



# GUIDED++: Enhancing Discrimination with Conjunctive Verification for Fine-Grained Open-Vocabulary Object Detection

Jiaming Li<sup>1,2</sup> · Zhijia Liang<sup>1</sup> · Shuangyin Liu<sup>4</sup> · Chengpei Tang<sup>5</sup> · Guanbin Li<sup>1,2,3</sup> 

Received: 15 December 2025 / Accepted: 25 July 2026 / Published online: 9 August 2026  
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2026

## Abstract

Fine-grained open-vocabulary detection (FG-OVD) challenges models to detect objects described by complex, compositional language. The prevalent strategy of encoding descriptive phrases into a single, holistic vector creates significant semantic ambiguity, hindering true compositional understanding. This ambiguity manifests as two opposing failures: attribute over-representation, where a visually dominant attribute leads to mislocalization, and attribute under-representation, where the model fails to enforce a conjunctive match across all specified attributes, leading to false positives. In this paper, we propose GUIDED++, a framework that systematically resolves both issues through a dual strategy. First, we employ architectural decoupling to separate coarse-grained localization from fine-grained discrimination, effectively mitigating attribute over-representation. Second, we introduce a novel conjunctive multi-attribute verification mechanism to explicitly combat attribute under-representation. This mechanism decomposes a complex query into its atomic attribute conditions, computes a distinct similarity score for each, and uses a minimal aggregation function to enforce a strict logical ‘AND’, ensuring an object satisfies all required attributes. Extensive experiments show that GUIDED++ significantly outperforms existing methods and establishes a new state-of-the-art on challenging FG-OVD benchmarks, demonstrating a more robust approach to compositional visual reasoning.

**Keywords** Fine-grained open vocabulary object detection · Open vocabulary object detection · Vision language models · Conjunctive verification

Communicated by Hong Liu.

✉ Guanbin Li  
liguanbin@mail.sysu.edu.cn

Jiaming Li  
lijm48@mail2.sysu.edu.cn

Zhijia Liang  
liangzhj56@mail2.sysu.edu.cn

Shuangyin Liu  
shuangyinliu@zhku.edu.cn

Chengpei Tang  
tchengp@mail.sysu.edu.cn

<sup>1</sup> School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, Guangdong, China

<sup>2</sup> Peng Cheng Laboratory, Shenzhen 518066, Guangdong, China

<sup>3</sup> Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, Guangdong, China

## 1 Introduction

A central goal of artificial intelligence is to build models that understand the world with the same compositional richness as human language. While Open-Vocabulary Object Detection (OVD) has democratized visual recognition by allowing models to detect unseen objects via free-form text, the true reliability of these systems is tested by their capacity to handle the combinatorial complexity of fine-grained descriptions. When a model fails to distinguish “a black car” from “a car with black tires”, it exhibits a failure of semantic alignment that fundamentally undermines its trustworthiness in open-world deployment. Fine-grained open-vocabulary detection (FG-OVD) thus emerges not merely as a harder

<sup>4</sup> College of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, Guangdong, China

<sup>5</sup> School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou 510225, Guangdong, China

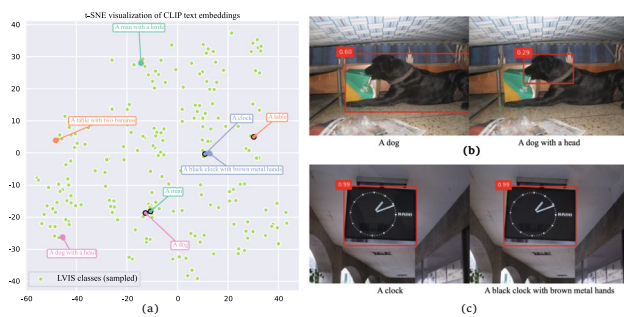


**Fig. 1** The challenge of fine-grained open-vocabulary detection (FG-OVD) stems from semantic ambiguity in text embeddings, leading to two critical failure modes. **a** A t-SNE visualization reveals this ambiguity at its source: fine-grained concepts are either scattered (e.g., A table with two bananas) or indistinguishable from their coarse parent (e.g., A black clock with brown metal hands). This results in **b** attribute over-representation, causing mislocalization where a salient attribute (hat) hijacks the detector’s focus from the main subject (dog), and **c** attribute under-representation, where the model fails to enforce conjunctive logic, incorrectly matching a clock despite the mismatched attribute (black vs. brown metal hands)

detection task, but as a critical frontier for ensuring that visual recognition systems remain faithful to user intent. By enforcing strict adherence to multi-attribute constraints, FG-OVD serves as a necessary crucible for validating a model’s reasoning capabilities against the ambiguity of natural language.

To generalize to unseen classes, existing OVD approaches typically leverage Vision-Language Models (VLMs) to encode complex descriptions into flat, holistic class embeddings. However, this holistic encoding compresses structured logic into a single vector, creating an entangled representation that obscures the distinction between subjects and modifiers. This lossy compression leads to two critical failures. First, attribute over-representation occurs when visually dominant attributes override the subject identity. For instance, the query “a dog with a head” often leads OVD models to erroneously localize the “head” rather than the “dog” (Fig. 1b). This spatial misalignment reflects a deeper distortion in the semantic topology. As feature representations become dominated by attributes, queries sharing the same subject fail to cluster cohesively. Instead, varying attributes fragment the subject’s embedding space (Fig. 1a), preventing the model from recognizing these fine-grained queries as variants of the same object class.

Conversely, addressing over-representation unmask a more insidious problem: attribute under-representation. When a class is defined by a conjunction of multiple attributes (e.g., “a black clock with a brown metal hands” as in Fig. 1c), the single-vector representation tends to average out their semantic contributions. A model might yield a high score for a “black clock with a brown metal hands” because the strong match on “black” and “clock” masks the critical mismatch on “with a brown metal hands”. This semantic averaging effect

demonstrates the model’s inability to enforce the strict, conjunctive logic inherent in fine-grained descriptions. Current methods fundamentally lack a mechanism to verify that an object satisfies all specified attributes simultaneously.

