

Towards Open-Vocabulary Multimodal 3D Object Detection with Attributes

Xiang, Xinhao; Peng, Kuan-Chuan; Lohit, Suhas; Jones, Michael J.; Zhang, Jiawei

TR2025-162 November 26, 2025

Abstract

3D object detection plays a crucial role in autonomous systems, yet existing methods are limited by closed-set assumptions and struggle to recognize novel objects and their attributes in real-world scenarios. We propose OVODA, a novel framework enabling both open-vocabulary 3D object and attribute detection with no need to know the novel class anchor size. OVODA uses foundation models to bridge the semantic gap between 3D features and texts while jointly detecting attributes, e.g., spatial relationships, motion states, etc. To facilitate such research direction, we propose OVAD, a new dataset that supplements existing 3D object detection benchmarks with comprehensive attribute annotations. OVODA incorporates several key innovations, including foundation model feature concatenation, prompt tuning strategies, and specialized techniques for attribute detection, including perspective-specified prompts and horizontal flip augmentation. Our results on both the nuScenes and Argoverse 2 datasets show that under the condition of no given anchor sizes of novel classes, OVODA outperforms the state-of-the-art methods in open-vocabulary 3D object detection while successfully recognizing object attributes. Our OVAD dataset is released here.

British Machine Vision Conference (BMVC) 2025

© 2025 MERL. This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Towards Open-Vocabulary Multimodal 3D Object Detection with Attributes

Xinhao Xiang^{1*}
xhxiang@ucdavis.edu

Kuan-Chuan Peng²
kpeng@merl.com

Suhas Lohit²
slohit@merl.com

Michael J. Jones²
mjones@merl.com

Jiawei Zhang¹
jiwzhang@ucdavis.edu

¹ University of California, Davis
Davis, CA, USA

² Mitsubishi Electric Research
Laboratories (MERL)
Cambridge, MA, USA

Abstract

3D object detection plays a crucial role in autonomous systems, yet existing methods are limited by closed-set assumptions and struggle to recognize novel objects and their attributes in real-world scenarios. We propose OVODA, a novel framework enabling both open-vocabulary 3D object and attribute detection with no need to know the novel class anchor size. OVODA uses foundation models to bridge the semantic gap between 3D features and texts while jointly detecting attributes, *e.g.*, spatial relationships, motion states, *etc.* To facilitate such research direction, we propose OVAD, a new dataset that supplements existing 3D object detection benchmarks with comprehensive attribute annotations. OVODA incorporates several key innovations, including foundation model feature concatenation, prompt tuning strategies, and specialized techniques for attribute detection, including perspective-specified prompts and horizontal flip augmentation. Our results on both the nuScenes and Argoverse 2 datasets show that under the condition of no given anchor sizes of novel classes, OVODA outperforms the state-of-the-art methods in open-vocabulary 3D object detection while successfully recognizing object attributes. Our OVAD dataset is released here[†].

1 Introduction

Autonomous systems require advanced 3D perception, but most rely on closed-set models limited to predefined object classes, failing with novel objects and complex scenes [0, 1, 15, 16, 20, 32, 34, 35, 39, 43, 52]. Beyond detection, understanding attributes such as motion and spatial relationships is critical. Multimodal foundation models offer promise via open-vocabulary, zero-shot learning [12, 21, 33, 40], but 3D data’s sparsity, complexity, and need for class-specific anchors [8]

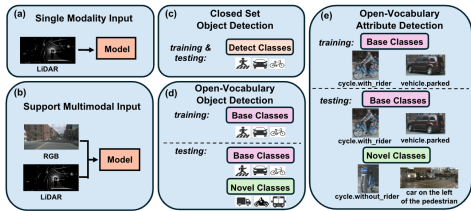


Figure 1: Relaxing the constraints of (a) single-modal (c) closed set LiDAR 3D object detection, OVODA can perform (b) multimodal open-vocabulary (d) object and (e) attribute detection.

Object/Attribute Detection Conditions	Method Categories						
	C_0	C_1	C_2	C_3	C_4	C_5	OVODA (ours)
support 3D object detection	✓	✗	✓	✓	✓	✓	✓
can detect attributes	✗	✗	✗	✗	✗	✗	✓
support OV object detection	✗	✓	✓	✓	✓	✓	✓
support multi-modal input	✗	✓	✓	✓	✗	✗	✓
need no novel class anchor size	✗	✗	✓	✓	✓	✓	✓
can detect complex events	✗	✗	✗	✗	✗	✓	✓

Table 1: OVODA addresses 3D object detection challenges which were formerly not fully considered in prior works (e.g., C_0 : [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100], C_1 : [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100], C_2 : [8], C_3 : [55, 56], C_4 : [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100], C_5 : [0, 50, 51, 52, 53, 54], C_6 : [0, 1, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100]). OV: open-vocabulary.

pose challenges. Attribute detection, vital in tasks like autonomous driving, is underexplored due to limited dataset annotations [10, 57, 62] and narrow focus on classification/localization [8, 8, 57, 62].

To address these challenges, we propose Open-Vocabulary Object Detection with Attributes (OVODA), a framework for open-vocabulary 3D object and attribute detection that uses foundation models (FM) for semantics while preserving 3D geometric precision. Unlike prior methods (e.g., [8]), OVODA detects novel classes and attributes without anchor size knowledge of the novel classes. OVODA combines temporal-spatial features, complex event generation, and semantic attribute alignment, using FM features and prompt tuning to unify 3D geometry and semantics.

OVODA integrates attribute detection into object detection, enabling unified recognition of novel objects and their attributes—spatial relations, motion states, and interactions. This is achieved via a complex event generation module that aligns 3D features with text in semantic space. Perspective-specific prompts and horizontal flip augmentation further improve accuracy under challenging conditions with varying viewpoints and object orientations. OVODA thus delivers comprehensive scene understanding by identifying objects, their relations, and behaviors.

To boost OVODA’s performance, we add two enhancements: (1) Combining FM and existing 3D detection backbone features for richer semantics understanding and precise localization; (2) Prompt tuning for task-specific FM adaptation. We also introduce two loss functions for learning novel attributes without annotations associated with them, making OVODA a robust effective solution for open-vocabulary 3D object and attribute detection. To facilitate attribute detection research, we propose the Open Vocabulary Attribute Detection (OVAD) dataset, a benchmark built on nuScenes [9] with 84384 instances labeled across 11 attribute classes. OVAD is the first benchmark to include detailed annotations on spatial relations, motion states, and interactions, enabling thorough evaluation of complex scene understanding and open-vocabulary 3D attribute detection in real outdoor scenes.

