SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

May 2026, Vol. 69, Iss. 5, 150106:1–150106:14
https://doi.org/10.1007/s11432-025-4853-5

Special Topic: Large Multimodal Models

# Multimodal 3D object detection for autonomous driving under vision-language supervision: a contrastive-learning perspective

Chunmian LIN[1,2,3], Wenze ZHANG[1,2,3], Yanyan CHEN[1,2,3], Lei YANG[4], Han JIANG[1,2,3], Daxin TIAN[1,2,3*], Xuting DUAN[1,2,3], Jianshan ZHOU[1,2,3] & Dongpu CAO[5]

[1]*State Key Laboratory of Intelligent Transportation System, Beijing 102206, China*
[2]*Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, Beijing 102206, China*
[3]*School of Transportation Science and Engineering, Beihang University, Beijing 102206, China*
[4]*School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore 569830, Singapore*
[5]*School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China*

Received 19 May 2025/Revised 26 September 2025/Accepted 8 December 2025/Published online 16 April 2026

**Abstract** Multimodal large language models (MLLMs) have been well acknowledged as the generalist across a broad spectrum of vision-language understanding tasks. Despite notable advancements, their potential for autonomous-driving perception remains largely underexplored. In response, we conduct an in-depth investigation of image-text-point interaction and propose a versatile paradigm of vision-language supervision (VLS) for 3D object detection, where multi-sensory proposals are primarily refined with meticulously-designed text-referred expression, and multimodal correspondences are further incorporated in a contrastive-learning manner. Moreover, VLS holds great advantages. (1) No complicated engineering. It could be seamlessly integrated into a camera-LiDAR 3D detector without troublesome hand-crafted engineering. (2) No extra computation. It provides auxiliary guidance only during training. (3) No additional data. It derives multimodal pairs from ground-truth label instead of a laborious annotation pipeline. Empirical study on publicly available KITTI and nuScenes benchmarks demonstrates the state-of-the-art detection performance against a wide span of counterparts, suggesting its effectiveness and advancement. We hope this work could pave a substantial path towards multimodal feature fusion and object detection for autonomous driving.

**Keywords** multimodal 3D detection, vision-language model, autonomous driving, contrastive learning, adapter

**Citation** Lin C M, Zhang W Z, Chen Y Y, et al. Multimodal 3D object detection for autonomous driving under vision-language supervision: a contrastive-learning perspective. Sci China Inf Sci, 2026, 69(5): 150106, https://doi.org/10.1007/s11432-025-4853-5

## 1 Introduction

The vigorous development of autonomous vehicles has driven an increased demand for comprehensive surrounding perception [1, 2]. Considering the inherent imaging principle, different sensors produce discrepant modalities with individual perceptual information, i.e., an image contains abundant semantic context but lacks object depth, while a point offers reliable geometry feature in a sparse format. Despite remarkable advancements in camera-based [3–8] and LiDAR-only 3D detection [9–13], single-modality deficiency still poses severe challenges for accurate and robust sensing.

Due to the complementary nature of sensors, camera-LiDAR 3D object detection has emerged with a broad spectrum of fusion strategies. Preliminarily, they are categorized into data-level [14, 15], feature-level [16–20], and proposal-level [21] fusion at different phases. Compared with time-consuming data fusion and error-accumulating proposal fusion, feature-level fusion typically adopts a two-stream modality-specific backbone for information encoding and combines region-of-interest (RoI) representations via a sum or concatenation operation, thus enabling an accuracy-speed balance. Recent advances on bird's-eye view (BEV) features [22, 23] and transformer [24] have established a new state-of-the-art benchmark, in which multi-source data is transformed into a unified format, and feature aggregation is conducted by the `QKV` attention mechanism [25, 26]. This benchmark has been viewed as the mainstream paradigm and has inspired a broad spectrum of explorations in architectural design and fusion manner [27–29]. Nonetheless, the more intuitive language signal is greatly overlooked in previous studies of 3D object detection.

* Corresponding author (email: dtian@buaa.edu.cn)

 

Multimodal large language model (MLLM) [30–32], or say, vision-language model (VLM), has marked a transformative advancement in artificial intelligence and demonstrated the profound potentials in complex vision-language dialogue [33, 34], interaction [35, 36], and understanding tasks [37, 38]. Contrastive language-image pre-training (CLIP) [39] pioneers the transferable visual feature directly from the raw language supervision by training from scratch on 400-million image-text pairs from the internet, being the go-to choice of vision encoder for building a multimodal foundation model. Subsequently, a wide array of pre-training recipes, architectural innovation, and fine-tuning adaptation has flourished successively, and its generalization has been confirmed in the realm of point cloud classification [40] and generation [41], scene understanding [42, 43], and grounding [44]. It is thus natural to ask how to unleash the emergence of CLIP for 3D object detection.

To this end, we delve into the first attempt to transfer the general knowledge from CLIP to the 3D perception domain and propose an elegant pipeline of vision-language contrastive-learning supervision (VLS) for multimodal 3D detection. More precisely, VLS is curated with a feature translator (FT), a feature adapter (FA), and feature registration (FR): the FT is responsible for aligning multi-sensor feature dimensions with the CLIP embedding space via channel transformation; the FA derives a batch of image-text and point-text pairs from the ground-truth (GT) annotations and adopts the frozen CLIP with lightweight adapters for constructing multimodal RoI correspondences under text-referred expression; and FR selects the most significant candidates for contrastive-learning loss calculation with GT boxes. In general, VLS contributes to several advantages.

- **No complicated engineering.** It could be seamlessly inserted into various multi-sensor 3D detectors to guide a better RoI feature for combination, alleviating tedious hand-crafted engineering.
- **No extra computation.** It only provides an auxiliary supervision for model training, without extra computation overhead during inference.
- **No additional data.** It derives image-text and point-text pairs directly from GT labels and does not conduct a time-consuming and laborious annotation pipeline.

An empirical study is conducted on the public-availably KITTI [1] and nuScenes [2] benchmarks, and without bells and whistles, our proposed method reports the state-of-the-art performance that substantially exceeds a broad range of counterparts, thus demonstrating its effectiveness and superiority. This work would open a new door toward multimodal feature fusion and object detection.

## 2 Related work

### 2.1 3D object detection

3D object detection for autonomous driving has attracted tremendous attention in both academia and industry, and we review the recent advances in camera-based, LiDAR-only, and camera-LiDAR 3D detection for simplicity. On the one hand, camera-based approaches have been widely studied due to their low cost, which lifts a 2D detector into 3D space and regresses object attributes in the perspective view [45], e.g., box coordinates, size, and angle. Contemporarily, multi-view object detection has been the mainstream paradigm, where the surrounding camera features are projected into BEV space via predicted depth distribution [5, 6, 46] or directly query image features by 3D-2D cross-attention operation [4, 7, 8]. Despite their outstanding results, the susceptible depth estimation poses a considerable challenge for robust perception. On the other hand, LiDAR-only detection has been popularized with its advantageous spatial geometry information, and current studies are mainly grouped into point-based [10, 47], voxel-based [13, 14], pillar-based [9], and range-based methods [48]. PointRCNN [47] is a two-stage point-based network that directly generates 3D candidates from the original point, followed by proposal refinement for semantic and spatial feature fusion at a fine-grained level. In contrast, a voxel-based detector quantizes the unordered point cloud into an even grid for efficiency. The fully sparse object detector (FSD) [13] formulates a fully-sparse detection pipeline on top of a general voxel encoder and a sparse instance recognition module, while the enhanced version FSDv2 [14] further introduces the concept of the virtual voxel to address the notorious empty voxel during quantization. Moreover, pillar-based and range-based approaches convert the raw points into 2D features and adopt an image backbone for feature encoding and object prediction. PointPillars [9] learns a point representation organized in vertical columns (pillars); RSN [48] predicts the foreground point from the range image and then detects potential objects via sparse convolution. However, the inherent sparsity of points unavoidably suffers from performance degradation in long-range or similar object detection.

