Global-Local Collaborative Inference with LLM for Lidar-Based Open-Vocabulary Detection

Xingyu Peng[1], Yan Bai[2], Chen Gao[4], Lirong Yang[2], Fei Xia[2], Beipeng Mu[2], Xiaofei Wang[2], and Si Liu[1,3]⋆

[1] Institute of Artificial Intelligence, Beihang University
[2] Meituan
[3] Hangzhou Innovation Institute, Beihang University
[4] University of Science and Technology Beijing
{pengxyai,liusi}@buaa.edu.cn
{baiyan02,yanglirong,wangxiaofei19}@meituan.com
{gaochen.ai,xia.fei09,mubeipeng}@gmail.com

Abstract. Open-Vocabulary Detection (OVD) is the task of detecting all interesting objects in a given scene without predefined object classes. Extensive work has been done to deal with the OVD for 2D RGB images, but the exploration of 3D OVD is still limited. Intuitively, lidar point clouds provide 3D information, both object level and scene level, to generate trustful detection results. However, previous lidar-based OVD methods only focus on the usage of object-level features, ignoring the essence of scene-level information. In this paper, we propose a Global-Local Collaborative Scheme (GLIS) for the lidar-based OVD task, which contains a local branch to generate object-level detection result and a global branch to obtain scene-level global feature. With the global-local information, a Large Language Model (LLM) is applied for chain-of-thought inference, and the detection result can be refined accordingly. We further propose Reflected Pseudo Labels Generation (RPLG) to generate high-quality pseudo labels for supervision and Background-Aware Object Localization (BAOL) to select precise object proposals. Extensive experiments on ScanNetV2 and SUN RGB-D demonstrate the superiority of our methods. Code is released at https://github.com/GradiusTwinbee/GLIS.

Keywords: Open-Vocabulary Detection · 3D Object Detection · Large Language Model

1 Introduction

As a basic function of machine perception, object detection has attracted much attention within computer vision communities. The traditional training pipeline for the detection model relies on elaborately labeled data, resulting in a limited number of classes that can be collected and annotated. In this way, the detection model fails to detect objects not belonging to the training object classes. In
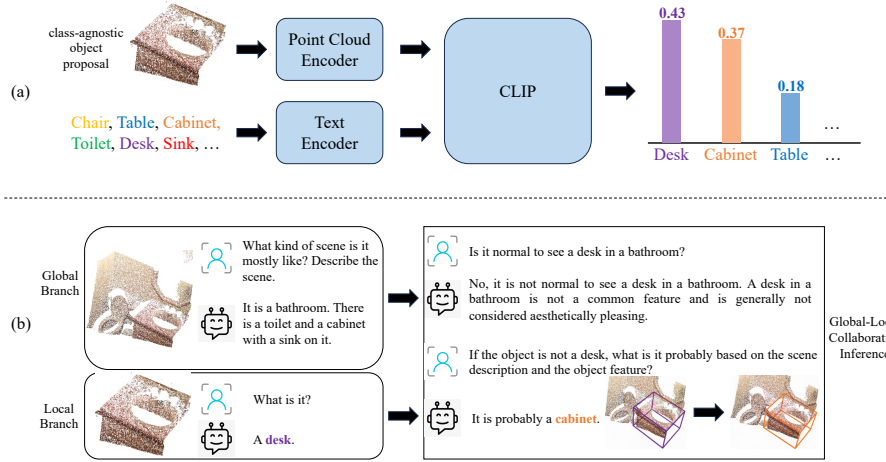

⋆ Corresponding author

Fig. 1: (a) The previous 3D OVD paradigm determines the class of an object proposal by comparing its point cloud feature with the class text features. As it only considers object-level/local information, the proposal may be wrongly recognized, *e.g.* mistaking the cabinet for a desk. (b) In contrast, we propose a Global-Local Collaborative Inference Scheme (GLIS) for 3D OVD, considering both the scene-level/global information and the object-level/local information. Additionally, we leverage the LLM to provide common sense for chain-of-thought inference, which can clarify the ambiguous object class step by step.

recent years, open-vocabulary detection has become a popular topic, which is hopeful to solve the problem. Generally, an open-vocabulary detection model requires no human-labeled data in the training stage and possesses the ability to detect any interested object in a given scene. However, although many open-vocabulary detection methods focus on 2D RGB images, the 3D point clouds, a data modality widely used in autonomous driving and robot vision, have not been thoroughly investigated. In this paper, we focus on the task of lidar-based open-vocabulary object detection, where only point clouds are utilized in the testing stage.

Compared to open-vocabulary detection for 2D RGB images, lidar-based open-vocabulary detection suffers from more difficulties. Firstly, point clouds have lower resolutions compared to 2D RGB images, leading to the loss of object details, *e.g.* material, texture, color, *etc*. Additionally, the quality of the point cloud could be easily affected by environmental conditions, thus noisy points may exist in the point cloud. Therefore, it is harder for a lidar-based detection model to recognize the class only based on the lidar information of a single object, emphasizing the necessity of environmental information in lidar-based OVD.

However, the dominant paradigm of current state-of-the-art lidar-based OVD methods only focuses on object-level features and neglects the importance of

scene-level information. As shown in Fig. 1a, the current paradigm determines the class of an object by comparing the object-level features and the text features of class names. Such a paradigm could be easily affected by the point cloud's low resolution and noises since it only considers local information.

To address this issue, we propose a Global-Local Collaborative Inference Scheme (GLIS) for the lidar-based OVD task, considering both the global scene-level information and the local object-level information. Additionally, the Large Language Model (LLM) is introduced to provide common sense for chain-of-thought inference. The common sense from LLM could help the model clarify the ambiguous object features, while chain-of-thought inference allows the model to refine the detection result step by step. For example, in the case of Fig. 1, the class-agnostic proposal, which is a cabinet in fact, is wrongly recognized as a desk. The previous paradigm (shown in Fig. 1a) does not perceive this error as it only considers local information. Contrastly, in GLIS (shown in Fig. 1b), such error could be corrected with the clues from global information. Specifically, with the scene being identified as a bathroom, the LLM recognizes that a desk is unlikely to be present, suggesting the object proposal might have been misclassified. Drawing on both local and global features, as well as insights from previous conversations, the mislabeled desk can be accurately reclassified as a cabinet. In summary, GLIS proactively employs global and local information within a chain-of-thought inference framework.

