

Article

Not peer-reviewed version

CLIP-Mono3D: End-to-End Open-Vocabulary Monocular 3D Object Detection via Semantic-Geometric Similarity

[Zichong Gu](#), [Shiyi Mu](#), Hanqi Lyu, [Shugong Xu](#)*

Posted Date: 16 March 2026

doi: 10.20944/preprints202603.1117.v1

Keywords: 3D monocular object detection; open-vocabulary learning; autonomous driving



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

CLIP-Mono3D: End-to-End Open-Vocabulary Monocular 3D Object Detection via Semantic-Geometric Similarity

Zichong Gu ¹, Shiyi Mu ¹, Hanqi Lyu ¹ and Shugong Xu ^{2,*}

¹ School of Communication and Information Engineering, Shanghai University, 99 Shangda Road, Shanghai 200444, China
² Xi'an Jiaotong-Liverpool University, Suzhou Industrial Park, Suzhou 215123, China
* Correspondence: shugong.xu@xjtlu.edu.cn

Abstract

Open-vocabulary 3D object detection (OV-3DOD) is crucial for real-world perception, yet existing monocular methods are often limited by predefined categories or heavy reliance on external 2D detectors. In this paper, we propose **CLIP-Mono3D**, an end-to-end one-stage transformer framework that directly integrates vision-language semantics into monocular 3D detection. By leveraging CLIP-derived semantic priors and grounding object queries in semantically salient regions, our model achieves robust zero-shot generalization to novel categories without requiring auxiliary 2D detectors. Furthermore, we introduce **OV-KITTI**, a large-scale benchmark extending KITTI with 40 new categories and over 7,000 annotated 3D bounding boxes. Extensive experiments on OV-KITTI, KITTI, and Argoverse demonstrate that CLIP-Mono3D achieves competitive performance in open-vocabulary scenarios.

Keywords: 3D monocular object detection; open-vocabulary learning; autonomous driving

1. Introduction

Monocular 3D object detection estimates the 3D bounding boxes of objects from a single RGB image, an approach that offers compelling cost and deployment advantages over LiDAR-based or multi-sensor fusion systems. This capability is particularly valuable in large-scale applications such as autonomous driving, robotics, and augmented reality, where hardware simplicity and scalability are paramount. However, despite its practical appeal, monocular 3D detection remains significantly more challenging than its 2D counterpart. This difficulty stems primarily from the inherently ill-posed nature of depth recovery from a single viewpoint and the scarcity of large-scale, semantically diverse 3D annotations.

Recent years have witnessed substantial progress driven by benchmark datasets such as KITTI [1], SUN RGB-D [2], ScanNet V2 [3], and nuScenes [4]. Yet, as summarized in Table 1, these datasets are largely constrained in semantic scope, typically encompassing only 9 to 23 object categories. Most of these categories are restricted to traffic-related agents, such as cars and pedestrians, or common indoor items. In contrast, modern 2D detection benchmarks like COCO and Objects365 cover hundreds of categories, enabling robust generalization across diverse visual concepts. This semantic gap severely limits the applicability of current 3D detectors in open-world environments, where systems must recognize and localize objects beyond a fixed, predefined set.

Table 1. Comparison of Common 3D Object Detection Datasets

Dataset	Classes	Scene	Modality	Object Range
KITTI 3D	9	Outdoor	RGB, PC	traffic agents
SUN RGB-D	20	Indoor	RGB-D	item
ScanNet V2	21	Indoor	RGB-D, PC	item
Nuscenes	23	Outdoor	RGB, PC	traffic agents
Objectron	9	Object-wise	RGB	item
OV-KITTI (ours)	49	Outdoor	RGB	traffic agents, item, animal

To bridge this gap, open-vocabulary 3D object detection has emerged as a promising research direction. The goal is to enable models to detect and localize objects described by arbitrary natural language prompts, including those unseen during training. Existing approaches, however, predominantly follow a two-stage paradigm. They first leverage a pre-trained 2D open-vocabulary detector to extract semantic proposals, which are then fed into a class-agnostic 3D detector for geometric regression, as illustrated in Figure 1(a). While effective, this design introduces several practical limitations: (i) it requires heavy reliance on external 2D detectors and their associated supervision; (ii) it involves multi-stage training pipelines that are difficult to optimize jointly; and (iii) many methods still depend on point cloud priors or depth estimators, undermining the simplicity of the monocular setting.

Recent attempts, such as OVMono3D [5], have sought to mitigate these issues by incorporating foundation models like SAM for segmentation priors or monocular depth estimators. Nevertheless, they remain fundamentally dependent on pre-trained 2D detectors and auxiliary data sources, preventing true end-to-end training from 3D supervision alone.

In this work, we present **CLIP-Mono3D**, a novel, end-to-end trainable framework for open-vocabulary monocular 3D object detection. Built upon the MonoDGP architecture [6], our method integrates semantic knowledge directly into the 3D detection pipeline by leveraging a pre-trained FG-CLIP visual-language encoder [7]. Unlike prior approaches, CLIP-Mono3D eliminates the need for an external 2D detector by fusing CLIP-derived visual-semantic features with geometric representations via cross-modal attention. By initializing detection queries using language embeddings, our model achieves zero-shot generalization to novel categories without additional 2D supervision, representing a significant step toward practical and generalizable monocular 3D perception.

To facilitate research in this under-explored direction, we further introduce **OV-KITTI**, a new benchmark that extends the original KITTI dataset with 40 additional object categories, including animals and everyday items. As shown in Table 1, OV-KITTI not only expands semantic coverage but also provides more diverse shape and scale priors, which help alleviate the depth ambiguity inherent in monocular setups. The dataset is carefully curated to ensure balanced distributions between base and novel categories, enabling fair evaluation of open-vocabulary generalization.

The significance of our work lies in three key contributions:

1. **Unified End-to-End Architecture:** We propose CLIP-Mono3D, an end-to-end framework that unifies semantic and geometric reasoning. By introducing a Cross-Modal Semantic-Geometric Fusion module, we inject fine-grained semantic clues into geometric features via a lightweight residual connection, enhancing semantic awareness without disrupting pre-trained geometric cues.
2. **Language-Aware Query Initialization:** We design a novel query initialization strategy that converts 2D semantic heatmaps into explicit 3D query positions. This mechanism significantly improves 3D center localization and recall for open-vocabulary objects compared to standard learned queries.
3. **OV-KITTI Benchmark and Evaluation:** We introduce OV-KITTI, a large-scale benchmark with controlled semantic and size distributions. Extensive experiments on OV-KITTI, KITTI, and Argoverse demonstrate that CLIP-Mono3D achieves competitive performance in both closed- and open-vocabulary settings, paving the way for deployment in truly open-world scenarios.

