Open-Vocabulary 3D Object Detection with Co-Distillation Discovery and Dual Guidance Robust Training

Shangbo Yuan[1,‡] Jie Xu[2,*] Xiaofeng Zhu[3] Na Zhao[2]

[1] School of Computer Science and Engineering,
University of Electronic Science and Technology of China, Chengdu, China
[2] Information Systems Technology and Design Pillar,
Singapore University of Technology and Design, Singapore
[3] School of Computer Science and Technology, Hainan University, Haikou, China

Abstract. Recently, open-vocabulary 3D object detection (3D-OVD) has gained increasing attention for its ability to detect unseen objects in 3D scenes. Existing approaches typically adopt a two-stage pipeline that first discovers novel objects using foundation models and then trains a 3D-OVD model based on these discovered objects. Although effective, this pipeline often suffers from inaccurate localization and mismatched classification during the discovery stage, which subsequently limits the performance of the model training stage. To address these limitations, we advocate for improving both the reliability of novel object discovery and the robustness of model training, and propose an innovative framework. Specifically, for reliable discovery, our co-distillation strategy distills high-quality novel objects by applying Hungarian matching over a comprehensive score that incorporates geometric consistency, structural objectness, and semantic certainty. To enhance robust model training, we further propose a dual-guidance learning scheme, incorporating a scene-awareness-guided uncertainty regularization for the regression head and an LLM-guided hierarchical alignment for the classification head, effectively mitigating the negative effects of imprecise 3D bounding boxes and semantic ambiguity. Extensive experiments on SUN RGB-D and ScanNetV2 demonstrate that our method achieves significant performance gains over state-of-the-art approaches. Code is available at https://github.com/shangboyuan/Co-3DGT.

Keywords: Robotic perception · Open-vocabulary · 3D object detection · Co-distillation · Robust training

1 Introduction

Open-vocabulary object detection (OVD) [12, 45, 47] aims to localize and recognize novel objects beyond the training categories in unseen scenes by employing

‡This work was carried out during Shangbo's visit to the IMPL Lab at SUTD.
*Corresponding author.

the unbounded vocabulary of vision-language foundation models [20, 21]. Recently, OVD has achieved remarkable progress in 2D object detection tasks, largely driven by the foundation models (*e.g.*, CLIP [28]) pre-trained on massive 2D image-text datasets [6, 11, 13, 22]. However, extending these advances to open-vocabulary 3D object detection (3D-OVD) [8, 10, 16] remains challenging due to the intrinsic gap between 2D images and 3D point clouds [44], as well as the limited availability of large-scale annotated 3D data, which hinders the training of 3D-specific foundation models.

To achieve 3D-OVD, existing methods typically follow a two-stage pipeline: they first leverage 2D foundation models to discover novel objects in 3D scenes, and then train a 3D detector using both ground-truth base objects and discovered novel objects to transfer localization and classification knowledge to the 3D domain. Depending on how novel object discovery stage is conducted, these methods can be categorized into two main types: *2D-Detection-based* and *3D-Proposal-based*. As illustrated in Figure 1a, 2D-Detection-based methods [24, 39] perform open-vocabulary object detection on 2D images associated with a 3D scene. Specifically, a pre-trained OV 2D detector is utilized to identify target objects (*e.g.*, a 'desk' alongside its confidence score) within the 2D frame, leveraging global 2D contexts for visual recognition. Subsequently, these methods back-project the accurately detected 2D bounding boxes (bboxes) into the 3D space to form and localize the identified 3D objects. In contrast, 3D-Proposal-based methods [3, 4], shown in Figure 1b, employ a class-agnostic 3D detector to generate 3D bboxes for all object proposals, which are projected onto the image plane to obtain corresponding 2D regions of interest. Vision-language models (*e.g.*, CLIP [28]) are then used to assign open-vocabulary class labels to the image crops of these detected 3D objects. During the training stage, both the 3D bboxes and semantic labels of the discovered novel objects are employed to train a 3D-OVD model, where 3D bboxes are regressed for localization and semantic label features from CLIP are aligned for classification.

Despite the effectiveness of the two-stage pipeline, its performance is still limited by noise in the discovered novel objects, which negatively impacts the subsequent training stage that relies on them as supervision. While both 2D-Detection-based and 3D-Proposal-based methods inevitably introduce noise during the discovery stage, they suffer from two different major types of errors. Specifically, the 2D-Detection-based strategy often yields inaccurate 3D bboxes because the noisy back-projection process encapsulates erroneous depth measurements and extraneous points from adjacent objects into the resulting 3D point set. Thus, calculating a 3D bbox from such a noisy 3D point set leads to imprecise localization. For example, in Figure 1a, the back-projected bbox of the desk is largely imprecise because its underlying point set includes noisy and unassociated points. On the other hand, 3D-Proposal-based approaches may assign incorrect semantic labels to 3D bboxes due to occlution, as shown in Figure 1b, the desk is partially occluded by the chair, causing the cropped image of the discovered proposal to be incorrectly classified as a chair by the CLIP model. Consequently, the discovery process directly bottlenecks the subsequent

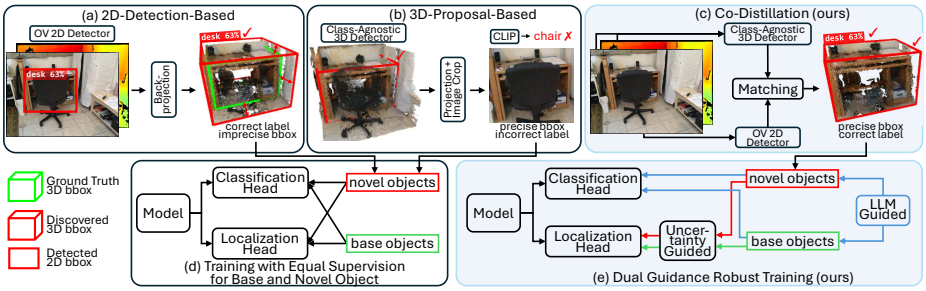

Fig. 1: Comparison between our Co-3DGT and two existing 3D-OVD paradigms. In the *Novel Object Discovery* stage: (a) 2D-Detection-based strategy [24, 39] back-projects 2D-OVD results into 3D space, often leading to imprecise 3D bounding boxes; (b) 3D-Proposal-based methods [3,4] project 3D proposals from class-agnostic 3D detectors onto 2D plane and crop images for open-vocabulary labeling, which may introduce semantic ambiguity; (c) We propose Co-Distillation, jointly considering consider geometric constraints alongside 2D semantic and 3D localization inofrmation for more reliable novel objects. In the *Training* stage: (d) Previous methods typically employ equal supervision for both base and novel objects; while (e) We propose Dual Guidance Robust Training, integrating uncertainty regularization for the regression head and leveraging an LLM to achieve hierarchical semantic alignment.

training stage, which relies on the bbox parameters and semantic labels from discovered novel objects for model optimization.

