned on the class-agnostic instance mask M_i, we aggregate only points inside the instance that have valid semantic features. Define the set of valid semantic points as $\Omega_i = \{j \mid M_{ij} = 1, \|f_j\|_2 > 0\}$. The instance semantic representation is computed as:

$$e_i = \mathrm{Norm}\left(\frac{\sum_{j\in\Omega_i} w_j f_j}{\sum_{j\in\Omega_i} w_j + \epsilon}\right), \tag{2}$$

where w_j denotes the valid support weight provided by the semantic cache. This instance-conditioned query restricts semantic aggregation to within the candidate mask, thereby reducing feature contamination caused by background and adjacent instances.

To obtain Open-Vocabulary class responses, we perform cosine matching between the instance representation e_i and class text embeddings t_c, and apply temperature scaling to obtain a 3D semantic distribution:

$$z_{i,c} = e_i^T t_c, \quad p_{i,c}^{3D} = \frac{\exp(z_{i,c}/T)}{\sum_{c'\in S_i}\exp(z_{i,c'}/T)}, \tag{3}$$

where T is a temperature parameter and S_i is the set of categories that currently receive 2D observation support for the instance. Restricting the category support range prevents the 3D semantic field from introducing new categories without 2D evidence in an unconstrained manner.

Considering that instance semantic responses are not always reliable, we evaluate them from two aspects: semantic sufficiency and class separability. The instance semantic coverage and class margin are defined respectively as:

$$\rho_i = \frac{|\Omega_i|}{\sum_j M_{ij}}, \quad \Delta_i = z_{i,(1)} - z_{i,(2)}, \tag{4}$$

where ρ_i denotes the proportion of points inside the instance that have valid semantic features, and $z_{i,(1)}$ and $z_{i,(2)}$ are the highest and second-highest class responses, respectively. Coverage measures whether semantic information is sufficient, and class margin measures whether the class decision is clear.

We normalize ρ_i and Δ_i through piecewise linear functions g_ρ and g_Δ, respectively, and compute instance-level reliability. Reliability further controls the participation strength of 3D semantic information:

$$r_i = \sqrt{g_\rho(\rho_i) g_\Delta(\Delta_i)}, \quad \alpha_i = \alpha_0 r_i, \quad p_{i,c} \propto \left(p_{i,c}^{2D}\right)^{1-\alpha_i} \left(p_{i,c}^{3D}\right)^{\alpha_i}, \tag{5}$$

where α_0 is the base semantic weight and $p_{i,c}^{2D}$ is the multi-view 2D class distribution. When instance semantic coverage is sufficient and class responses are clear, 3D semantics obtain a higher weight;

when coverage is insufficient or classes are ambiguous, its contribution automatically decreases, thus avoiding unreliable semantics from damaging the original prediction.

By querying the point-level 3D semantic field conditioned on instance candidates, and jointly establishing an instance-level reliability control mechanism via semantic coverage and class margin, general point-level features are transformed into reliable and controllable instance-level Open-Vocabulary semantic representations. This module mainly addresses the granularity mismatch between point-level semantics and instance-level prediction, as well as erroneous semantic enhancement caused by sparse semantic coverage and class ambiguity.

3.2 SUPERPOINT-CONSTRAINED VISION-LANGUAGE MASK REFINEMENT

To supplement local regions missed by the initial instance mask while avoiding absorbing distant same-class objects or adjacent instances solely by semantic similarity, we propose a superpoint-constrained vision-language mask refinement method. It takes the instance semantic representation generated by the first module as semantic guidance and 3D superpoint adjacency relations as spatial constraints, performing local, restricted mask expansion for each instance candidate separately.

For instance M_i^0, we first compute the instance center e_{i0} from its internal valid semantic points. Then, we build a superpoint graph according to the adjacency relations between superpoints in the triangular mesh: $G = (V, \mathcal{E})$.

Unlike global semantic expansion, we only take superpoints touched by the original instance and their limited-hop neighbors as candidate regions. For a candidate point j, compute its similarity to the instance center:

$$r_{ij} = \cos(f_j, e_i). \tag{6}$$

A candidate point is added to the instance mask only when $r_{ij} \geq \tau$ and its superpoint belongs to the local candidate region. To avoid unstable expansion caused by small-scale noise, when the number of valid seed points inside the instance is below a minimum threshold, we directly retain the original mask.