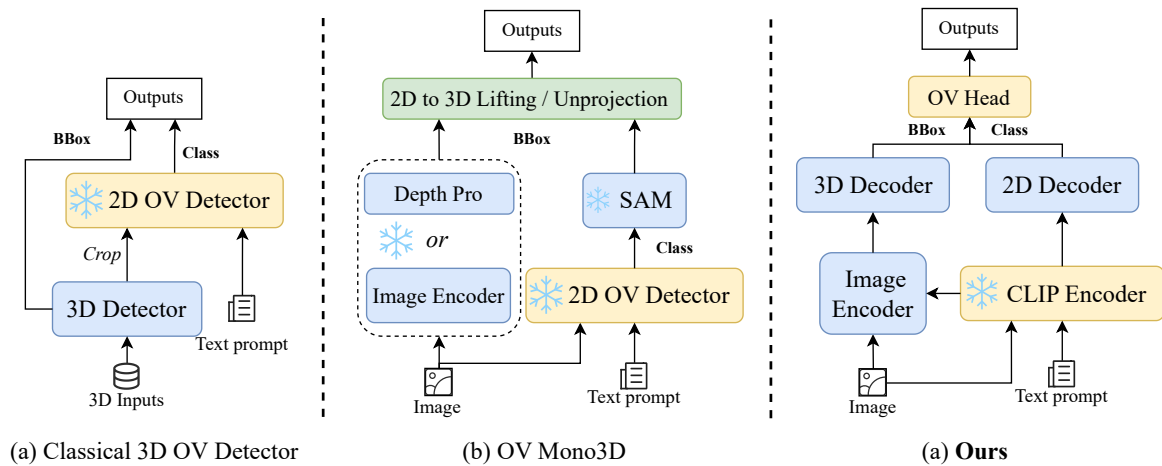


Figure 1. Comparison with existing open-vocabulary 3D detectors. (a) Traditional paradigm using pretrained 2D detectors. (b) Monocular extensions like OVMono3D. (c) Our end-to-end CLIP-Mono3D, which fuses CLIP semantics directly into a 3D DETR framework without reliance on 2D detectors or multi-stage training.

2. Related Work

2.1. Monocular 3D Object Detection

Monocular 3D object detection localizes 3D bounding boxes from single RGB images, but its development lags behind point cloud and multi-view methods due to ill-posed depth estimation. Current approaches include geometry-based methods and depth-assisted techniques. Geometry-based methods leverage geometric consistency: some reduce the task to 2D detection with 3D-2D projection for center calculation [8,9], while others directly regress 3D boxes via keypoints [10,11]. Enhanced variants incorporate geometric priors like ground planes [12], shape awareness [13], uncertainty modeling [14], and height cues [15]. To address depth perception limitations, depth-assisted methods utilize pseudo-LiDAR representations [16,17], LiDAR-supervised depth estimation [18], frustum discretization for BEV mapping [19], and depth pre-training [20]. Transformer-based designs further integrate visual/depth features [21] and employ depth-guided decoding [22].

2.2. Open-Vocabulary Object Detection

Open-vocabulary object detection handles both seen and unseen categories to accommodate real-world diversity. With CLIP’s rise, 2D open-vocabulary detection has advanced through knowledge distillation from image-text pairs [23], pseudo-label generation [24,25], pretrained encoders (CLIP [26–30] or BERT [31–33]), and generative modeling [34,35]. Transformer-based methods dominate for their effective integration of text encoders.

2.3. Open-Vocabulary 3D Object Detection

Extending to 3D space, this task detects unseen categories with higher complexity. Early works used unsupervised clustering for open-set detection [36]. Recent CLIP-inspired approaches include pseudo-label methods where OV3DET [37] trains class-agnostic detectors with 2D-generated labels then aligns 3D, image and language features. Cross-modal alignment techniques combine 3D geometric and 2D semantic priors without 2D detectors [38], or enrich datasets via cross-domain alignment [39]. Knowledge propagation frameworks cyclically transfer 2D semantic knowledge to 3D and geometric cues to 2D [40]. Point cloud-language alignment further bridges modalities [41,42].

3. Method

3.1. Task Definition and Overall Architecture

Open-vocabulary monocular 3D object detection aims to estimate 3D bounding boxes $\mathcal{D} = \{\mathcal{O}_i\}_{i=1}^N$ from a single RGB image $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$, guided by a set of arbitrary text prompts $\mathcal{T} = \{\tau_1, \dots, \tau_K\}$. Each object $\mathcal{O}_i$ is parameterized by its 3D center (x_i, y_i, z_i) , dimensions (l_i, w_i, h_i) , orienta-

tion θ_i , and a semantic similarity vector $\mathbf{s}_i \in [0, 1]^K$. The similarity score for each prompt is computed as:

$$\mathbf{s}_i[k] = \sigma\left(\phi_{\text{img}}(\mathbf{p}_i)^\top \phi_{\text{text}}(\tau_k)\right), \quad (1)$$

where σ denotes the sigmoid function, ϕ_{img} and ϕ_{text} represent the image and text encoders (e.g., CLIP), and $\mathbf{p}_i$ denotes the region-level features associated with the i -th object. This formulation enables the detection of novel categories by replacing fixed-set classification with open-ended semantic matching.

As illustrated in Figure 2, CLIP-Mono3D extends the MonoDGP architecture [6]. A ResNet-50 backbone first extracts multi-scale features $\{\mathbf{F}_i\}_{i=0}^3$, which are enhanced by a Region Segmentation Head (RSH) to produce $\mathbf{F}_V$. These features are fused with CLIP-derived semantics (Sec. 3.2) and processed by a depth predictor to generate $\mathbf{F}_D$. Dual transformer encoders independently process these streams, followed by a 2D visual decoder and a 3D depth-guided decoder. To provide explicit geometric priors, language-aware queries (Sec. 3.3) are initialized from CLIP similarity maps. Final predictions are refined via geometric depth correction and scored against text embeddings.

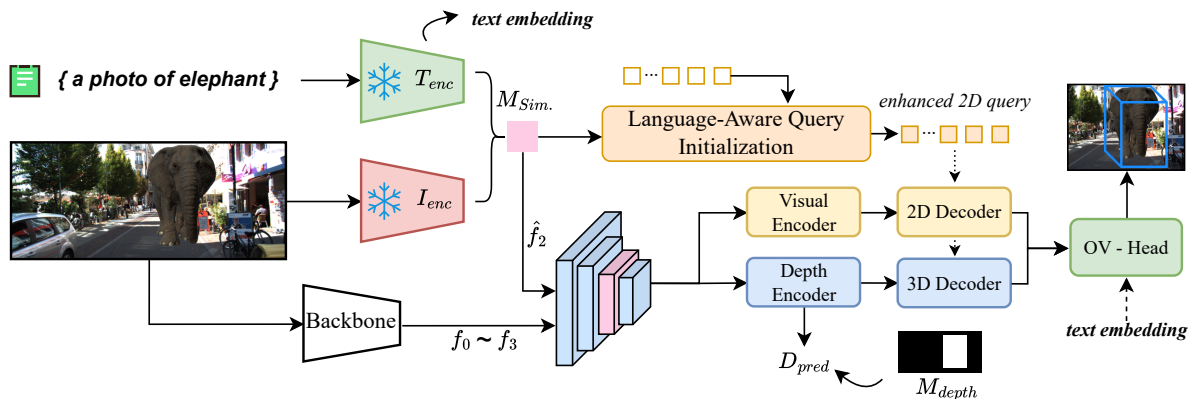


Figure 2. Pipeline of CLIP-Mono3D. Visual features are fused with the CLIP similarity matrix M_{sim} to produce multimodal feature maps. These maps are then passed to the visual encoder and depth encoder. The depth encoder outputs depth predictions through a depth predictor and is supervised via an object-wise depth map M_{depth} . Next, a query enhancer based on the CLIP similarity matrix initializes language-aware queries. Finally, regression outputs are handled by the 2D and 3D decoders and the open vocabulary detection head outputs the final prediction boxes with the high similarity to the text embeddings.

3.2. Cross-Modal Feature Fusion

To bridge the semantic gap between vision and language modalities, we introduce a lightweight cross-modal fusion mechanism that injects text-guided spatial priors into the visual backbone. Given an input image $\mathcal{I}$ and prompts $\mathcal{T}$, we extract dense visual features $\mathbf{F}_{\text{img}} \in \mathbb{R}^{14 \times 14 \times d}$ and global textual embeddings $\mathbf{F}_{\text{txt}} \in \mathbb{R}^{K \times d}$ using a frozen CLIP encoder. Freezing CLIP’s parameters is essential to preserve its zero-shot generalization and prevent the loss of rich semantic knowledge during 3D detection training.