To resolve these dual challenges, we propose GUIDED++, an evolution of our prior work, GUIDED. While GUIDED initiated the decoupling of subject and attribute to mitigate attribute over-representation, it did not fully address the more insidious problem of attribute under-representation, as its matching still relied on a holistic score. The core motivation of this work is therefore to introduce a mechanism for strict conjunctive logic. To this end, we re-architect the FG-OVD process around a dual strategy: architectural decoupling and conjunctive verification. First, we employ architectural decoupling to separate coarse-grained object detection from fine-grained discrimination. This allows the detector to focus on the stable subject without interference from dominant attributes during the coarse-grained object detection, effectively neutralizing attribute over-representation. To retain essential context while preventing attribute dominance, this stage is further enhanced by an Attribute Embedding Fusion module, which selectively injects attribute cues via attention mechanisms. Second, to explicitly counter attribute under-representation, we introduce a novel conjunctive multi-attribute verification mechanism for fine-grained discrimination. Unlike the holistic matching in prior methods, we decompose the query into atomic, single-attribute phrases and derive the final fine-grained score from the minimum of these individual scores. This “all-or-nothing” verification enforces a logical ‘AND’, ensuring that an object is considered a match only if it satisfies every defining attribute.

The GUIDED++ pipeline operationalizes this strategy through a three-stage process: (1) An offline Category Name Parsing stage uses an LLM to generate coarse and atomic queries; (2) The Coarse-Grained Detection stage generates proposals for subjects; and (3) The conjunctive multi-attribute verification applies our conjunctive verification logic to these proposals. Extensive experiments on challenging FG-OVD benchmarks demonstrate that GUIDED++ sets a new state-of-the-art, significantly outperforming existing methods. Our contributions are:

- We identify attribute over-representation and under-representation as key limitations in existing OVD methods, which compromises the trustworthiness of visual recognition.
- We propose GUIDED++, a novel framework that enhances robustness by addressing these limitations through an architectural decoupling strategy and an attribute embedding fusion module.

- We introduce a conjunctive multi-attribute verification mechanism, which departs from holistic matching by decomposing complex queries and using a minimal aggregation strategy to enforce strict logical adherence.
- Our method establishes a new state-of-the-art on FG-OVD benchmarks, demonstrating significant advantages in resolving conjunctive hard negatives compared to existing approaches.

This paper presents a substantial extension of our conference work, GUIDED (Li et al., 2025). The matching paradigm in GUIDED, despite its decoupled architecture, still computed a single, holistic similarity score, which proved insufficient for queries requiring strict logical conjunction. While it mitigated attribute over-representation, it was still vulnerable to the more subtle challenge of attribute under-representation. The core contribution in this work is therefore the introduction of the conjunctive multi-attribute verification (CMAV) mechanism, which fundamentally shift the matching paradigm from holistic approximation to strict logical enforcement. To achieve this, we designed a new training mechanism that explicitly teaches the model this conjunctive reasoning. Finally, we validate GUIDED++ through a more extensive set of experiments that specifically probe its ability to resolve conjunctive hard negatives, demonstrating a significant advancement.

## 2 Related Work

### 2.1 Open Vocabulary Object Detection

Open-vocabulary object detection (OVD) re-frames object detection from a closed-set classification problem to an open-set vision-language alignment task. This allows detectors to recognize categories not seen during training by leveraging textual descriptions. The evolution of OVD can be broadly understood through two major paradigms. The initial and most direct approach was to transfer the powerful zero-shot capabilities of pre-trained Vision-Language Models (VLMs) like CLIP (Radford et al., 2021) to the detection task. This was primarily achieved in knowledge distillation and fine-tuning adaptation. Methods like ViLD (Gu et al., 2021) treat the VLM as a frozen teacher model. The detector's region-level features are aligned with the VLM's image-level or text-level embeddings, thereby distilling the VLM's semantic knowledge into the detector (Li et al., 2024b; Wu et al., 2023a; Li et al., 2024a; Zhao et al., 2024a; Du et al., 2022; Kim et al., 2024; Wang et al., 2024a; Xi et al., 2025; Ma et al., 2025c). Recently, CAKE (Ma et al., 2025b) refined this by proposing a category-aware distillation process to extract more specific knowledge and mitigate the suppres-

sion of novel objects. Other works fine-tune the VLM itself to better suit detection, for instance, by creating region-text pairs for contrastive learning (Zhong et al., 2022; Yao et al., 2023; Minderer et al., 2022; Cheng et al., 2024; Yao et al., 2024; Zhao et al., 2024b; Wang et al., 2023). Alternatively, methods like F-VLM (Kuo et al., 2022) bypass distillation by directly pooling frozen VLM features at inference time, fusing them with the detector's category scores via a geometric mean. While effective for standard OVD, F-VLM focuses strictly on homogeneous category-level fusion, where both branches predict coarse object classes. It does not address the extraction or verification of fine-grained attributes, which requires a heterogeneous coarse-to-fine division of labor. The fundamental limitation of this paradigm is the granularity mismatch. Since VLMs like CLIP are pre-trained on image-level objectives, their features are not inherently optimized for the precise, region-level localization and classification required by detection. A second paradigm emerged to directly tackle this granularity mismatch by pre-training models on object-level data. Methods like GLIP (Li et al., 2022) pioneered this by unifying object detection and phrase grounding tasks on a massive scale, using a diverse mix of detection, grounding, and image-text datasets. This creates a vision-language space that is inherently object-aware. This paradigm has been further strengthened by integrating modern architectures. Grounding DINO (Liu et al., 2023) combines this object-centric approach with the end-to-end DETR framework for improved efficiency and performance. More recently, the field has begun leveraging Large Language Models (LLMs) for enhanced supervision. LAMI-DETR (Du et al., 2024) further leverages language models to generate conceptual relationships, enhancing the detector's reasoning. LLMDet (Fu et al., 2025) demonstrates that co-training with an LLM that generates detailed, image-level captions can significantly improve the detector's open-vocabulary ability. Despite the significant progress across these paradigms, a common and persistent challenge remains: the inability to robustly distinguish between fine-grained categories defined by subtle attributes. This limitation has catalyzed the emergence of a dedicated subfield, namely FG-OVD.