In addition, to limit the refinement range, we set a maximum mask expansion ratio and an upper limit on the number of newly added points:

$$|M_i^{'}| \leq \gamma |M_i^0|, \quad |M_i^{'} \setminus M_i^0| \leq K_{\text{add}}. \tag{7}$$

If the number of points satisfying the similarity condition exceeds the upper limit, candidate points are selected in descending order of similarity. The final $M_i^{'}$ is simultaneously constrained by semantic similarity and local geometric topology, so that instance regions occluded or not covered by initial candidates can be completed while reducing cross-object propagation. After mask refinement, we recompute the instance semantic representation so that subsequent class prediction reflects the updated mask region.

3.3 RESIDUAL GATED FUSION

After the above two modules, the model can obtain a set of semantically enhanced prediction results. However, semantic enhancement affects different classes inconsistently. For classes with clear vision-language representations, the enhancement branch usually improves prediction quality; for classes with similar appearance, sparse observations, or unstable semantic features, the enhancement operation may lead to class confusion or mask degradation.

Therefore, we further propose a class-aware residual gated fusion module. On an independent validation set or calibration set, we evaluate the base branch and the semantic enhancement branch separately, and generate a fixed class gating table according to the performance difference of each class.

Let the performance of class c on the two branches be AP_c^{base} and AP_c^{imp}, respectively. Then the class performance gain is:

$$\Delta AP_c = AP_c^{\text{imp}} - AP_c^{\text{base}}. \tag{8}$$

According to a minimum effective gain threshold δ, the class gate is defined as:

$$g_c = \begin{cases} 1, & \Delta AP_c \geq \delta, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

When $g_c = 1$, it indicates that the enhancement branch can bring stable gains on class c, and the model selects the enhancement branch; otherwise, the base branch result is retained. For instance i, the final prediction can be expressed as:

$$\dot{P}_i = g_{c_i} P_i^{\text{imp}} + (1 - g_{c_i}) P_i^{\text{base}}, \tag{10}$$

where c_i is the instance class and P_i may include the instance mask, class label, and confidence.

From the residual perspective, the change produced by the enhancement branch relative to the base branch is:

$$R_i = P_i^{\text{imp}} - P_i^{\text{base}}, \tag{11}$$

so the final prediction can also be written as:

$$\dot{P}_i = P_i^{\text{base}} + g_{c_i} R_i. \tag{12}$$

This form shows that class gating does not force all classes to accept the same semantic enhancement, but activates residual correction only when there is a positive class gain; for two branches with consistent candidate masks, the module can directly select the confidence or complete prediction, while for inconsistent candidate sets, instances are first matched by 3D mask IoU and then class-level gating is applied to matched instances, with unmatched ones retaining the original branch results. To avoid test-set information leakage, the class gating table is computed only from an independent calibration set and remains fixed during testing, so that CAR-GF transfers the class-specific benefits of the semantic field and mask refinement modules to the final prediction without retraining the backbone network, while avoiding uniform and uncontrollable performance degradation on difficult classes.

3.4 Final Prediction

For each 3D instance candidate, this paper first obtains the instance-level class distribution through semantic field query, then updates the mask range according to superpoint topology and vision-language similarity, and finally uses the class-aware gating mechanism to select the result of the base branch or the enhancement branch. Finally, each instance is represented by a triplet $\dot{Y}_i = (M_i, c_i, s_i)$, where M_i is the 3D instance mask, c_i is the class label, and s_i is the final confidence. All instance-class candidates are globally ranked by confidence, and the final Open-Vocabulary 3D instance segmentation results are output through a foreground Top-K strategy.

Table 1: Open-Vocabulary 3D instance segmentation results of different methods on the ScanNet200 validation set. Mask3D serves as a closed-set supervised reference, while other methods are evaluated under the Open-Vocabulary setting.