Semantically relevant regions are identified by computing a spatial similarity map $\mathcal{S} \in \mathbb{R}^{14 \times 14}$ through the aggregation of cosine similarities across all text tokens:

$$\mathcal{S}(u, v) = \frac{1}{K} \sum_{k=1}^K \frac{\mathbf{F}_{\text{img}}^{(u,v)} \cdot \mathbf{F}_{\text{txt}}^{(k)}}{\|\mathbf{F}_{\text{img}}^{(u,v)}\| \|\mathbf{F}_{\text{txt}}^{(k)}\|}. \quad (2)$$

This map functions as a soft attention mask, highlighting regions aligned with the linguistic input. Unlike post-hoc filtering methods, our approach integrates this signal early in the feature hierarchy, allowing semantic guidance to inform all downstream components.

The similarity map $\mathcal{S}$ is bilinearly upsampled to match the dimensions of the intermediate feature map $\mathbf{F}_V^{(2)} \in \mathbb{R}^{C \times H_2 \times W_2}$. It is then processed by a convolutional module Γ_{fuse} —comprising two 3×3

convolutions with ReLU activations—to refine its spatial structure and match the channel dimensions. The final fused feature map is obtained via an additive residual connection:

$$\mathbf{F}_V^{\text{enh}(2)} = \mathbf{F}_V^{(2)} + \Gamma_{\text{fuse}}(\text{Upsample}(\mathcal{S})). \quad (3)$$

This residual design ensures that primary geometric and structural information remains intact, which is critical for accurate 3D localization. This early fusion strategy creates a "semantic spotlight" that benefits both the RSH and the subsequent query initialization stage (Sec. 3.3).

3.3. Language-Aware Query Initialization

Standard DETR-like architectures typically initialize object queries as content-agnostic learnable embeddings, which can lead to slow convergence in complex scenes. To address this, we propose a language-aware query initialization strategy that transforms 2D semantic probability maps into explicit 3D geometric anchors.

As shown in Figure 3, the initialization involves three steps. First, we interpret the CLIP similarity map $\mathcal{S}$ as a spatial probability distribution and identify N_q locations $\mathcal{P} = \{(u_k, v_k)\}_{k=1}^{N_q}$ with the highest activations. These serve as candidate 2D centers for potential 3D objects.

Second, we employ `F.grid_sample` to sample local descriptors $\mathbf{P}_{\text{keypoint}} \in \mathbb{R}^{N_q \times C}$ from the enhanced feature map $\mathbf{F}_V^{\text{enh}(2)}$ at coordinates $\mathcal{P}$. This differentiable operation ensures that each query is initialized with features specific to its corresponding object region.

Third, a global semantic prior is distilled from these descriptors via a two-layer MLP:

$$\mathbf{Q}_{\text{prior}} = W_2 \delta \left(W_1 \cdot \frac{1}{N_q} \sum_{k=1}^{N_q} \mathbf{P}_{\text{keypoint}}^{(k)} \right). \quad (4)$$

The base queries $\mathbf{Q} \in \mathbb{R}^{N_q \times 2d}$ consist of a content component $\mathbf{Q}_c$ and a positional component $\mathbf{Q}_p$. We use sine-cosine positional encodings of $\mathcal{P}$ to initialize $\mathbf{Q}_p$, while $\mathbf{Q}_c$ is enhanced by the global prior:

$$\mathbf{Q}_{\text{content}} = \mathbf{Q}_c + \mathbf{Q}_{\text{prior}}, \quad \mathbf{Q} = [\mathbf{Q}_{\text{content}}; \mathbf{Q}_p]. \quad (5)$$

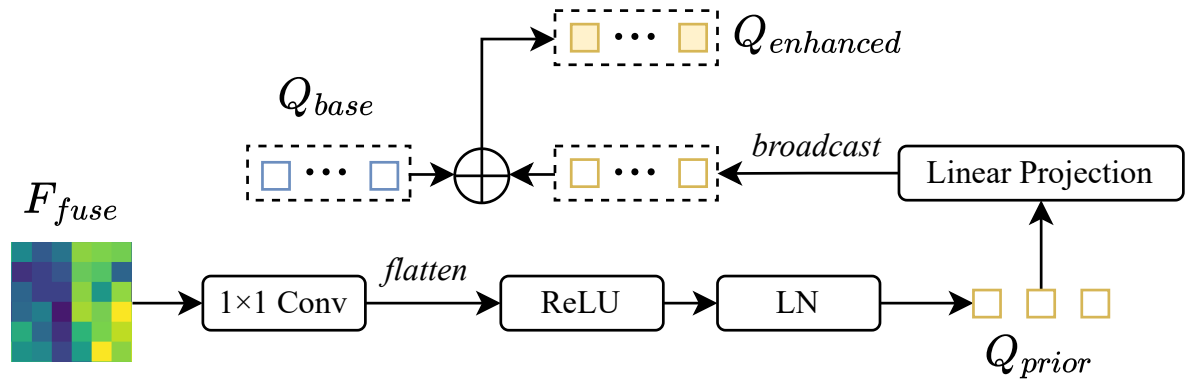


Figure 3. Language-aware query initialization. Top- N_q keypoints from the CLIP similarity map are used to sample and aggregate feature priors.

By grounding queries in semantically verified regions, we transform the detection process from exhaustive spatial searching to targeted localization. This "semantic priming" accelerates training convergence and reduces attention to background clutter. Importantly, the CLIP-derived prior provides a strong inductive bias for open-world concepts. During inference, if no text is provided, the system reverts to base queries ($\mathbf{Q}_{\text{prior}} = \mathbf{0}$) to maintain backward compatibility.

3.4. Open-Vocabulary Detection Head and Loss

The detection head enables open-vocabulary classification by computing semantic similarity between decoder outputs and projected text embeddings. For each output $\mathbf{h} \in \mathbb{R}^{d_h}$, we project it into the CLIP feature space: $\mathbf{v}' = \text{norm}(\mathbf{W}_v \mathbf{h})$, where $\mathbf{W}_v \in \mathbb{R}^{512 \times d_h}$ and $\text{norm}(\cdot)$ denotes ℓ^2 normalization. Similarly, text embeddings $\mathbf{t}_{\text{CLIP}}$ are projected via a learnable matrix $\mathbf{W}_t$. This dual-projection design mitigates the domain gap between the detector's internal representations and CLIP's pre-trained embeddings.

The similarity score is computed as a temperature-scaled dot product, $s = \tau \cdot \mathbf{v}'^\top \mathbf{t}'$, where the learnable temperature $\tau = \exp(\gamma)$ controls the distribution concentration. To handle background regions, we introduce a learnable background embedding $\mathbf{t}_{bg}$ appended to the text embeddings.

Training follows a bipartite matching strategy using the Hungarian algorithm. The matching cost incorporates both geometric and semantic terms:

$$C_{\text{match}} = \lambda_g (\|\Delta \mathbf{c}_{3d}\|_1 + \|\Delta \mathbf{b}_{2d}\|_1 + (1 - \text{GIoU})) - \lambda_s s. \quad (6)$$

Integrating the similarity score s into the matching process ensures that predictions are assigned based on both spatial accuracy and semantic coherence, significantly enhancing generalization to unseen classes.