Our experiments on nuScenes [9] and Argoverse 2 [42] show that OVODA beats the state-of-the-art (SOTA) in open-vocabulary 3D object detection when the novel class anchor sizes are unavailable, while also detecting novel attributes. This marks a key advance in 3D scene understanding with applications like autonomous driving and robotics. Our contributions include:

1. We propose OVODA, a novel open-vocabulary multimodal 3D object detector from multi-view input to detect complex events (including attributes) without needing novel class anchor size.
2. We propose concatenation with foundation model features, prompt tuning strategies, two novel loss functions, and two attribute-specific techniques (perspective-specified prompt, horizontally flip augmentation), to improve the open-vocabulary object and attribute detection performance.
3. Proposing the OVAD dataset for open-vocabulary attribute detection, we show that under the condition of no predefined novel class anchor sizes, OVODA outperforms the SOTA methods

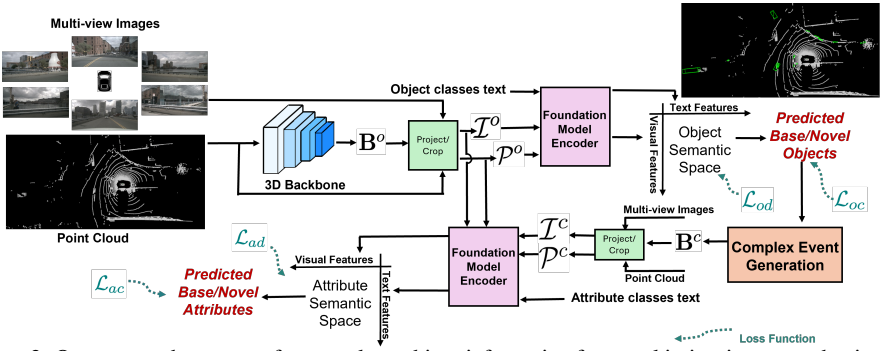


Figure 2: Our proposed OVODA framework combines information from multi-view images and point clouds, align object and attribute text with vision features using a common foundation model encoder in order to discover and localize complex open-vocabulary events that include multiple objects and attributes. We show the figure using information from a single time instant for simplicity. In our method, we aggregate information over multiple temporal instances enabling better complex event discovery involving motion attributes.

on the open-vocabulary 3D object detection task on both the nuScenes and Argoverse 2 datasets.

2 Related work

Contrasting OVODA with prior works in Tab. 1, we categorize them by properties and elaborate.

Open-vocabulary (OV) 3D object detection. OV detection uses language models to classify both seen and novel classes, offering greater flexibility than traditional open-set or zero-shot learning. While most 2D OV methods (*e.g.*, C_1) use pretrained vision-language (VL) models like CLIP [63] and GLIP [41], conventional 3D detectors (C_0) rely on closed-set supervision. Recent works (C_3 , C_4) adapt VL features to 3D perception via multi-modal embeddings [18], but struggle to encode spatial cues from sparse point clouds and focus mainly on indoor scenes. OVODA overcomes these limits in diverse and dynamic outdoor settings with complex spatial relationships.

Complex event extraction captures object relationships, temporal dynamics, and context, unlike basic object detection. Events like “person behind car” or “person sitting” involve multiple objects or attributes. Prior methods (C_5) use handcrafted features or models like HMMs [27], CNN- [6, 63], RNN- [47], and Transformer-based [2, 66] models, but struggle with generalization and efficiency. Multimodal [10] and graph-based [49] methods add richer context but face alignment or manual setup issues. OVODA is the first to detect complex events in OV 3D settings by jointly predicting objects and attributes.

Attribute detection in 3D outdoor scenes. Detecting attributes like motion and spatial attributes in 3D outdoor scenes is vital but hard due to sparse sensor data and complex scenes. Prior methods (C_4) extend detectors to predict attributes [30] or fuse modalities [61], but need fine-grained cues or precise calibration. Graph-based approaches [12, 68] need manual graph design and struggle with 3D sparsity. OVODA uses foundation models [12, 40] to incorporate semantics, enabling open-vocabulary attribute detection with better generalization.

3 Our proposed method — OVODA

3.1 Framework overview

We define the OV detection setup with object and attribute labels available for base classes $\mathbb{C}^b$ and $\mathbb{C}^{ba}$ during training. At inference, the goal is to detect both base and novel object classes $\mathbb{C}^n$ and attributes $\mathbb{C}^{na}$. Complex events refer to (a) an object with an attribute or (b) two objects with a spatial relation. As shown in Fig. 2, OVODA is the first OV 3D detector to jointly recognize objects and complex events. It processes 3D point clouds and multi-view images via a spatiotemporal extractor (STE) to produce single-object proposals (Sec. 3.2), and builds on 3DETR [46] for transformer-based localization and classification. We enable OV detection by aligning vision, point cloud, and text features using a frozen OneLLM [14] model, allowing recognition of novel classes in the FM’s semantic space. The CEG module (Sec. 3.3) proposes spatially related object pairs for complex event detection, which are matched in attribute space for novel attribute recognition. OVODA is trained end-to-end with joint losses for both tasks (Sec. 3.4).

3.2 Single-object proposal

Single-object proposal extraction. We begin with a set of base-class objects in the training dataset with ground truth 3D object bounding box: $\mathbf{D}^b = \{o_j = (c_j, B_j) \mid c_j \in \mathbb{C}^b\}$, where c_j is a class in the set of the base object classes ($\mathbb{C}^b$) and B_j is the corresponding ground truth bounding box. We train an initial class-agnostic 3D object proposer f_{det} using $\mathbf{D}^b$ by minimizing an object box regression loss inspired by 3DETR [46]. By focusing exclusively on objectness score prediction and box regression, we avoid the limitations of class-specific training that have been shown to hinder novel object detection in OV 2D detection [44]. The trained f_{det} generates hidden features $\mathbf{F}_{\text{det}}$ and object proposals $\mathbf{B}^o$ which is characterized by objectness scores and precise 3D localization parameters.

Single-object proposal generation. $\mathbf{B}^o$ guides the generation of single-object instances from multimodal visual inputs (point cloud $\mathcal{P}$ and multi-view images $\mathcal{I}$) for subsequent text-visual alignment and class prediction. We crop the object instances from the 3D boxes $\mathcal{P}^o$ in $\mathcal{P}$ via $\mathcal{P}^o = \text{Crop}(\mathbf{B}^o, \mathcal{P})$. For $\mathcal{I}$, we first project $\mathbf{B}^o$ onto the 2D image plane using the camera matrices M : $\mathbf{B}_{2D}^o = \text{Proj}(\mathbf{B}^o, M)$, and then we crop the corresponding objects in the image using $\mathcal{I}^o = \text{Crop}(\mathbf{B}_{2D}^o, \mathcal{I})$. All cropped image features are concatenated together with an additional encoding representing view direction.

Novel object class discovery. To recognize novel object classes, we use the frozen OneLLM [14] to align image and point cloud features of each object proposal B_j^o —denoted by $V_j^o = [\mathcal{I}_j^o; \mathcal{P}_j^o]$ —with a super-class vocabulary T^o containing both base ($\mathbb{C}^b$) and novel ($\mathbb{C}^n$) classes. The prediction distribution over C classes is: $\mathbf{P}_j^o = \{p_{j,1}^o, p_{j,2}^o, \dots, p_{j,C}^o\} = \text{Softmax}(V_j^o \cdot \mathbf{F}_{T^o})$, where $\mathbf{F}_{T^o}$ is the textual embedding of T^o , $\cdot$ denotes the dot product, and the distribution $\mathbf{P}_j^o$ serves as the OV semantic priors. The final predicted class is $c_j^* = \text{argmax}_c \mathbf{P}_j^o$, and the resulting single-object detections are: $\mathbf{D}^s = \left\{ (c_j^*, B_j^o) \mid c_j^* \in \mathbb{C}^b \cup \mathbb{C}^n \right\}$. We classify B_j^o as a novel object if it meets:

$$\begin{aligned} \mathbf{O}^{\text{disc}} = \left\{ B_j^o \mid \forall B_i^b \in \mathbf{B}^b, \text{IoU}_{3D}(B_j^o, B_i^b) < \theta^b, Q_j^o > \theta^o, \right. \\ \left. p_{j,c_j^*}^o > \theta^s, B_j^o \in \mathbf{B}^o, c_j^* \notin \mathbb{C}^b \right\}, \end{aligned} \quad (1)$$

where $\mathbf{B}^b$ is the set of proposals classified as base classes, Q_j^o is the 3D objectness score, and $\theta^b = 0.2$, $\theta^o = 0.8$, $\theta^s = 0.5$ are thresholds for IoU, objectness, and semantic confidence.