To support high-quality global-local collaborative inference, both the local branch and the global branch are elaborately designed. For the local branch, we propose Reflected Pseudo Labels Generation (RPLG) to generate high-quality pseudo labels for training. Additionally, Background-Aware Object Localization (BAOL) is proposed to generate precise object proposals. For the global branch, the model is learned to generate scene descriptions following the demonstration of MiniGPT-v2 [3].

In summary, our contributions are as follows.

- We propose a lidar-based open-vocabulary detection method, GLIS, which is the first work to explore the interactions of the global scene-level information and local object-level information in this field.
- We introduce LLM to conduct chain-of-thought inference in the lidar-based open-vocabulary detection pipeline, where the common sense extracted from LLM is utilized to refine the detection result.
- To further improve the detection performance of GLIS, we propose Reflected Pseudo Labels Generation (RPLG) and Background-Aware Object Localization (BAOL), which can alleviate the influence of noises in training and testing.

Superior performance on ScanNetV2 [7] and SUN RGB-D [38] demonstrates the effectiveness of our methods.

2   Related Work

2.1   Open Vocabulary Detection

2D Open Vocabulary Detection. Many works have been done to deal with 2D OVD [3,8,10,12,13,18,21,29,34,35,45,51,52,54,55,61,62]. Some works [55,62] utilize image-text pairs to conduct open-vocabulary detection. Other works [54, 61] try to use pre-aligned text-image embedding space (*e.g.*, CLIP [33]) to improve the detection performance. For example, RegionCLIP [61] utilizes CLIP to match captions with image regions, achieving fine-grained alignment between images and texts. Recently, with the prosperity of Generative Pre-training Transformer (GPT), some works [3, 18, 21] also try to conduct 2D OVD via image grounding. GLIP [18] proposes a scheme to unify the pre-training of object detection and image grounding. Grounding DINO [21] utilizes grounded pre-training to equip the closed-set detector DINO [58] with open-vocabulary detection abilities. Minigpt-v2 [3] builds a unified LLM for completing various vision-language tasks including object detection and image caption.

3D Open Vocabulary Detection. As point cloud plays an important role in autonomous driving and robot vision, some researchers are paying attention to 3D OVD [2,25,26,43,56,59]. OV-3DET [25,26] utilizes pre-trained 2D detectors to generate 2D pseudo labels from RGB images, which are then converted to 3D pseudo labels via projection. Besides, a debiased contrastive learning strategy is proposed in OV-3DET to bridge connections between texts, images and point clouds. Unlike OV-3DET, CoDA [2] trains a based detector with limited annotations for localization and discovers novel objects with 3D box geometry priors and 2D semantic open-vocabulary priors. FM-OV3D [56] obtains knowledge from pre-trained foundation models to improve the detection performance. OpenSight [59] explores 3D OVD in outdoor scenes.

However, all previous methods for lidar-based OVD only focus on the usage of object-level features, ignoring the value of the scene-level information.

2.2   3D Object Detection

As a basic technique for 3D world perception, lidar-based 3D object detection has achieved great progress in recent years [4,5,9,11,15,17,19,24,27,28,30–32,36, 37,39,42,46,47,49,50,53,60,63]. Some works [4,49] project 3D point clouds onto the Bird's Eye View, so they can be processed by 2D CNNs. Other works [17, 19,37,50] design networks for direct feature extraction from point clouds. For example, PointRCNN [37] proposes a two-stage detection pipeline to handle proposal generation and refining respectively. With the rise of Transformer [41], transformer-based 3D detectors [24,30] are also devised. Group-Free [24] utilizes the attention mechanism of Transformer to directly learn the contribution of each point to a single object. 3DETR [30] builds an end-to-end transformer-based detector for 3D object detection.

However, these works are focused on closed-set lidar-based 3D object detection. In contrast, our GLIS is targeted for 3D open-vocabulary object detection.

2.3   Large Language Models

Trained on numerous texts from the Internet in a self-supervised way, Large Language Models (LLMs) acquire the ability to generate human-like natural language [1, 40]. For example, GPT-3 [1] can chat with users or write essays according to instructions. To apply pre-trained LLMs in specific areas, many fine-tuning methods are proposed, *e.g.* P-Tuning [22,23] and Lora [14]. There are also works [16,44] trying to improve the inference ability of LLMs. For example, [44] utilizes the technique of Chain-of-Thought Prompting (CoT) to help LLMs conduct reasoning step by step. To introduce the advantages of LLMs from texts to other modals, Multi-Modal Large Language Models (MLLMs) [20, 57] are proposed. For example, LLaVA [20] equips LLMs with visual encoders for visual and language understanding.

In this paper, we devise a global-local collaborative inference scheme to utilize the inference ability of LLMs in lidar-based OVD.

3   Methods

3.1   Notation and Preliminaries

Generally, a training sample for lidar-based OVD is consisted of three parts: the point cloud P, the image I, and the projection matrix M. The point cloud P is a set of 3D points $\{(x_i, y_i, z_i)\}_{i=1}^{N_p}$, where N_p is the points number. The 2D RGB image $I \in \mathbb{R}^{h \times w \times 3}$ is in pair with P. Following the lidar-based OVD experiment setup [25, 26, 56, 59], images are only utilized during the training stage, while test is purely based on point clouds. The projection matrix M is used for bounding box conversion between 2D and 3D. It should be noted that OVD is an unsupervised task, and no ground-truth label is provided in the training stage.