Camera-LiDAR 3D detection is recognized as an effective solution to promote robust perception for autonomous vehicle, and previous studies are roughly categorized into data-level [14, 15], feature-level [16–20], and proposal-level fusion [21]. PointPainting [14] is a sequential data fusion pipeline by projecting LiDAR points into the output of an

image segmentation network and appending the class scores to each point, while PointAugmenting [15] decorates the point cloud with the corresponding point-wise convolution feature extracted by a pretrained 2D detector. Conversely, CLOCs [21] develops a low-complexity multi-sensor fusion network that produces better 2D-3D object candidates by leveraging semantic and geometric consistency and performs the combination before the post-processing operation. Considering the high-computational data fusion and accumulated-error proposal fusion, the feature-based method achieves satisfactory detection accuracy with an acceptable inference latency, thus receiving widespread attention from academia [49, 50]. Recent hot-spot interest lies in projecting multi-view images into a unified BEV space and adopting a cross-attention mechanism for modality-specific interaction [18, 22–24]. MV2DFusion [19] proposes a query-based fusion by constructing image and point query generators, which combine object semantics without biasing toward one modality. SimpleBEV [20] performs cascaded depth estimation and rectification assisted by LiDAR points, and an auxiliary camera-BEV branch is introduced under the guidance of camera configuration during model training. In a nutshell, camera-LiDAR fusion has flourished and demonstrated promising detection performance, but how to formulate a multimodal detection paradigm with an intuitive text description has remained underexplored.

### 2.2 Multimodal large language model

MLLM has experienced explosive growth on top of the recent success of large language model (LLM) [51], which integrates the pretrained visual model for vision-language comprehension and reasoning capabilities. The representative creations, i.e., Flamingo [30] and OpenFlamingo [33], incorporate novel gated cross-attention layers with a frozen LLM, conditioned on visual input. LLaVA [32] expands GPT-4 to generate multimodal knowledge by instruction-following data, while BLIP-2 [52] designs a querying transformer (Q-Former) to bridge the gap between vision and language. To unleash the power of visual-language representation learning, CLIP [39] pioneers a simple and scalable multimodal pre-training manner in which the transferable visual concept is learned directly from language supervision by jointly training image and text encoders to predict the correct pairs from a batch of samples, suggesting inspirational zero-shot adaptation to the downstream tasks. Followed by this, CLIP2Point [40] extends CLIP to 3D point classification by an image depth contrastive pre-training approach; CLIP2Scene [42] launches a semantic-driven cross modal contrastive learning framework that transfers CLIP knowledge from an image-text pretrained model to a 3D point network; ULIP [43] learns a unified representation of image, text, and 3D point by pre-training with object triplets from three modalities; and GLIP [44] presents a language-image pre-trained model that exploits massive image-text pairs to learn object-level, language-aware, and semantic-rich knowledge for visual detection and grounding. Another research line concentrates on parameter-efficient fine-tuning techniques, e.g., prompt-tuning [53], adapter [54], and low-rank [55], to explore their potential for zero-shot generality with marginal computational overhead. Despite their impressive performance over a wide range of visual applications, how to transfer CLIP to 3D object detection remains an open issue.

## 3 Methodology

### 3.1 Prerequisites

**Camera-LiDAR 3D detector.** Given the multi-view images $\mathcal{I} = \{I_1, I_2, \ldots, I_m\} \in \mathbb{R}^{H\times W\times 3}$ and point cloud $\mathcal{P} = \{p_1, p_2, \ldots, p_n\} \in \mathbb{R}^4$ as inputs, a common pipeline[1)] of multi-sensor 3D object detection is streamlined as the dual-branch framework where a modality-specific backbone $\{\mathcal{M}_\mathcal{I}, \mathcal{M}_\mathcal{P}\}$ is utilized to encode image and point features $\{\mathcal{F}_{I_i}, \mathcal{F}_{p_i}\}$, respectively, and a fusion module $\mathcal{M}_\mathcal{F}$ is then curated for key-point or RoI information aggregation $\mathcal{F}_{(\mathcal{I},\mathcal{P})}$, followed by the task head with object category $\mathcal{C}$ and 3D bounding-box coordinate $\mathcal{B} = (x, y, z, w, h, l, r) \in \mathbb{R}^7$.

**CLIP.** Conditioned on a batch of $N$ image-text pairs from the web, CLIP learns a multimodal transferable feature space by jointly training an image encoder $E_\mathcal{I}$ and text encoder $E_\mathcal{T}$, to maximize the cosine similarity of $N$ image-text embeddings $(\mathcal{O}_\mathcal{I}, \mathcal{O}_\mathcal{T})$ while minimizing the cosine similarity of the $N^2 - N$ incorrect pairs. The whole model is optimized by a symmetric cross-entropy loss over these scores, to measure multi-class $N$-pair similarity during contrastive learning.

1) For simplicity, we only discuss the feature-level fusion paradigm here.

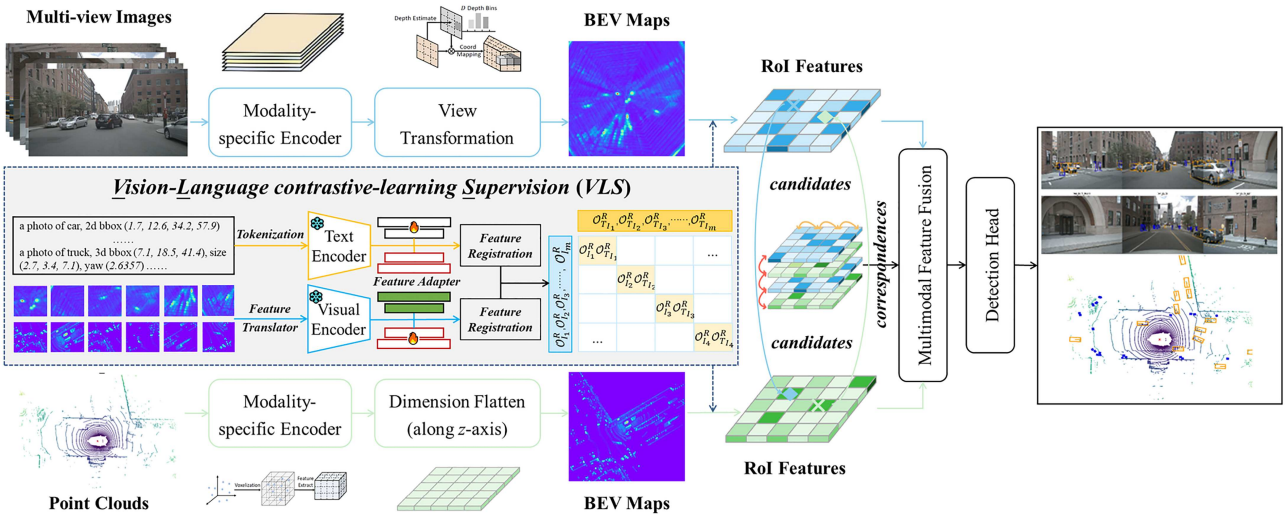


**Figure 1** (Color online) Architectural overview of a multimodal 3D detector with VLS. After modality-specific feature encoding, CLIP with a fine-tuning adapter is introduced during training, which constructs image-text and point-text pairs from GT label and refines multiple correspondences in a contrastive learning manner. Finally, the aligned candidates are fed into the multimodal feature fusion and detection head for prediction.

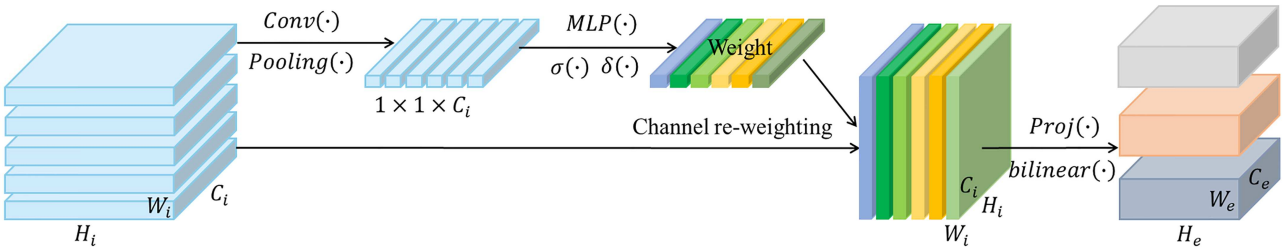


**Figure 2** (Color online) The overall architecture of FT, which experiences global average pooling to capture the contextual information and then re-weights channel semantics by the activated probability. Finally, projection and bilinear interpolation are utilized to adapt to the embedding space.

### 3.2 Overall architecture

In this work, we conduct an in-depth investigation of the CLIP-based 3D detector, and pioneer a simple yet versatile paradigm of VLS for multimodal 3D detection. The architectural overview is depicted in Figure 1. On the basis of image-point feature map $\{\mathcal{F}_{I_i}, \mathcal{F}_{p_i}\}$ from a camera-LiDAR detector, VLS is designed to provide auxiliary guidance to encourage the relevant instance cue closer while refining the RoI candidate with the text-referred expression. As expected, multimodal representation aligns with the targeted region under the guidance of a referred object, and we only introduce this assisted supervision for model training, thus alleviating the additional computational overhead during inference. More technical details are elaborated as follows.