Moreover, the inherent errors stemming from both 2D and 3D detectors remain largely unavoidable in practical open-vocabulary scenarios. Specifically, 2D OVDs can also introduce semantic misclassifications (i.e., classification noise) [34], while class-agnostic 3D detectors inevitably yield imprecise spatial localization (i.e., bounding box noise) [7]. However, during the subsequent training stage, existing methods typically treat the high-quality ground-truth annotations of base objects and the noisy pseudo-labels of discovered novel objects with *equal supervision* as presented in Figure 1d. Without accounting for the severe quality disparity between these two sources of annotation, localization accuracy of the model is impaired by the imprecise bboxes, and semantic errors from the 2D models are inadvertently propagated into the trained model. This indiscriminate training strategy consequently degrades the feature representations for both base and novel categories, leading to suboptimal detection performance.

To address the above challenges, we propose open-vocabulary 3D object detection with Co-Distillation Discovery and Dual Guidance Training (Co-3DGT), a unified framework that enhances both the *reliability of novel 3D object discovery* and the *robustness of model training*. As illustrated in Figure 1c, our *co-distillation discovery* module jointly leverages a 2D open-vocabulary detector and a class-agnostic 3D detector to discover novel objects from both spatial and semantic perspectives in parallel. Specifically, we measure the spatial IoU scores between semantic 2D proposals and class-agnostic 3D proposals in a scene, and then combine the 2D class confidence and 3D objectness scores to jointly solve

a scene-level matching problem. This objective expects to discover the spatial consistency among 3D boxes with semantic constraints, to co-distill reliable 3D novel objects.

To mitigate the impact of inevitable noise in discovered objects, we introduce the *dual-guidance training strategy* as shown in Figure 1e, which comprises (i) scene-awareness-guided uncertainty regularization and (ii) large language model (LLM)-guided hierarchical alignment. Concretely, (i) we exploit the scene-aware uncertainty from base classes to regularize the uncertainties for novel classes, thus dynamically weighting the 3D bbox regression loss. This prevents the model from becoming overconfident in noise 3D bboxes, thereby ensuring stable and robust localization performance. Meanwhile, (ii) we propose an LLM-guided hierarchical labeling to enrich label supervision instead of directly using given category labels. It derives two levels of super-category labels from the original category names via an LLM to capture shared semantic and geometric properties among related classes. Then, hierarchical semantic alignment aligns 3D object features with CLIP embeddings across multiple semantic levels, making the semantic classification less sensitive to noise labels. This increases more discriminative supervision from super-categories to the original supervision, and thus helps the model to achieve robust classification.

In summary, our work makes four contributions as follows:

- We identify the performance bottleneck in the widely adopted two-stage 3D-OVD pipeline and propose a unified framework that jointly reduces noise in novel object discovery and enhances training robustness.
- We propose a Co-Distillation Discovery strategy that enforces 3D-2D consistency via a joint optimization formulation that leverages spatial and semantic scores to selectively distill reliable novel 3D objects.
- We introduce a Dual Guidance Robust Training scheme consisting of scene-awareness-guided uncertainty regularization for robust uncertainty-weighted regression, and LLM-guided hierarchical alignment for improved classification via label-augmented semantic feature alignment.
- Extensive experiments on SUN RGB-D and ScanNetV2 demonstrate that our method achieves state-of-the-art performance, with accuracy improvements of 4.71% and 9.82% AP_{25} in novel-class object detection, respectively.

2 Related Work

3D Object Detection focuses on identifying and localizing objects in 3D space from point clouds or RGB-D data [14, 31, 40, 48, 49]. VoteNet [26] introduce an end-to-end point-based framework using Hough voting to generate anchor-free object proposals. Specifically, it employs PointNet++ [27] to extract point cloud features, where seed points vote for object centers. 3DETR [25] adopts a Transformer architecture with self-attention for global context modeling, effectively eliminating the need for hierarchical aggregation and heuristic voting schemes. Instead, it uses learnable queries and a Hungarian matching loss for direct, end-to-end detection. More recently, Cubify-Anything [19] introduced the

Cubify Transformer (CuTR) model alongside the Cubify-Anything 1M (CA-1M) dataset, further advancing indoor 3D object detection by leveraging Vision Transformer (ViT)-based [9] architectures.

Open-Vocabulary Object Detection seeks to generalize object detections beyond their seen training categories to recognize and localize previously unseen object classes by leveraging the powerful representational capabilities of vision-language pretraining models [1, 15, 36, 37]. The seminal work of CLIP [28] establishes the foundational paradigm for this research field by learning semantically aligned vision-language representations through contrastive learning on web-scale image-text pairs. Based on this, Detic [50] integrate zero-shot recognition into detectors like Faster R-CNN [29] by replacing the fixed-vocabulary classification heads with CLIP-based classifiers that can dynamically accommodate arbitrary text object categories during inference without requiring additional fine-tuning. Grounding DINO [23] achieves tighter vision-language integration by unifying detection and phrase grounding via language-guided query mechanisms, enabling the model to learn robust associations between textual descriptions and visual regions without requiring expensive box-level supervision annotations during the pretraining phase.

Open-Vocabulary 3D Object Detection has emerged to overcome the scalability limitations of traditional closed-set detectors, which rely on exhaustive and costly 3D annotations [2, 35, 41, 51, 52]. OV-3DET [24] pioneers 3D-OVD without requiring 3D annotations by leveraging pretrained 2D detectors for semantic transfer. CoDA [3] introduces a collaborative framework that integrates novel object discovery with cross-modal alignment, which jointly exploits 3D geometry to discover potential objects and 2D semantics for their identification, enhancing the model's ability to distinguish unknown from known instances. Addressing the challenge of semantic ambiguity, INHA [17] advances this by seeding 3D novel object proposals from 2D OVD results and introducing a multi-level hierarchical alignment across instance, category, and scene contexts to reduce semantic inconsistency. More recently, OV-Uni3DETR [39] extends open-vocabulary detection to multimodal and outdoor environments, unifying representations across point clouds and RGB images within a transformer-based architecture, demonstrating robust performance and flexibility for various modality combinations.

3 Method

Problem Definition. A 3D scene is represented by point cloud $\mathcal{P} = \{\mathbf{P} \in \mathbb{R}^{N\times 3}\}$ and corresponding RGB-D images ($I^{\text{rgb}} \in \mathbb{R}^{3\times H\times W}, I^{\text{depth}} \in \mathbb{R}^{1\times H\times W}$) with known camera intrinsics $\mathbf{K} \in \mathbb{R}^{3\times 3}$, rotation and translation extrinsics ($\mathbf{R} \in \mathbb{R}^{3\times 3}, \mathbf{t} \in \mathbb{R}^3$). Each 3D object in the scene is parameterized as $(\mathbf{b}, c)$, where $\mathbf{b}$ denotes the bbox defined as $\mathbf{b} = (x, y, z, l, w, h, \theta)$. Here, (x, y, z) is the center coordinate, (l, w, h) are the dimensions, and θ is the orientation around the vertical axis. The category label is $c \in \mathcal{C}$, where $\mathcal{C}$ is the set of object categories.