For matched pairs $\mathcal{M}$, we apply a contrastive loss:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{|\mathcal{M}|} \sum_{(i,j) \in \mathcal{M}} \log \frac{\exp(s_{ij}/\tau)}{\sum_k \exp(s_{ik}/\tau)}. \quad (7)$$

The total objective combines this with geometric regression losses $\mathcal{L}_g$:

$$\mathcal{L}_{\text{total}} = \lambda_c \mathcal{L}_{\text{contrast}} + \sum \lambda_g \mathcal{L}_g. \quad (8)$$

This co-design ensures that gradients from the contrastive loss are applied to the most semantically relevant predictions, creating a robust optimization cycle that yields a model both geometrically precise and semantically aware.

4. OV-KITTI Benchmark

Most existing monocular 3D object detection models rely on the KITTI dataset for training and evaluation. However, this dataset presents several critical limitations. KITTI contains only nine object categories, and prior works predominantly evaluate models on the "Car" category due to the scarcity of other instances. In open-vocabulary or zero-shot learning settings, a common practice involves training on "Car" and "Cyclist" classes while attempting to transfer detection capabilities to the "Pedestrian" category. Nevertheless, as illustrated in Figure 4(c), the 3D bounding box dimensions of the "Car" category differ significantly from those of "Pedestrian" and "Cyclist". This discrepancy leads to an inherent bias where the detection of "Pedestrian" largely relies on knowledge spillover from "Cyclist" rather than benefiting from the rich feature learning associated with the dominant "Car" category. We argue that this imbalance in both category support and scale distribution within KITTI hinders the development of generalizable 3D detection models.

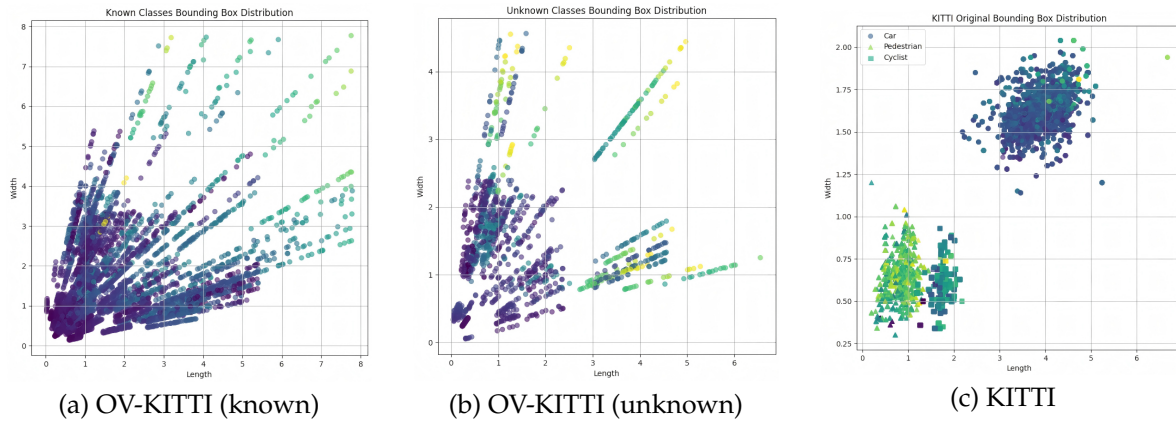


Figure 4. Comparison of bbox size distribution across datasets. The axes represent dimensions of 3D boxes. OV-KITTI exhibits a more balanced size distribution between known and unknown categories compared to KITTI.

To overcome these limitations, we introduce **OV-KITTI**, an augmented benchmark based on the KITTI dataset designed specifically for open-vocabulary monocular 3D detection. We enrich the original dataset with 40 additional object categories sourced from Objaverse [43]. These categories encompass a diverse set of animals and household items. The new objects are carefully selected to avoid semantic overlap with existing traffic participants in KITTI, such as cars and pedestrians, thereby enabling precise supervision and evaluation.

4.1. Dataset Construction

The construction of OV-KITTI follows a systematic multi-step pipeline:

(1) **Bounding Box Design:** For each new category, we define physically plausible size constraints for its 3D bounding box. During rendering, the actual box dimensions are randomly sampled within this predefined range to ensure variability. (2) **Scene Integration:** Using Blender, we render 3D meshes into real driving scenes from KITTI. Objects are placed on the ground plane with random rotation, scaling, and translation. We explicitly enforce constraints to ensure no physical overlap with existing objects in the scene. (3) **Stereo Rendering:** We generate stereo-consistent left and right views for each modified scene. Annotations are provided in the standard KITTI format to ensure seamless compatibility with existing detection frameworks. (4) **Balance Control:** Special care is taken to balance the 3D box size distributions between known and unknown classes, as shown in Figure 4(a) and (b). This step reduces potential size-related biases and supports a fair evaluation of model generalization capabilities.

4.2. Category Statistics

We divide the 40 new classes into *known* (training) and *unknown* (testing) categories. The known set comprises 32 classes, consisting of 19 household items and 13 animals, while the unknown set contains 8 classes, including 5 items and 3 animals, for zero-shot evaluation. This split ensures diversity and scale balance across training and testing phases. The full category list and instance counts are provided in Table 2. As shown in Figure 5, OV-KITTI contains high-quality renderings of diverse objects under varied scales and contexts. This provides a challenging yet realistic testbed for evaluating open-vocabulary 3D detection performance.

Table 2. Training and Test Set Distribution of OV-KITTI

Category	Class Names	Quantity
Train	air conditioner, alligator, fireplug, bathtub, bed, volleyball, camel, cat, chair, cow, deer, baby buggy, sofa, hippopotamus, horse, lion, oven, pillow, refrigerator, rhinoceros, sculpture, suitcase, telephone booth, television set, wheelchair, zebra, automatic washer, goat, hog, snowman, soccer ball, elephant	5936
Test	bench, dog, giraffe, guitar, table, tiger, bucket, wardrobe	1545



Figure 5. Examples of OV-KITTI. Diverse objects such as animals and household items are rendered with random scales and positions on real KITTI scenes.

5. Experiments

5.1. Experimental Setup

5.1.1. Datasets and Metrics

We evaluate on KITTI and OV-KITTI. For KITTI, we use standard splits (3712 train, 3769 val). For OV-KITTI, we use 5936 training samples (32 classes) and 1545 test samples (8 classes). Original categories are excluded from open-vocabulary evaluation. We use AP at 40 recall as the metric: AP_{3D} and AP_{BEV} for closed-vocabulary; AP_{15} per category for open-vocabulary.

5.1.2. Settings

We follow MonoDETR-style augmentation and FG-CLIP image preprocessing. The model is trained on a single NVIDIA RTX 4090D for 200 epochs with batch size 8 using PyTorch.

5.2. Closed-Vocabulary Results

As shown in Table 3, our method achieves state-of-the-art performance in the standard monocular 3D object detection task. On the KITTI val set, our approach attains 31.40% AP_{3D} under the “easy” difficulty level, surpassing previous methods by a clear margin: +2.84 percentage points over MonoDETR [22] and +1.72 over MonoDGP [6]. The improvements are consistent across moderate and hard difficulty levels, demonstrating enhanced robustness in detecting occluded and distant objects. In BEV detection, our model also sets a new benchmark with 39.47% AP_{BEV} (easy), outperforming both baselines by over 1.5 points.

Table 3. Comparisons with SOTA monocular methods. Best results are bolded; *: reproduced in our setting.