### 3.3 Vision-language contrastive-learning supervision

In general, a pipeline of VLS includes three folds: the FT is prepared to align the feature dimension with CLIP embedding space; the FA provides batch-wise image-text or point-text pairs derived from the GT labels, thus constructing multimodal correspondences on the domain-specific samples via the frozen CLIP with an efficient adapter; and FR is further responsible for selecting the most significant candidates from the similarity matrix and calculating the contrastive-learning loss with GT boxes.

**FT.** It is unfeasible to feed modality-encoded feature $\mathcal{F}_{I_i}$ or $\mathcal{F}_{p_i}$ directly into CLIP due to dimension discrepancy. The nãive solution of brutal shape transformation might damage the intrinsic structure. In response, we design an FT module for adaptive weight rectification and global semantic exploitation, as shown in Figure 2. Taking $\mathcal{F}_{I_i} \in \mathbb{R}^{H_i \times W_i \times C_i}$ as example, it operates with $3 \times 3$ convolution and averages along the channel dimension via

global average pooling

$$\mathcal{F}_{I_i}^P = \frac{1}{H_i \times W_i} \sum_{h}^{H_i} \sum_{w}^{W_i} \mathrm{Conv}\left(\mathcal{F}_{I_i}(h,w)\right). \tag{1}$$

Then, we introduce the gating mechanism with a multi-layer perceptron and an activation function to produce channel-wise weight matrix $\mathcal{W}$, hence highlighting the most important information

$$\mathcal{W} = \delta\left(\mathbf{W}_2 \sigma\left(\mathbf{W}_1 \mathcal{F}_{I_i}^P\right)\right). \tag{2}$$

Finally, the feature map is rescaled to the original one by channel re-weighting, and we adapt it to the CLIP embedding space via projection and bilinear interpolation

$$\mathcal{F}_{I_i}^R = \mathcal{W} \cdot \mathrm{Conv}\left(\mathcal{F}_{I_i}(h,w)\right), \tag{3}$$

$$\mathcal{F}_{I_i}^T = \mathbf{B}\left(\mathbf{W}_3 \mathcal{F}_{I_i}^R\right) \in \mathbb{R}^{H_c \times W_c \times C_c}, \tag{4}$$

where $(\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3)$ implies the linear weight for projection, $\sigma(\cdot)$ and $\delta(\cdot)$ denote the ReLU and sigmoid functions, respectively, $\mathbf{B}$ implies bilinear interpolation operation, and $(H_c, W_c, C_c)$ is the desired dimension. In this way, the translated $\left(\mathcal{F}_{I_i} \to \mathcal{F}_{I_i}^T\right)$ and $\left(\mathcal{F}_{p_i} \to \mathcal{F}_{p_i}^T\right)$ maps are in accordance with the embedding space rationally, and we further enforce VLS with the adoption of CLIP and adapters.

**FA.** Although CLIP has demonstrated great advantages in zero-shot transferring capability to downstream tasks, full-parameter fine-tuning with relatively few samples on the target dataset is prone to suffer from the overfitting issue and performance drop. Therefore, on top of pretrained CLIP, a recipe of FA is curated with task-related object prompt and parameter-efficient fine-tuning module, which provides image/point-text pairs via text-referred guidance and finds the multimodal RoI correspondences for better fusion.

On the one hand, we extend the prompt template of CLIP with the semantic category and geometric location derived from 2D-3D GT annotations, which is expressed as $\mathbf{P}^{\mathrm{2D}}$ and $\mathbf{P}^{\mathrm{3D}}$:

`"a photo of a [category], 2d bbox [2D coordinate]"`

`"a photo of a [category], 3d bbox [3D coordinate], size [scale], yaw [orientation]"`

for image-text and point-text expressions, respectively, where `[category]` denotes the instance class, `[2D coordinate]` $= (x_1, y_1, x_2, y_2)$ stands for the left-top and right-bottom corner coordinates of 2D box, `[3d coordinate]`$= (x, y, z)$ indicates the centric coordinate of 3D box, `[scale]`$= (w, h, l)$ implies the 3D size, and `[orientation]`$= (r)$ is the heading angle. More details can be referred to Figure 3. In this way, both semantic and spatial information are encapsulated to guide modality-specific feature interaction across the de-facto object region, and we would evaluate its effectiveness in the empirical study.

On the other hand, an additional adapter is fine-tuned on both visual and text encoders of CLIP, in which a new feature is encoded and blended with the frozen one in a residual style. As described in Figure 4, a batch of image/point-text pairs is constructed from the GT annotation, respectively, and we feed them into a frozen CLIP to produce the pretrained feature correspondences $\left(\mathcal{O}_{I_i}, \mathcal{O}_{T_{I_i}}\right)$ and $\left(\mathcal{O}_{p_i}, \mathcal{O}_{T_{p_i}}\right)$:

$$\mathcal{O}_{I_i} = E_{\mathcal{I}}\left(\mathcal{F}_{I_i}^T\right) \in \mathbb{R}^{N \times D_I},\ \mathcal{O}_{p_i} = E_{\mathcal{I}}\left(\mathcal{F}_{p_i}^T\right) \in \mathbb{R}^{M \times D_P}, \tag{5}$$

$$\mathcal{O}_{T_{I_i}} = E_{\mathcal{T}}\left(\texttt{Tokenizer}\left(\mathbf{P}^{\mathrm{2D}}\right)\right) \in \mathbb{R}^{L_{\mathrm{2D}} \times D_I},\ \mathcal{O}_{T_{p_i}} = E_{\mathcal{T}}\left(\texttt{Tokenizer}\left(\mathbf{P}^{\mathrm{3D}}\right)\right) \in \mathbb{R}^{L_{\mathrm{3D}} \times D_P}. \tag{6}$$

Subsequently, adapter contains two linear layers $\mathbf{W}_{\mathrm{a1}}$ and $\mathbf{W}_{\mathrm{a2}}$ followed by the visual-text encoder, which outputs the fine-tuning map correspondences $\left(\mathcal{O}_{I_i}^A, \mathcal{O}_{T_{I_i}}^A\right)$ and $\left(\mathcal{O}_{p_i}^A, \mathcal{O}_{T_{p_i}}^A\right)$:

$$\mathcal{O}_{I_i}^A = \mathbf{W}_{\mathrm{a2}}\left(\mathbf{W}_{\mathrm{a1}} \mathcal{O}_{I_i}\right) \in \mathbb{R}^{N \times D_I},\ \mathcal{O}_{p_i}^A = \mathbf{W}_{\mathrm{a2}}\left(\mathbf{W}_{\mathrm{a1}} \mathcal{O}_{p_i}\right) \in \mathbb{R}^{M \times D_P}, \tag{7}$$

$$\mathcal{O}_{T_{I_i}}^A = \mathbf{W}_{\mathrm{a2}}\left(\mathbf{W}_{\mathrm{a1}} \mathcal{O}_{T_{I_i}}\right) \in \mathbb{R}^{L_{\mathrm{2D}} \times D_I},\ \mathcal{O}_{T_{p_i}}^A = \mathbf{W}_{\mathrm{a2}}\left(\mathbf{W}_{\mathrm{a1}} \mathcal{O}_{T_{p_i}}\right) \in \mathbb{R}^{L_{\mathrm{3D}} \times D_P}. \tag{8}$$