We record a bounding box as $(x, y, z, l, w, h, \theta)$, where (x, y, z) is the center, l, w, h are length, width, height respectively, and θ is the heading angle. We also use a 3D backbone to extract local feature $f_{loc} \in \mathbb{R}^{N_q \times D_p}$ and global feature $f_{glob} \in \mathbb{R}^{1 \times D_p}$ from the point cloud, where N_q is the query number of Transformer in 3D backbone, and D_p is the 3D feature dimension.

3.2   Overview

For lidar-based open-vocabulary object detection, the object-level information and scene-level information are both useful. In this way, we propose a **G**lobal-**L**ocal Collaborative **I**nference **S**cheme (GLIS) for lidar-based open-vocabulary object detection. The whole GLIS is comprised of a training pipeline and an inference pipeline.

Fig. 2 shows the training pipeline of GLIS. Initially, the point cloud is input into the 3D backbone to extract local feature f_{loc} and global feature f_{glob}. Since the unsupervised OVD task lacks supervisory signals, both local and global branches contain model training and pseudo-label acquisition. For the
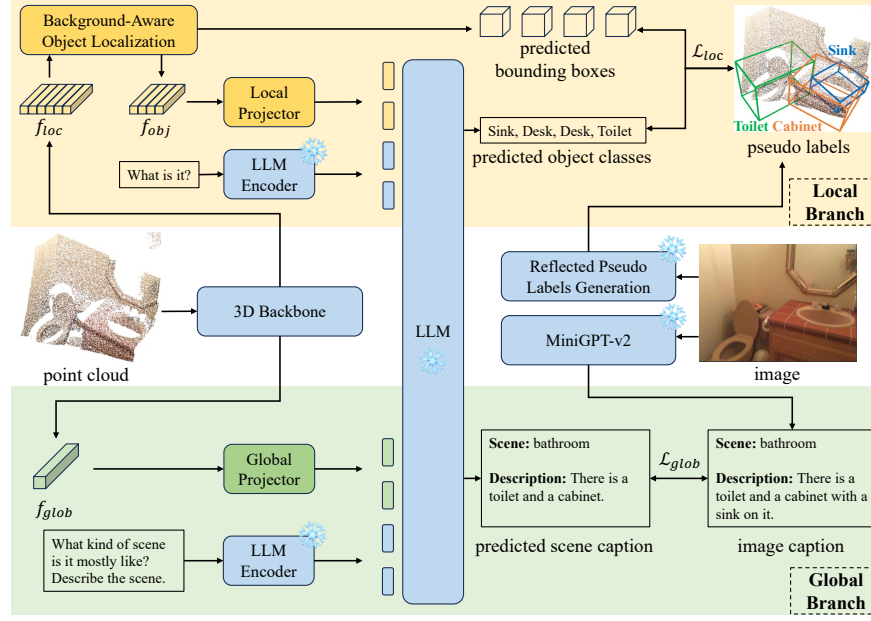
Fig. 2: The training pipeline of GLIS.

local branch, the bounding boxes are extracted by the Background-Aware Object Localization (BAOL) module, while the object classes are predicted by the LLM. Note that a Reflected Pseudo Labels Generation (RPLG) module is designed to generate pseudo labels, which are used as the supervised signals for loss computing. For the global branch, the scene caption is generated by the LLM based on the global feature. The scene caption is supervised by the image caption from MiniGPT-v2 [3]. Note that the snow symbol in Fig. 2 means that model parameters are fixed during training.

The inference pipeline of GLIS is presented in Fig. 3. Firstly, the LLM generates a description of the scene and predicts what type the scene is. Then a preliminary detection result is formed based on the local feature. Finally, chain-of-thought prompts are used to guide the LLM's inference, where the detection result is refined accordingly. The detailed process will be introduced in section Sec. 3.5.

3.3   Local Object Localization and Classification

The local branch is responsible for the generation of preliminary detection results. Firstly, the Background-Aware Object Localization (BAOL) module extracts object proposals $\{b_i, f_{obj}^i\}_{i=1}^{N_{obj}}$ from the local feature f_{loc}:

$$\{b_i, f_{obj}^i\}_{i=1}^{N_{obj}} = \text{BAOL}(f_{loc}), \tag{1}$$

where b_i is the i-th object bounding box, f_{obj}^i is the i-th object feature, N_{obj} is the object number. Then object classes $\{c_i\}_{i=1}^{N_{obj}}$ are predicted by LLM:

$$c_i = \text{LLM}(T_{loc}, \text{Local-Projector}(f_{obj}^i)), \tag{2}$$

where c_i is the predicted class of i-th object and T_{loc} is the LLM prompt. In practice, we use *"What is it?"* as T_{loc}. The Local Projector is a linear layer, which aligns object features $\{f_{obj}^i\}_{i=1}^{N_{obj}}$ to the LLM embedding space. The training of BAOL and Local Projector is supervised by the pseudo labels from the Reflected Pseudo Labels Generation (RPLG) module. In the following paragraph, we will introduce the proposed RPLG and BAOL in detail.

Reflected Pseudo Labels Generation (RPLG). As an unsupervised task, OVD lacks off-the-shelf labels for model training. To deal with this problem, previous works [25, 26, 56, 59] firstly obtain 2D pseudo labels from the image by 2D open-vocabulary detectors. Then these 2D pseudo labels are converted to 3D pseudo labels via the 2D-to-3D projection matrix. However, due to the limited detection ability, 2D open-vocabulary detectors may generate false labels, which may further confuse the OVD training. To alleviate this issue, we propose a reflected pseudo labels generation scheme to reduce the noise in pseudo labels. Specifically, Detectron2 [45] is used to generate 2D labels from images and CLIP [33] is further adopted to play the role of reflection, *i.e.*, checks the correctness of these labels.