Detector improvement strategy. We propose two key enhancements for f_{det} to improve OV detection performance: (1) Feature Fusion with FM: We augment f_{det} 's encoded features with visual features extracted from OneLLM [14]. This multi-dimensional feature fusion strategy enriches the detector's input representation, enabling more comprehensive learning from limited data. (2) Prompt Tuning Integration: We incorporate learnable visual prompts at the FM's input layer, facilitating task-specific adaptation of the pretrained FM.

3.3 Complex event generation (CEG)

OVODA extends beyond traditional OV 3D object detection to be able to detect complex events via jointly predicting objects and their attributes. We consider some of the most critical outdoor scene attributes, including spatial relationships between objects, motion states, and human-traffic participant interactions. To this end, different from conventional methods focusing only on single-object detection, OVODA includes a novel complex event proposal generation method to extract complex inter-object contextual knowledge which is crucial to support downstream complex event detection.

Complex event visual proposal generation. OVODA addresses three critical outdoor scene attributes: spatial relationships, motion states, and human-traffic participant interactions. Effective attribute detection needs both spatial and temporal context beyond single objects. The spatial attribute detection relies on relative positional and orientational relationships between objects, while temporal attribute (*e.g.*, motion state) detection relies on the temporal information across multiple timestamps. To meet these needs, we construct complex event proposals by concatenating: (1) Non-spatial attribute features: Temporal sequence of single-object proposals. (2) Spatial attribute features: Proposals generated from two nearby single-object proposals. We define $\mathbf{D}^o$ as the set of detected single-object proposals: $\mathbf{D}^o = \left\{ \left(c_j^o, B_j^o \right) \mid c_j^o \in \mathbb{C}^b \cup \mathbb{C}^n \right\}$. To generate non-spatial attribute proposals $\mathbf{B}^n$, we concatenate current single-object proposals with those generated in the past T timestamps in total, to incorporate temporal sequence information into the proposals (*i.e.* $\mathbf{B}^n = (\mathbf{B}^o, \mathbf{B}^{o-1}, \dots, \mathbf{B}^{o-T})$). The spatial attribute proposals $\mathbf{B}^s$ are generated by combining nearby single-object proposals:

$$\mathbf{B}^s = \left\{ B_{ij}^s = \left(\text{Comb}(B_i^o, B_j^o) \mid \text{Dist}(B_i^o, B_j^o) \leq \theta^d \right) \right\}, \quad (2)$$

where θ^d is the distance threshold of 15 meters. The Comb operation creates a larger two-object proposal B_{ij}^s by merging selected single-object proposals B_i^o and B_j^o . B_{ij}^s 's spatial extent is defined by:

$$\left[\begin{array}{l} x_{\min} = \text{Min}(B_{ix}^o, B_{jx}^o), \quad x_{\max} = \text{Max}(B_{ix}^o, B_{jx}^o), \\ y_{\min} = \text{Min}(B_{iy}^o, B_{jy}^o), \quad y_{\max} = \text{Max}(B_{iy}^o, B_{jy}^o), \\ z_{\min} = \text{Min}(B_{iz}^o, B_{jz}^o), \quad z_{\max} = \text{Max}(B_{iz}^o, B_{jz}^o) \end{array} \right] \quad (3)$$

The final set of generated complex event visual proposals $\mathbf{B}^c$ is the concatenation of the two feature groups (*i.e.* $\mathbf{B}^c = \mathbf{B}^n \cup \mathbf{B}^s$.) Following the same Crop and Proj operations in Sec. 3.2, we extract corresponding visual proposals in image ($\mathcal{I}^c$) and point cloud ($\mathcal{P}^c$) modalities. We perform data augmentation by horizontal flipping (*i.e.*, $\mathcal{I}_{\text{flip}}^c = \text{Flip}(\mathcal{I}^c)$) for more robust detection. For each proposal, we perform inference by using $\mathcal{I}^c$ and $\mathcal{I}_{\text{flip}}^c$ in turn as input. We average the prediction scores from these two to get the final prediction result.

Complex event text proposal generation. For each complex visual proposal B^c , we generate the corresponding text which can be fed into OneLLM's text encoder. For non-spatial attribute text

proposals c^n , the text is directly from the predicted object class (i.e., $c_j^n = \mathbb{T}(c_j^*) \mathbb{T}(\text{NSA})$, where $\mathbb{T}(\cdot)$ is the function translating the input class label or spatial attribute to the corresponding text), NSA is one of the non-spatial attributes. For spatial attribute text proposals c^s , their text are generated based on the constituent single proposals and their relative spatial configuration. Specifically, if B_i^o is combined with B_j^o , its corresponding text is:

$$c_{ij}^s = \text{"From the perspective of } \mathbb{T}(c_i^o), \mathbb{T}(c_i^o) \mathbb{T}(\text{SA}) \mathbb{T}(c_j^o)\text{"} \quad (4)$$

where SA is one of the four spatial attributes: *in front of*, *behind*, *on the left of*, *on the right of*. The classes are derived from relative coordinates of the two single proposals by mathematical definitions, making them inherently objective. The perspective-based prefix ensures unique and distinguishable text features while establishing clear spatial relationships. The final set of generated complex event text proposals c^c is defined as: $c^c = c^n \cup c^s$. Overall, the set of generated complex events are denoted as: $\mathbf{D}^c = \left\{ \left(c_j^c, B_j^c \right) \mid c_j^c \in \mathbb{C}^{ba} \cup \mathbb{C}^{na} \right\}$.

Novel attribute class discovery. We use OneLLM [44] to align these proposals from image, point and text modalities. Firstly, we generate visual feature V^A by concatenating from image encoder feature of $\mathcal{I}^c$ or $\mathcal{I}_{flip}^c$, and point cloud encoder feature of $\mathcal{P}^c$ from OneLLM. Then, we use OneLLM’s text encoder to extract the text proposals c^c to get the text features $\mathbf{F}_{TA}$. V^A and $\mathbf{F}_{TA}$ are then be aligned in the attribute semantic space where we compute the distance between them: $\mathbf{P}_j^A = \left\{ p_{j,1}^A, p_{j,2}^A, \dots, p_{j,E}^A \right\} = \text{Softmax} \left(V_j^A \cdot \mathbf{F}_{TA} \right)$. Where E is the total number of attribute classes. The distribution $\mathbf{P}_j^A$ serves as the OV semantic priors. The final predicted attribute class e_j^* for each $B_j^c \in \mathbf{B}^c$ is decided by the maximum probability among $\mathbf{P}_j^A$, i.e., $e_j^* = \arg\max_e \mathbf{P}_j^A$. The set of detected complex event proposals is denoted as: $\mathbf{D}^c = \left\{ \left(e_j^*, B_j^c \right) \mid e_j^* \in \mathbb{C}^{ba} \cup \mathbb{C}^{na} \right\}$. We determine whether a complex event proposal involves a novel attribute or not by the following criteria:

$$\mathbf{A}^{disc} = \left\{ B_j^c \mid \forall B_i^{ba} \in \mathbf{B}^{ba}, \text{IoU}_{3D} \left(B_j^c, B_i^{ba} \right) < \theta^b, \right. \\ \left. p_{j,e^*}^A > \theta^a, B_j^c \in \mathbf{B}^c, e^* \notin \mathbb{C}^{ba} \right\}, \quad (5)$$

where $\mathbf{B}^{ba} \in \mathbf{B}^c$ denotes attribute proposals in $\mathbf{B}^c$ predicted as one of the base attribute classes. θ^a is the threshold for attribute semantic scores set to 0.5. $\mathbb{C}^{ba}$ is the set of base attribute classes.