After that, we append a residual connection for information combination and adjust the degree of maintaining the frozen knowledge dynamically via a learnable adapter ratio $(\alpha, \beta)$:

$$\mathcal{O}_{I_i}^R = \alpha \mathcal{O}_{I_i}^A + (1-\alpha)\,\mathcal{O}_{I_i},\ \mathcal{O}_{p_i}^R = \alpha \mathcal{O}_{p_i}^A + (1-\alpha)\,\mathcal{O}_{p_i}, \tag{9}$$

$$\mathcal{O}_{T_{I_i}}^R = \beta \mathcal{O}_{T_{I_i}}^A + (1-\beta)\,\mathcal{O}_{T_{I_i}},\ \mathcal{O}_{T_{p_i}}^R = \beta \mathcal{O}_{T_{p_i}}^A + (1-\beta)\,\mathcal{O}_{T_{p_i}}. \tag{10}$$

| | |
|---|---|
| **$P_{2D}$**:<br>① a photo of ***Van***, 2d bbox (***0.00 142.54 336.56 374.00***)<br>② a photo of ***Car***, 2d bbox (***222.78 166.78 454.29 308.44***)<br>③ a photo of ***Car***, 2d bbox (***658.66 166.95 697.19 201.46***)<br>④ a photo of ***Car***, 2d bbox (***615.97 168.04 651.95 189.21***) | **$P_{2D}$**:<br>⑤ a photo of ***Pedestrian***, 2d bbox (***592.18 142.30 623.62 224.85***)<br>⑥ a photo of ***Pedestrian***, 2d bbox (***654.37 147.30 678.72 222.86***)<br>⑦ a photo of ***Pedestrian***, 2d bbox (***660.51 145.73 706.89 262.13***)<br>⑧ a photo of ***Pedestrian***, 2d bbox (***549.49 129.31 578.58 225.29***) |
| **$P_{3D}$**:<br>① a photo of ***Van***, 3d bbox (***-3.71 1.72 5.77***), size (***1.70 1.85 4.35***), yaw (***1.47***)<br>② a photo of ***Car***, 3d bbox (***-3.82 1.66 10.96***), size (***1.82 1.70 4.35***), yaw (***1.58***)<br>③ a photo of ***Car***, 3d bbox (***3.24 1.32 34.46***), size (***1.56 1.57 2.62***), yaw (***-1.38***)<br>④ a photo of ***Car***, 3d bbox (***1.72 1.13 51.25***), size (***1.68 1.43 4.19***), yaw (***1.82***) | **$P_{3D}$**:<br>⑤ a photo of ***Pedestrian***, 3d bbox (***0.05 0.92 14.98***), size (***0.60 1.71 0.52***), yaw (***1.27***)<br>⑥ a photo of ***Pedestrian***, 3d bbox (***1.26 0.89 15.27***), size (***0.46 1.59 0.67***), yaw (***1.30***)<br>⑦ a photo of ***Pedestrian***, 3d bbox (***1.14 1.17 10.61***), size (***0.45 1.68 0.75***), yaw (***1.22***)<br>⑧ a photo of ***Pedestrian***, 3d bbox (***-0.83 0.91 14.81***), size (***0.63 1.95 0.74***), yaw (***1.38***) |

**Figure 3** (Color online) Samples of semantic-geometric prompt derived from GT annotation.

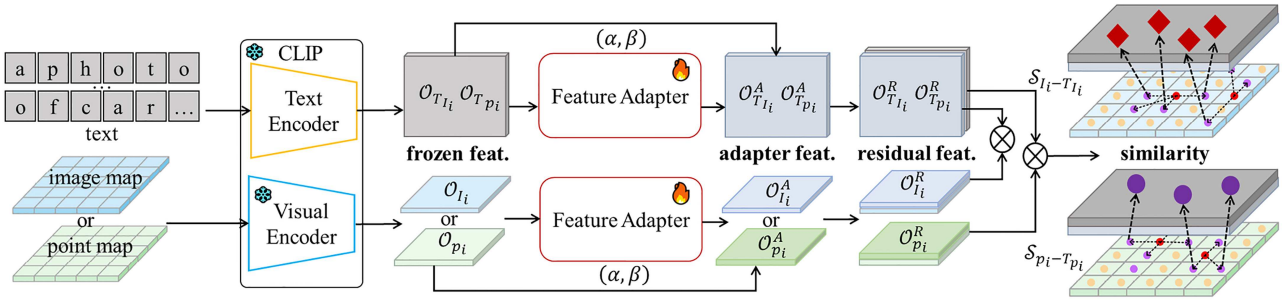


**Figure 4** (Color online) Diagram of the FA. It derives image-text and point-text pairs from the GT annotations, respectively, and produces the frozen and fine-tuned features with a group of lightweight adapters. We finally perform a residual-style summation and calculate the cosine similarity to find the multimodal correspondences.

Lastly, cosine similarity $\left(\mathcal{S}_{I_i-T_{I_i}}, \mathcal{S}_{p_i-T_{p_i}}\right)$ is calculated between each normalized image/point-text pair:

$$\begin{aligned} \mathcal{S}_{I_i-T_{I_i}} &= \frac{s \cdot \mathrm{Norm}(\mathcal{O}^R_{I_i}) \bigotimes \mathrm{Norm}(\mathcal{O}^R_{T_{I_i}})^{-1}}{||\mathrm{Norm}(\mathcal{O}^R_{I_i})|| \cdot ||\mathrm{Norm}(\mathcal{O}^R_{T_{I_i}})^{-1}||} \in \mathbb{R}^{N \times L_{2\mathrm{D}}}, \\ \mathcal{S}_{p_i-T_{p_i}} &= \frac{s \cdot \mathrm{Norm}(\mathcal{O}^R_{p_i}) \bigotimes \mathrm{Norm}(\mathcal{O}^R_{T_{p_i}})^{-1}}{||\mathrm{Norm}(\mathcal{O}^R_{p_i})|| \cdot ||\mathrm{Norm}(\mathcal{O}^R_{T_{p_i}})^{-1}||} \in \mathbb{R}^{M \times L_{3\mathrm{D}}}, \end{aligned} \tag{11}$$

where $s$ is the scaled factor, $\mathrm{Norm}(\cdot)$ implies normalization, $||\cdot||$ presents the vector norm, and $\bigotimes$ denotes a matrix-multiplication operation. Arguably, the introduction of a semantic-geometric prompt could be advantageous to locate the object candidates better via text-referred expression. More importantly, the adapter could further exploit general prior stored in the frozen CLIP and the freshly-updated knowledge originated from the target domain, being helpful to multimodal RoI feature interaction and refinement.

**FR.** For each sample, the RoI number is ordinarily larger than GT instances in the context of object detection, that is, neither $\mathcal{S}_{I_i-T_{I_i}}$ nor $\mathcal{S}_{p_i-T_{p_i}}$ is a symmetric square matrix ($N \neq L_{2\mathrm{D}}, M \neq L_{3\mathrm{D}}$), which might be causing the difficulty in optimization. Here, we develop an FR strategy to choose the most relevant candidates for contrastive-loss calculation. More precisely, a similarity matrix is activated element-wise by a softmax function, and we sort the probability values in descending order. According to GT objects, a new candidate matrix is constructed by

**Table 1** (Color online) Quantitative results of 3D detection on the KITTI val split. Our method is marked in gray. We signify the best result with bold-black font and accuracy gains with bold-blue font in brackets. And Mod. refers to the Moderate level.

| Method | Car ($AP_{R40}$ % ↑) | | | Pedestrian ($AP_{R40}$ % ↑) | | | Cyclist ($AP_{R40}$ % ↑) | | | mAP % ↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| | Easy | Mod. | Hard | Easy | Mod. | Hard | Easy | Mod. | Hard | |
| EPNet [57] | 88.76 | 78.65 | 78.32 | 66.74 | 59.29 | 54.82 | 83.88 | 65.60 | 62.70 | 70.96 |
| 3D-CVF [58] | 89.67 | 79.88 | 78.47 | – | – | – | – | – | – | – |
| CLOCs [21] | 89.49 | 79.31 | 77.36 | 62.88 | 56.20 | 50.10 | 87.57 | 67.92 | 63.67 | 70.50 |
| F-PointNet [59] | 83.76 | 70.92 | 63.65 | 70.00 | 61.32 | 53.59 | 77.15 | 56.49 | 53.37 | 65.58 |
| CAT-Det [60] | 90.12 | 81.46 | 79.15 | 74.08 | 66.35 | 58.92 | 87.64 | 72.82 | 68.20 | 75.42 |
| FocalsConv [61] | 92.26 | 85.32 | 82.95 | – | – | – | – | – | – | – |
| LoGoNet [62] | 92.04 | 85.04 | 84.31 | 70.20 | 63.72 | 59.46 | **91.74** | **75.35** | 72.42 | 77.14 |
| CL3D [49] | 90.32 | 83.20 | 78.91 | – | – | – | – | – | – | – |
| CL3D *w*.VLS | 91.89$_{(+1.57)}$ | 84.35$_{(+1.15)}$ | 79.87$_{(+0.96)}$ | – | – | – | – | – | – | – |
| 3D-DFM [50] | 91.94 | 84.90 | 82.48 | – | – | – | – | – | – | |
| 3D-DFM *w*.VLS | 93.07$_{(+1.13)}$ | 85.84$_{(+0.94)}$ | 83.31$_{(+0.83)}$ | – | – | – | – | – | – | – |
| VFF [56] | 92.31 | 85.51 | 82.92 | 73.26 | 65.11 | 60.03 | 89.40 | 73.12 | 69.86 | 76.94 |
| VFF *w*.VLS | **93.20**$_{(+0.89)}$ | **86.98**$_{(+1.47)}$ | **84.72**$_{(+1.80)}$ | **74.20** | **66.83** | **61.46** | 91.19 | 74.44 | **72.67** | **78.41** |

preserving the larger elements via the top-$k$ indices:

$$\begin{aligned} \mathrm{CM}_{I_i-T_{I_i}} &= \mathrm{TopK}\left(\mathrm{softmax}\left(\mathcal{S}_{I_i-T_{I_i}}\right)\right) \in \mathbb{R}^{L_{2\mathrm{D}}\times L_{2\mathrm{D}}}, \\ \mathrm{CM}_{p_i-T_{p_i}} &= \mathrm{TopK}\left(\mathrm{softmax}\left(\mathcal{S}_{p_i-T_{p_i}}\right)\right) \in \mathbb{R}^{L_{3\mathrm{D}}\times L_{3\mathrm{D}}}. \end{aligned} \tag{12}$$