Method	KITTI (3D)			KITTI (BEV)			OV-KITTI	
	Easy	Mod.	Hard	Easy	Mod.	Hard	AP_{25}	AP_{50}
MonoDTR [21]	21.99	15.39	12.73	33.33	25.35	21.68	–	–
GUPNet [14]	22.76	16.46	13.72	31.07	22.94	19.75	10.87	2.53
MonoCon [44]	26.33	19.01	15.98	–	–	–	12.15	2.70
MonoCD [45]	26.45	19.37	16.38	34.60	24.96	21.51	12.21	2.74
FD3D [46]	28.22	20.23	17.04	36.98	26.77	23.16	–	–
MonoDETR* [22]	28.56	19.92	16.17	37.59	26.78	22.79	12.76	3.05
MonoDGP* [6]	29.68	21.95	18.58	37.87	27.54	24.63	13.33	3.94
Ours	31.40	22.98	19.61	39.47	28.72	24.93	14.44	4.88
Gain	+1.72	+1.03	+1.03	+1.60	+1.18	+0.30	+1.11	+0.94

These gains can be attributed to our improved 2D feature representation enriched with CLIP-derived semantics, which provides stronger cues for depth estimation and spatial localization. Unlike purely geometric reasoning methods, our integration of semantic context allows the detector to better disambiguate scale and position, especially in low-texture or cluttered scenes.

On the OV-KITTI dataset, while the focus shifts toward open-vocabulary generalization, we still report strong closed-vocabulary performance. Our AP_{25} reaches 14.44%, exceeding MonoDGP by 1.11%, which indicates that our enhancements do not compromise detection accuracy on seen classes. This balance between specialization and generalization is crucial for real-world deployment, where systems must handle both known and emerging object categories.

As shown in Table 4, we compare the efficiency metrics with baseline methods. It is important to note that while our framework incorporates a frozen CLIP image encoder of approximately 150 M parameters to extract semantic features, these parameters do not require gradient updates. Consequently, our trainable parameter count of 43.48 M is nearly identical to that of MonoDGP at 43.33 M. This minimal increase of approximately 0.15 M confirms that the performance gains stem from our effective semantic-geometric alignment design rather than a brute-force increase in model capacity. Although the integration of the visual encoder introduces a marginal latency increase, the inference speed of 51ms per frame remains sufficient for real-time autonomous driving applications.

Table 4. Efficiency comparison. Note that the parameter count refers to **trainable parameters**.

Method	#Param (M)	FLOPs (G)	Latency (ms)
MonoDETR*	39.82	47.2	33
MonoDGP*	43.33	55.19	36
Ours	43.48	72.52	51

5.3. Open-Vocabulary Results

The true strength of our method lies in its ability to detect objects from previously unseen categories, a key challenge in open-vocabulary 3D detection. As shown in Table 5, our method achieves

an overall 5.81% AP on the eight unseen categories in OV-KITTI, outperforming the fine-tuned variant OVMono3D⁺ by +1.21%.

Table 5. Open-vocabulary 3D detection performance on OV-KITTI (AP_{15}). ⁺ denotes fine-tuning (FT) on the base training set.

Methods	bench	table	tiger	bucket	wardrobe	guitar	giraffe	dog	Overall
G-DINO-3D	1.33	2.25	5.48	0.03	7.67	0.00	0.85	0.85	2.31
Gain from FT	+1.63	+2.02	+7.63	+0.64	+1.64	+0.13	+2.44	+0.78	+2.11
G-DINO-3D ⁺	2.96	4.27	13.11	0.67	9.31	0.13	3.29	1.63	4.42
OVMono3D	2.08	3.85	8.96	0.13	8.71	0.04	2.17	0.18	3.27
Gain from FT	+1.19	+0.61	+7.02	+0.49	+2.42	+0.15	+1.61	+2.14	+1.33
OVMono3D ⁺	3.27	4.46	15.98	0.62	11.13	0.19	3.78	2.32	4.60
Ours	3.70	4.74	18.71	0.81	10.70	0.26	4.82	2.76	5.81
Gain	+0.43	+0.28	+2.73	+0.19	-0.43	+0.07	+1.04	+0.44	+1.21

A critical observation is that fine-tuning on the known training set provides substantial gains for baseline methods, particularly for geometrically complex categories. As shown in the “Gain from FT” rows, OVMono3D⁺ improves by a remarkable +7.02 points on “tiger” and +1.61 on “giraffe” compared to its non-fine-tuned version. This confirms that exposure to 3D shape priors during fine-tuning is essential for accurate localization of categories largely absent from standard 3D datasets.

Despite this significant boost from fine-tuning, our method still achieves superior performance across most categories. We outperform OVMono3D⁺ by +2.73 on “tiger” and +1.04 on “giraffe”, demonstrating that our direct integration of vision-language semantics enables better generalization to novel biological shapes, without relying on cascaded 2D detectors or fine-tuning heuristics. For smaller or less frequent objects like “bucket” and “dog”, our method also shows consistent improvements.

The only exception is “wardrobe” where OVMono3D⁺ retains a slight edge. We attribute this to its stronger pre-trained knowledge of household items and the relatively simple geometric structure of wardrobes, which are more easily captured by existing 2D open-vocabulary detectors. Overall, our results validate that grounding 3D detection directly in language semantics, rather than relying on fine-tuned 2D detectors, leads to superior generalization, especially for structurally diverse and underrepresented categories like animals.

5.4. Visualization of Cross-Modal Alignment

To elucidate how language priors guide the detection process, we visualize the CLIP-based similarity maps $\mathcal{S}$ and their influence on query initialization in Figure 6. As shown in the middle column, $\mathcal{S}$ effectively serves as a semantic attention mechanism, highlighting regions aligned with class-specific prompts such as “car” and “wheelchair”. Notably, for the rare class “wheelchair”, which lacks 3D annotations during training, the model identifies plausible candidates by leveraging semantic cues (e.g., human silhouettes with wheels), demonstrating robust open-vocabulary generalization.

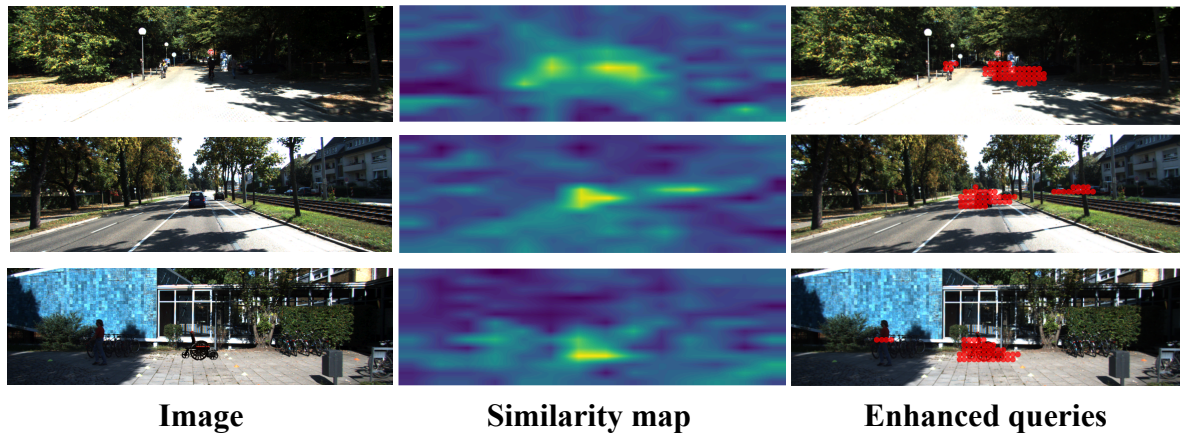


Figure 6. Visualization of similarity map and enhanced queries. The label prompt of each row corresponds to KITTI classes: “car” and “wheelchair”. The similarity map (middle) highlights regions matching the text description, and the top- N_q keypoints (right) are used to initialize language-aware object queries.