3.4 Overall optimization

Open-vocabulary object losses. To transfer knowledge from LiDAR to images, following [9], we enforce V_j^O and $\mathbf{F}_{det}$ to be the same by using a class-agnostic L1 loss to minimize their feature distance: $\mathcal{L}_{od} = \sum_{j=1}^N \|V_j^O - \mathbf{F}_{det}\|_1$, where N is the number of object proposals. $\mathcal{L}_{od}$ effectively reduces cross-modal gaps, enhancing feature alignment across diverse scenes, including background regions. $\mathcal{L}_{od}$ is independent of class annotations, as it needs no ground-truth box class labels.

Like [9], we also use a loss to promote discriminative classification by maximizing the matching score of the ground-truth class while minimizing scores for other classes: $\mathcal{L}_{oc} = \sum_{j=1}^N f \left(B_j^{disc}, \mathbf{B}^b \right) \cdot CE \left(\mathbf{P}_j^{disc}, h_j^O \right)$, where B_j^{disc} is the j -th object proposal of $\mathbf{O}^{disc}$. $\mathbf{P}_j^{disc} = \text{Softmax} \left(V_j^{disc} \cdot \mathbf{F}_{TO} \right)$, where V_j^{disc} is the visual features of B_j^{disc} . $CE(\cdot)$ denotes the cross-entropy loss. The function $f(x)$ is to check if B_j^{disc} is within $\mathbf{B}^b$, returning 1/0 when the answer is yes/no. $\mathbf{P}_j^{disc}$ is the probability of $\mathbf{O}^{disc}$. h_j^O is the ground truth one-hot vector for B_j^{disc} .

Open-vocabulary attribute losses. Similar to object losses, we align V_j^A with $\mathbf{F}_c$ to transfer knowledge from LiDAR to images, where $\mathbf{F}_c$ is the 3D backbone features of $\mathbf{B}^c$. We enforce V_j^A and $\mathbf{F}_c$ to be the same via a class-agnostic L1 loss to minimize their feature distance: $\mathcal{L}_{ad} = \sum_{j=1}^N \|V_j^A - \mathbf{F}_c\|_1$.

In addition, we propose to use a contrastive loss to ensure correct base attribute classification by maximizing the matching score of the ground-truth attribute class while minimizing the scores for other attribute classes: $\mathcal{L}_{ac} = \sum_{j=1}^N g(A_j^{disc}, \mathbf{B}^{ba}) \cdot CE(\mathbf{P}_j^{disc,a}, h_j^A)$, where A_j^{disc} is the j -th complex event proposal of $\mathbf{A}^{disc}$. $\mathbf{P}_j^{disc,a} = \text{Softmax}(V_j^{disc,a} \cdot \mathbf{F}_{TA})$, where $V_j^{disc,a}$ is the visual features of A_j^{disc} . The function $g(x)$ is to check if A_j^{disc} is within $\mathbf{B}^{ba}$, returning 1/0 when the answer is yes/no. h_j^A is the one-hot ground truth attribute vector for A_j^{disc} .

These aforesaid four losses jointly improve feature alignment, making the 3D features of novel objects/attributes more discriminative, thus enhancing the model’s ability to detect novel objects/attributes. The final loss function $\mathcal{L}$ is defined as: $\mathcal{L} = w_{od}\mathcal{L}_{od} + w_{oc}\mathcal{L}_{oc} + w_{ad}\mathcal{L}_{ad} + w_{ac}\mathcal{L}_{ac}$, where w ’s represent the weights balancing each loss to ensure comparable ranges.

4 The OVAD dataset

OVODA is the first method to perform complex event detection in OV 3D object detection by jointly detecting objects and key outdoor attributes: spatial relations, motion state, and presence of people among traffic participants. Existing datasets lack full annotations—*e.g.*, nuScenes [9] includes motion and presence of people but not spatial attributes. To fill this gap, we propose a novel attribute dataset, OVAD, for comprehensive attribute training and detection evaluation.

Built on nuScenes, OVAD retains its attribute annotations. From 28,130 nuScenes time instances, 5,000 were sampled, yielding 170,149 object annotations, filtered to 84,384 annotations across 10 target classes. To create spatial attribute annotations, we selected pairs of object annotations with a distance within [0m, 15m]: $\mathbf{B}^{\text{OVAD}} = \{B_{ij}^{\text{OVAD}} = (\text{Comb}(B_i, B_j) \mid \text{Dist}(B_i, B_j) \leq 15\text{m})\}$. The Comb operation creates a larger two-object proposal B_{ij}^{OVAD} by merging the selected two nearby ground truth single-object proposals B_i and B_j . The spatial extent of B_{ij}^{OVAD} is defined in a similar way as Eq. 3. The 15-meter spatial threshold follows real-world traffic constraints and practical sensor limitations to ensure reliable sampling. For the ground-truth label of B^{OVAD} (*i.e.*, c_{ij}^{OVAD}), its text is generated based on the constituent single proposals and their relative spatial configuration. Specifically, if B_i is combined with B_j , its corresponding text is: $c_{ij}^{\text{OVAD}} = \mathbb{T}(B_i) \mathbb{T}(\text{SA}) \mathbb{T}(B_j)$.

5 Experiments

Dataset & metrics. We evaluate on nuScenes [9] and Argoverse 2 [42] using mean Average Precision (mAP), nuScenes Detection Score (NDS) for object detection, success rate (SR) for attribute detection, and AP_N (mean AP computed only over novel classes). In nuScenes settings, we use 10 object classes, 7 non-spatial, and 4 spatial attribute classes from OVAD; in Argoverse 2 settings, we use 8 object classes for consistency with nuScenes. Our OV settings for object and attribute detection are detailed in Tab. 2 and 3, respectively. Additional details about our dataset settings, metrics settings, vocabulary settings, and implementation are in the supplement.

dataset setting	base object class	novel object class
N_{b6n4}	Car, Construction vehicles, Trailer, Barrier, Bicycle, Pedestrian	Truck, Bus, Motorcycle, Traffic cone
N_{b3n7}	Car, Bicycle, Pedestrian	Construction vehicles, Trailer, Barrier, Truck, Bus, Motorcycle, Traffic cone
N_{b0n10}	$\emptyset$	Car, Construction vehicles, Trailer, Barrier, Bicycle, Pedestrian Truck, Bus, Motorcycle, Traffic cone
A_{b4n4}	Regular Vehicle, Trailer, Bicycle, Pedestrian	Truck, Bus, Motorcycle, Construction cone

Table 2: For fair comparison, we follow [8] and evaluate on nuScenes using 3 OV settings (N_{b6n4} , N_{b3n7} , N_{b0n10}). We select the same object classes in Argoverse 2 and split them into base and novel classes (A_{b4n4}) for consistency with nuScenes.

dataset setting	base attribute class	novel attribute class
<i>OVAD</i>	with rider, sitting lying down, parked, in front of, behind	without rider, standing, moving, on the left of

Table 3: The OV settings on attribute classes in nuScenes. Our OVAD dataset includes all the attributes in nuScenes and 4 spatial classes. The attributes colored with teal/violet/brown are the attributes exclusively associated with cycle/pedestrian/vehicle classes. We included all providing attribute annotations in nuScenes.

method	CFM	prompt tuning	need no predefined anchor size for each novel class?	N_{b6n4}			N_{b3n7}			N_{b0n10}		
				mAP	NDS	AP _{N20}	mAP	NDS	AP _{N20}	mAP	NDS	AP _{N20}
Find n' Propagate [8]	✗	✗	✗	44.95	47.87	33.65	37.38	40.28	18.46	N/A	N/A	16.72
CoDAv2 [8]	✗	✗	✓	27.35	29.48	12.63	18.73	20.14	8.74	4.32	5.82	1.37
OVODA	✗	✗	✓	30.24	31.54	14.24	20.46	21.83	9.31	4.57	6.96	2.16
OVODA	✓	✓	✓	32.25	31.85	14.72	21.03	22.14	10.23	4.70	7.05	2.39

Table 4: OV object detection results on nuScenes. Acronym: CFM: concatenating foundation model features.