Afterward, we obtain one-to-one matching relation between the GT and object proposal, and contrastive-learning loss based on the cross-entropy function is computed to maximize the similarity of positive samples on the diagonal while minimizing the remaining negative ones

$$\begin{aligned} \mathcal{L}_{I_i-T_{I_i}} &= \mathrm{CE}\left(\mathrm{CM}_{I_i-T_{I_i}}, \mathrm{GT}_{2\mathrm{D}}\right), \\ \mathcal{L}_{p_i-T_{p_i}} &= \mathrm{CE}\left(\mathrm{CM}_{p_i-T_{p_i}}, \mathrm{GT}_{3\mathrm{D}}\right), \end{aligned} \tag{13}$$

where $\mathrm{CE}(\cdot)$ symbolizes cross-entropy function, $\mathrm{GT}_{2\mathrm{D}}$ and $\mathrm{GT}_{3\mathrm{D}}$ denote 2D and 3D GT labels.

## 4 Experiments

### 4.1 Implementation details

**Dataset.** KITTI [1], which has been the long-renowned dataset for autonomous-driving perception, provides 7481 and 7518 samples across three object categories (i.e., Car, Pedestrian, and Cyclist) for training and test, respectively. Following the official setting, we report the average precision of 40 interpolated recall points $\mathrm{AP}_{R40}$ on Easy, Moderate, and Hard levels to measure the detection performance.

nuScenes [2] is one of the most prevalent benchmarks for autonomous driving. It is acquired by six surrounding cameras and one 32-beam LiDAR from different countries. The whole dataset covers 1000 scenes across 10 different categories with 1.4 million 3D boxes, and we split 700 samples for training, 150 examples for validation, and 150 data for testing. As commonly done, mean average precision (mAP) and nuScenes detection score (NDS) metrics over 10 classes are employed for performance evaluation.

**Experimental setup.** In our experiment, we select camera-LiDAR detectors across a broad span, e.g., CL3D [49], 3D-DFM [50], VFF [56], BEVFusion [22], CMT [26], and SparseFusion [27], as the baselines and incorporate with VLS for performance evaluation. The pretrained CLIP with a ViT-B/32 backbone is utilized, featuring a 512-dim embedding space, 49408-size vocabulary, and 77-length context, respectively, and the dimensionality of the visual and text adapters is set to 1024. Unless otherwise specified, all experiments are conducted on 4 NVIDIA A6000 GPUs, and the overall objective comprises detection loss $\mathcal{L}_{det}$ and contrastive-learning loss ($\mathcal{L}_{I_i-T_{I_i}}$ and $\mathcal{L}_{p_i-T_{p_i}}$). Notably, we follow the default hyper-parameter settings of the individual detector, thus incurring admissible training overhead and time, i.e., CL3D [49] with 80 epochs for approximately 3 days.

### 4.2 Experimental analysis

**Results on KITTI.** VLS is integrated with various detectors for performance evaluation, and the quantitative results on the KITTI val set are tabulated in Table 1. Obviously, it presents a remarkable performance improvement

**Table 2** (Color online) Quantitative results of 3D detection on the KITTI test split. Our method is marked in gray. We signify the best result with bold-black font and accuracy gains with bold-blue font in brackets. And Mod. refers to the Moderate level.

| Method | Car ($AP_{R40}$ % ↑) | | |
|---|---|---|---|
| | Easy | Mod. | Hard |
| PointPainting [14] | 82.11 | 71.70 | 67.08 |
| EPNet [57] | 89.81 | 79.28 | 74.59 |
| 3D-CVF [58] | 89.20 | 80.05 | 73.11 |
| CLOCs [21] | 88.94 | 80.67 | 77.15 |
| SFD [16] | 91.73 | 84.76 | 77.92 |
| LoGoNet [62] | 91.80 | 85.06 | **80.74** |
| CL3D [49] | 87.45 | 80.28 | 76.21 |
| CL3D *w.*VLS | 88.58$_{(+1.13)}$ | 82.90$_{(+2.78)}$ | 78.57$_{(+2.36)}$ |
| 3D-DFM [50] | 87.75 | 80.93 | 76.12 |
| 3D-DFM *w.*VLS | 91.03$_{(+3.28)}$ | 84.36$_{(+3.43)}$ | 80.78$_{(+4.66)}$ |
| VFF [56] | 89.50 | 82.09 | 79.29 |
| VFF *w.*VLS | **92.09**$_{(+2.59)}$ | **85.45**$_{(+3.36)}$ | 80.68$_{(+1.39)}$ |

**Table 3** (Color online) Quantitative results of 3D detection on the nuScenes val split. Our method is marked in gray. We signify the best result with bold-black font and accuracy gains with bold-blue font in brackets.

| Method | NDS ↑ | mAP ↑ |
|---|---|---|
| UVTR [63] | 0.702 | 0.654 |
| TransFusion [25] | 0.713 | 0.675 |
| BEVFusion [23] | 0.721 | 0.696 |
| DeepInteraction [18] | 0.726 | 0.699 |
| BEVFusion4D [28] | 0.735 | 0.720 |
| AutoalignV2 [29] | 0.712 | 0.671 |
| FusionFormer [64] | 0.741 | 0.714 |
| BEVFusion [22] | 0.714 | 0.685 |
| BEVFusion *w.*VLS | 0.741$_{(+0.027)}$ | 0.719$_{(+0.034)}$ |
| CMT [26] | 0.729 | 0.703 |
| CMT *w.*VLS | 0.755$_{(+0.026)}$ | 0.732$_{(+0.029)}$ |
| SparseFusion [27] | 0.728 | 0.704 |
| SparseFusion *w.*VLS | **0.747**$_{(+0.019)}$ | **0.736**$_{(+0.032)}$ |

on multimodal 3D detection. For Car detection, VLS offers 1.57%, 1.15% and 0.96% $AP_{R40}$ increases against the baseline CL3D in the Easy, Mod., and Hard levels, respectively. When combined with VFF, it also makes substantial progress on Pedestrian and Cyclist detection at three difficulties and improves the overall mAP by 1.47%, which demonstrates the effectiveness of our proposed VLS. As for other counterparts, VFF *w.*VLS outperforms the top-tier LoGoNet [62] over 1.17% mAP and reports stable accuracy gains among all classes, further suggesting its superiority.

Furthermore, detection results on the KITTI test set are listed in Table 2, and we only report Car $AP_{R40}$ for comparison. Apparently, CL3D *w.*VLS and 3D-DFM *w.*VLS exhibit a marked upgrade over the baseline; the best-performing VFF *w.*VLS achieves 92.09%, 85.45%, and 80.68% Car $AP_{R40}$ and surpasses the other alternatives by a visible margin, particularly in the Easy and Mod. levels. These findings conformably manifest the considerable advantage of our proposed VLS in multimodal 3D detection.

**Results on nuScenes.** We combine VLS with BEVFusion [22], CMT [26], and SparseFusion [27] to measure the detection performance on the nuScenes val split. As shown in Table 3, VLS introduces a notable performance enhancement with respect to different baselines, e.g., a growth of 0.027 NDS and 0.034 mAP for BEVFusion [22], and a rising of 0.026 NDS and 0.029 mAP for CMT [26]. In particular, SparseFusion [27] *w.*VLS provides a competitive 0.747 NDS and 0.736 mAP among 10 categories, which exceeds all competitors by a considerable margin.