### 2.2 Fine-Grained Open Vocabulary Object Detection

Fine-grained open-vocabulary detection (FG-OVD) (Bianchi et al., 2024a) seeks to detect objects defined by compositional descriptions. The dominant paradigm in this area has focused on text embedding refinement, attempting to encode complex attribute information directly into a single feature vector. HAFGOVD (Ma et al., 2025a) exemplifies this strategy, which proposes masking the attention maps of a Vision-Language Model (VLM) to generate attribute-highlighted text embeddings. Other methods follow a similar philosophy, such as crafting hierarchy-aware sentences (Liu et al., 2024; Bianchi

et al., 2024b), fine-tuning an additional projection layer to enhance the embedding's fine-grained capacity.

However, these approaches share a fundamental flaw. They compress a structured, multi-attribute query into a single, holistic vector. This compression is inherently lossy and creates what we term semantic ambiguity, which manifests in two opposing failure modes. On one hand, forcing attribute awareness into a single vector often leads to attribute over-representation, where a visually salient attribute hijacks the representation, causing severe mislocalization. A recent work (Dou et al., 2025) recognizes the need for decomposition by learning attributes in a multi-label fashion. However, by retaining a unified detection framework, it still tasks a single model with the conflicting goals of localization and verification and, critically, does not introduce an explicit mechanism to enforce a conjunctive match. Therefore, existing methods, even those recognizing decomposition, fail to fully resolve the semantic ambiguity inherent in FG-OVD. Our work, detailed in the following sections, introduces a framework designed to explicitly address both attribute over-representation through architectural decoupling and attribute under-representation via a novel conjunctive verification logic.

### 3 Approach

This paper focuses on Fine-grained open-vocabulary detection (FG-OVD), which aims to detect both the base class in training set and the unseen fine-grained classes with specific attributes. To this end, we propose GUIDED++, an enhanced framework based on our conference work, GUIDED (Li et al., 2025). The core philosophy of GUIDED++ is to re-architect FG-OVD from a single, holistic matching task into a structured, sequential verification process. Instead of solving the entangled problem of localization and multi-attribute discrimination simultaneously, our framework decomposes it into coarse-grained object detection and fine-grained discrimination. This is achieved through a three-stage pipeline, built upon a DETR-based architecture like LaMI-DETR, as illustrated in Fig. 2. GUIDED++ first uses an LLM to parse queries into coarse subjects and atomic attributes. A detector then localizes objects with the embeddings of the coarse subjects, while dynamically fusing attribute-specific semantics through an attribute embedding fusion module. Finally, our proposed conjunctive verification mechanism scores these proposals by taking the minimum value of their similarities to each atomic attribute, enforcing a logical 'AND' to ensure all conditions are met.

#### 3.1 Category Name Parsing

The first step in operationalizing architectural decoupling and conjunctive verification is to decompose complex, compositional category names into specialized prompts. This offline pre-processing stage uses a powerful LLM (e.g., LLaMA 3.3-70B (Grattafiori et al., 2024)) to generate two distinct sets of queries for each fine-grained category, each tailored to combat one of the failure modes identified in our introduction.

First, to enable architectural decoupling and mitigate attribute over-representation, we generate a single, robust coarse-grained subject query. This query isolates the core object identity, providing a stable target for the initial localization stage. Second, to facilitate our conjunctive verification mechanism and counteract attribute under-representation, we generate a set of atomic, single-attribute queries. Each of these queries combines the coarse subject with just one of its original attributes. For instance, for a full category name "A black plastic television set with a logo pattern", we phrase it into subject "A television set" and the atomic attribute set ("A black television set", "A plastic television set", "A television set with a logo pattern"). This design is central to our verification logic, as it forces the model to check for the attribute on the subject, preventing spurious matches where an attribute is present in the region but belongs to a different object. The specific prompts used for subjects and atomic queries are detailed in Fig. 3 and Fig. 7 in Appendix, respectively. This parsing method can also be applied to handle implicit attributes. For a class like "Siberian Husky," it can infer the coarse subject "a dog" and generate atomic queries based on its world knowledge of the breed's defining features.

Formally, for each fine-grained category  $c$  in the vocabulary  $\mathcal{C}$ , this process yields a single coarse-grained subject query and a corresponding set of  $n_c$  atomic attribute queries. We then use a pre-trained, frozen text encoder  $\mathcal{T}$  to generate a coarse-grained text embedding  $\{\mathbf{e}_c \in \mathbb{R}^d\}_{c \in \mathcal{C}}$  and a set of atomic attribute embeddings  $\{\hat{\mathbf{e}}_{c,i} \in \mathbb{R}^d\}_{i=1}^{n_c}$ , where  $d$  is the embedding dimension. These decomposed embeddings form the foundation for the subsequent detection and scoring stages.

#### 3.2 Coarse-Grained Object Detection

The goal of the coarse-grained object detection (CGOD) stage is to generate a set of object proposals for all coarse-grained categories present in the image. To this end, we employ a detector building upon LAMI-DETR (Du et al., 2024), which is based on the architecture of DINO.