Specifically, initial 2D labels $\{\bar{b}_{2D}^i, \bar{c}_i\}_{i=1}^{\bar{N}}$ are generated by Detectron2, where $\bar{b}_{2D}^i$ is the i-th 2D bounding box, $\bar{c}_i$ is the i-th object class, $\bar{N}$ is the number of 2D labels. Then corresponding patches $\{p_i\}_{i=1}^{\bar{N}}$ are cropped from image I according to $\{\bar{b}_{2D}^i\}_{i=1}^{\bar{N}}$. To apply CLIP for label checking, we set two prompt templates:

$$T^+(class): \textit{"This is a \{class\}.",}$$
$$T^-(class): \textit{"This is not a \{class\}.".}$$

These templates, together with patches $\{p_i\}_{i=1}^{\bar{N}}$ and classes $\{\bar{c}_i\}_{i=1}^{\bar{N}}$, are sent into CLIP to calculate confidence scores:

$$[\phi_i^+, \phi_i^-] = \text{Softmax}(\text{CLIP}(T^+(\bar{c}_i), p_i), \text{CLIP}(T^-(\bar{c}_i), p_i)), \tag{3}$$

where ϕ_i^+ is the confidence score that p_i belongs to class c_i and vice versa. We keep labels with ϕ_i^+ higher than a predefined threshold ϕ_{CLIP}, creating new 2D labels $\{\tilde{b}_{2D}^i, \tilde{c}_i\}_{i=1}^{\tilde{N}}$. These 2D labels are converted to 3D pseudo labels $\{\tilde{b}_i, \tilde{c}_i\}_{i=1}^{\tilde{N}}$ via projection matrix M, where $\tilde{N}$ is the number of 3D pseudo labels.

Background-Aware Object Localization (BAOL). Due to noises within the point cloud, the detector may confuse foreground objects with the background and outputs false object proposals. To select high-quality object proposals, we propose background-aware object localization . Basically, class-agnostic object proposals $\{\hat{b}_i, \hat{o}_i\}_{i=1}^{N_q}$ are extracted from f_{loc} via a series of prediction

heads, where $\hat{b}_i$ is the i-th bounding box, and $\hat{o}_i$ is the confidence of the i-th proposal. Then proposals whose confidence is below a threshold ϕ_{obj} are removed, forming the class-agnostic detection result $\{b_i, o_i\}_{i=1}^{N_{obj}}$, where N_{obj} is the number of objects.

In closed-vocabulary detection, the prediction of confidence can be learned from groundtrurh labels. However, in OVD, we have to design the supervision signals manually. Preliminarily, we conduct bipartite matching between the proposals $\{\hat{b}_i\}_{i=1}^{N_q}$ and the pseudo labels $\{\tilde{b}_i\}_{i=1}^{\tilde{N}}$ based on IoU. Then all proposals are labeled via:

$$y_i = \begin{cases} 1 & \exists j, \hat{b}_i \text{ is matched with } \tilde{b}_j, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $y_i = 1$ means the i-th proposal is positive, *i.e.*, is a foreground object, and vice versa. However, such label assignment can be inaccurate in two situations: (i) The matched proposal has a low IoU with the matched label, which means the proposal is not accurate enough; (ii) Two different proposals refer to the same object, yet only one of the proposal could be labeled as foreground object, which may cause confusion. To tackle these situations, we further modify the label rule. For situation (i), if a proposal has a IoU below ϕ_{low} with the matched pseudo bounding box, then it will be labeled as negative. For situation (ii), if the IoU of the two proposals are higher than ϕ_{high}, then both of them will be labeled as positive. For simplicity, we still use $\{y_i\}_{i=1}^{N_q}$ to denote the refined labels.

3.4   Global Scene Understanding and Description

The global branch predicts scene type s (*e.g.*, bedroom, kitchen, *etc.*) and generates scene description d based on global feature f_{glob}. Specifically, we prompt LLM with the following text:

T_{glob}: *"What kind of scene is it mostly like? Describe the scene."*.

Besides, we use the global projector to align global feature f_{glob} to the LLM embedding space. The whole process could be represented as

$$s, d = \text{LLM}(T_{glob}, \text{Global-Projector}(f_{glob})). \tag{5}$$

For supervision, we utilize the 2D-based MiniGPT-v2 [3] to generate scene type label $\tilde{s}$ and scene description label $\tilde{d}$ from the paired image, which are used for loss computation with the LLM answer.

3.5   Global-Local Collaborative Inference with LLM

In this section, we will introduce the inference pipeline of our proposed GLIS, which is presented in Fig. 3. Firstly, as shown in Fig. 3a, global feature f_{glob} and local feature f_{loc} are extracted from 3D backbone. Then class-agnostic object proposals, as well as their confidence scores and 3D features f_{obj}, are generated by the BAOL module.
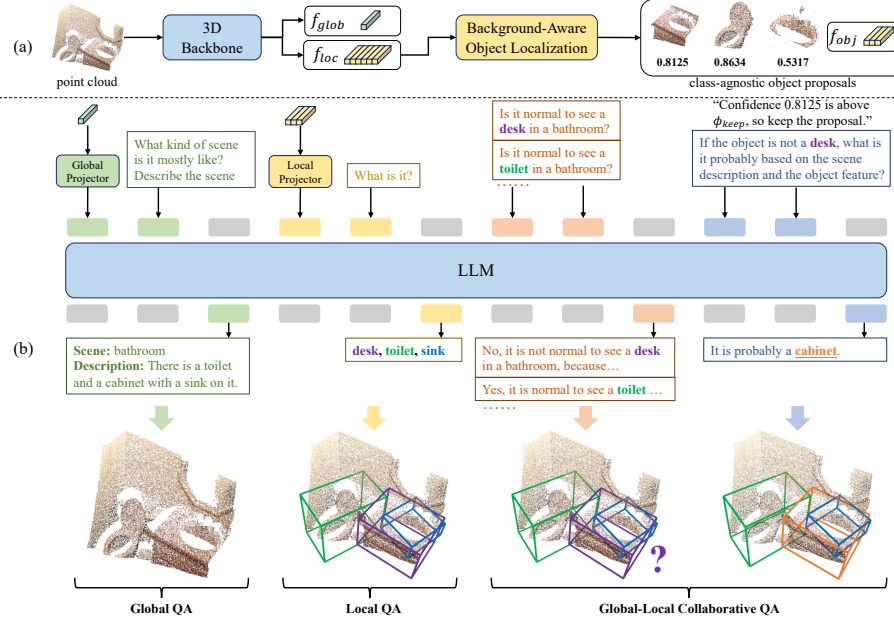
Fig. 3: The inference pipeline of GLIS.