Unlike closed-set 3D object detection, open-vocabulary 3D object detection is only provided with base objects $\mathcal{O}^{\text{base}} = \{\mathbf{b}_i, c_i \mid c_i \in \mathcal{C}^{\text{base}}\}$, while the goal is to

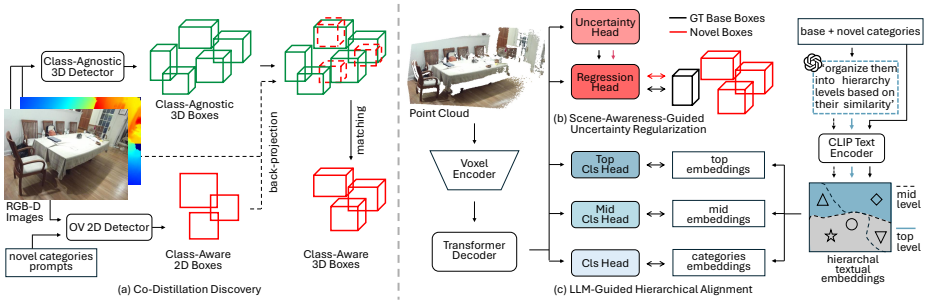

Fig. 2: Framework Overview of our Co-3DGT. In *Co-Distillation Discovery* stage: (a) a class-agnostic 3D detector generates 3D boxes, while a 2D open-vocabulary detector produces class-aware 2D boxes which are back-projected and to co-distill reliable class-aware 3D boxes. In *Dual Guidance Robust Training* stage: (b) *Scene-Awareness-Guided Uncertainty Regularization* exploits the scene-aware uncertainty from base classes to regularize the uncertainties for novel classes, dynamically weighting the 3D bbox regression. (c) *LLM-Guided Hierarchical Alignment* leverage LLM to form super-categories that increase structural semantic supervision beyond original categories, promoting the model to achieve robust classification.

detect and recognize objects from the combined category set $\mathcal{O} = \mathcal{O}^{\text{base}} \cup \mathcal{O}^{\text{novel}}$. To achieve this, a discovery stage is employed to identify potential novel objects $\mathcal{O}^{\text{novel}} = \{\mathbf{b}_j, c_j \mid c_j \in \mathcal{C}^{\text{novel}}\}$ within the scene. These discovered instances serve as additional pseudo-labels that expand the category coverage beyond $\mathcal{C}^{\text{base}}$. In the training stage, both the annotated base objects $\mathcal{O}^{\text{base}}$ and the discovered novel objects $\mathcal{O}^{\text{novel}}$ are jointly utilized as supervision. This enables the model to localize and recognize both base and novel objects.

Framework Overview. As illustrated in Figure 2, we propose Co-3DGT to enhance the reliability of novel object discovery and the robustness of model training. Our framework consists of two stages. In the novel object discovery stage, we propose a scene-level matching problem based co-distillation discovery strategy (Sec. 3.1) that can jointly distill high-quality novel objects. In the model training stage, we propose a dual guidance robust training scheme, where a scene-awareness-guided uncertainty regularization (Sec. 3.2) and an LLM-guided hierarchical alignment (Sec. 3.3) are designed to mitigate negative effects of the unavoidable imprecise 3D bboxes and semantic ambiguity, respectively.

3.1 Co-Distillation Discovery

We observe that the 2D-Detection-based discovery exhibits stronger performance in capturing semantically accurate object categories but suffers from producing noise 3D bboxes, while the 3D-Proposal-based discovery is conducive to predicting precise 3D bboxes but is prone to yielding noise semantic labels. Motivated by this, we propose a *co-distillation discovery* approach to avoid the drawbacks of previous two pipelines, and to fully exploit the respective advantages of 3D

localization and 2D semantic recognition. That is, as shown in Figure 2a, we first obtain semantically labeled 2D bounding boxes by leveraging the rich categorical knowledge of open-vocabulary 2D detectors, alongside class-agnostic 3D object proposals generated in parallel. We then introduce a bipartite matching problem, to jointly distill high-quality novel object annotations from both 2D semantic and 3D geometric perspectives.

Specifically, to lift 2D detections to 3D space, we follow the back-projection operation used in previous methods [24,39]: $P_k = (KR_t)^{-1} \cdot (d_k \cdot \mathbf{x}_k)$. The pixel set $\{\mathbf{x}_k\} = \{[u_k, v_k, 1]^\top\}$ within a 2D OVD result is back-projected to obtain the 3D point set $\{P_k\}$ using the camera intrinsic matrix K, camera extrinsics R_t, and the corresponding depth value $\{d_k\}$. The resulting 3D bbox is then defined as the minimal cuboid enclosing the 3D point set $\{P_k\}$.

Following this, let $\mathcal{O}^{\text{2D}} = \{(\mathbf{b}_i^{\text{2D}}, c_i^{\text{2D}}, s_i^{\text{2D}})\}_{i=1}^{M}$ denote the set of back-projected 3D objects derived from the 2D OVD results across all views. Here, $\mathbf{b}_i^{\text{2D}}$ represents the i-th 3D bbox lifted from the 2D OVD result, while c_i^{2D} and s_i^{2D} indicate its predicted semantic category label and associated confidence score, respectively. Similarly, let $\mathcal{O}^{\text{3D}} = \{(\mathbf{b}_j^{\text{3D}}, s_j^{\text{3D}})\}_{j=1}^{N}$ represent the set of class-agnostic 3D proposals by the class-agnostic detector, where for the j-th proposal, $\mathbf{b}_j^{\text{3D}}$ denotes the 3D bbox and s_j^{3D} is the foreground probability.

To achieve co-distillation discovery, we formulate the association between M 2D bounding boxes and N 3D proposals as a bipartite matching problem. To solve this matching problem via the Hungarian algorithm, we define the cost matrix $\mathbf{C} \in \mathbb{R}^{M \times N}$ as:

$$C_{ij} = -\text{IoU}_{ij} - 0.1(s_j^{\text{3D}} + s_i^{\text{2D}}). \tag{1}$$

Here, IoU_{ij} denotes the Intersection over Union (IoU) between the i-th 3D bbox lifted from the 2D OVD result and the j-th class-agnostic 3D proposal. To filter out geometrically unfeasible pairs, we establish a minimal IoU threshold $\tau = 0.1$. Specifically, if $\text{IoU}_{ij} < \tau$, we prevent their assignment by setting the corresponding cost $C_{ij} = \infty$.