In LaMI-DETR, an input image is first processed by a frozen image backbone (ConvNeXt (Liu et al., 2022) from a pre-trained CLIP) to produce a multi-scale feature map  $\mathbf{F}_{\text{conv}}$ . This feature map is then fed into a learnable Transformer

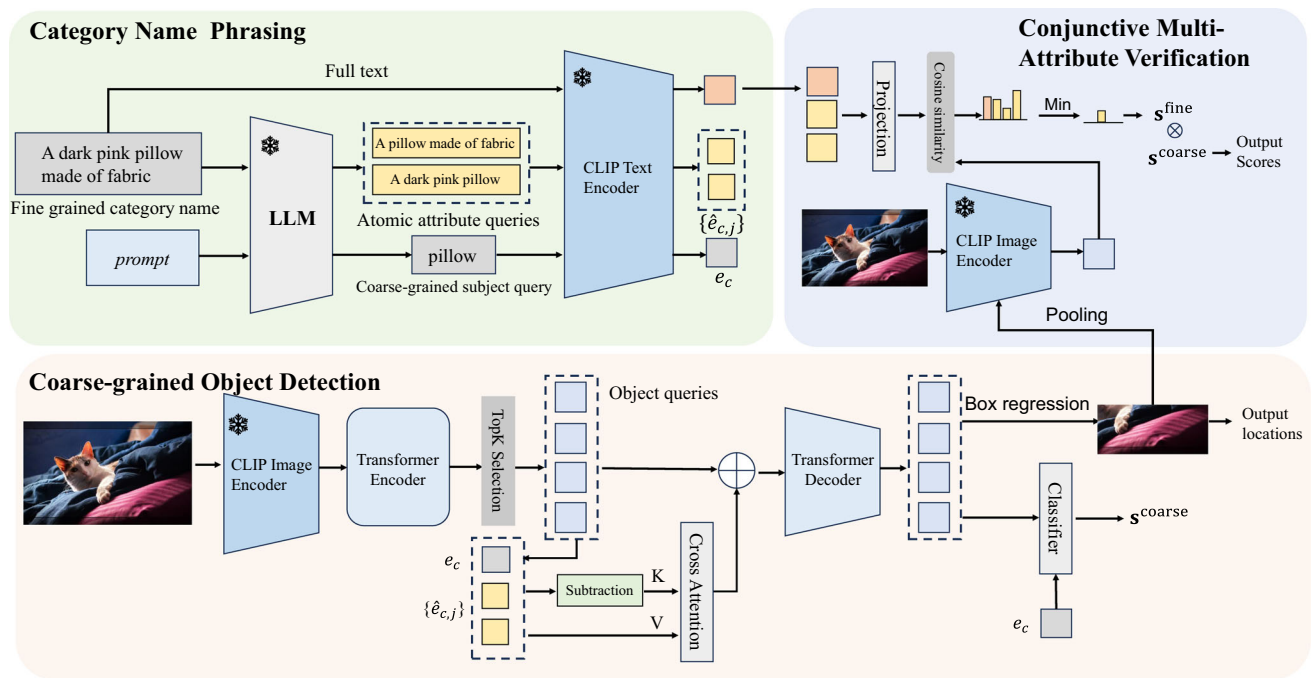


**Fig. 2** An overview of the GUIDED++ pipeline. (1) Category name parsing: An LLM decomposes the full query into a coarse subject for localization and a set of atomic attributes for verification. (2) Coarse-grained detection: To mitigate attribute over-representation, the detector uses only the coarse subject embedding to generate robust object proposals. Our attribute embedding fusion module provides subtle

context without hijacking localization. (3) Conjunctive multi-attribute verification: This novel stage explicitly counteracts attribute under-representation. Instead of using a single, averaged vector, it computes separate similarity scores against each atomic attribute embedding. The final score is the minimum of these scores, enforcing a logical ‘AND’ to ensure all attributes are present

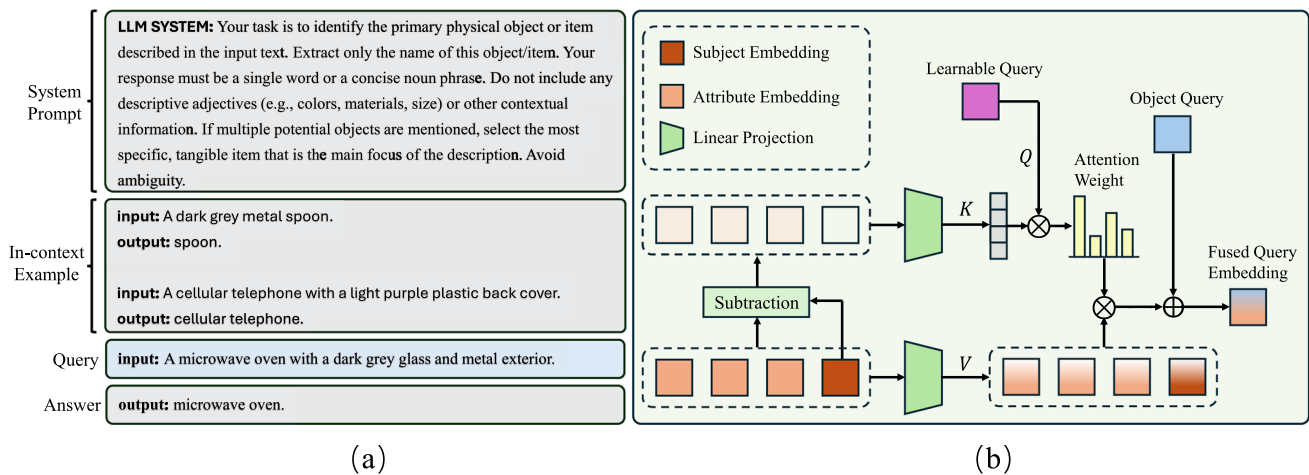


**Fig. 3** **a** Illustration of the prompt for phrasing the subject in category names. **b** The architecture of the AEF module

encoder to yield a refined feature map  $\mathbf{F}_{\text{enc}}$ , which captures rich contextual information. The refined feature map is subsequently input to the decoder and the classifier to generate the detection predictions, which are elaborated below.