With the extracted local and global information, we can conduct global-local collaborative inference (GLCI) with LLM and refine the detection result accordingly. Generally, as shown in Fig. 3b, chain-of-thought prompts are used to guide the LLM's inference. The whole inference process could be divided into three stages, *i.e.*, global QA, local QA, and global-local collaborative QA. In global QA, the LLM is asked to predict the scene type and describe the scene based on the global feature f_{glob}. In local QA, f_{obj} is utilized to predicted the classes of the class-agnostic object proposals.

After finishing global QA and local QA, a preliminary detection result is given. This detection result will be refined step by step in global-local collaborative QA. Specifically, for each predicted class, the following prompt template is used to check its rationality: *"Is it normal to see a {class} in a {scene type} ?"*. If a class c is thought as reasonable to exist in the predicted scene type, then all predicted objects belonging to class c are reserved, *e.g.*, toilet in Fig. 3b. Contrastly, those objects that are not likely to be in the scene will be further inspected. Specifically, for each unreasonable object, if its confidence is below a predefined threshold $\phi_{keep} = 0.75$, it will be automatically removed, and vice versa. For example, in Fig. 3b, the object proposal of desk is kept as its confidence is beyond ϕ_{keep}. To obtain the correct class of this proposal, we further prompt LLM with following template: *"If the object is not a desk, what is it probably based on the scene description and the object feature?"*. Then the LLM

utilizes the information from both local and global to give the answer. Finally, the object class is corrected according to the LLM's answer.

3.6   Training Objectives

The training loss contains four parts, *i.e.*, bounding box regression loss $\mathcal{L}_{bbox}$, confidence prediction loss $\mathcal{L}_{conf}$, object classifiction loss $\mathcal{L}_{cls}$, and scene understanding loss $\mathcal{L}_{scene}$. We use the regression loss function in 3DETR [30] to compute $\mathcal{L}_{bbox}$, which could be represented as:

$$\mathcal{L}_{bbox} = \text{RegressionLoss}(\{b_i\}_{i=1}^{N_{obj}}, \{\tilde{b}_i\}_{i=1}^{\tilde{N}}). \tag{6}$$

The confidence prediction loss is defined as:

$$\mathcal{L}_{conf} = -\frac{1}{N_q} \sum_{i=1}^{N_q} [y_i \log \hat{o}_i + \lambda_{conf}(1 - y_i) \log(1 - \hat{o}_i)], \tag{7}$$

where λ_{conf} is a balanced factor. As object classes are predicted by LLM, we compute $\mathcal{L}_{cls}$ by maximizing the possibility of label text tokens [48]. Specifically, assuming the label text is a sequence of tokens $t = (w_1, w_2, \cdots, w_l)$ and the predicted possibility for each token is $p(t) = [p(w_1), p(w_2), \cdots, p(w_l)]$, the text loss is defined as:

$$\mathcal{L}_{text}(p(t)) = -\sum_{i=1}^{l} \log p(w_i). \tag{8}$$

In this way, $\mathcal{L}_{cls}$ could be computed as:

$$\mathcal{L}_{cls} = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \mathcal{L}_{text}(p_{loc}(\tilde{c}_i)), \tag{9}$$

where p_{loc} is the LLM predicted token possibility in the local branch. Similarly, $\mathcal{L}_{scene}$ could be computed via:

$$\mathcal{L}_{scene} = \mathcal{L}_{text}(p_{glob}(\tilde{s})) + \mathcal{L}_{text}(p_{glob}(\tilde{d})), \tag{10}$$

where p_{glob} is the LLM predicted token possibility in the global branch.

In summary, the total loss is computed as

$$\mathcal{L} = \lambda_1 \mathcal{L}_{bbox} + \lambda_2 \mathcal{L}_{conf} + \lambda_3 \mathcal{L}_{cls} + \lambda_4 \mathcal{L}_{scene}, \tag{11}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are balanced factors.

4   Experiments

4.1   Datasets and Metrics

We conduct experiments on two datasets: ScanNetV2 [7] and SUN RGB-D [38].

Table 1: Comparisons with other methods on ScanNetV2

Method	mAP_{25}^{10cls}	toilet	bed	chair	sofa	dresser	table	cabinet	bookshelf	pillow	sink
OV-3DETIC [25]	12.65	48.99	2.63	7.27	18.64	2.77	14.34	2.35	4.54	3.93	21.08
FM-OV3D [56]	21.53	62.32	41.97	22.24	31.80	1.89	10.73	1.38	0.11	12.26	30.62
OV-3DET [26]	24.36	57.29	42.26	27.06	31.50	8.21	14.17	2.98	5.56	23.00	31.60
CoDA [2]	28.76	68.09	44.04	28.72	44.57	3.41	20.23	5.32	0.03	27.95	45.26
GLIS (ours)	**30.94**	73.90	39.69	39.51	44.41	6.09	25.38	5.92	8.31	25.63	43.51

Method	mAP_{25}^{20cls}	bathtub	refrigerator	desk	nightstand	counter	door	curtain	box	lamp	bag
OV-3DET [26]	18.02	56.28	10.99	19.72	0.77	0.31	9.59	10.53	3.78	2.11	2.71
CoDA [2]	19.32	50.51	6.55	12.42	15.15	0.68	7.95	0.01	2.94	0.51	2.02
GLIS (ours)	**20.83**	53.21	4.76	20.79	7.62	0.09	0.95	7.79	3.32	3.73	1.93

ScanNetV2 is a widely used 3D object detection and semantic segmentation dataset, which has 1513 scenes with over 200 object classes.