For candidates that satisfy this minimal IoU, their exact assignment is driven by minimizing the cost function. The IoU_{ij} term serves as the primary geometric alignment metric, penalizing size mismatches and location misalignments, thereby encouraging tight spatial correspondence between cross-modal candidates. The 3D foreground probability s_j^{3D} and the 2D confidence score s_i^{2D} quantify the spatial objectness of the 3D proposal and the semantic reliability of the 2D prediction, respectively. To preserve IoU_{ij} as the primary selection criterion, we apply a weight coefficient of 0.1 to both confidence scores. Together, this composite cost drives the assignment of cross-modal pairs that are simultaneously geometrically consistent and semantically reliable, facilitating high-quality transfer of categorical knowledge from the 2D domain to the 3D representation.

By minimizing the total cost via the Hungarian algorithm, we obtain the optimal assignment set $\mathcal{H}$ of the matched pairs (i, j). Consequently, our co-distillation strategy discovers and extracts more reliable novel objects to form

the final distilled set as follows:

$$\mathcal{O}^{\text{novel}}_{\text{Co-D}} = \left\{ \left(\mathbf{b}^{\text{3D}}_j, c^{\text{2D}}_i \right) \mid (i,j) \in \mathcal{H} \right\}. \tag{2}$$

Here, $\mathbf{b}^{\text{3D}}_j$ denotes the bbox parameters of the selected 3D proposal, and c^{2D}_i represents the open-vocabulary semantic category transferred from its paired 2D detection. By pairing high-quality 3D geometry with holistically inferred 2D semantics, this formulation effectively constructs a more reliable pseudo-label set for novel classes, facilitating the subsequent learning process.

Given the base and discovered novel objects $\mathcal{O} = \mathcal{O}^{\text{base}} \cup \mathcal{O}^{\text{novel}}_{\text{Co-D}}$, we train a 3D detector to transfer the geometric localization and semantic classification knowledge to the 3D-OVD model. Considering that both the class-agnostic 3D detector and the 2D open-vocabulary detector might introduce inevitable error, we propose the *Dual Guidance Robust Training* scheme to achieve the robust 3D-OVD model with imperfect localization and classification supervision. It consists of *Scene-Awareness-Guided Uncertainty Regularization* and *LLM-Guided Hierarchical Alignment* that will be introduced in following sections.

3.2 Scene-Awareness-Guided Uncertainty Regularization

Inspired by the uncertainty estimation in deep learning [18, 30], we employ the uncertainty-weighted regression loss to train the model as shown in Figure 2b. Specifically, the 3D bbox supervision for base objects (with manual annotation) and novel objects (from discovery stage) are derived from different processes, and they usually exhibit different noise levels. To this end, we treat base and novel objects as two distinct supervision sources during training and design specialized regression loss functions for each. For base objects, the bboxes $\mathbf{b}_i \in \mathcal{O}^{\text{base}}$ are manually annotated, we adopt the following regression loss [30]:

$$\mathcal{L}^{\text{base}}_{\text{Reg}} = \sum_i \left\lfloor \frac{\sigma_i}{\sqrt{2}} \right\rfloor^{\frac{1}{2}} \left(\frac{\sqrt{2}}{\sigma_i}, \left| \hat{\mathbf{b}}_i - \mathbf{b}_i \right| + \log \sigma_i \right), \tag{3}$$

where $\hat{\mathbf{b}}_i$ is the i-th prediction of the 3D bbox of the model for the object $\mathbf{b}_i \in \mathcal{O}^{\text{base}}$, and σ_i is the estimated uncertainty for the i-th object obtained by neural networks [18]. $\lfloor \cdot \rfloor$ denotes the stop-gradient operation.

In this formulation, the regression loss is driven by three key components. We first employ two complementary weighting terms, where the uncertainty weighting term $\sqrt{2}/\sigma_i$ enables the model to focus on reliable bboxes with low σ_i, and the adaptive re-weighting term $\lfloor \sigma_i/\sqrt{2} \rfloor^{\frac{1}{2}}$ prevents the model from ignoring difficult samples. Next, $\log \sigma_i$ acts as a regularization term, which explicitly penalizes the network for predicting an arbitrarily large uncertainty σ_i to reduce loss. Finally, the interplay among these three components effectively maintains balanced supervision across varying uncertainty levels for base boxes.

For novel objects, the bboxes $\mathbf{b}_j \in \mathcal{O}^{\text{novel}}_{\text{Co-D}}$ are generated by the class-agnostic 3D detector, and consequently contain additional localization errors. The uncertainty term $\log \sigma_i$ in Eq. (3) is learned primarily from precise manual annotations of base objects. As a result, the learned uncertainty estimates may

not adequately capture the elevated noise level of novel bboxes. To address this issue, we improve Eq. (3) and propose the scene-awareness-guided uncertainty regularization as follows:

$$\mathcal{L}_{\text{Reg}}^{\text{novel}} = \sum_j \left(\frac{\sqrt{2}}{\sigma_j} \left| \hat{\mathbf{b}}_j - \mathbf{b}_j \right| \left| \log \sigma_j - \mathbb{E}\left[\lfloor \log \sigma_i \rfloor \right] \right| \right), \tag{4}$$

where $\mathbb{E}\left[\lfloor \log \sigma_i \rfloor\right]$ represents the scene-aware inherent uncertainty predicted by the model, derived from the expectation of the stop-gradient log-uncertainty of base bboxes. To handle scenes without base objects, we maintain a moving average $\mathbb{E}'$ term updated using the expectations from base-containing scenes via a standard momentum update (decay rate of 0.9) during training.

Here, the term $|\log \sigma_j - \mathbb{E}[\lfloor \log \sigma_i \rfloor]|$ serves as a soft regularization that guides the uncertainty of novel objects toward the scene-aware uncertainty. The motivation of our regression loss is that we hope the bbox uncertainty of novel objects remains at the same level as that of base objects $\mathbb{E}\left[\lfloor \log \sigma_{\text{base}} \rfloor\right]$, which can be supported by the classical theory of maximum mean discrepancy (MMD). This prevents both underestimation (when $\log \sigma_j \leq \mathbb{E}\left[\lfloor \log \sigma_i \rfloor\right]$) and unbounded overestimation (when $\log \sigma_j > \mathbb{E}\left[\lfloor \log \sigma_i \rfloor\right]$).Notably, we omit the adaptive re-weighting term $\left\lfloor \sigma/\sqrt{2} \right\rfloor^{\frac{1}{2}}$ from Eq. (3) since applying this term to novel bboxes with higher noise levels could amplify erroneous signals.

Overall, the 3D bbox regression loss with our scene-awareness-guided uncertainty regularization is:

$$\mathcal{L}_{\text{Reg}} = \mathcal{L}_{\text{Reg}}^{\text{base}} + \mathcal{L}_{\text{Reg}}^{\text{novel}}. \tag{5}$$

This scene-awareness-guided uncertainty regularization leverages scene-aware uncertainty estimates from base objects as an anchor to robustly guide the uncertainty prediction for novel objects within the same scene.

3.3 LLM-Guided Hierarchical Alignment

To achieve the robustness against inaccurate semantic information form the given category labels, we first propose an LLM-guided hierarchical labeling to enrich label supervision, and then conduct the hierarchical semantic alignment to achieve the reliable classification as shown in Figure 2c.