The right column illustrates the top- N_q keypoints selected for query initialization. These points concentrate on the object’s spatial extent, confirming that our language-aware sampling effectively localizes semantically meaningful regions. These keypoints provide informed feature priors that prime the object queries, enabling the decoder to focus on relevant content from the initial stages.

This visualization validates two key advantages: (1) the similarity map bridges language semantics and visual geometry for zero-shot grounding; (2) language-aware initialization mitigates monocular depth ambiguity by providing semantically grounded spatial anchors.

5.5. Ablation Studies

To evaluate the contribution of each proposed component, we conduct comprehensive ablation studies on the OV-KITTI benchmark. Unless otherwise specified, performance is measured by AP_{25} for seen classes (CV) and AP_{15} for unseen categories (OV).

Core Components and Query Initialization. As summarized in Table 6, each module consistently improves performance. While feature fusion enhances the base representation, the language-aware query initialization provides the most significant gain for open-vocabulary generalization (+1.12% OV AP). This suggests that grounding queries in semantic regions from the early decoding stages is more effective for novel objects than relying on generic learnable embeddings.

Table 6. Ablation study on core components and visual-language encoders, highlighting the impact of freezing CLIP parameters.

Method	CV	OV
Baseline	13.33	3.94
+ Feature Fusion	14.22	4.69
+ Query Init.	14.44	5.81
Text Encoder	CV	OV
CLIP (Frozen)	14.05	5.32
FineCLIP (Frozen)	14.31	5.63
FG-CLIP (Finetuned)	14.15	5.45
FG-CLIP (Frozen)	14.44	5.81

Regarding the specific initialization design in Table 7, we find that using 50 queries with a global prior aggregated via mean-pooling yields optimal results. This global context informs each query of the overall semantic landscape, preventing them from over-focusing on isolated, potentially noisy local regions. In contrast, max-pooling over-emphasizes the single most salient activation, leading to poorer generalization for diverse scenes. We also compare VLM backbones: FG-CLIP (Frozen) achieves the highest accuracy, while fine-tuning it via LoRA [47] leads to a performance drop. This indicates that

the 3D dataset’s scale is insufficient to update the VLM without causing catastrophic forgetting of its pre-trained open-world semantic knowledge.

Table 7. Design of language-aware query initialization. NQ represents the number of object queries; Agg. indicates the pooling method (Mean or Max) used to distill global context from keypoints.

Config	NQ	Agg.	CV	OV
(a) Learnable	50	-	13.33	3.94
(b) Top-K	50	-	13.95	4.80
(c) + Global	25	Mean	14.28	5.41
(d) + Global	50	Mean	14.44	5.81
(e) + Global	75	Mean	14.39	5.65
(f) + Global	50	Max	14.09	5.17

Feature Fusion Strategy and Domain Robustness. Table 8 explores fusion architectures. We find that additive residual connections at *res2* outperform simple concatenation. This additive design acts as a calibrated semantic spotlight that preserves the geometric integrity of intermediate features, whereas concatenation may introduce noise that disrupts sensitive 3D regressions. Injecting semantic guidance early in the feature hierarchy (stage 2) is crucial for informing downstream depth estimation and query sampling.

Table 8. Comparison of cross-modal fusion strategies across different backbone stages and architectural designs.

Fusion	Loc.	CV	OV
(a) None	-	13.33	3.94
(b) Cat	<i>res2</i>	13.67	4.18
(c) Add (1L)	<i>res2</i>	13.99	4.36
(d) Add (2L)	<i>res2</i>	14.22	4.69
(e) Add (3L)	<i>res2</i>	14.14	4.58
(f) Add (2L)	<i>res3</i>	13.92	4.41

To investigate whether the model overfits to synthetic artifacts, we conduct a data-mixing study in Table 9. We progressively introduce real-world KITTI objects into the training pipeline. Crucially, to ensure a fair comparison, the Seen AP is evaluated exclusively on the original synthetic classes. The results show that adding real-world instances with distinct lighting and textures leads to negligible fluctuations in Unseen AP. This stability confirms that CLIP-Mono3D learns generalized semantic concepts rather than low-level texture priors. The drop observed in Row (3) is likely due to the extreme class imbalance introduced by real-world Cars, which biases the optimization away from rare open-vocabulary classes.

Table 9. Robustness evaluation under real-synthetic data mixing. Seen AP is evaluated exclusively on the original 32 synthetic categories to ensure a fair comparison against the baseline.

Training Data	Seen	Unseen
(1) Pure OV	15.32	5.81
(2) + Ped./Cyc.	15.51	5.76
(3) + Real Car	15.18	4.97

5.6. Visualization

To provide intuitive insights into the performance of our CLIP-Mono3D model, we present qualitative visualizations of detection results on both the KITTI and OV-KITTI datasets, as shown in Figures 7 and 8.

Figure 7 visualizes detection results on KITTI. Our method exhibits higher recall for distant vehicles compared to MonoDETR. While sharing a similar architecture with MonoDGP, the integration

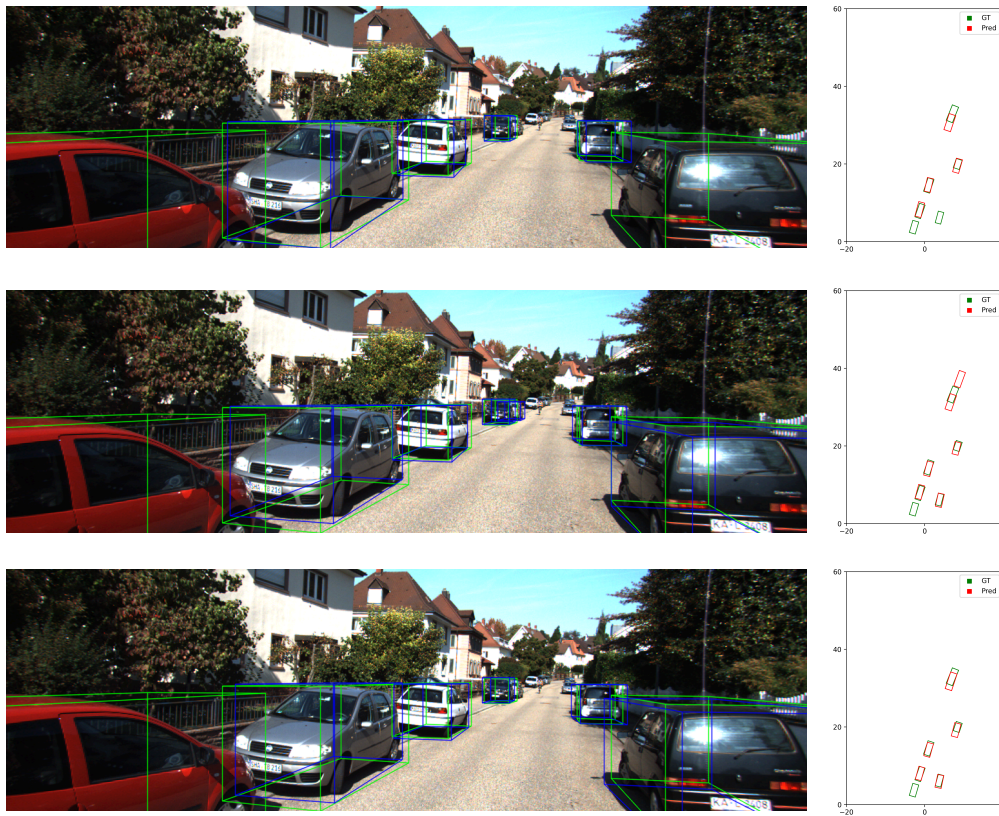


Figure 7. Detection visualization on KITTI. Left: front view with GT (blue) and predictions (red). Right: BEV view with GT (green) and predictions (red).

of CLIP-derived semantics improves 2D detection performance, which subsequently leads to more accurate distance regression. These results confirm that semantically enriched 2D features provide robust spatial priors for 3D localization, especially for long-range objects where geometric cues are often ambiguous.