5.1 Experimental results

3D open-vocabulary object detection. Tab. 4 shows the results of OV 3D object detection on nuScenes across three experimental settings. Prior works [22, 24, 25, 26] focus on indoor datasets, leaving Find n' Propagate [8] as the only outdoor 3D OV baseline. To expand comparisons, we adapt CoDAv2 [8] for nusenes datasets by modifying only the dataloader. The low baseline score reflects the significant difficulty of the task. The result shows that OVODA outperforms CoDAv2 when novel class anchor size is unavailable, and that by concatenating FM features (CFM) and prompt tuning, OVODA outperforms CoDAv2 in mAP for the N_{b6n4} , N_{b3n7} , and N_{b0n10} settings, respectively. In the N_{b0n10} setting (detecting completely unseen classes), our OVODA achieves an **11.4%** performance improvement. Although Find n' Propagate [8] performs well in its original setting, it needs predefined per-class anchor sizes to decode box geometry for novel classes (C_2 in Tab. 1), which we argue is an impractical and unfair advantage in the OV 3D object detection task. In contrast, OVODA reaches competitive performance with no such constraints, showing greater flexibility and applicability. Tab. 5 shows OVODA's result on Argoverse 2. We only use the degraded version of OVODA (without CFM and prompt tuning) as the baseline because the degraded OVODA already outperforms CoDAv2 on nuScenes. Tab. 4 and 5 show OVODA's generalizability and adaptability across different datasets. We show qualitative comparison in the supplement.

Complex event detection. OVODA outperforms other OV 3D detectors by jointly predicting objects and attributes for complex event detection. Tab. 6 shows the results under the N_{b6n4} object and OVAD attribute settings. We evaluate in two modes: using the predicted objects (the last column) and ground-truth objects (the second to the last column) for isolated attribute evaluation. OVODA shows strong performance in mAP, NDS, AP_{N20}, and success rates (SR), with the concatenation of FM features (CFM) and prompt tuning (PT) yielding consistent gains. Since foundation models are frozen, OVODA could reach 6.77% SR for full pipeline and runs at **27 FPS** on an NVIDIA RTX A6000, confirming real-time capability. Qualitative results are in Fig. 3 and the supplement.

method	CFM	prompt tuning	mAP	AP _{N20}
OVODA	✗	✗	17.24	8.23
OVODA	✓	✓	27.43	12.34

Table 5: Performance comparison of different methods on Argoverse 2 under the A_{b6n4} setting. Acronyms: CFM: concatenating foundation model features.

method	CFM	PT	mAP	NDS	AP _{N20}	SR (%) (AD only)	SR (%) (AD & OD)
OVODA	✗	✗	30.24	31.54	14.24	16.35	4.23
	✓	✗	31.34	31.63	14.42	22.74	5.56
	✓	✓	32.25	31.85	14.72	25.90	6.77

Table 6: Both the object and attribute detection results as well as the ablation study of our framework on the nuScenes dataset, under N_{b6n4} object setting and $OVAD$ attribute setting. Acronyms: SR: success rate; AD: attribute detection; OD: object detection; CFM: concatenating foundation model features; PT: prompt tuning.

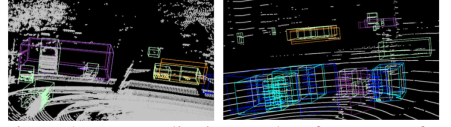


Figure 3: Two qualitative results of OVODA for 3D complex event detection in nuScenes dataset. All ground truth annotations of single object are rendered in light green. The ground truth annotations are rendered in light purple/yellow/light blue for the car-car/pedestrian-pedestrian/others complex events, the predicted bounding boxes are rendered in purple/orange/blue for the car-car/pedestrian-pedestrian/others complex events. Examples of car-car/pedestrian-pedestrian/others complex events can be: a car in front of the car/a pedestrian on the left of the pedestrian/a cyclist behind a car.

5.2 Ablation study

Foundation model augmentation & prompt tuning. We evaluate the impact of FM feature concatenation (CFM) and prompt tuning (PT) through the ablation study in Tab. 6, where both CFM and PT contribute to OVODA’s final performance. CFM enhances feature representation by using rich semantic embeddings, while PT enables better task adaptation. This synergistic effect is particularly evident in the second to the last column (attribute detection only), where OVODA with CFM and PT yields a 9.55% absolute gain in SR over the degraded OVODA without CFM and PT.

Adapting attribute detection model. We evaluate the impact of different FMs (CLIP [33], CogVLM [40], and OneLLM [44]) on OVODA’s performance. OneLLM provides FM encoders in text, image, and point cloud modalities for OVODA to use, while CLIP and CogVLM only provide FM encoders in text and image modalities. Tab. 7 shows that OneLLM significantly outperforms the other alternatives across all metrics. This represents

substantial gain over CLIP and CogVLM. Even in novel class detection (AP_{N20}), the performance gap is pronounced. We attribute OneLLM’s superior performance over CLIP and CogVLM to the additional FM encoder for point cloud which is particularly suitable for the LiDAR input modality.

Using clearer descriptions & augmenting by horizontally flipped visual features. We evaluate the efficacy of perspective-specified prompts (PSP) and horizontally flip augmentation (HFA) via the ablation study in Tab. 8. Using CLIP as the FM, we find that: (1) PSP alone improves mAP by 0.41% and attribute detection SR by 0.63%. This gain supports the benefit of providing more descriptive prompts that help the model better distinguish between objects based on perspective, enhancing its accuracy. (2) HFA alone yields larger gains, with mAP improving by 0.53% and attribute detection SR improving by 1.23%. These results show that HFA makes OVODA’s detection more robust. (3) Jointly using PSP and HFA achieves even better performance than using each individually. Finally, integrating PSP and HFA with OneLLM instead of CLIP achieves the best performance across all metrics. These results suggest that viewpoint prefixes in prompts could

method	foundation model	mAP	NDS	AP _{N20}	SR (%) (AD only)	SR (%) (for both AD & OD)
OVODA	CLIP [33]	21.93	21.54	11.45	16.20	3.22
	CogVLM [40]	25.34	26.81	12.84	19.84	5.39
	OneLLM [44]	32.25	31.85	14.72	25.90	6.77

Table 7: Ablation study using different foundation models on nuScenes ($N_{b6n4}/OVAD$ for object/attribute detection settings). Acronyms: SR: success rate; AD: attribute detection; OD: object detection.

ensure spatial direction remains disambiguated. OVODA with OneLLM, when augmented with PSP and HFA, provides a stronger baseline for the OV 3D object and attribute detection task, likely due to OneLLM’s richer, more contextually aware embeddings.