LLM-Guided Hierarchical Labeling. To mitigate the possible noise supervision from the given category labels, we employ an LLM to infer functional, geometric, and conceptual relationships among base and novel categories, to obtain the label-augmented supervision. Specifically, we define three hierarchies for each category label, *i.e.*, top, mid, bottom. The bottom level contains the original fine-grained category labels $\mathcal{C}^{\text{bot}} = \mathcal{C}^{\text{base}} \cup \mathcal{C}^{\text{novel}}$, while the top and mid levels $\mathcal{C}^{\text{mid}}$, $\mathcal{C}^{\text{top}}$ are super-categories inferred by the LLM. For example, hierarchical

labels from $\mathcal{C}^{\text{bot}}$ is constructed as:

$$\begin{cases} \mathcal{C}^{\text{top}} : \text{‘furniture’, ‘appliance’, ‘accessory’, ...} \\ \mathcal{C}^{\text{mid}} : \text{‘seating’, ‘storage’, ‘domestic’, ‘tools’, ...} \\ \mathcal{C}^{\text{bot}} : \text{‘chair’, ‘table’, ‘pillow’, ‘door’, ‘fridge’, ...} \end{cases} \tag{6}$$

In our method, the mid level corresponds to functionally and geometrically similar super-categories, and the top level represents abstract ones. The top and mid levels are less sensitive to semantic ambiguity from visually similar objects, which provides structural information for subsequent semantic alignment.

Hierarchical Semantic Alignment. For the j-th semantic label, we use CLIP to encode its corresponding hierarchical labels c_j^{top}, c_j^{mid}, c_j^{bot} to obtain the textual embeddings $\mathbf{h}_j^{(1)}$, $\mathbf{h}_j^{(2)}$, $\mathbf{h}_j^{(3)}$, respectively. These embeddings are subsequently aligned with the 3D-OVD model which also produce hierarchical semantic features $\mathbf{f}_i^{(1)}$, $\mathbf{f}_i^{(2)}$, $\mathbf{f}_i^{(3)}$.

To obtain the predicted probability that the i-th sample belongs to the j-th category at level l, we first compute the cosine similarity score and then map it to the interval (0,1) using a sigmoid function:

$$p_{ij}^{(l)} = \text{sigmoid}\left(\frac{\langle \mathbf{f}_i^{(l)}, \mathbf{h}_j^{(l)} \rangle}{\|\mathbf{f}_i^{(l)}\| \|\mathbf{h}_j^{(l)}\|}\right), \tag{7}$$

and then adopt binary cross-entropy (BCE) loss:

$$\mathcal{L}_{\text{bce}}^{(l)} = -\frac{1}{N} \sum_{j}^{|\mathcal{C}^{(l)}|} \sum_{i}^{N} \left[y_{ij}^{(l)} \cdot \log(p_{ij}^{(l)}) + (1 - y_{ij}^{(l)}) \cdot \log(1 - p_{ij}^{(l)}) \right] \tag{8}$$

where $y_{ij}^{(l)}$ denotes the indicator variable, e.g., $y_{ij}^{(l)} = 1$ if the i-th object belongs to the j-th category at the l-th level, otherwise $y_{ij}^{(l)} = 0$.

The overall hierarchical alignment loss is formulated as:

$$\mathcal{L}_{\text{Align}} = \mathcal{L}_{\text{bce}}^{(1)} + \mathcal{L}_{\text{bce}}^{(2)} + \mathcal{L}_{\text{bce}}^{(3)}. \tag{9}$$

In the hierarchical semantic alignment, the bottom level preserves fine-grained category distinctions and supervises the learning of detailed object features, while the mid and top levels mitigate the impact of noise labels by improving robustness against visually similar or ambiguous objects.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our proposed method on two challenging large-scale indoor 3D object detection datasets: SUN RGB-D [33] and ScanNetV2 [5]. The

SUN RGB-D dataset contains about 10,335 RGB-D scenes across diverse indoor environments. Following common practice, we use 5,285 training samples covering 46 object categories, where the 10 most frequent categories are treated as base classes and the remaining 36 as novel classes. The ScanNetV2 dataset [5] consists of more than 1500 indoor scenes with 200 object categories. From the 1,201 training scenes in ScanNetV2, we select the 10 most frequent classes as base classes and the next 50 as novel classes. In addition, following OV-3DET [24], we adopt a setting where no ground-truth annotations are provided, and evaluate the top 20 most frequent unseen classes on ScanNetV2.

Baselines. We compare our method with a series of 3D-OVD methods that take point clouds as input. These baselines include Det-PointCLIP [43], Det-PointCLIPv2 [42], Det-CLIP^2 [46], OV-3DET [24], CoDA [3], INHA [17], CoDAv2 [4], and OV-Uni3DETR [39].

Evaluation Metrics. We adopt the mean Average Precision (mAP) and mean Average Recall (mAR) at an IoU threshold of 0.25 for evaluation. We report the mAP for unseen classes, seen classes, and all classes, denoted as $\mathrm{AP}_{25}^{\mathrm{novel}}$, $\mathrm{AP}_{25}^{\mathrm{base}}$, and $\mathrm{AP}_{25}^{\mathrm{mean}}$, respectively. Similarly, the corresponding mAR are denoted as $\mathrm{AR}_{25}^{\mathrm{novel}}$, $\mathrm{AR}_{25}^{\mathrm{base}}$, and $\mathrm{AR}_{25}^{\mathrm{mean}}$. In addition, under the same setting with no ground-truth annotations following OV-3DET, we adopt the AP_{25} as the metric for the top 20 most frequent unseen classes.

Implementation Details. We adopt the backbone of Uni3DETR [38], including its voxel encoder and transformer decoder, as our 3D detection model. In line with common approaches [3, 17, 24, 39, 43, 46], we utilize CLIP [28] to transform category labels into textual features. Following previous work [17, 24, 39], we leverage Detic as the 2D open-vocabulary detector. Additionally, we employ CuTR [19] as the 3D class-agnostic detector and ChatGPT-5 [32] to infer category relationships and generate hierarchical labels. In terms of computational cost, our full pipeline requires approximately 47 minutes and 169 minutes for novel object discovery on SUN RGB-D and ScanNetV2, respectively, using a single RTX 3090 GPU. The LLM-guided hierarchical alignment introduces negligible time cost, requiring only a single API call for the entire dataset (approximately $20 \sim 30$ seconds) to generate the two-level hierarchical labels. Although the discovery stage is moderately slower than OV-Uni3DETR (25 minutes and 103 minutes with single RTX 3090 GPU on SUN RGB-D and ScanNetV2, respectively), our approach substantially reduces the overall training time. Using four RTX 3090 GPUs, the training stage takes 490 and 309 minutes on SUN RGB-D and ScanNetV2, respectively (compared to 716 and 621 minutes for OV-Uni3DETR), since our model operates only on point cloud inputs, in contrast to OV-Uni3DETR which integrates both point cloud and image modalities.