**Coarse-grained OVD classifier.** Conventional OVD frameworks typically leverage text embeddings from pre-trained VLMs (e.g., CLIP) to initialize the classification weights, thereby enabling zero-shot generalization to unseen cate-

gories. However, as highlighted in our introduction, directly employing holistic embeddings of complex, fine-grained descriptions inherently introduces semantic ambiguity and attribute over-representation. To circumvent this, we construct the classifier using exclusively the subject embeddings  $\{e_c\}_{c \in \mathcal{C}}$ , rather than the full phrase embeddings. This strategy effectively decouples the subject identity from attribute

ambiguity. Formally, we denote this subject-aligned classifier as  $\mathcal{C}(\{\mathbf{e}_c\}_c, \cdot)$ .

**Attribute Embedding Fusion.** Following the query selection strategy in DINO, we estimate the classification score of each pixel in  $\mathbf{F}_{\text{enc}}$  with the classifier  $\mathcal{C}(\{\mathbf{e}_c\}_c, \cdot)$ . Subsequently, we select the top- $K$  pixel features with the highest classification scores to serve as initial object queries  $\{\mathbf{q}_j\}_{j=1}^K$ . Each query  $\mathbf{q}_j$  is associated with a predicted category  $c_j$ . To further guide the decoder, we enrich these queries with specific attribute information using an Attribute Embedding Fusion (AEF) module. For each query  $\mathbf{q}_j$  matched with coarse class  $c_j$ , we fuse it with its corresponding coarse-grained text embeddings  $\mathbf{e}_{c_j}$  and the set of all possible attribute embeddings  $\{\hat{\mathbf{e}}_{c_j,i}\}$  for that class. This is achieved via a cross-attention layer:

$$\mathbf{q}_j^f = \mathbf{q}_j + \text{ATTN}(\mathbf{q}_j^l, \mathbf{K}_j, \mathbf{V}_j), \quad (1)$$

where  $\mathbf{q}_j^l$  is a learnable content query. The key and value matrices are constructed to highlight attribute-specific semantics:

$$\begin{aligned} \mathbf{K}_j &= \{\mathbf{e}_{c_j} - \mathbf{e}_{c_j}\} \cup \{\hat{\mathbf{e}}_{c_j,i} - \mathbf{e}_{c_j}\}_i, \\ \mathbf{V}_j &= \{\mathbf{e}_{c_j}\} \cup \{\hat{\mathbf{e}}_{c_j,i}\}_i. \end{aligned} \quad (2)$$

The overview of AEF is shown in Fig. 3. Using the difference between attribute and coarse embeddings as keys encourages the attention mechanism to focus on the semantic difference provided by each attribute.

The fused queries  $\{\mathbf{q}_j^f\}$  are then fed into the transformer decoder, which refines them and outputs  $K$  predictions. Each prediction  $j$  consists of a bounding box  $\mathbf{b}_j$  and a final prediction embedding  $\mathbf{p}_j$ . The coarse-grained confidence score vector  $\mathbf{s}_j^{\text{coarse}}$  is computed as:

$$\mathbf{s}_j^{\text{coarse}} = \text{Sigmoid}(m_{\text{coarse}} \cdot \text{CLS}(\{\mathbf{w}_c\}, \mathbf{p}_j)), \quad (3)$$

where  $m_{\text{coarse}}$  is a hyper parameter for scaling.

### 3.3 Conjunctive Multi-attribute Verification

The final stage of our framework is to discriminate the robust proposals from CGOD at a fine-grained level. This requires a mechanism that explicitly counteracts attribute under-representation. To this end, we introduce the conjunctive multi-attribute verification (CMAV) module.

Concretely, for each proposal box  $\mathbf{b}_j$ , we extract its visual features  $\mathbf{v}_j$  by applying RoIAlign to the backbone feature map  $\mathbf{F}_{\text{conv}}$ . Concurrently, we process the set of atomic attribute queries for each class  $c$ . To adapt this space for the specific demands of fine-grained discrimination, the CLIP text encoder  $\mathcal{T}_{\text{CLIP}}$  is augmented with a lightweight, learnable linear projection layer, creating the refined encoder  $\mathcal{T}'$ .

This refined encoder produces a set of refined atomic attribute embeddings  $\{\tilde{\mathbf{e}}_{c,i}\}_{i=1}^{n_c}$  for the corresponding atomic attribute queries. We also compute a refined full text embedding  $\mathbf{e}'_c$  for the full text query by the refined text encoder.

Then we compute a distinct score for each atomic attribute and the full-text query. For a given proposal  $\mathbf{b}_j$  with visual feature and for each class  $c$  with its set of  $n_c$  refined atomic attribute embeddings  $\{\tilde{\mathbf{e}}_{c,i}\}_{i=1}^{n_c}$ , we first compute the individual attribute matching logits,

$$\mathbf{I}_{j,c,i}^{\text{attr}} = m_{\text{fine}} \cos(\mathbf{v}_j, \tilde{\mathbf{e}}_{c,i}). \quad (4)$$

Each logit represents the evidence that the proposal  $\mathbf{b}_j$  satisfies the specific attribute of class  $c$ . Subsequently, we also compute a full matching logit with  $\mathbf{e}'_c$ ,

$$\mathbf{I}_{j,c}^{\text{full}} = m_{\text{fine}} \cos(\mathbf{v}_j, \mathbf{e}'_c). \quad (5)$$

The full matching logit represents the global matching with the full text. While atomic attribute verification excels at distinguishing semantically close classes, it is sensitive to noise from a single attribute when comparing semantically distant classes that differ across multiple attributes. The full matching logit mitigates this by offering a holistic signal that robustly captures the overall semantic distance, ensuring reliable discrimination even against semantically distant classes.