Method	3D proposals from 2D masks	SAM for 3D mask labeling	mAP	mAP50	mAP25	head	common	tail
Mask3D	×	×	26.9	36.2	41.4	39.8	21.7	17.9
OVIR-3D	√	×	13.0	24.9	32.3	14.4	12.7	11.7
OpenMask3D	√	×	15.4	19.9	23.1	17.1	14.1	14.9
Open3DIS	√	√	23.7	29.4	32.8	27.8	21.1	21.8
OpenYOLO3D	×	×	24.7	31.7	36.2	27.8	24.3	21.6
Ours	×	×	**26.1**	**33.4**	**39.6**	**29.9**	**24.1**	**24.1**

4 Experiments

4.1 Experimental Settings

Datasets and evaluation metrics. Experiments are carried out on two benchmarks, ScanNet200 (Rozenberszki et al., 2022) and Replica (Straub et al., 2019). As in previous studies (Takmaz et al., 2023; Nguyen et al., 2024; Boudjoghra et al., 2025; Jung et al., 2025), evaluation is performed on the validation split. Category lists defined by each dataset are adopted as the input prompts. For quantitative comparison, we follow ScanNet (Dai et al., 2017) to design the evaluation metrics.

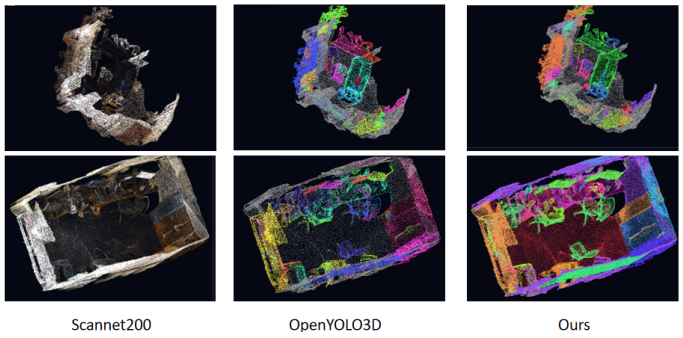

Figure 4: Qualitative comparison of scenes ScanNet200. Compared with OpenYOLO3D, our method obtains more complete instance regions and clearer boundaries between adjacent instances, demonstrating the complementary roles of instance-level semantic enhancement and local topological constraints.

Implementation details. Following prior OV-3DIS methods (Takmaz et al., 2023; Boudjoghra et al., 2025), we use pretrained Mask3D (Schult et al., 2023) to generate class-agnostic point-based masks. We also avoid using DBSCAN (Ester et al., 1996) to prevent inference speed degradation. All experiments are conducted on a single NVIDIA 4090 GPU.

4.2 Results Analysis

Open-Vocabulary 3D instance segmentation on ScanNet200. We summarize the quantitative evaluation on ScanNet200 in Table 1. Both OpenYOLO3D and our method are significantly faster than other approaches that rely on SAM (Kirillov et al., 2023) for 2D segmentation and CLIP (Radford et al., 2021) for prompt label assignment. Compared with OpenYOLO3D, our method improves the mAP of the tail category, demonstrating the improvement of our model in rare-object recognition. In addition, the gains on head categories and overall performance confirm the impact of sparse observations or boundary contamination on Open-vocabulary 3D segmentation. This highlights a promising direction for future work: improving instance mask quality while preserving computational efficiency.

Table 2: Open-Vocabulary 3D instance segmentation results of different methods on the Replica dataset.

Method	3D proposals from 2D masks	SAM for 3D mask labeling	mAP	mAP50	mAP25	Time/scene(s)
OVIR-3D	×	×	10.5	12.8	27.5	52.8
OpenMask3D	√	×	15.4	19.9	23.1	547.5
Open3DIS	√	√	23.7	29.4	32.8	189.5
OpenYOLO3D	×	×	23.7	28.7	34.8	16.6
Box-Guided	×	×	24.0	31.8	37.4	43.7
Ours	×	×	**25.1**	**32.4**	**39.6**	29.9

Cross-dataset generalization. The Replica experiment directly uses the 3D candidate network pretrained on the ScanNet200 training set, without retraining for Replica. Under this cross-dataset setting, our method still achieves the best overall performance, indicating that instance-based reliability semantic enhancement is not only effective on a single dataset but also maintains good generalization ability when the scene distribution changes.