SUN RGB-D is a large 3D object detection and scene understanding dataset, which contains 10335 samples with around 800 object classes.

Metrics. We use the mean Average Precision (mAP) at the IoU threshold of 0.25 to evaluate the detection performance. For a fair comparison, we evaluate our GLIS on the top-20 object classes in ScanNetV2 and SUN RGB-D respectively, following OV-3DET [26]. We also report the results on top-10 classes for comparison with methods like [25, 56]. The metrics are notated as mAP_{25}^{20cls} and mAP_{25}^{10cls} respectively for distinguishing.

4.2 Implementation Details

The training process contains two phases: 1) training the 3D backbone and the bounding box prediction heads; 2) training the object confidence prediction head, as well as the local projector and global projector. The training of phase 1 lasts for 400 epochs with a total batch size of 32 (*i.e.*, a single batch size 4×8 GPUs). The training of phase 2 lasts for 50 epochs with a total batch size of 16. The base learning rate is set as $1e-4$. The loss balanced factors are $\lambda_{conf} = 0.2, \lambda_1 = 4, \lambda_2 = 10, \lambda_3 = 1, \lambda_4 = 1$, respectively. The mentioned thresholds are $\phi_{CLIP} = 0.5, \phi_{obj} = 0.1, \phi_{low} = 0.25, \phi_{high} = 0.6, \phi_{keep} = 0.75$, respectively. We choose the model architecture of 3DETR [30] as our 3D backbone and bounding box prediction heads. We use LLaMA [40] as the LLM backbone, which is initialized by the checkpoint vicuna-7b-v1.5-16k [6]. All experiments are conducted on 8 A800 GPUs.

4.3 Main Results

ScanNetV2. As shown in Tab. 1, our proposed GLIS greatly improves the open-vocabulary detection performance on ScanNetV2. Compared to previous sota method CoDA [2], mAP_{25}^{10cls} is raised from 28.76% to 30.94% and mAP_{25}^{20cls} is raised from 19.32% to 20.83%. Our methods also significantly improve the detection precision of many classes, *e.g.*, chair is improved by 10.79%, toilet is improved by 5.81%, and table is improved by 5.15%.

Table 2: Comparisons with other methods on SUN RGB-D

Method	mAP_{25}^{10cls}	toilet	bed	chair	bathtub	sofa	dresser	scanner	fridge	lamp	desk
OV-3DETIC [25]	13.03	43.97	6.17	0.89	45.75	2.26	8.22	0.02	8.32	0.07	14.60
FM-OV3D [56]	21.47	55.00	38.80	19.20	41.91	23.82	3.52	0.36	5.95	17.40	8.77
OV-3DET [26]	**31.06**	72.64	66.13	34.80	44.74	42.10	11.52	0.29	12.57	14.64	11.21
GLIS (ours)	30.83	69.88	63.83	34.78	49.62	40.78	10.73	1.49	8.37	16.40	12.44

Method	mAP_{25}^{20cls}	table	stand	cabinet	counter	bin	bookshelf	pillow	microwave	sink	stool
OV-3DET [26]	20.46	23.31	2.75	3.40	0.75	23.52	9.83	10.27	1.98	18.57	4.10
GLIS (ours)	**21.45**	19.17	13.84	2.75	0.59	22.22	12.65	15.78	5.30	27.62	0.84

Table 3: Ablation study on ScanNetV2

Method	mAP_{25}^{10cls}		mAP_{25}^{20cls}	
Base Model	28.59		19.36	
+RPLG	28.80	+0.21	19.51	+0.15
+BAOL	29.75	+0.95	20.03	+0.52
+GLCI	**30.94**	+1.19	**20.83**	+0.80

SUN RGB-D. The detection precision on SUN RGB-D is reported in Tab. 2. When tested on 20 classes, mAP_{25}^{20cls} is improved from 20.46% to 21.45%, demonstrating the effectiveness of our methods. Specifically, stand is improved by 11.09%, sink is improved by 9.05%, and bathtub is improved by 4.88%, *etc.* Our methods also achieve competitive performance compared to OV-3DET [26] when evaluated on 10 classes.

4.4   Ablation Study

To analyze the effect of each module in our proposed GLIS, we conduct an ablation study on ScanNetV2. Detailed results are presented in Tab. 3

Base Model. Base model is different from GLIS in three aspects: i) In RPLG, all pseudo labels from Detectron2 are reserved. ii) In BAOL, all local features are reserved (*i.e.*, $f_{obj} = f_{loc}$). iii) Global inference is not performed.

Effect of RPLG. With RPLG added to the Base Model, the detection precision increases. For example, mAP_{25}^{10cls} is improved from 28.59% to 28.80%. Such improvement demonstrates that RPLG can recognize false detection in MiniGPT-v2 and improve the pseudo labels' quality.

Effect of BAOL. BAOL further refines the detection performance. For instance, mAP_{25}^{10cls} is lifted from 28.80% to 29.75% and mAP_{25}^{20cls} is lifted from 19.51% to 20.03%. This proves that BAOL can overcome the disturbance of noises in point clouds, resulting in better localization for interested objects.
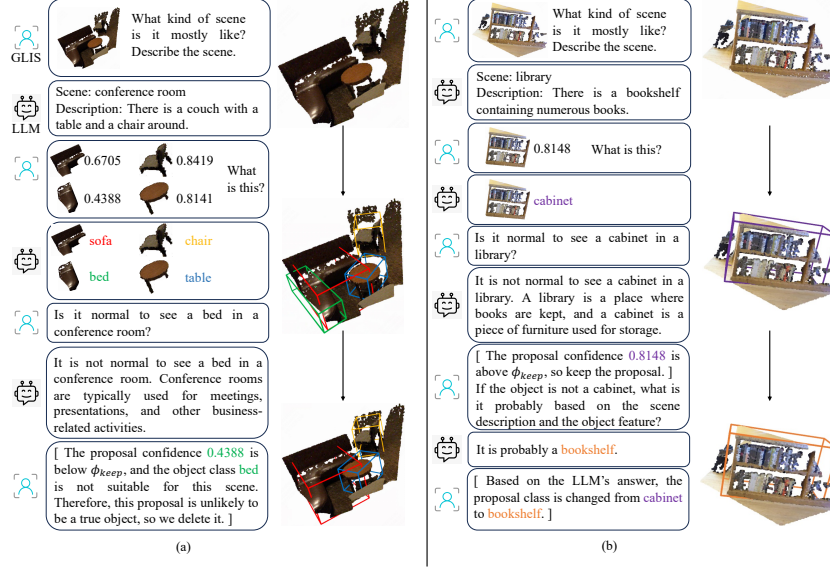
Fig. 4: Visualizations of GLIS. The score of each proposal is the confidence that the proposal is truly a foreground object. These proposals, as well as their confidence scores, are generated by BAOL.