During inference, we enforce a conjunctive match by aggregating these logits with a minimum function. To this end, we aggregate these logits by,

$$\mathbf{I}_{j,c}^{\text{fine}} = \min(\mathbf{I}_{j,c}^{\text{full}} \cup \{\mathbf{I}_{j,c,i}^{\text{attr}}\}_i), \quad (6)$$

This design dynamically leverages the precision of atomic checks for fine distinctions and the stability of the global signal for coarse ones. We concatenate the logits for all classes to form the fine-grained logit vector  $\mathbf{I}_j^{\text{fine}}$ . Then, we obtain the final finegrained scores over all classes for the proposal by applying a SoftMax function to  $\mathbf{I}_j^{\text{fine}}$ :

$$\mathbf{s}_j^{\text{fine}} = \text{Softmax}(\mathbf{I}_j^{\text{fine}}). \quad (7)$$

This “all-or-nothing” approach directly involves semantic averaging. A high final score is only possible if the object proposal satisfies every specified attribute, effectively preventing the model from ignoring a critical mismatch, no matter how strong the match is on other attributes.

Given the fine-grained matching scores  $\mathbf{s}_j^{\text{fine}}$  and the coarse-grained confidence scores, the final scores for proposal  $\mathbf{b}_j$  are estimated by,

$$\mathbf{s}_j^{\text{final}} = (\mathbf{s}_j^{\text{coarse}})^{(\alpha)} (\mathbf{s}_j^{\text{fine}})^{(1-\alpha)}, \quad (8)$$

where  $\alpha$  is a hyperparameter that balances the magnitudes of the two scores. While similar geometric mean formulations have been used in prior works like F-VLM (Kuo et al., 2022) for category-level score ensembling, our motivation differs. Specifically, we employ it as a strict logical ‘AND’ gate for coarse-to-fine verification, ensuring the proposal satisfies both the coarse subject identity and all fine-grained attributes.

### 3.4 Training Strategy

A key challenge is training a model to succeed at the conjunctive inference. Directly optimizing through the min function is problematic. Its gradient is non-smooth and, more critically, it would route the learning signal to only one weakest attribute per sample. This would starve other embeddings of supervision, leading to poor convergence. To address this, we propose a novel training strategy that serves as a more stable and effective proxy objective. The training proceeds in two stages.

**Coarse-grained detector pre-training.** We first train the detector along with the CGOD module on a large-scale standard OVD dataset (e.g., COCO base categories). The model is trained using the standard set-based loss from Lami-Detr, which includes classification and box regression losses. The supervision is provided by the coarse-grained subject labels extracted from the category name parsing. The attribute fusion modules are active during this stage, using attributes extracted from base class names by an LLM as in LaMI-DETR, to accustom the model to attribute-aware learning.

**End-to-end fine-tuning for FG-OVD.** In the second stage, we fine-tune the entire detector and the learnable linear projection layer in the refined encoder  $\mathcal{T}'$  on the FG-OVD dataset. The primary objective of this stage is to enhance the model’s discriminative capability, specifically to resolve the semantic ambiguity between fine-grained classes that differ by only a few attributes. To this end, we introduce a training mechanism that explicitly forces the model to identify and penalize these subtle yet critical attribute mismatches.

Given the atomic attribute embeddings  $\{\tilde{e}_{c,i}\}_{i=1}^{n_c}$  of a class  $c$  and proposal  $b_j$ , we then identify the subset of these attributes that do not match the ground-truth class associated with the proposal. Let  $\{\mathbf{I}_{j,c,i}^{\text{diff}}\}_i$  denote the similarity logits corresponding to this subset of distinguishing attributes. We aggregate them into a single error signal by averaging,

$$\mathbf{I}_{j,c}^{\text{diff}} = \text{Mean}(\{\mathbf{I}_{j,c,i}^{\text{diff}}\}_i). \quad (9)$$

Using ‘Mean’ provides a stable, comprehensive error signal from all distinguishing attributes, avoiding the instability of focusing on a single attribute during training.

Then, to balance the training between the holistic embedding and the numerous atomic ones, we employ a stochastic

sampling strategy. The final logit  $\mathbf{I}_{j,c}$  used for training is sampled via a Bernoulli trial,

$$\mathbf{I}_{j,c} = \begin{cases} \mathbf{I}_{j,c}^{\text{diff}} & \text{with probability } p, \\ \mathbf{I}_{j,c}^{\text{full}} & \text{with probability } 1 - p. \end{cases} \quad (10)$$

This strategy prevents the model from overfitting to a single descriptive modality of atomic queries. Furthermore, by alternating between  $\mathbf{I}_{j,c}^{\text{diff}}$  and  $\mathbf{I}_{j,c}^{\text{full}}$  for the same supervised signal, the training process compels the model to project both full-text and atomic attribute representations into a coherent semantic space. Finally, we introduce a new binary cross-entropy loss with the final logits  $\mathbf{I}_j$ ,

$$\begin{aligned} \mathbf{s}_j^{\text{attr}} &= \text{Softmax}(\mathbf{I}_j), \\ L_{\text{fine}} &= - \sum_j \text{gt}_j \log((\mathbf{s}_j^{\text{coarse}})^{\alpha} (\mathbf{s}_j^{\text{attr}})^{(1-\alpha)}). \end{aligned} \quad (11)$$

This strategy ensures that both the full-text representation and the atomic attribute representations are independently and robustly optimized. It compels the model to project both embedding types into a coherent semantic space, preparing them for the strict min comparison at inference time.

### 3.5 Inference Process

The inference process of GUIDED++ follows a sequential, multi-stage pipeline designed to first robustly localize objects and then perform strict fine-grained verification. The overall workflow is summarized below.

**Offline preparation:** Before processing any images, all category names in the target vocabulary are pre-processed using an LLM, as detailed in Sect. 3.1. This step, performed only once, generates the cached sets of coarse subject embeddings and atomic attribute embeddings required for the subsequent online stages.