As shown in Table 3, without 3D semantic fusion, mAP, mAP50, and mAP25 are 25.00%, 33.11%, and 37.29%, respectively. As w increases, the three metrics generally rise steadily, reaching 25.59%, 33.93%, and 38.18% at $w = 0.40$, improving by 0.60, 0.81, and 0.89 percentage points over the baseline, respectively. The mAP of head, common, and tail categories increases from 29.4%, 22.2%, and 23.2% to 29.9%, 22.4%, and 24.1%, respectively, with the tail category gaining the most. The

Table 3: Parameter analysis of the semantic fusion weight w in the reliability-aware 3D semantic field.

Semantic fusion weight w	mAP	mAP50	mAP25	Head AP	Common AP	Tail AP
0 (baseline)	24.995	33.112	37.289	29.4	22.2	23.2
0.02	25.023	33.149	37.332	29.4	22.2	23.3
0.05	25.229	33.241	37.420	29.5	22.2	23.4
0.10	25.418	33.601	37.601	29.5	22.3	23.8
0.15	25.279	33.501	37.699	29.6	22.3	23.8
0.20	25.353	33.614	37.856	29.8	22.3	23.9
0.25	25.457	33.716	37.964	29.9	22.3	24.1
0.30	25.503	33.814	38.081	29.9	22.4	24.1
0.40	25.590	33.926	38.182	29.9	22.4	24.1

Table 4: Ablation study on the ScanNet200 dataset.

	mAP	mAP50	mAP25
R1	0.247	31.7	36.2
R2	0.250	33.0	37.8
R3	25.9	33.5	38.7
R4	26.3	33.4	39.6

actual average fusion strength in the experiment is only 0.185, lower than the set base weight of 0.40, indicating that semantic coverage and class margin can perform instance-adaptive suppression of the enhancement magnitude, reducing interference of low-reliability semantics on the original prediction.

From Table 4, it can be seen that introducing superpoint mask refinement on the base model can correct local missing regions; further adding the reliability-aware semantic field improves overall accuracy metrics. The complete model retains effective enhancement and rolls back damaged classes through class-aware gating, finally achieving 26.3% mAP and 39.6% mAP25. This result indicates that the three modules play roles at the levels of instance semantics, local geometry, and class decision, respectively, and form a complementary relationship in the complete framework.

5 Conclusion

This paper addresses the core contradiction in Open-Vocabulary 3D instance segmentation that semantic information is rich but of varying quality, and proposes the reliability-aware semantic field framework OV-RASF. Rather than indiscriminately projecting 2D semantics into 3D space, we treat candidate instances as units of semantic organization and decision-making, so that vision-language information, local geometric structure, and class-level priors jointly govern instance representation, mask boundaries, and prediction selection. This design avoids assuming all semantic evidence is equally trustworthy, and instead controls the scope and strength of semantic enhancement while retaining an efficient 2D detection path. Experiments on ScanNet200 and Replica show consistent gains, especially on common and tail categories, validating the effectiveness of instance-level reliability modeling and branch complementary utilization. The framework requires no backbone retraining and offers good interpretability and transferability.

Limitations and future work. The current method is still affected by the coverage quality of initial 3D candidates, the completeness of the precomputed semantic field, and the representativeness of the calibration set; future work can further study online semantic field construction, cross-dataset reliability calibration, and learnable dynamic gating to improve generalization ability in complex open scenarios.

AI USE STATEMENT

Generative AI tools were employed solely for language refinement and LaTeX formatting support. All AI-assisted outputs were carefully examined by the authors, who remain fully responsible for the final manuscript, including its text, claims, and artifacts.

ETHICS STATEMENT

This work uses publicly available datasets (ScanNet200 and Replica) and does not involve human subjects. The proposed method is intended for 3D scene understanding and may benefit applications such as service robotics and autonomous agents. We do not foresee any direct negative societal impacts from this work.

REPRODUCIBILITY STATEMENT

We provide implementation details in Section 4.1, including datasets, evaluation metrics, and model configurations. The method uses pretrained Mask3D and YOLO-World without additional training. For reproducibility, we describe the semantic field construction, superpoint-constrained mask refinement, and class-aware residual gated fusion in Sections 3.1–3.3. The class gating table is computed on an independent calibration set and fixed during testing. We will release the code and configuration files as supplementary materials upon acceptance.