6 Conclusion

We propose OVODA, a novel framework enabling open-vocabulary (OV) multimodal 3D object detection with attribute detection, requiring no novel class information. It uses foundation model features and prompt tuning to bridge 3D features and text descriptions, while jointly detecting attributes like spatial relationships and motion states. We introduce OVAD with comprehensive attribute annotations for evaluating OV attribute detection. On nuScenes and Argoverse 2, OVODA outperforms SOTA OV 3D object detection methods while successfully detecting object attributes.

method	PSP	HFA	FM	mAP	NDS	AP _{N20}	SR (%) (AD only)	SR (%) (for both AD & OD)
OVODA	✗	✗	CLIP	21.93	21.54	11.45	16.20	3.22
	✓	✗	CLIP	22.34	24.35	11.94	16.83	3.49
	✗	✓	CLIP	22.46	23.43	12.32	17.43	4.03
	✓	✓	CLIP	24.37	25.83	13.02	19.42	4.92
	✓	✓	OneLLM	32.25	31.85	14.72	25.90	6.77

Table 8: Ablation study of using foundation models (FM) (CLIP [5] & OneLLM [14]) on nuScenes under the N_{b6n4} object detection (OD) and OVAD attribute detection (AD) setting. Acronyms: PSP: perspective specified prompt; HFA: horizontally flip augmentation; FM: foundation model; SR: success rate.

References

- [1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning (ICML)*, 2021. URL <https://arxiv.org/abs/2102.05095>.
- [3] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [4] Yang Cao, Yihan Zeng, Hang Xu, and Dan Xu. CoDA: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3D object detection. In *Neural Information Processing Systems (NeurIPS)*, 2023.
- [5] Yang Cao, Yihan Zeng, Hang Xu, and Dan Xu. Collaborative novel object discovery and box-guided cross-modal alignment for open-vocabulary 3D object detection. *arXiv preprint arXiv:2406.00830*, 2024. URL <https://arxiv.org/abs/2406.00830>.
- [6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL <https://arxiv.org/abs/1705.07750>.
- [7] Yu Du, Fangyun Wei, Zihe Zhang, Miaoqing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *IEEE/CVF*

- Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL <https://arxiv.org/abs/2203.14940>.
- [8] Djamahl Etchegaray, Zi Huang, Tatsuya Harada, and Yadan Luo. Find n' propagate: Open-vocabulary 3D object detection in urban environments. In *European Conference on Computer Vision (ECCV)*, 2024. URL <https://arxiv.org/abs/2403.13556>.
- [9] Lue Fan, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Fully sparse 3D object detection. In *Neural Information Processing Systems (NeurIPS)*, 2022.
- [10] Valentin Gabeur, Chen Sun, KarteeK Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision (ECCV)*, 2020.
- [11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. doi: 10.1109/CVPR.2012.6248074.
- [12] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024. URL <https://arxiv.org/abs/2309.16650>.
- [13] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations (ICLR)*, 2022.
- [14] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. OneLLM: One framework to align all modalities with language. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL <https://arxiv.org/abs/2312.03700>.
- [15] Deepti Hegde, Suhas Lohit, Kuan-Chuan Peng, Michael J. Jones, and Vishal M. Patel. Equivariant spatio-temporal self-supervision for LiDAR object detection. In *European Conference on Computer Vision (ECCV)*, 2025.
- [16] Deepti Hegde, Suhas Lohit, Kuan-Chuan Peng, Michael J. Jones, and Vishal M. Patel. Multimodal 3D object detection on unseen domains. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, 2025.
- [17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9 (8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL <https://doi.org/10.1162/neco.1997.9.8.1735>.
- [18] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. ConceptFusion: Open-set multimodal 3D mapping. In *Robotics: Science and Systems (RSS)*, 2023. URL <https://arxiv.org/abs/2302.07241>.

- [19] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as composition of spatio-temporal scene graphs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL <https://arxiv.org/abs/1912.06992>.
- [20] Kaidong Li, Tianxiao Zhang, Kuan-Chuan Peng, and Guanghui Wang. PF3Det: A prompted foundation feature assisted visual LiDAR 3D detector. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, 2025.
- [21] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL <https://arxiv.org/abs/2112.03857>.
- [22] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yin hao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. OpenShape: Scaling up 3D shape representation towards open-world understanding. In *Neural Information Processing Systems (NeurIPS)*, 2023.
- [23] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, 2024.
- [24] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary 3D detection via image-level class and debiased cross-modal contrastive learning. *arXiv preprint arXiv:2311.03079*, 2022. URL <https://arxiv.org/abs/2207.01987>.
- [25] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary point-cloud object detection without 3D annotation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [26] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3D object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. URL <https://arxiv.org/abs/2109.08141>.
- [27] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *European Conference on Computer Vision (ECCV)*, ECCV'10, page 392–405, 2010. ISBN 3642155510.
- [28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Neural Information Processing Systems (NeurIPS)*, 2019.
- [29] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3D scene understanding with open vocabularies. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

- [30] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL <https://arxiv.org/abs/1612.00593>.
- [31] Charles R. Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas. Frustum PointNets for 3D object detection from RGB-D data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. URL <https://arxiv.org/abs/1711.08488>.
- [32] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3D object detection in point clouds. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. URL <https://arxiv.org/abs/1904.09664>.
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of Machine Learning Research (PMLR)*, 2021. URL <https://arxiv.org/abs/2103.00020>.
- [34] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. PointRCNN: 3D object proposal generation and detection from point cloud. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. URL <https://arxiv.org/abs/1812.04244>.
- [35] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: Point-voxel feature set abstraction for 3D object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [36] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. URL <https://arxiv.org/abs/1904.01766>.
- [37] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Sheng Zhao, Shuyang Cheng, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset, 2020.
- [38] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015. URL <https://arxiv.org/abs/1412.0767>.
- [39] Haiyang Wang, Chen Shi, Shaoshuai Shi, Meng Lei, Sen Wang, Di He, Bernt Schiele, and Liwei Wang. DSVT: Dynamic sparse voxel transformer with rotated sets. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [40] Weihang Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. CogVLM: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2024. URL <https://arxiv.org/abs/2311.03079>.

- [41] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3D scene graphs for language-grounded robot navigation. In *Robotics: Science and Systems (RSS)*, RSS2024, July 2024. doi: 10.15607/rss.2024.xx.077. URL <http://dx.doi.org/10.15607/RSS.2024.XX.077>.
- [42] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Neural Information Processing Systems (NeurIPS)*, 2024.
- [43] Hai Wu, Chenglu Wen, Wei Li, Xin Li, Ruigang Yang, and Cheng Wang. Transformation-equivariant 3D object detection for autonomous driving. In *The Thirty-Seven AAAI Conference on Artificial Intelligence (AAAI)*, 2023.
- [44] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, and Dacheng Tao. Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. URL <https://arxiv.org/abs/2306.15880>.
- [45] Xinhao Xiang and Jiawei Zhang. FusionViT: Hierarchical 3D object detection via LiDAR-camera vision transformer fusion. *arXiv preprint arXiv:2311.03620*, 2023. URL <https://arxiv.org/abs/2311.03620>.
- [46] Xinhao Xiang, Simon Dräger, and Jiawei Zhang. 3DiffFusionDet: Diffusion model for 3D object detection with robust LiDAR-camera fusion. *arXiv preprint arXiv:2311.0374*, 2023. URL <https://arxiv.org/abs/2311.03742>.
- [47] Xinhao Xiang, Simon Dräger, and Jiawei Zhang. EffiPerception: An efficient framework for various perception tasks. *arXiv preprint arXiv:2403.12317*, 2024. URL <https://arxiv.org/abs/2403.12317>.
- [48] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph R-CNN for scene graph generation. In *European Conference on Computer Vision (ECCV)*, 2018. URL <https://arxiv.org/abs/1808.00191>.
- [49] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary DETR with conditional matching. In *European Conference on Computer Vision (ECCV)*, page 106–122, 2022. ISBN 9783031200779. doi: 10.1007/978-3-031-20077-9_7. URL http://dx.doi.org/10.1007/978-3-031-20077-9_7.
- [50] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. URL <https://arxiv.org/abs/2011.10678>.
- [51] Dongmei Zhang, Chang Li, Ray Zhang, Shenghao Xie, Wei Xue, Xiaodong Xie, and Shanghang Zhang. FM-OV3D: Foundation model-based cross-modal knowledge blending for open-vocabulary 3D detection. In *The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, 2024.
- [52] Hu Zhang, Jianhua Xu, Tao Tang, Haiyang Sun, Xin Yu, Zi Huang, and Kaicheng Yu. OpenSight: A simple open-vocabulary framework for LiDAR-based object detection. In *European Conference on Computer Vision (ECCV)*, 2024.