4.2 Main Results

As shown in Table 1, our method achieves the highest overall performance on both SUN RGB-D and ScanNetV2 benchmarks. Notably, our model significantly improves detection of novel categories, boosting $\mathrm{AP}_{25}^{\mathrm{novel}}$ to 14.37% on SUN RGB-D and 21.91% on ScanNetV2, which corresponds to considerable gains of $+4.71\%$

Table 1: Comparison of AP_{25} (%) and AR_{25} (%) for open-vocabulary 3D object detection on SUN RGB-D and ScanNetV2 datasets.

Method	SUN RGB-D						ScanNetV2					
	AP_{25}^{novel}	AP_{25}^{base}	AP_{25}^{mean}	AR_{25}^{novel}	AR_{25}^{base}	AR_{25}^{mean}	AP_{25}^{novel}	AP_{25}^{base}	AP_{25}^{mean}	AR_{25}^{novel}	AR_{25}^{base}	AR_{25}^{mean}
Det-PointCLIP [43]	0.09	5.04	1.17	21.98	65.03	31.33	0.13	2.38	0.50	33.38	54.88	36.96
Det-PointCLIPv2 [42]	0.12	4.82	1.14	21.33	63.74	30.55	0.13	1.75	0.40	32.60	54.52	36.25
Det-CLIP^2 [46]	0.88	22.74	5.63	22.21	65.04	31.52	0.14	1.76	0.40	34.26	56.22	37.92
3D-CLIP [28]	3.61	30.56	9.47	21.47	63.74	30.66	3.74	14.14	5.47	32.15	54.15	35.81
CoDA [3]	6.71	38.72	13.66	33.66	66.42	40.78	6.54	21.57	9.04	43.36	61.00	46.30
INHA [17]	8.91	42.17	16.18	51.34	78.65	57.23	7.79	25.10	10.68	55.10	71.60	57.85
CoDAv2 [4]	9.17	42.04	16.31	43.16	71.64	49.35	9.12	23.35	11.49	64.00	72.16	65.36
OV-Uni3DETR [39]	9.66	48.29	18.06	–	–	–	12.09	30.47	15.15	–	–	–
Co-3DGT (ours)	**14.37**	**49.85**	**22.63**	**57.70**	**79.36**	**63.08**	**21.91**	**32.83**	**23.75**	**66.30**	**74.21**	**69.33**

Table 2: Detailed comparison of per-category AP_{25} (%) for open-vocabulary 3D object detection on ScanNetV2 dataset across 20 unseen categories.

Method	Mean	toilet	bed	chair	sofa	dresser	table	cabinet	bookshelf	pillow	sink
OV-3DET [24]	18.02	57.29	42.26	27.06	31.50	8.21	14.17	2.98	5.56	23.00	31.60
CoDA [3]	19.32	68.09	44.04	28.72	44.57	3.41	20.23	5.32	0.03	27.95	45.26
CoDAv2 [4]	22.72	77.24	43.96	15.05	53.27	11.37	13.96	1.42	0.11	34.42	44.38
OV-Uni3DETR [39]	25.33	86.05	50.49	28.11	31.51	18.22	24.03	6.58	12.17	29.62	54.63
Co-3DGT (ours)	**36.15**	92.75	70.42	81.27	60.64	12.50	39.15	15.92	12.56	37.58	60.59

Method	bathtub	refrigerator	desk	nightstand	counter	door	curtain	box	lamp	bag
OV-3DET [24]	56.28	10.99	19.72	0.77	0.31	9.59	10.53	3.78	2.11	2.71
CoDA [3]	50.51	6.55	12.42	15.15	0.68	7.95	0.01	2.94	0.51	2.02
CoDAv2 [4]	55.60	24.41	20.67	20.72	0.28	13.54	0.92	4.16	4.37	9.20
OV-Uni3DETR [39]	63.73	14.41	30.47	2.94	1.00	19.02	19.90	12.70	5.58	13.46
Co-3DGT (ours)	69.52	34.57	24.23	13.15	7.01	15.99	16.86	15.99	38.61	4.35

and +9.82% over the previous state-of-the-art approach (OV-Uni3DETR). In comparison, the improvements on base categories are more modest, with increases of +1.56% on SUN RGB-D and +2.36% on ScanNetV2. Overall, our model achieves a AP_{25}^{mean} of 22.63% on SUN RGB-D and 23.75% on ScanNetV2, surpassing OV-Uni3DETR by +4.57% and +8.60%, respectively. These results demonstrate that our approach effectively enhances robustness to unseen categories, while maintaining strong performance on base categories.

Additionally, in the same annotation-free setting following OV-3DET [24], no ground-truth annotated base objects are available. As a result, our scene-awareness-guided uncertainty regularization becomes inapplicable. In this scenario, our model is trained using a simplified variant that retains only the LLM-guided hierarchical alignment with the novel objects discovered by Co-Distillation. Our method still achieves 36.15% AP_{25}^{mean} as shown in Table 2, improving upon OV-Uni3DETR by +10.82%. This improvement confirms the reliability of our Co-Distillation discovery and the robustness of the LLM-guided hierarchical alignment.

Moreover, Figure 3 presents qualitative examples of 3D-OVD results by our method in various indoor scenes on SUN RGB-D and ScanNetV2. Each subfigure illustrates detected 3D objects within reconstructed indoor scenes, indicating our method's effectiveness to localize and recognize both base and novel objects.

Table 3: Ablation study of the Co-Distillation module on SUN RGB-D and ScanNetV2. Models are evaluated using $AP_{25}^{novel,disc}$ (%) and $AR_{25}^{novel,disc}$ (%) to measure pseudo-label quality on the training samples, and $AP_{25}^{novel,det}$ (%) and $AR_{25}^{novel,det}$ (%) to assess final detection performance.

Method	Novel Object Discovery				Training Result			
	SUN RGB-D		ScanNetV2		SUN RGB-D		ScanNetV2	
	$AP_{25}^{novel,disc}$	$AR_{25}^{novel,disc}$	$AP_{25}^{novel,disc}$	$AR_{25}^{novel,disc}$	$AP_{25}^{novel,det}$	$AR_{25}^{novel,det}$	$AP_{25}^{novel,det}$	$AR_{25}^{novel,det}$
2D-Detection-based	6.21	16.10	3.93	23.44	9.59	52.99	11.92	57.85
3D-Proposal-based	8.49	12.02	4.57	20.28	9.88	52.15	12.71	58.38
Co-distil(scratch)	10.45(+1.96)	19.72(+7.70)	9.35(+4.76)	28.46(+8.18)	11.76(+1.88)	54.26(+2.11)	16.68(+3.97)	63.70(+5.32)
Co-distil(CuTR)	10.97(+2.48)	24.33(+12.31)	10.46(+5.89)	32.29(+12.01)	12.69(+2.81)	55.41(+3.26)	18.74(+6.03)	64.19(+5.81)

Table 4: Ablation study for dual guidance robust model training. AP_{25}^{mean} (%) on SUN RGB-D and ScanNetV2 are reported.