REFERENCES

Mohamed El Amine Boudjoghra, Angela Dai, Jean Lahoud, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Open-yolo 3d: Towards fast and accurate open-vocabulary 3d instance segmentation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16901–16911, 2024.

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

Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7010–7019, 2023.

Herbert Edelsbrunner and John Harer. Persistent homology—a survey. *Contemporary Mathematics*, 453:257–282, 2008.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 226–231. AAAI Press, 1996.

Louis Geist, Loic Landrieu, and Damien Robert. Ez-sp: Fast and lightweight superpoint-based 3d segmentation. *arXiv preprint arXiv:2512.00385*, 2025.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision (ECCV)*, pp. 540–557. Springer, 2022.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, and 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, Yokohama, 2024. IEEE. doi: 10.1109/ICRA57147.2024.10610243.

Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. 3d concept learning and reasoning from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9202–9212, 2023.

Tung-Wei Hui, Chen Change Loy, and Xiaoou Tang. Superpoint network for point cloud over-segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5510–5519, 2021.

Ayush Jain, Pushkal Katara, Nikolaos Gkanatsios, Adam W. Harley, Gabriel Sarch, Kriti Aggarwal, Vishal Chaudhary, and Katerina Fragkiadaki. Odin: A single model for 2d and 3d segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4867–4876, 2020.

Sanghun Jung, Jing Zheng, Kai Zhang, Nan Qiao, Anthony Y. C. Chen, Lu Xia, Cheng Liu, Yizhou Sun, Xin Zeng, Heng-Wen Huang, Byron Boots, Min Sun, and Cheng-Hao Kuo. Details matter for indoor open-vocabulary 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9627–9637, 2025.

Prannay Kaul, Weidi Xie, and Andrew Zisserman. Multi-modal classifiers for open-vocabulary object detection. In *International Conference on Machine Learning*, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, and et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4558–4567, 2018.

Feng Liang, Bichen Wu, Xiaoliang Dai, and et al. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7061–7070, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, and et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, and et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 38–55, 2024b.

Shiyang Lu, Haonan Chang, Eric P. Jing, and et al. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In *Proceedings of the 7th Conference on Robot Learning*, pp. 1610–1620. PMLR, 2023.

Khoi Nguyen, Dasith de Silva Edirimuni, Ghulam Mubashar Hassan, and Ajmal Mian. Retrieving objects from 3d scenes with box-guided open-vocabulary instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open vocabulary 3d instance segmentation with 2d mask guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4018–4028, 2024.

Phuc Nguyen, Minh Luu, Anh Tran, Cuong Pham, and Khoi Nguyen. Any3dis: Class-agnostic 3d instance segmentation by 2d mask tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 3636–3645, 2025.

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, and et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 815–824, 2023.

Cuong Pham, Truong Vu, and Khoi Nguyen. Lp-ovod: Open-vocabulary object detection by linear probing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 779–788, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Damien Robert, Hugo Raguet, and Loic Landrieu. Efficient 3d semantic segmentation with superpoint transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 125–141, Cham, 2022. Springer.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

Julian Straub, Thomas Whelan, Lingni Ma, and et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *Advances in Neural Information Processing Systems*, 36:68367–68390, 2023.

Vibashan VS, Ning Yu, Chen Xing, Can Qin, Mingfei Gao, Juan Carlos Niebles, Vishal M. Patel, and Ran Xu. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23539–23549, 2023.

Thang Vu, Kookhoi Kim, Tung M. Luu, Xuan Thanh Nguyen, and Chang D. Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Luting Wang, Yi Liu, Penghao Du, Zihan Ding, Yue Liao, Qiang Qi, Bing Chen, and Songtao Liu. Object-aware distillation pyramid for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11186–11196, 2023.

Yanmin Wu, Jiarui Meng, Haoran Li, Chenming Wu, Yujun Shi, Xinhua Cheng, Chunhui Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, and Juyong Zhang. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21469–21480, 2025.

Lewei Yao, Jianhua Han, Xiaodan Liang, and et al. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23497–23506, 2023.

Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, pp. 106–122. Springer, 2022.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Linjie H. Li, Lu Zhou, Xiyang Dai, Lu Yuan, Yin Li, and et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16793–16803, 2022.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, and et al. Detecting twenty-thousand classes using image-level supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 350–368, 2022.