Moreover, we further evaluate the methods on the nuScenes test split, as elaborated in Table 4. With the help of vision-language guidance, both BEVFusion and SparseFusion receive a substantial promotion and showcase the first-class detection performance. Especially, CMT *w.*VLS reports the state-of-the-art performance with an NDS of 0.760 and a mAP of 0.735, advancing its baseline by 0.019 NDS and 0.015 mAP, respectively. Moreover, it consistently outperforms competing rivals across all metrics, achieving gains of 0.009 NDS/mAP over FusionFormer [64] and 0.013 NDS and 0.002 mAP over BEVFusion4D. These findings congruously uncover the generality and advancement of multimodal feature supervision.

**Table 4** (Color online) Quantitative results of 3D detection on the nuScenes test split. Our method is marked in gray. We signify the best result with bold-black font and accuracy gains with bold-blue font in brackets.

| Method | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| PointPainting [14] | 0.610 | 0.541 | 0.380 | 0.260 | 0.541 | 0.293 | 0.131 |
| UVTR [63] | 0.711 | 0.671 | 0.306 | 0.245 | 0.351 | 0.225 | 0.124 |
| TransFusion [25] | 0.717 | 0.689 | 0.259 | 0.243 | 0.359 | 0.288 | 0.127 |
| BEVFusion [23] | 0.733 | 0.713 | 0.250 | 0.240 | 0.359 | 0.254 | 0.132 |
| DeepInteraction [18] | 0.734 | 0.708 | 0.257 | 0.240 | 0.325 | 0.245 | 0.128 |
| BEVFusion4D [28] | 0.747 | 0.733 | – | – | – | – | – |
| AutoAlignV2 [29] | 0.724 | 0.684 | **0.245** | **0.233** | 0.311 | 0.258 | 0.133 |
| FusionFormer [64] | 0.751 | 0.726 | 0.267 | 0.236 | 0.286 | 0.225 | **0.105** |
| BEVFusion [22] | 0.729 | 0.702 | 0.261 | 0.239 | 0.329 | 0.260 | 0.134 |
| BEVFusion *w*.VLS | 0.748 (+0.019) | 0.730 (+0.028) | 0.281 | 0.245 | **0.250** | 0.205 | 0.191 |
| CMT [26] | 0.741 | 0.720 | 0.279 | 0.235 | 0.308 | 0.259 | 0.112 |
| CMT *w*.VLS | **0.760** (+0.019) | 0.735 (+0.015) | 0.253 | 0.241 | 0.272 | **0.193** | 0.118 |
| SparseFusion [27] | 0.738 | 0.720 | 0.258 | 0.243 | 0.329 | 0.265 | 0.131 |
| SparseFusion *w*.VLS | 0.753 (+0.015) | **0.742** (+0.022) | 0.277 | 0.247 | 0.258 | 0.215 | 0.190 |

**Qualitative results.** On the one hand, we make a deeper exploration of where the model attends under the guidance of VLS and provide an intuitive heatmap visualization for clarification. As exhibited in Figure 5, BEVFusion *w*.VLS could pay more attention to the instance region while suppressing the background information, which demonstrates its effectiveness in multimodal alignment and fusion.

On the other hand, the visualization of detection results is accomplished by BEVFusion *w*.VLS on the nuScenes val set for a transparent analysis. As presented in Figure 6, this set covers various driving scenes under different physical conditions, e.g., a crowded street, a construction site, a rainy day, and a low-illumination situation. Nonetheless, BEVFusion *w*.VLS realizes a prominent detection performance across various object categories and scales, implicitly suggesting its advancement and robustness.

### 4.3 Ablation study

Unless specifically stated, BEVFusion [22] is chosen as the fundamental detector, and we conduct an ablation study on the nuScenes val set to analyze the performance contribution of each component in vision-language contrastive-learning supervision.

**Analysis of object prompt.** Conditioned on the remaining architecture of VLS, we investigate the impact of the semantic-geometric prompt derived from GT label and the comparison with different variants, as shown in Table 5. It is observed that the detector can benefit from the semantic class, that is, the CLIP template, and geometric information might be much more favorable for 3D detection due to the considerable performance promotion. With the adoption of semantic and geometric cues, our proposed task-related object prompt further boosts the detection accuracy. These findings verify the effectiveness of prompt design and the importance of spatial attribute for 3D detection task.

**Analysis of fine-tuning technique.** We explore the effect of CLIP with various fine-tuning techniques, e.g., CoOp [53], LoRA [55], and an adapter, for multimodal 3D detection. The baseline is to retrain a frozen CLIP in a full-parameter manner, and it reports a slight accuracy drop, as depicted in Table 6, i.e., 0.712 versus 0.714 in NDS and 0.680 versus 0.685 in mAP. We speculate that the general knowledge from pretrained CLIP is inappropriate for autonomous-driving perception due to the severe domain gap. It is insightful that the parameter-efficient fine-tuning technique could provide the uncommon performance gains, and CLIP with the adapter presents 0.029 NDS and 0.039 mAP boosts, respectively, demonstrating the advantage of our proposed method.

**Analysis of feature adapter.** We further probe into the location and number of adapter module and choose a frozen CLIP as the baseline. As illustrated in Table 7, inserting adapter into the visual or text encoder is profitable for maintaining the pretrained priors and blending with the domain-specific knowledge; both visual and text encoders with adapter lead to the distinguished NDS and mAP risings, which conforms to our intuition. Moreover, we make a thorough inquiry of the optimal number of adapter in Table 8. 1-layer adapter offers insufficient transferring ability and network capacity for information fusion, while excess layers (e.g., 3-layer or 4-layer) might suffer from the overfitting problem and accuracy degradation. Empirically, CLIP with 2-layer adapter is the perfect choice for vision-text integration across domains and eventually displays a promising performance advancement.

**Analysis of contrastive learning.** Finally, we study the effectiveness of the contrastive-learning paradigm for

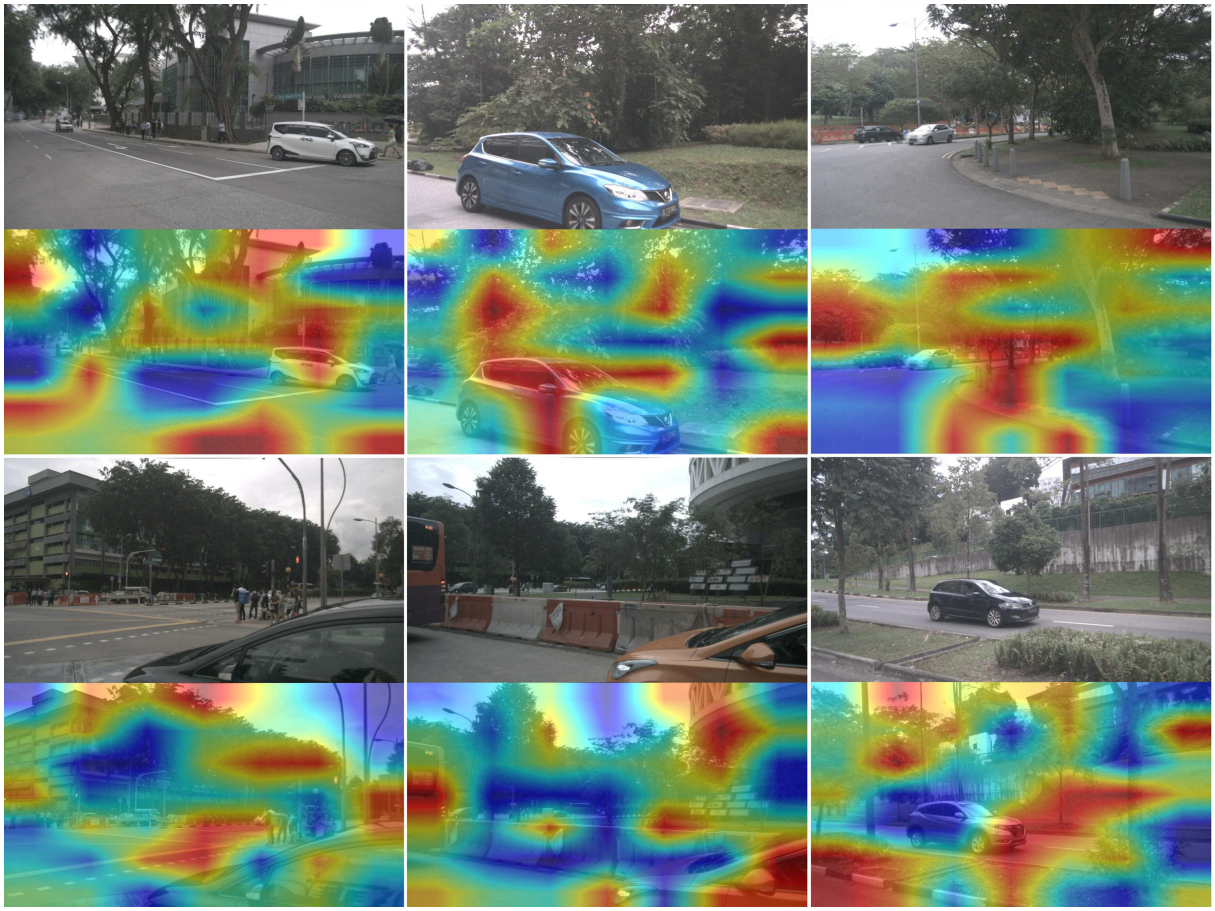

**Figure 5** (Color online) Visualization of feature heatmaps activated by BEVFusion [22] on the nuScenes val split. Under the guidance of VLS, the model could concentrate on object semantics and details while suppressing the background region.