Figure 8 compares CLIP-Mono3D with OVMono3D on the OV-KITTI benchmark. While OVMono3D occasionally yields more precise 2D boxes, our framework demonstrates superior 3D geometric reasoning. Specifically, our depth-guided transformer and language-aware queries enable more accurate depth estimation (row 2, “bench”), enhanced bounding box integrity for complex shapes (row 3, “giraffe”), and significantly reduced missed detections for small objects (row 4, “dog”), highlighting the robustness of our end-to-end semantic-geometric alignment in open-world scenarios.

5.7. Real-World Scenario Verification

To verify generalization in unconstrained real-world environments, we extend our evaluation to the standard KITTI and Argoverse datasets. For KITTI, we adopt a cross category split by training solely on Car and Pedestrian and evaluating zero shot performance on Cyclist. For Argoverse, we train on common categories and evaluate on seven distinct novel classes.

Settings. To adapt to the complex nature of the Argoverse dataset, we modified our query initialization strategy. Specifically, we shifted from the global mean aggregation used in OV-KITTI to a discrete spatial assignment paradigm. This ensures that queries are initialized at distinct spatial locations, effectively preventing multiple queries from collapsing onto a single salient object in complex scenes.

Results. As shown in Table 10, on the KITTI dataset, despite the extreme scarcity of training categories limiting the learning of generalized 3D shape priors, our end-to-end framework still achieves competitive performance. It is worth noting that the baseline OVMono3D utilizes an external open vocabulary 2D detector for strong 2D localization, whereas our method operates seamlessly without

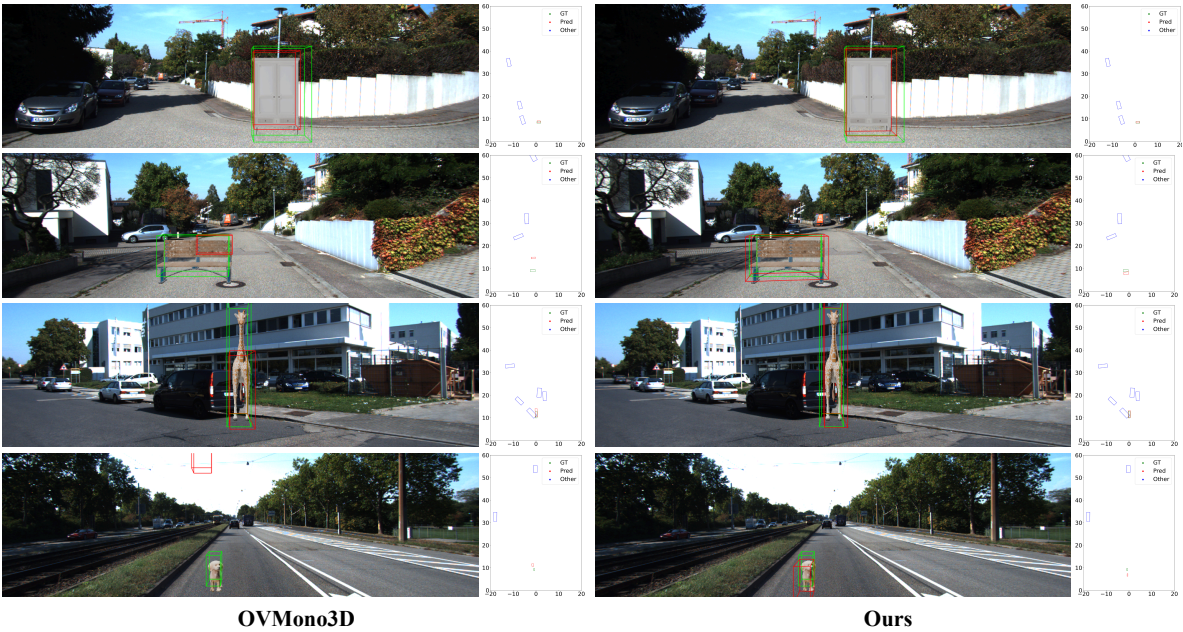


Figure 8. Qualitative comparison on OV-KITTI. Our method improves detection and localization, especially for distant objects.

explicit 2D bounding box guidance. On the larger scale Argoverse dataset detailed in Table 11, our method demonstrates robust generalization. Supported by the updated initialization strategy, our model successfully guides geometric reasoning for novel objects, achieving promising results across various categories and obtaining a higher overall average precision compared to the baseline.

To further illustrate the practical performance, we provide qualitative comparisons in Figure 9. In the first column, our method successfully detects the occluded moped while maintaining the detection completeness of the surrounding objects. In the third column, our model manages to recall the challenging animal class in the distance without missed detections. We acknowledge that the geometric regression for highly irregular novel classes like animals still requires further improvement and exhibits certain deviations. This is primarily due to the unreasonable category distribution inherent in the dataset and the extreme difficulty of these highly crowded scenes.



Figure 9. Qualitative comparison on Argoverse. Green boxes represent unseen categories, and red boxes indicate seen categories.

Table 10. Zero shot evaluation on KITTI ‘Cyclist’. Models are trained only on Car and Pedestrian. Metric: AP_{15} for 3D and BEV for Easy and Moderate settings.

Method	Cyclist AP_{3D}		Cyclist AP_{BEV}	
	Easy	Mod.	Easy	Mod.
OVMono3D	6.70	9.05	4.35	5.27
Ours	6.51	7.82	4.54	4.99

Table 11. Open vocabulary performance on Argoverse. Evaluated on 7 unseen categories.

Method	Animal	Bus	Emer. Veh.	Moped	Motor.	M.Cyc	Stroller	Overall
OVMono3D	1.56	23.60	4.93	3.03	14.11	14.49	0.35	8.87
Ours	2.91	19.88	4.82	5.96	14.37	17.12	0.72	9.40
Gain	+1.35	-3.72	-0.11	+2.93	+0.26	+2.63	+0.37	+0.53

6. Conclusion

We present CLIP-Mono3D, an end-to-end open-vocabulary monocular 3D detector that directly integrates vision-language semantics without relying on pretrained 2D detectors. To facilitate comprehensive evaluation, we introduce OV-KITTI, a new benchmark with balanced category distributions. Furthermore, we validate our framework on real-world datasets including KITTI and Argoverse. Extensive experiments demonstrate that our method achieves competitive performance in both closed- and open-vocabulary settings, showing robust generalization across diverse environments. We hope this work inspires further research in semantic-geometric fusion for 3D perception.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in CLIP-Mono3D at <https://github.com/ZC0102-shu/CLIP-Mono3D>.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012; pp. 3354–3361.
- Song, S.; Lichtenberg, S.P.; Xiao, J. Sun rgb-d: A rgb-d scene understanding benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015; pp. 567–576.
- Dai, A.; Chang, A.X.; Savva, M.; Halber, M.; Funkhouser, T.; Nießner, M. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017; pp. 5828–5839.
- Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020; pp. 11621–11631.
- Yao, J.; Gu, H.; Chen, X.; Wang, J.; Cheng, Z. Open Vocabulary Monocular 3D Object Detection. *arXiv* **2024**, arXiv:2411.16833.
- Pu, F.; Wang, Y.; Deng, J.; Yang, W. MonoDGP: Monocular 3D Object Detection with Decoupled-Query and Geometry-Error Priors. *arXiv* **2024**, arXiv:2410.19590.
- Xie, C.; Wang, B.; Kong, F.; Li, J.; Liang, D.; Zhang, G.; Leng, D.; Yin, Y. FG-CLIP: Fine-Grained Visual and Textual Alignment. *arXiv* **2025**, arXiv:2505.05071.
- Qin, Z.; Li, X. Monoground: Detecting monocular 3d objects from the ground. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022; pp. 3793–3802.
- Shi, X.; Ye, Q.; Chen, X.; Chen, C.; Chen, Z.; Kim, T.K. Geometry-based distance decomposition for monocular 3d object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021; pp. 15172–15181.