- [53] Shiyu Zhao, Zhixing Zhang, Samuel Schuster, Long Zhao, Vijay Kumar B. G, Anastasis Sathopoulos, Manmohan Chandraker, and Dimitris Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *European Conference on Computer Vision (ECCV)*, 2022.
- [54] Chao Zhou, Yanan Zhang, Jiaxin Chen, and Di Huang. OcTr: Octree-based transformer for 3D object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [55] Chenming Zhu, Wenwei Zhang, Tai Wang, Xihui Liu, and Kai Chen. Object2Scene: Putting objects in context for open-vocabulary 3D detection. *arXiv preprint arXiv:2311.03079*, 2023. URL <https://arxiv.org/abs/2309.09456>.
- [56] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. PointCLIP V2: Prompting CLIP and GPT for powerful 3D open-world learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Towards Open-Vocabulary Multimodal 3D Object Detection with Attributes

Supplementary Material

1 Vocabulary Settings

Following the rule of open-vocabulary setting [5, 8, 14], we designed the vocabulary sets for the object and attribute detection, which are used during training. Following prior work [8, 14], the vocabulary sets contain all existing base and novel classes, as well as additional classes. The size of the vocabulary set determines the size of the text feature during the text-visual feature alignment, which further decides the size of spawned semantic space. At testing time, the vocabulary set is replaced only by the union of all the existing base and novel classes. Tab. 9 and Tab. 10 show the object and attribute vocabulary sets we used, respectively.

dataset	vocabulary set for object detection
nuScenes	Car, Construction vehicles, Trailer, Barrier, Bicycle Pedestrian, Truck, Bus, Motorcycle, Traffic cone Animal, Ambulance, Police, Pushable pullable object Debris, Bicycle rack
Argoverse 2	Regular Vehicle, Trailer, Bicycle, Pedestrian Truck, Bus, Motorcycle, Construction cone Animal, Bollard, Sign, Large vehicle Wheeled device, Stroller, Railed vehicle

Table 9: The vocabulary sets we used for open-vocabulary object detection on different datasets. Following the rule of open-vocabulary setting [5, 8, 14], our object vocabulary set contains all existing classes in base and novel objects as well as all rest object classes defined in nuScenes and Argoverse 2 dataset.

dataset	vocabulary set for attribute detection
OVAD	with rider, without rider, moving, standing, sitting lying down, parked, moving, stopped, in front of, behind, on the left of, on the right of

Table 10: The vocabulary set we used for open-vocabulary attribute detection on our OVAD dataset. The attributes colored with teal/violet/brown are the attributes exclusively associated with cycle/pedestrian/vehicle classes. Following the rule of open-vocabulary setting [5, 8, 14], our attribute vocabulary set contains all existing classes in base and novel attribute as well as all the rest of attribute classes defined in the OVAD dataset.

2 More Details on Dataset & Metrics

In nuScenes [9], we follow the official split of nuScenes, which contains 1000 driving scenes captured in complex urban environments, divided into 700 for training, 150 for validation, and 150 for testing. In Argoverse 2 [42], we use a similar train/val/test split (700/150/150) as nuScenes. As detailed in Section 1 of the supplementary, our vocabulary set includes all base and novel classes plus additional dataset-defined labels, following the OV setting protocols from [6, 8, 42]. During training, the full vocabulary is used for text embedding; during testing, only base and novel classes are retained.

In the task of Object Detection, we report:

- **mAP**: mean Average Precision with standard 3D IoU thresholds (0.2 for Argoverse 2, nuScenes follows official setup).
- **NDS**: nuScenes Detection Score, combining mAP with additional metrics such as translation, scale, orientation, velocity, and attribute accuracy.
- **AP_N**: mean AP computed only over novel classes to evaluate generalization under OV settings.

In the task of Attribute Detection, we use **Success Rate (SR)**:

- **SR (AD only)**: measures the percentage of correctly classified attribute labels among all attribute-annotated proposals.
- **SR (AD & OD)**: measures success rate conditioned on correct object category and localization.

The classification threshold is set to 0.5 for all attribute categories unless otherwise specified.

3 Implementation details

Implementing OVODA in PyTorch [48], we use AdamW optimizer with $\beta_1=0.9$, $\beta_2=0.95$, and weight decay of 0.1. We set the number of object queries to 128 for both nuScenes and Argoverse 2. Initially, we train a base 3DETR model for 20 epochs using only class-agnostic distillation. Then, the model continues to be trained for 20 epochs. The hyper-parameters used during training follow the default 3DETR configuration specified in [6, 46].

4 More Qualitative Results

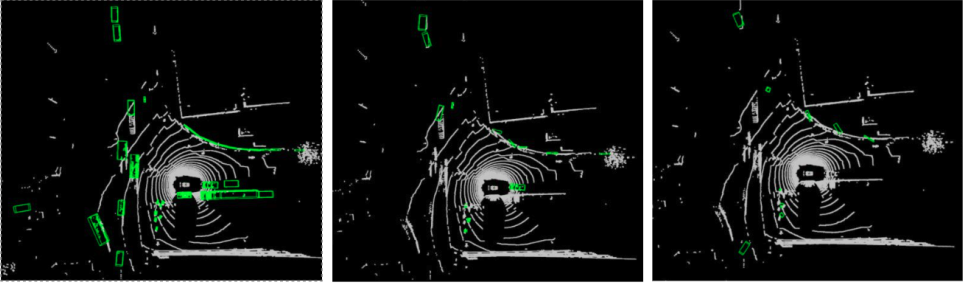


Figure 6: Qualitative comparison of OVODA (middle) versus CoDAv2 [9] (right) with the ground truth (left).

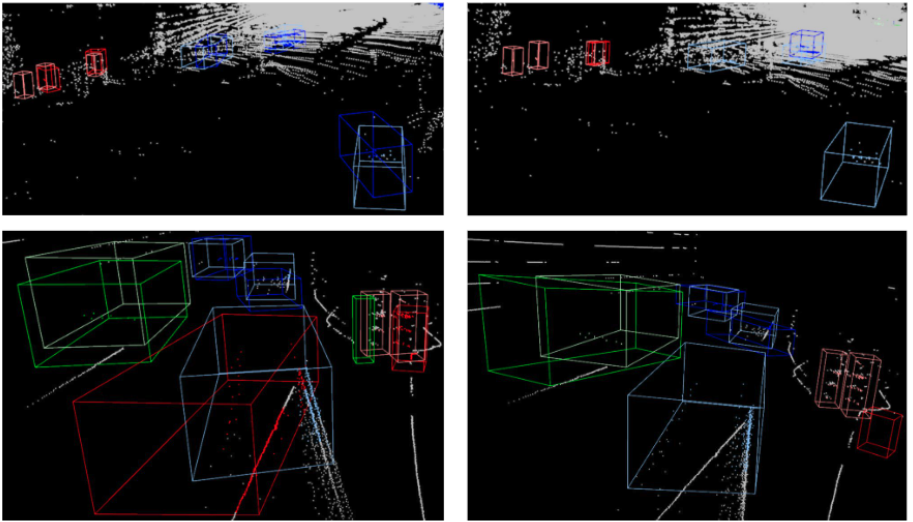


Figure 7: The qualitative comparison of OVODA (left) versus CoDAv2 [9] (right) for 3D single object detection in nuScenes dataset. The ground truth annotations are rendered in light blue/light red/light green for the class car/pedestrian/others, the predicted bounding boxes are rendered in blue/red/green for the class car/pedestrian/others.