**Table 5** (Color online) Ablation study on various semantic and geometric prompts. Our method is marked in gray. We signify the best result with bold-black font and accuracy gains with bold-blue font in brackets.

| *w*.prompt | | NDS ↑ | mAP ↑ |
|---|---|---|---|
| Semantic | Geometric | | |
| | | 0.714 | 0.685 |
| ✓ | | 0.720 (+0.006) | 0.696 (+0.010) |
| | ✓ | 0.726 (+0.012) | 0.703 (+0.018) |
| ✓ | ✓ | **0.741** (+0.027) | **0.719** (+0.034) |

**Table 6** (Color online) Ablation study on different fine-tuning techniques. Our method is marked in gray. We signify the best result with bold-black font and accuracy gains with bold-blue font in brackets.

| *w*.fine-tuning | | | NDS ↑ | mAP ↑ |
|---|---|---|---|---|
| CoOP | LoRA | Adapter | | |
| | | | 0.712 | 0.680 |
| ✓ | | | 0.724 (+0.012) | 0.699 (+0.019) |
| | ✓ | | 0.732 (+0.020) | 0.708 (+0.028) |
| | | ✓ | **0.741** (+0.029) | **0.719** (+0.039) |

multimodal fusion. For generalization, BEVFusion [22] and CMT [26] are selected as the baselines, and we compare the contributions of contrastive learning with a simple fine-tuned detector. As shown in Table 9, the introduction of text prompt with fine-tuned technique could provide a considerable performance upgrade, e.g., 0.012 NDS and 0.014 mAP gains for BEVFusion, 0.007 NDS and 0.017 mAP for CMT, while contrastive learning surpasses the counterpart by a substantial margin. This advancement demonstrates its superiority in vision-language-point feature combination.

**Figure 6** (Color online) Visualization of detection results achieved by BEVFusion [22] *w*.VLS on nuScenes val set across a diversity of driving scenarios and weather conditions, i.e., driving on the crowded street in the rainy day, and robust perception can be distinctly observed.

**Table 7** (Color online) Ablation study on adapter location. Our method is marked in gray. We signify the best result with bold-black font and accuracy gains with bold-blue font in brackets.

| *w*.adapter | | NDS ↑ | mAP ↑ |
|---|---|---|---|
| Vision ($E_{\mathcal{I}}$) | Text ($E_{\mathcal{T}}$) | | |
| | | 0.712 | 0.680 |
| ✓ | | 0.732$_{(+0.020)}$ | 0.708$_{(+0.028)}$ |
| | ✓ | 0.726$_{(+0.014)}$ | 0.703$_{(+0.023)}$ |
| ✓ | ✓ | **0.741**$_{(+0.029)}$ | **0.719**$_{(+0.039)}$ |

**Table 8** Ablation study on the optimal number of adapter. Our method is marked in gray, and we signify the best result with bold-black font.

| *w*.adapter num | | | | NDS ↑ | mAP ↑ |
|---|---|---|---|---|---|
| 1-layer | 2-layer | 3-layer | 4-layer | | |
| ✓ | | | | 0.721 | 0.695 |
| | ✓ | | | **0.741** | **0.719** |
| | | ✓ | | 0.734 | 0.711 |
| | | | ✓ | 0.723 | 0.702 |

**Table 9** (Color online) Ablation study on contrastive-learning paradigm. We signify the best result with bold-black font and accuracy gains with bold-blue font in brackets. o: original; ft: fine-tuned technique; cl: contrastive-learning pipeline.

| Method | BEVFusion [22] | | | CMT [26] | | |
|---|---|---|---|---|---|---|
| | *w*.o | *w*.ft | *w*.cl | *w*.o | *w*.ft | *w*.cl |
| NDS ↑ | 0.714 | 0.726$_{(+0.012)}$ | **0.741**$_{(+0.027)}$ | 0.729 | 0.736$_{(+0.007)}$ | **0.755**$_{(+0.026)}$ |
| mAP ↑ | 0.685 | 0.699$_{(+0.014)}$ | **0.719**$_{(+0.034)}$ | 0.703 | 0.720$_{(+0.017)}$ | **0.732**$_{(+0.029)}$ |

## 5 Discussion

Despite impressive performance, an in-depth investigation of model robustness, prompt generalization, and performance gain is left for future work.

• **Robustness to label noise.** Label noise is indeed a common challenge for 3D detection, where imperfect bounding boxes or class labels may propagate noise into the generated text prompts. Despite the inherent robustness

of CLIP pretrained on large-scale web data with noise, how the noisy label influences detection performance deserves to be studied in future work.

• **Generalization to prompt design.** In our implementation, a fixed-template prompt is designed from GT box with object semantic and geometric attributes. We argue that the intrinsic generalization of CLIP would be beneficial for unseen object or scenario detection, and the adapter module would absorb general knowledge and specific pattern when encountering a new category. An in-depth investigation of open-vocabulary or open-world 3D detection would be a promising direction in the future.

• **Classification or localization gain.** The proposed VLS incorporates semantic category with geometric localization for referred object detection, and several experiments have been conducted to validate its effectiveness, e.g., various NDS metric and semantic-geometric ablation studies. Nevertheless, the disentanglement between classification and localization gains from VLS is not yet fully conclusive, and we consider this important issue in the subsequent study.

## 6 Conclusion

In this work, we propose a general paradigm of VLS for multimodal 3D object detection. It comprises an FT to align multi-sensor feature dimension with CLIP embedding space, an FA to construct multimodal RoI correspondence via text-referred expression, and an FR to select the most significant candidate for contrastive-learning loss calculation with GT boxes. Empirical study on the KITTI and nuScenes benchmarks demonstrates its effectiveness and advancement: when incorporated with various detectors, e.g., CL3D, 3D-DFM, and BEVFusion, our proposed method further provides substantial performance gains and outperforms its counterparts by a remarkable margin across a wide span of driving scenarios. Ablation study further verifies the contribution of each component. We hope this work could shed light on multimodal feature fusion and object detection in the future.

**Acknowledgements** This work was supported by National Natural Science Foundation of China (Grant Nos. 24B20127, U2433202, 52402394, T2588101), Young Elite Scientists Sponsorship Program by China Association for Science and Technology (CAST) (Grant No. 2023QNRC001), China Postdoctoral Science Foundation (Grant No. 2025T181118), Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (Grant No. JYB2025XDXM123), Open Projects Program of State Key Laboratory of Multimodal Artificial Intelligence Systems, and Open Foundation of the State Key Laboratory of Precision Space-time Information Sensing Technology (Grant No. STL2023-B-11-01(D)).