Variants	SUN RGB-D	ScanNetV2
Training w/o dual guidance	20.62	20.15
SGUR ✓	22.17 (+1.55)	23.12 (+2.97)
LGHA ✓	21.35 (+0.73)	22.73 (+2.58)
SGUR ✓ + LGHA ✓	22.63 (+2.01)	23.75 (+3.60)

Table 5: Ablation study for LLM selection in LGHA. AP_{25}^{mean} (%) on SUN RGB-D and ScanNetV2 are reported.

Models	SUN RGB-D	ScanNetV2
ChatGPT-5	22.63	23.75
Gemini-3-Flash-Thinking	21.89	23.84
Qwen3.5-397B-A17B	22.71	23.34
Llama-3.3-70B-Instruct	21.42	23.51

4.3 Ablation Study

In Tables 3, 4 and 5, we conduct ablation studies to verify the effectiveness of each component in our framework.

Effectiveness of Co-Distillation Discovery. We first evaluate the performance of different discovery pipelines on the training set, focusing on the instance-level precision and recall for novel categories (left side of Table 3). Compared to the 2D-Detection-based and 3D-Proposal-based pipelines, our Co-Distillation discovery pipeline yields notable improvements. Specifically, compared to the 3D-Proposal-based pipeline, Co-distil(scratch) improves precision by +0.97% and recall by +3.86% on SUN RGB-D, alongside gains of +4.97% and +8.18% on ScanNetV2. Incorporating the pretrained class-agnostic 3D detector, CuTR, further enhances discovery capability. Co-distil(CuTR) improves precision by +2.99% and recall by +10.72% on SUN RGB-D, over the 3D-Proposal-based baseline, and achieves +5.89% and +12.01% respective gains on ScanNetV2.

Subsequently, we analyze how the discovery results from Co-Distillation with CuTR translate to the final detection performance for novel classes (right side of Table 3), where our proposed dual-guidance training strategies are not applied. Better quality in novel object discovery directly contributes to higher final training results. Relying solely on the baseline training method without our robust training methods, Co-distil(CuTR) achieves 12.69% AP_{25}^{novel} and 55.41% AR_{25}^{novel} on SUN RGB-D, as well as 18.74% AP_{25}^{novel} and 64.19% AR_{25}^{novel} on ScanNetV2. These consistent improvements across both the training set and the final detection results verify that our Co-Distillation effectively mines high-quality novel objects to boost model performance.

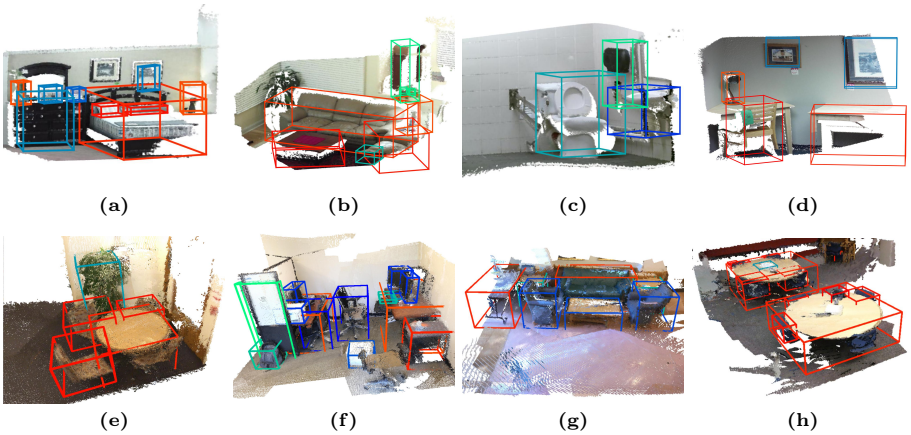

Fig. 3: Qualitative examples on SUN RGB-D (the first row) and ScanNetV2 (the second row) to visualize the 3D-OVD results by our method. The red and orange colored bboxes correspond to different base-class objects, while blue and green colored bboxes represent different novel-class objects.

Effectiveness of Dual Guidance Robust Training. Based on the results of Co-Distil(CuTR) discovery, we further ablate our proposed dual guidance robust training strategies in Table 4, including scene-awareness-guided uncertainty regularization (SGUR) and LLM-guided hierarchical alignment (LGHA). Here, we focus on the performance improvements across all classes (AP_{25}^{mean}). On SUN RGB-D, applying SGUR alone improves the performance by +1.55% (achieving 22.17% AP_{25}^{mean}), while LGHA alone provides a +0.73% gain (reaching 21.35% AP_{25}^{mean}). When combined, they bring a total improvement of +2.01%, yielding a final result of 22.63% AP_{25}^{mean}. We observe a similar trend on ScanNetV2, where SGUR and LGHA contribute gains of +2.97% and +2.58%, respectively. Their combination achieves 23.75% AP_{25}^{mean}, surpassing the baseline by +3.60%. These results demonstrate the complementary roles of SGUR and LGHA in mitigating noise and enhancing overall training robustness.

Influence of Different LLMs. In Table 5, we investigate the impact of utilizing different LLMs to generate the hierarchical class labels required for LGHA. Our results indicate that changing the underlying LLM yields similar performances of AP_{25}^{mean} with minimal fluctuations. This demonstrates that our LGHA strategy is robust to the choice of the LLM, as the alignment relies primarily on the rich structural and semantic hierarchy of the generated class labels rather than the specific capabilities or phrasing of a particular LLM.

5 Conclusion

In this work, we propose an innovative framework for 3D-OVD, focusing on reliable novel object discovery and robust model training. For reliable discovery,

we introduce a Hungarian matching-based co-distillation strategy to generate high-quality novel objects by integrating 2D semantic and 3D geometric information. For robust model training, we introduce a dual guidance training method including scene-awareness-guided uncertainty regularization for robust localization and LLM-guided hierarchical alignment for mitigating semantic ambiguity in classification. Experiments show significant performance gains over state-of-the-art methods. In future work, we plan to extend the proposed framework toward real-time 3D-OVD by incorporating temporal fusion to enable efficient perception for robotic navigation in indoor environments.

Acknowledgments

This research was supported in part by the National Key Research & Development Program of China under Grant 2022YFA1004100, and in part by the Ministry of Education, Singapore, under its MOE Academic Research Fund Tier 2 (MOE-T2EP20124-0013).