Fig. 6 and Fig. 7 show the class-agnostic and class-specific qualitative comparison between OVODA and CoDAv2, respectively. Both qualitative results show that OVODA’s prediction is closer to the ground truth.

We show more qualitative comparisons between OVODA (**ours**) (left) versus CoDAv2 [9] (right) for 3D single object detection on the nuScenes dataset in Fig. 8, where OVODA’s prediction is closer to the ground truth compared with CoDAv2. We show more qualitative results of OVODA (**ours**) for 3D complex event detection on the nuScenes dataset in Fig. 9, where OVODA can successfully detect complex events.

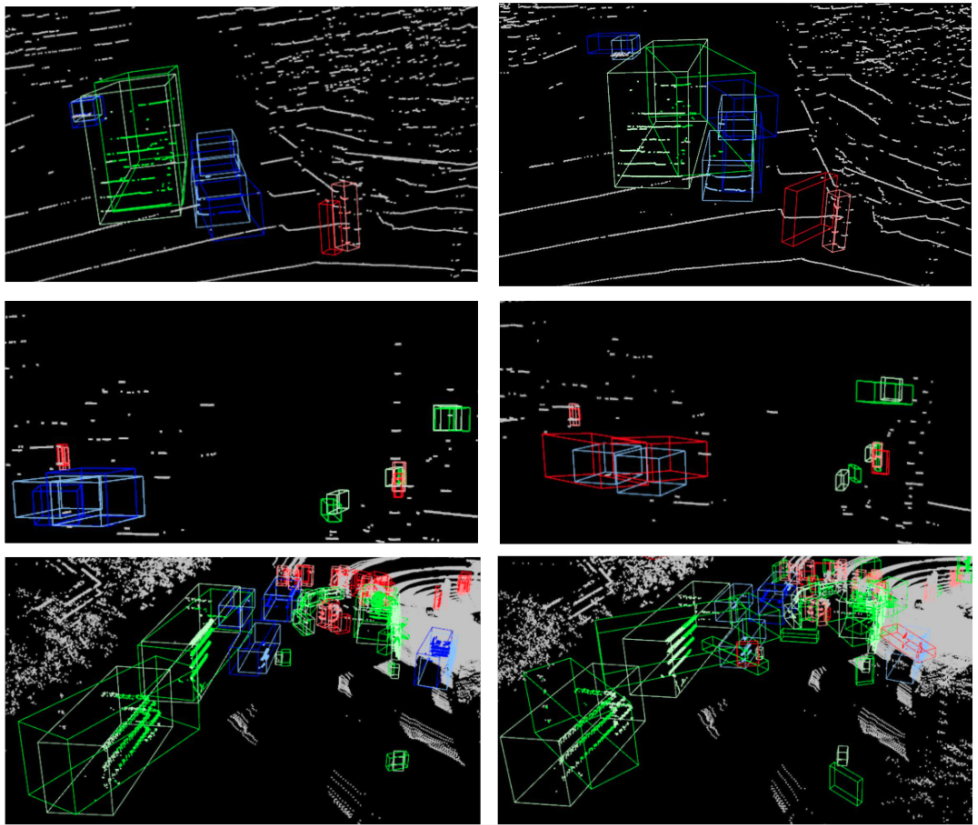


Figure 8: More qualitative comparison of OVODA (left) versus CoDAv2 [1] (right) for 3D single object detection in nuScenes dataset. The ground truth annotations are rendered in light blue/light red/light green for the class car/pedestrian/others, the predicted bounding boxes are rendered in blue/red/green for the class car/pedestrian/others.

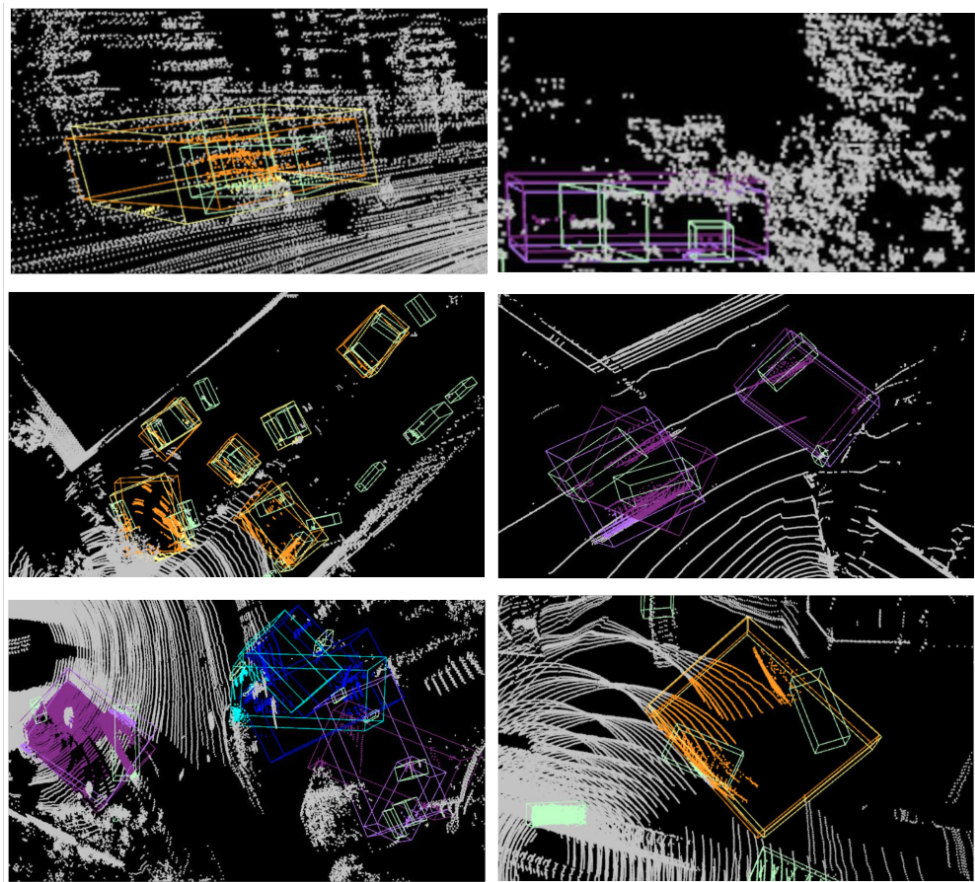


Figure 9: More qualitative results of OVODA for 3D complex event detection in nuScenes dataset. All ground truth annotations of single object are rendered in light green. The ground truth annotations are rendered in light purple/yellow/light blue for the car-car/pedestrian-pedestrian/others complex events, the predicted bounding boxes are rendered in purple/orange/blue for the car-car/pedestrian-pedestrian/others complex events. Examples of car-car/pedestrian-pedestrian/others complex events can be: a car in front of the car/a pedestrian on the left of the pedestrian/a cyclist behind a car.