**Online detection and verification:** For a given input image, the online process proceeds as follows:

1. **Coarse proposal generation:** The image is first processed by the coarse-grained object detection (CGOD) stage (Sec. 3.2). This stage utilizes the cached **coarse subject embeddings** to produce a set of high-recall object proposals  $\{b_j\}$  and their corresponding coarse scores  $\{s_j^{\text{coarse}}\}$ .
2. **Fine-grained scoring:** Each proposal  $b_j$  is then evaluated by the conjunctive multi-attribute verification (CMAV) module (Sec. 3.3). This module computes a fine-grained score  $s_j^{\text{fine}}$  by enforcing a logical ‘AND’ match against the cached atomic attribute embeddings.
3. **Score fusion and final output:** The final confidence score for each proposal,  $s_j^{\text{final}}$ , is computed by fusing the coarse

and fine-grained scores as per Eq. (7). Standard Non-Maximum Suppression (NMS) is then applied to these final scores to yield the final detection results.

This structured process ensures that the computationally efficient CGOD stage filters the search space, allowing the more precise CMAV stage to focus only on promising candidates, thus operationalizing our framework's core design principle.

## 4 Experiments

### 4.1 Datasets and Evaluation Metrics

Our method's performance is rigorously evaluated on two state-of-the-art fine-grained open-vocabulary detection benchmarks, FG-OVD and 3F-OVD. For training, we utilize the base classes from the large-scale LVIS dataset.

**FG-OVD.** The fine-grained open-vocabulary detection (FG-OVD) dataset (Bianchi et al., 2024a) is specifically designed to comprehensively evaluate a model's fine-grained discrimination capabilities. It augments the part-level annotations of the PACO dataset (Ramanathan et al., 2023) with rich natural-language captions. Each ground-truth object is paired with a positive caption and up to ten challenging hard negatives. These negatives are systematically generated by substituting one to three attribute words (e.g., color, material) while preserving the sentence structure, thereby creating scenarios of high semantic ambiguity. This design provides a controlled testbed for compositional reasoning. The dataset is partitioned into eight evaluation tracks: four difficulty-based splits (Trivial, Easy, Medium, Hard), where the number of substituted attributes in negatives decreases, and four attribute-focused subsets (Color, Material, Pattern, Transparency). Following the official protocol, we report the mean Average Precision (mAP) (Lin et al., 2014) averaged across all eight tracks to provide a holistic performance measure.

**3F-OVD.** The recently proposed Fine-Grained, Free-form, and Free-category Open-vocabulary Detection (3F-OVD) benchmark (Liu et al., 2025) presents a more demanding testbed, focusing on long-caption queries and extremely subtle inter-class differences. Built upon the NEU-171K corpus, it comprises over 145k images and 676k bounding boxes across 719 fine-grained classes. The benchmark is split into two distinct domains: vehicles (NEU-171K-C), with 598 car models, and retail products (NEU-171K-RP), with 121 classes. Its primary challenge lies in its pronounced long-tailed distribution and the minimal visual distinctions between classes, such as the subtle head-lamp shape that separates a Ford Focus from a Ford Fiesta. Each class is described by a single, sentence-length caption, which is reused across all instances, testing the model's robustness

to varied contexts. Consistent with the benchmark's setup, we report mAP across both domains.

**LVIS.** LVIS is a large-scale object detection dataset known for its long-tailed distribution of categories. Following the standard open-vocabulary detection protocol established in prior works (Du et al., 2024), we partition the dataset into 866 common and frequent categories, designated as base classes for training, and 335 rare categories, which are held out as novel classes for zero-shot evaluation. In the coarse-grained detector pre-training stage, our model is trained on the base classes of the LVIS (Large Vocabulary Instance Segmentation) dataset (Gupta et al., 2019). Our primary use of LVIS in this work is to leverage its diverse base classes for training our detector's foundational visual representations.

### 4.2 Implementation Details

Our implementation is built upon the LaMI-DETR code-base (Du et al., 2024). The detector's ConvNeXt-Large backbone is initialized from the OpenCLIP pre-trained model (ConvNeXt-Large-D-320) (Liu et al., 2022; Ilharco et al., yyy), and we employ the Llama3.3-70B-Chat model for the offline category name parsing. Our training follows a two-stage pipeline: we first pre-train the model on the LVIS base classes for 85,200 iterations, and subsequently fine-tune it on the FG-OVD training set for an additional 3,000 iterations. During inference, we follow the practice of GroundingDINO (Liu et al., 2023) and retain the top  $k=900$  object proposals ranked by their coarse-grained scores. For evaluation on 3F-OVD, all class captions are processed concurrently. For key hyperparameters in GUIDED++,  $\alpha$  is 0.6 and  $p$  is 0.5. Follow the scaling factor setting in LaMI-DETR, temperature scaling factor  $m_{\text{fine}}$  and  $m_{\text{coarse}}$  for logits, are both set to 100. All other settings, unless specified, are inherited from the original LaMI-DETR configuration.

### 4.3 Experimental Results

**Comparison on the FG-OVD benchmark.** We conduct extensive experiments on the FG-OVD benchmark, with detailed results presented in Table 1. Specifically, we compare GUIDED++ against the existing OVD methods, including OWL-ViT (Minderer et al., 2022), Detic (Zhou et al., 2022), ViLD (Gu et al., 2021), Grounding DINO (Liu et al., 2023), CORA (Wu et al., 2023b), and OV-DINO (Wang et al., 2024b). As shown in the table, GUIDED++ establishes a new state-of-the-art across multiple base detectors. Note that the attribute embedding fusion module is only applied to the DETR-based framework, LaMI-DETR. When integrated with LaMI-DETR, it achieves an overall mAP of 68.2%, a remarkable +25.0% absolute improvement over the baseline, and a significant +1.8% gain over our prior work, GUIDED. A detailed breakdown of the results reveals