Effect of GLCI. As shown in Tab. 3, GLCI significantly improves the detection performance. Specifically, mAP_{25}^{10cls} is increased by 1.19% and mAP_{25}^{20cls} is increased by 0.80%. Such results show that GLIS successfully utilizes the information from both local and global, and refines the detection result with effective inference.

4.5   Visualizations

To further show the effectiveness of our proposed GLIS, we present visualization results in Fig. 4. The visualization mainly exhibits the dialogue between GLIS and the LLM. Besides, the detection result of each stage is presented, which changes according to the dialogue.

As shown in Fig. 4a, LLM recognizes the scene as a conference room based on global information. Then, four class-agnostic object proposals are generated by BAOL. The four proposals are recognized as a sofa, a chair, a bed, and a table respectively by LLM. Note that the score (*e.g.*, 0.6705) of each proposal is the confidence that the proposal is truly a foreground object. So far, an initial detection result has formed. Subsequently, the global-local collaborative inference is conducted with predefined questions. For example, the question *"Is it normal to see a bed in a conference room?"* is used to guide the LLM's inference. Then LLM outputs the answer to the question, *i.e.*, *"It is not normal to see a bed in a conference room."*, and gives reasonable explanations. Based on
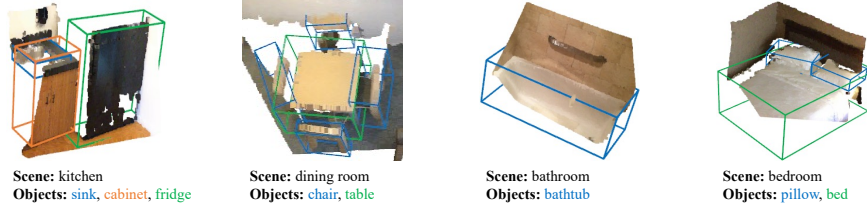
Scene: kitchen
Objects: sink, cabinet, fridge

Scene: dining room
Objects: chair, table

Scene: bathroom
Objects: bathtub

Scene: bedroom
Objects: pillow, bed

Fig. 5: Visualizations of detection results.

the LLM's answer, our scheme automatically recognizes the bed as false detection. Subsequently, as the confidence of the bed is below the predefined threshold $\phi_{keep} = 0.75$, the wrongly detected bed is automatically deleted from the result. Such example shows that GLIS can effectively conduct inference with global and local information, extracting common sense from LLM to eliminate wrongly detected objects.

Fig. 4b exhibits another example. Initially, LLM recognizes the scene as a library, and generates scene description *"There is a bookshelf containing numerous books."*. Only one object proposal is generated from BAOL, and the proposal is recognized as a cabinet. Then, our scheme automatically asks the LLM with the question *"Is it normal to see a cabinet in a library?"*, and LLM answers that *"It is not normal to see a cabinet in a library."*. As the proposal confidence 0.8148 is above the threshold $\phi_{keep} = 0.75$, the proposal is not removed. To correct the predicted class of this proposal, our scheme asks LLM to change the object class according to the scene description and the object feature. Based on local and global information, LLM corrects the object class from cabinet to bookshelf.

More visualizations of detection results are presented in Fig. 5.

5   Conclusion and Discussion

In this paper, we propose GLIS for lidar-based open-vocabulary detection task, which is the first to introduce global-local collaboration in this area. The proposed GLIS extracts object-level information from the local branch and scene-level information from the global branch. With local and global information, a pre-trained LLM is utilized for inference to refine the detection result. To further improve the performance, RPLG and BAOL are devised to ameliorate the object-level information. Experiments on ScanNetV2 and SUN RGB-D demonstrate the effectiveness of GLIS, which achieves state-of-the-art results.

Limitation. The limitation of GLIS exists due to the noises within the point cloud and the false pseudo labels generated from the 2D image. Though we have proposed methods to alleviate the influence, these noises may still confuse the model. These limitations could inspire our future work.

Acknowledgements

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33** (2020) 5
2. Cao, Y., Yihan, Z., Xu, H., Xu, D.: Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. Advances in Neural Information Processing Systems **36** (2024) 4, 11
3. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023) 3, 4, 6, 8
4. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3d object detection network for autonomous driving. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1907–1915 (2017) 4
5. Cheng, B., Sheng, L., Shi, S., Yang, M., Xu, D.: Back-tracing representative points for voting-based 3d object detection in point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021) 4
6. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023) (2023) 11
7. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2017) 3, 10
8. Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G.: Learning to prompt for open-vocabulary object detection with vision-language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14084–14093 (2022) 4
9. Fan, L., Xiong, X., Wang, F., Wang, N., Zhang, Z.: Rangedet: In defense of range view for lidar-based 3d object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2918–2927 (2021) 4
10. Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., Ma, L.: Promptdet: Towards open-vocabulary detection using uncurated images. In: European Conference on Computer Vision. pp. 701–717. Springer (2022) 4
11. Fu, J., Gao, C., Wang, Z., Yang, L., Wang, X., Mu, B., Liu, S.: Eliminating cross-modal conflicts in bev space for lidar-camera 3d object detection. arXiv preprint arXiv:2403.07372 (2024) 4
12. Gao, C., Chen, J., Liu, S., Wang, L., Zhang, Q., Wu, Q.: Room-and-object aware knowledge reasoning for remote embodied referring expression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3064–3073 (2021) 4