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. NeurIPS **35**, 23716–23736 (2022)
2. Cao, S., Li, C., Xu, J., Li, T., Zhao, N.: Late-decoupled 3d hierarchical semantic segmentation with semantic prototype discrimination based bi-branch supervision. arXiv preprint arXiv:2511.16650 (2025)
3. Cao, Y., Yihan, Z., Xu, H., Xu, D.: Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. NeurIPS **36**, 71862–71873 (2023)
4. Cao, Y., Zeng, Y., Xu, H., Xu, D.: Collaborative novel object discovery and box-guided cross-modal alignment for open-vocabulary 3d object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(11), 10475–10489 (2025)
5. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR. pp. 5828–5839 (2017)
6. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
7. Deng, J., Lu, J., Zhang, T.: Quantity-quality enhanced self-training network for weakly supervised point cloud semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(5), 3580–3596 (2025)
8. Deng, J., He, T., Jiang, L., Wang, T., Dayoub, F., Reid, I.: 3d-llava: Towards generalist 3d lmms with omni superpoint transformer. In: CVPR. pp. 3772–3782 (2025)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)

10. Etchegaray, D., Huang, Z., Harada, T., Luo, Y.: Find n'propagate: Open-vocabulary 3d object detection in urban environments. In: ECCV. vol. 15098, pp. 133–151 (2024)
11. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2), 303–338 (2010)
12. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: ICLR (2022)
13. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: CVPR. pp. 5356–5364 (2019)
14. Han, Y., Zhao, N., Chen, W., Ma, K.T., Zhang, H.: Dual-perspective knowledge enrichment for semi-supervised 3d object detection. In: AAAI. vol. 38, pp. 2049–2057 (2024)
15. Huang, J., Zhang, J., Jiang, K., Lu, S.: Open-vocabulary object detection via language hierarchy. In: NeurIPS. vol. 37, pp. 124951–124978 (2024)
16. Jiang, L., Shi, S., Schiele, B.: Open-vocabulary 3d semantic segmentation with foundation models. In: CVPR. pp. 21284–21294 (2024)
17. Jiao, P., Zhao, N., Chen, J., Jiang, Y.G.: Unlocking textual and visual wisdom: Open-vocabulary 3d object detection enhanced by comprehensive guidance from text and image. In: ECCV. vol. 15106, pp. 376–392 (2024)
18. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? NeurIPS **30**, 5574–5584 (2017)
19. Lazarow, J., Griffiths, D., Kohavi, G., Crespo, F., Dehghan, A.: Cubify anything: Scaling indoor 3d object detection. In: CVPR. pp. 22225–22233 (2025)
20. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML. pp. 12888–12900 (2022)
21. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: CVPR. pp. 10965–10975 (2022)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014)
23. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: ECCV. vol. 15105, pp. 38–55 (2024)
24. Lu, Y., Xu, C., Wei, X., Xie, X., Tomizuka, M., Keutzer, K., Zhang, S.: Open-vocabulary point-cloud object detection without 3d annotation. In: CVPR. pp. 1190–1199 (2023)
25. Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3d object detection. In: ICCV. pp. 2906–2917 (2021)
26. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: ICCV. pp. 9277–9286 (2019)
27. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. NeurIPS **30**, 5099–5108 (2017)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149 (2016)

30. Seitzer, M., Tavakoli, A., Antic, D., Martius, G.: On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In: ICLR (2022)
31. Sheng, H., Cai, S., Zhao, N., Deng, B., Liang, Q., Zhao, M.J., Ye, J.: Ct3d++: Improving 3d object detection with keypoint-induced channel-wise transformer. International Journal of Computer Vision **133**(7), 4817–4836 (2025)
32. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
33. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: CVPR. pp. 567–576 (2015)
34. Tran, P.V.: Simltd: Simple supervised and semi-supervised long-tailed object detection. In: CVPR. pp. 4672–4681 (2025)
35. Wang, J., Zhao, N.: Uncertainty meets diversity: A comprehensive active learning framework for indoor 3d object detection. In: CVPR. pp. 20329–20339 (2025)
36. Wang, J., Chen, B., Kang, B., Li, Y., Xian, W., Chen, Y., Xu, Y.: Ov-dquo: Open-vocabulary detr with denoising text query training and open-world unknown objects supervision. In: AAAI. pp. 7762–7770 (2025)
37. Wang, X., Li, S., Kallidromitis, K., Kato, Y., Kozuka, K., Darrell, T.: Hierarchical open-vocabulary universal image segmentation. NeurIPS **36**, 21429–21453 (2023)
38. Wang, Z., Li, Y.L., Chen, X., Zhao, H., Wang, S.: Uni3detr: Unified 3d detection transformer. NeurIPS **36**, 39876–39896 (2023)
39. Wang, Z., Li, Y., Liu, T., Zhao, H., Wang, S.: Ov-uni3detr: Towards unified open-vocabulary 3d object detection via cycle-modality propagation. In: ECCV. vol. 15105, pp. 73–89 (2024)
40. Wu, Y., Wang, K., Pan, Y., Zhao, N.: Ccf: Complementary collaborative fusion for domain generalized multi-modal 3d object detection. In: CVPR. pp. 18745–18754 (2026)
41. Xu, J., Zhao, N.: Stream3D: Streaming zero-shot 3d instance segmentation with multi-view noise mask filtering and manifold refining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings. pp. 327–337 (2026)
42. Yao, L., Han, J., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, H.: Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In: CVPR. pp. 23497–23506 (2023)
43. Yao, L., Han, J., Wen, Y., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, C., Xu, H.: Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. NeurIPS **35**, 9125–9138 (2022)
44. Yuan, S., Xu, J., Hu, P., Zhu, X., Zhao, N.: Graph smoothing for enhanced local geometry learning in point cloud analysis. In: AAAI. vol. 40, p. 12250–12258 (2026)
45. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: CVPR. pp. 14393–14402 (2021)
46. Zeng, Y., Jiang, C., Mao, J., Han, J., Ye, C., Huang, Q., Yeung, D.Y., Yang, Z., Liang, X., Xu, H.: Clip2: Contrastive language-image-point pretraining from real-world point cloud data. In: CVPR. pp. 15244–15253 (2023)
47. Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Yang, J., Zhang, L.: A simple framework for open-vocabulary segmentation and detection. In: ICCV. pp. 1020–1031 (2023)
48. Zhao, N., Chua, T.S., Lee, G.H.: Sess: Self-ensembling semi-supervised 3d object detection. In: CVPR. pp. 11079–11087 (2020)
49. Zhao, N., Lee, G.H.: Static-dynamic co-teaching for class-incremental 3d object detection. In: AAAI. vol. 36, pp. 3436–3445 (2022)

50. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: ECCV. vol. 13669, pp. 350–368 (2022)
51. Zhu, X., Zhou, H., Xing, P., Zhao, L., Xu, H., Liang, J., Hauptmann, A., Liu, T., Gallagher, A.: Open-vocabulary 3d semantic segmentation with text-to-image diffusion models. In: ECCV. vol. 15087, pp. 357–375 (2024)
52. Zhu, Y., Qian, J., Yang, J., Xie, J., Zhao, N.: Few-shot incremental 3d object detection in dynamic indoor environments. In: CVPR. pp. 18786–18795 (2026)