10. Liu, Z.; Wu, Z.; Tóth, R. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020; pp. 996–997.
11. Wang, T.; Zhu, X.; Pang, J.; Lin, D. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021; pp. 913–922.
12. Yao, H.; Chen, J.; Wang, Z.; Wang, X.; Han, P.; Chai, X.; Qiu, Y. Occlusion-Aware Plane-Constraints for Monocular 3D Object Detection. *IEEE Trans. Intell. Transp. Syst.* **2023**.
13. Liu, Z.; Zhou, D.; Lu, Feixiang.; Fang, J.; Zhang, L. Autoshape: Real-time shape-aware monocular 3d object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021; pp. 15641–15650.
14. Lu, Y.; Ma, X.; Yang, L.; Zhang, T.; Liu, Y.; Chu, Q.; Yan, J.; Ouyang, W. Geometry uncertainty projection network for monocular 3d object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021; pp. 3111–3121.
15. Yang, L.; Yu, K.; Tang, T.; Li, J.; Yuan, K.; Wang, L.; Zhang, X.; Chen, P. Bevhight: A robust framework for vision-based roadside 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023; pp. 21611–21620.
16. Wang, Y.; Chao, W.L.; Garg, D.; Hariharan, B.; Campbell, M.; Weinberger, K.Q. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019; pp. 8445–8453.
17. Ma, X.; Liu, S.; Xia, Z.; Zhang, H.; Zeng, X.; Ouyang, W. Rethinking pseudo-lidar representation. In Proceedings of the European Conference on Computer Vision (ECCV), 2020; pp. 311–327.
18. Chong, Z.; Ma, X.; Zhang, H.; Yue, Y.; Li, H.; Wang, Z.; Ouyang, W. Monodistill: Learning spatial features for monocular 3d object detection. *arXiv* **2022**, arXiv:2201.10830.
19. Rukhovich, D.; Vorontsova, Anna.; Konushin, A. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022; pp. 2397–2406.
20. Park, D.; Ambrus, R.; Guizilini, V.; Li, J.; Gaidon, A. Is pseudo-lidar needed for monocular 3d object detection? In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021; pp. 3142–3152.
21. Huang, K.C.; Wu, T.H.; Su, H.T.; Hsu, W.H. Monodtr: Monocular 3d object detection with depth-aware transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022; pp. 4012–4021.
22. Zhang, R.; Qiu, H.; Wang, T.; Guo, Z.; Cui, Z.; Qiao, Y.; Li, H.; Gao, P. MonoDETR: Depth-guided transformer for monocular 3D object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023; pp. 9155–9166.
23. Ma, Z.; Luo, G.; Gao, J.; Li, L.; Chen, Y.; Wang, S.; Zhang, C.; Hu, W. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022; pp. 14074–14083.
24. Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L.H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. Regionclip: Region-based language-image pretraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022; pp. 16793–16803.
25. Zareian, A.; Rosa, K.D.; Hu, D.H.; Chang, S.F. Open-vocabulary object detection using captions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021; pp. 14393–14402.
26. Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; Misra, I. Detecting twenty-thousand classes using image-level supervision. In Proceedings of the European Conference on Computer Vision (ECCV), 2022; pp. 350–368.
27. Yao, L.; Han, J.; Liang, X.; Xu, D.; Zhang, W.; Li, Z.; Xu, H. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023; pp. 23497–23506.
28. Wu, X.; Zhu, F.; Zhao, R.; Li, H. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023; pp. 7031–7040.
29. Zang, Y.; Li, W.; Zhou, K.; Huang, C.; Loy, C.C. Open-vocabulary detr with conditional matching. In Proceedings of the European Conference on Computer Vision (ECCV), 2022; pp. 106–122.

30. Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; Shan, Y. Yolo-world: Real-time open-vocabulary object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024; pp. 16901–16911.
31. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, Chunyuan.; Yang, J.; Su, Hang.; et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In Proceedings of the European Conference on Computer Vision (ECCV), 2024; pp. 38–55.
32. Zhao, T.; Liu, P.; He, X.; Zhang, L.; Lee, K. Real-time Transformer-based Open-Vocabulary Detection with Efficient Fusion Head. *arXiv* **2024**, arXiv:2403.06892.
33. Li, L.H.; Zhang, P.; Zhang, H.; Yang, J.; Li, Chunyuan.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, Lei.; Hwang, J.N.; et al. Grounded language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022; pp. 10965–10975.
34. Chuang, L.; Yi, J.; Lizhen, Q.; Zehuan, Y.; Jianfei, C. Generative region-language pretraining for open-ended object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.
35. Yao, L.; Pi, R.; Han, J.; Liang, X.; Xu, Hang.; Zhang, Wei.; Li, Zhenguo.; Xu, Dan. DetCLIPv3: Towards Versatile Generative Open-vocabulary Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024; pp. 27391–27401.
36. Cen, J.; Yun, P.; Cai, J.; Wang, M.Y.; Liu, M. Open-set 3d object detection. In Proceedings of the 2021 International Conference on 3D Vision (3DV), 2021; pp. 869–878.
37. Lu, Y.; Xu, C.; Wei, X.; Xie, X.; Tomizuka, M.; Keutzer, K.; Zhang, S. Open-vocabulary point-cloud object detection without 3d annotation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023; pp. 1190–1199.
38. Cao, Y.; Zeng, Y.; Xu, Hang.; Xu, Dan. Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *36*.
39. Zhu, C.; Zhang, W.; Wang, T.; Liu, X.; Chen, K. Object2scene: Putting objects in context for open-vocabulary 3d detection. *arXiv* **2023**, arXiv:2309.09456.
40. Wang, Z.; Li, Y.; Liu, T.; Zhao, H.; Wang, S. OV-Uni3DETR: Towards Unified Open-Vocabulary 3D Object Detection via Cycle-Modality Propagation. *arXiv* **2024**, arXiv:2403.19580.
41. Xue, L.; Gao, M.; Xing, C.; Martín-Martín, R.; Wu, J.; Xiong, C.; Xu, R.; Niebles, J.C.; Savarese, S. 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 (CVPR), 2023; pp. 1179–1189.
42. Zhang, R.; Guo, Z.; Zhang, W.; Li, K.; Miao, X.; Cui, B.; Qiao, Yu.; Gao, Peng.; Li, Hongsheng. Pointclip: Point cloud understanding by clip. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022; pp. 8552–8562.
43. Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; Farhadi, A. Objaverse: A universe of annotated 3d objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023; pp. 13142–13153.
44. Liu, X.; Xue, N.; Wu, T. Learning auxiliary monocular contexts helps monocular 3d object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, 2022; pp. 1810–1818.
45. Yan, L.; Yan, P.; Xiong, S.; Xiang, X.; Tan, Y. Monocd: Monocular 3d object detection with complementary depths. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024; pp. 10248–10257.
46. Wu, Z.; Gan, Y.; Wu, Y.; Wang, R.; Wang, Xiaoquan.; Pu, Jian. Fd3d: Exploiting foreground depth map for feature-supervised monocular 3d object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, 2024; pp. 6189–6197.
47. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), 2022.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.