13. Gupta, A., Narayan, S., Joseph, K., Khan, S., Khan, F.S., Shah, M.: Ow-detr: Open-world detection transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9235–9244 (2022) 4

14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021) 5

15. Huang, T., Liu, Z., Chen, X., Bai, X.: Epnet: Enhancing point features with image semantics for 3d object detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. pp. 35–52. Springer (2020) 4

16. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022) 5

17. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12697–12705 (2019) 4

18. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022) 4

19. Li, Z., Wang, F., Wang, N.: Lidar r-cnn: An efficient and universal 3d object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7546–7555 (2021) 4

20. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024) 5

21. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023) 4

22. Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., Tang, J.: P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 61–68 (2022) 5

23. Liu, X., Ji, K., Fu, Y., Tam, W.L., Du, Z., Yang, Z., Tang, J.: P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv preprint arXiv:2110.07602 (2021) 5

24. Liu, Z., Zhang, Z., Cao, Y., Hu, H., Tong, X.: Group-free 3d object detection via transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2949–2958 (2021) 4

25. Lu, Y., Xu, C., Wei, X., Xie, X., Tomizuka, M., Keutzer, K., Zhang, S.: Open-vocabulary 3d detection via image-level class and debiased cross-modal contrastive learning. arXiv preprint arXiv:2207.01987 (2022) 4, 5, 7, 11, 12

26. Lu, Y., Xu, C., Wei, X., Xie, X., Tomizuka, M., Keutzer, K., Zhang, S.: Open-vocabulary point-cloud object detection without 3d annotation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1190–1199 (2023) 4, 5, 7, 11, 12

27. Luo, J., Fu, J., Kong, X., Gao, C., Ren, H., Shen, H., Xia, H., Liu, S.: 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16454–16463 (2022) 4

28. Ma, X., Zhang, Y., Xu, D., Zhou, D., Yi, S., Li, H., Ouyang, W.: Delving into localization errors for monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4721–4730 (2021) 4

29. Ma, Z., Luo, G., Gao, J., Li, L., Chen, Y., Wang, S., Zhang, C., Hu, W.: Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14074–14083 (2022) 4

30. Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2906–2917 (2021) 4, 10, 11

31. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9277–9286 (2019) 4

32. Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J.: Frustum pointnets for 3d object detection from rgb-d data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 918–927 (2018) 4

33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763 (2021) 4, 7

34. Rahman, S., Khan, S., Barnes, N.: Improved visual-semantic alignment for zero-shot object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11932–11939 (2020) 4

35. Rahman, S., Khan, S.H., Porikli, F.: Zero-shot object detection: Joint recognition and localization of novel concepts. International Journal of Computer Vision **128**, 2979–2999 (2020) 4

36. Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10529–10538 (2020) 4

37. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 770–779 (2019) 4

38. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 567–576 (2015) 3, 10

39. Sun, P., Wang, W., Chai, Y., Elsayed, G., Bewley, A., Zhang, X., Sminchisescu, C., Anguelov, D.: Rsn: Range sparse net for efficient, accurate lidar 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5725–5734 (2021) 4

40. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 5, 11

41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 4

42. Wang, Y., Fathi, A., Kundu, A., Ross, D.A., Pantofaru, C., Funkhouser, T., Solomon, J.: Pillar-based object detection for autonomous driving. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. pp. 18–34. Springer (2020) 4

43. Wang, Z., Li, Y., Liu, T., Zhao, H., Wang, S.: Ov-uni3detr: Towards unified open-vocabulary 3d object detection via cycle-modality propagation. arXiv preprint arXiv:2403.19580 (2024) 4

44. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022) 5

45. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019) 4, 7

46. Xie, Q., Lai, Y.K., Wu, J., Wang, Z., Lu, D., Wei, M., Wang, J.: Venet: Voting enhancement network for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3712–3721 (2021) 4

47. Xie, Q., Lai, Y.K., Wu, J., Wang, Z., Zhang, Y., Xu, K., Wang, J.: Mlcvnet: Multi-level context votenet for 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10447–10456 (2020) 4

48. Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., Lin, D.: Pointllm: Empowering large language models to understand point clouds. arXiv preprint arXiv:2308.16911 (2023) 10

49. Yang, B., Luo, W., Urtasun, R.: Pixor: Real-time 3d object detection from point clouds. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 7652–7660 (2018) 4

50. Yang, Z., Sun, Y., Liu, S., Jia, J.: 3dssd: Point-based 3d single stage object detector. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11040–11048 (2020) 4

51. 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 (2023) 4

52. 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. Advances in Neural Information Processing Systems **35** (2022) 4

53. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11784–11793 (2021) 4

54. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Open-vocabulary detr with conditional matching. In: European Conference on Computer Vision. pp. 106–122 (2022) 4

55. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14393–14402 (2021) 4

56. Zhang, D., Li, C., Zhang, R., Xie, S., Xue, W., Xie, X., Zhang, S.: Fm-ov3d: Foundation model-based cross-modal knowledge blending for open-vocabulary 3d detection. arXiv preprint arXiv:2312.14465 (2023) 4, 5, 7, 11, 12

57. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023) 5

58. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022) 4

59. Zhang, H., Xu, J., Tang, T., Sun, H., Yu, X., Huang, Z., Yu, K.: Opensight: A simple open-vocabulary framework for lidar-based object detection. arXiv preprint arXiv:2312.08876 (2023) 4, 5, 7

60. Zhang, Z., Sun, B., Yang, H., Huang, Q.: H3dnet: 3d object detection using hybrid geometric primitives. In: European Conference on Computer Vision. pp. 311–329 (2020) 4

61. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., 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) 4
62. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: European Conference on Computer Vision. pp. 350–368 (2022) 4
63. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4490–4499 (2018) 4