**References**

1 Geiger A, Lens Z, Urtasun R. Are we ready for autonomous driving? The kitti vision benchmark suite. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2012. 3354–3361
2 Caesar H, Bankiti V, Lang A H, et al. NuScenes: a multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 11621–11631
3 Reading C, Harakeh A, Chae J, et al. Categorical depth distribution network for monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 11621–11631
4 Wang Y, Guizilini V C, Zhang T Y, et al. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries. In: Proceedings of the 5th Conference on Robot Learning, 2022. 180–191
5 Huang J J, Huang G, Zhu Z, et al. BEVDet: high-performance multi-camera 3D object detection in bird-eye-view. 2021. ArXiv:2112.11790
6 Liu Y F, Wang T C, Zhang X Y, et al. PETR: position embedding transformation for multi-view 3D object detection. In: Proceedings of the European Conference on Computer Vision, 2022. 531–548
7 Li Z Q, Wang W H, Li H Y, et al. BEVFormer: learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: Proceedings of the European Conference on Computer Vision, 2022. 1–18
8 Wang Z T, Huang Z H, Fu J H, et al. Object as query: lifting any 2D object detector to 3D detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 3791–3800
9 Lang A H, Vora S, Caesar H, et al. PointPillars: fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 12697–12705
10 Shi S S, Guo C X, Jiang L, et al. PV-RCNN: point-voxel feature set abstraction for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 10529–10538
11 Chen Y L, Yu Z D, Chen Y K, et al. FocalFormer3D: focusing on hard instances for 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 8394–8405
12 Fan L, Wang F, Wang N Y, et al. Fully sparse 3D object detection. In: Proceedings of the International Conference on Neural Information Processing Systems, 2022. 351–363
13 Fan L, Wang F, Wang N, et al. FSD V2: improving fully sparse 3D object detection with virtual voxels. IEEE Trans Pattern Anal Mach Intell, 2025, 47: 1279–1292
14 Vora S, Lang A H, Helou B, et al. PointPainting: sequential fusion for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 4604–4612

15 Wang C W, Ma C, Zhu M, et al. PointAugmenting: cross-modal augmentation for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 11794–11803

16 Wu X P, Peng L, Yang H H, et al. Sparse fuse dense: towards high quality 3d detection with depth completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 5418–5427

17 Wu H, Wen C L, Shi S S, et al. Virtual sparse convolution for multimodal 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 21653–21662

18 Yang Z Y, Chen J Q, Miao Z W, et al. Deepinteraction: 3D object detection via modality interaction. In: Proceedings of the International Conference on Neural Information Processing Systems, 2022. 1992–2005

19 Wang Z T, Huang Z H, Gao Y L, et al. MV2DFusion: leveraging modality-specific object semantics for multi-modal 3d detection. 2024. ArXiv:2408.05945

20 Zhao Y, Gong Z, Zheng P R, et al. SimpleBEV: improved lidar-camera fusion architecture for 3D object detection. 2024. ArXiv:2411.05292

21 Pang S, Morris D, Radha H, et al. CLOCs: camera-lidar object candidates fusion for 3D object detection. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2020. 10386–10393

22 Liu Z J, Tang H T, Amini A, et al. BEVFusion: multi-task multi-sensor fusion with unified bird's-eye view representation. In: Proceedings of the IEEE International Conference on Robotics and Automation, 2023. 1–8

23 Liang T T, Xie H W, Yu K C, et al. BEVFusion: a simple and robust lidar-camera fusion framework. In: Proceedings of the International Conference on Neural Information Processing Systems, 2022. 10421–10434

24 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of the International Conference on Neural Information Processing Systems, 2017. 6000–6010

25 Bai X Y, Hu Z Y, Zhu X G, et al. TransFusion: robust lidar-camera fusion for 3D object detection with transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022. 1090–1099

26 Yan J J, Liu Y F, Sun J J, et al. Cross modal transformer: towards fast and robust 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 18268–18278

27 Xie Y C, Xu C F, Rakotosaona M J, et al. SparseFusion: fusing multi-modal sparse representations for multi-sensor 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 17591–17602

28 Cai H X, Zhang Z Y, Zhou Z Y, et al. BEVFusion4D: learning lidar-camera fusion under bird's-eye-view via cross-modality guidance and temporal aggregation. 2023. ArXiv:2303.17099

29 Chen Z H, Li Z Y, Zhang S Q, et al. Deformable feature aggregation for dynamic multi-modal 3D object detection. In: Proceedings of the European Conference on Computer Vision, 2022. 628–644

30 Alayrac J B, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning. In: Proceedings of the International Conference on Neural Information Processing Systems, 2022. 23716–23736

31 Google Gemini Team. Gemini: a family of highly capable multimodal models. 2023. ArXiv:2312.11805

32 Liu H T, Li C Y, Wu Q Y, et al. Visual instruction tuning. In: Proceedings of the International Conference on Neural Information Processing Systems, 2023. 34892–34916

33 Awadalla A, Gao I, Gardner J, et al. OpenFlamingo: an open-source framework for training large autoregressive vision-language models. 2023. ArXiv:2308.01390

34 Wang W H, Lv Q S, Yu W M, et al. CogVLM: visual expert for pretrained language models. 2023. ArXiv:2311.03079

35 Chen Z, Wu J N, Wang W H, et al. InternVL: scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 24185–24198

36 Reid M, Savinov N, Teplyashin D, et al. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. 2024. ArXiv:2403.05530

37 Zhu D Y, Chen J, Shen X Q, et al. MiniGPT-4: enhancing vision-language understanding with advanced large language models. In: Proceedings of the International Conference on Neural Information Processing Systems, 2023. 34892–34916

38 Lu H Y, Liu W, Zhang B, et al. DeepSeek-VL: towards real-world vision-language understanding. 2024. ArXiv:2403.05525

39 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning, 2021. 8748–8763

40 Huang T Y, Du B W, Yang Y H, et al. CLIP2Point: transfer clip to point cloud classification with image-depth pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 22157–22167

41 Qi Z Y, Fang Y, Sun Z Y, et al. GPT4Point: a unified framework for point-language understanding and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 26417–26427

42 Chen R N, Liu Y Q, Kong L D, et al. CLIP2Scene: towards label-efficient 3D scene understanding by clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 7020–7030

43 Xue L, Gao M F, Xing C, et al. ULIP: learning a unified representation of language, images, and point clouds for 3D understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 1179–1189

44 Li L N, Zhang P C, Zhang H T, et al. Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 10965–10975

45 Zhou X Y, Wang D Q, Krähenbühl P. Objects as Points. 2019. ArXiv:1904.07850

46 Philion J, Fidler S. Lift, splat, shoot: encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In: Proceedings of

the European Conference on Computer Vision, 2020. 194–210
47 Shi S S, Wang X G, Li H S. PointRCNN: 3D object proposal generation and detection from point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 770–779
48 Sun P, Wang W Y, Chai Y N, et al. RSN: range sparse net for efficient, accurate lidar 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 5725–5734
49 Lin C, Tian D, Duan X, et al. CL3D: camera-LiDAR 3D object detection with point feature enhancement and point-guided fusion. IEEE Trans Intell Transp Syst, 2022, 23: 18040–18050
50 Lin C, Tian D, Duan X, et al. 3D-DFM: anchor-free multimodal 3-D object detection with dynamic fusion module for autonomous driving. IEEE Trans Neural Netw Learn Syst, 2023, 34: 10812–10822
51 Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. 2023. ArXiv:2303.08774
52 Li J N, Li D X, Savarese S, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the International Conference on Machine Learning, 2023. 19730–19742
53 Zhou K, Yang J, Loy C C, et al. Learning to prompt for vision-language models. Int J Comput Vis, 2022, 130: 2337–2348
54 Gao P, Geng S, Zhang R, et al. CLIP-adapter: better vision-language models with feature adapters. Int J Comput Vis, 2023, 132: 581–595
55 Zanella M, Ayed I B. Low-rank few-shot adaptation of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop, 2024. 1593–1603
56 Li Y W, Qi X J, Chen Y K, et al. Voxel field fusion for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 1120–1129
57 Huang T T, Liu Z, Chen X W, et al. EPNet: enhancing point features with image semantics for 3D object detection. In: Proceedings of the European Conference on Computer Vision, 2020. 35–52
58 Yoo J H, Kim Y, Kim J, et al. 3D-CVF: generating joint camera and lidar features using cross-view spatial feature fusion for 3D object detection. In: Proceedings of the European Conference on Computer Vision, 2020. 720–736
59 Qi C R, Liu W, Wu C X, et al. Frustum pointNets for 3D object detection from RGB-D data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. 918–927
60 Zhang Y N, Chen J X, Huang D. Cat-det: contrastively augmented transformer for multi-modal 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 908–917
61 Chen Y K, Li Y W, Zhang X Y, et al. Focal sparse convolutional networks for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 5428–5437
62 Li X, Ma T, Hou Y N, et al. LoGoNet: towards accurate 3D object detection with local-to-global cross-modal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 17524–17534
63 Li Y W, Chen Y L, Qi X J, et al. Unifying voxel-based representation with transformer for 3D object detection. In: Proceedings of the International Conference on Neural Information Processing Systems, 2022. 18442–18455
64 Hu C Y, Zheng H, Li K, et al. FusionFormer: a multi-sensory fusion in bird's-eye-view and temporal consistent transformer for 3D object detection. 2023. ArXiv:2309.05257