a performance pattern that strongly validates our core thesis. On the ‘Easy’, ‘Medium’, and ‘Trivial’ subsets, the performance of GUIDED++ is comparable to GUIDED. These subsets are characterized by classes with large semantic distances between them, differing in their main subject or multiple attributes. In such cases, the strict conjunctive logic of GUIDED++ offers little additional benefit when the categories are already easy to distinguish. The true strength of GUIDED++’s new conjunctive multi-attribute verification mechanism is unleashed on the most challenging, attribute-dense subsets. It achieves its most significant gains over GUIDED on ‘Hard’ (+2.1%), ‘Pattern’ (+5.4%), and ‘Transparency’ (+3.6%). These are precisely the scenarios where models must distinguish objects based on a conjunction of subtle attributes, the exact failure point of that our new mechanism is designed to solve. Furthermore, GUIDED++ provides a far greater performance leap than methods like HA-FGOVD (+3.4%). Notably, HA-FGOVD even underperforms simple fine-tuning. By strictly freezing the visual detector and only updating text embeddings, it struggles to capture subtle visual nuances that simple fine-tuning can directly learn. This confirms that a dedicated verification mechanism, rather than merely refining text embeddings, is necessary to resolve conjunctive ambiguity. Even powerful general-purpose models like OV-DINO (29.6% mAP) falter, underscoring their inherent limitation in processing compositional queries. Finally, the versatility of the GUIDED++ framework is confirmed by its consistent and significant performance boosts when applied to diverse detector architectures. This proves that GUIDED++ represents a generalizable paradigm for enhancing compositional reasoning in FG-OVD.

**Comparison on 3F-OVD dataset.** We further test GUIDED++ on the demanding 3F-OVD benchmark, where models must distinguish classes based on extremely subtle visual cues (e.g., ‘Ford Focus’ vs. ‘Ford Fiesta’). This scenario is a quintessential example of the attribute under-representation problem. The results, presented in Table 2, starkly reveal the limitations of existing paradigms. Conventional OVD methods like Detic and GroundingDINO, which rely on holistic, single-vector representations, fail almost completely. It validates that their semantic averaging mechanism cannot handle such fine-grained distinctions. In this challenging setting, GUIDED++ demonstrates a significant relative improvement over all baselines. The performance gain over our prior work (GUIDED) is particularly noteworthy, as it directly stems from the newly introduced conjunctive multi-attribute verification mechanism. This demonstrates that our verification logic forces the model to find and confirm that one critical, subtle attribute, preventing it from being overlooked. Notably, these results are achieved via zero-shot transfer from models trained on LVIS and FG-OVD, without any fine-tuning on the 3F-OVD

**Table 2** mAP evaluation results on the 3FOVD benchmark(%)

| Method        | NEU-171K-C           | NEU-171K-RP          |
|---------------|----------------------|----------------------|
| Detic         | $6.6 \times 10^{-4}$ | $2.2 \times 10^{-2}$ |
| Vild          | $3.8 \times 10^{-4}$ | $1.1 \times 10^{-2}$ |
| GroundingDino | $1.3 \times 10^{-3}$ | $7.6 \times 10^{-4}$ |
| LaMI-DETR     | $9.0 \times 10^{-4}$ | $2.3 \times 10^{-1}$ |
| + GUIDED      | $7.2 \times 10^{-3}$ | $2.7 \times 10^{-1}$ |
| + GUIDED++    | $1.3 \times 10^{-2}$ | $3.2 \times 10^{-1}$ |

dataset itself. This not only underscores the generalization capabilities of GUIDED++ but also proves that our framework provides a more principled and robust solution to the core challenge of compositional reasoning in the wild.

#### 4.4 Ablation Study and Analysis

**Ablations of our main components.** In Table 3, we conducted an ablation study to evaluate the main components in our proposed GUIDED++ framework. Specifically, while fine-tuning the LaMI-DETR baseline on FG-OVD (& Fine-tune) provides a solid starting point (50.9% mAP), it is far from optimal, highlighting the need for a more principled architecture. Furthermore, we validate our architectural decoupling. Removing the Coarse-grained OVD Classifier (w/o COC) and forcing the detector to directly localize fine-grained classes causes a massive 9.3% mAP drop. This proves that separating localization from discrimination is critical to mitigate attribute over-representation. The results also demonstrate the necessity of the conjunctive multi-attribute verification (CMAV) module. Removing it (& GUIDED++ w/o CMAV)—results in a 5.8% performance loss. Critically, this drop is most severe on the attribute-dense subsets. This provides direct evidence that CMAV successfully resolves the attribute under-representation problem in precisely the scenarios it was designed for. Finally, removing the projection layer leads to a 3.3% performance decrease, confirming its role in refining attribute sensitivity. In summary, these results demonstrate that the full power of GUIDED++ comes from the synergy of its two core strategies: architectural decoupling against over-representation and conjunctive verification against under-representation.

**Ablations of our CMAV module.** We performed a detailed ablation of our CMAV module to validate its core design choices, with results in Table 4. The cornerstone of CMAV is the min operator, which enforces strict conjunctive logic. Replacing it with a mean operator reintroduces the very semantic averaging we aim to solve, causing a significant 2.3% mAP drop. This confirms that averaging allows strong attribute matches to mask weak ones, while min enforces the necessary “all-or-nothing” verification. The synergy between

**Table 1** mAP evaluation results on FG-OVD benchmark (%). The performance in 'Average' is the average of performance over the 8 sub-datasets. 'Trans.' denotes the performance on the Transparency subset.

'Finetune' denotes finetune the LaMI-DETR on the FG-OVD training set. Here \* denotes the results from our reproduction

| Detector       | Hard | Medium | Easy | Trivial | Color | Material | Pattern | Trans | Average      |
|----------------|------|--------|------|---------|-------|----------|---------|-------|--------------|
| OWL-ViT(B/16)  | 26.2 | 39.8   | 38.4 | 53.9    | 45.3  | 37.3     | 26.6    | 34.1  | 37.7         |
| OWLv2(B/16)    | 25.3 | 38.5   | 40.0 | 52.9    | 45.1  | 33.5     | 19.2    | 28.5  | 35.4         |
| OWLv2(L/14)    | 25.4 | 41.2   | 42.8 | 63.2    | 53.3  | 36.9     | 23.3    | 12.2  | 37.3         |
| Detic          | 11.5 | 18.6   | 18.6 | 69.7    | 21.5  | 38.8     | 30.1    | 24.6  | 29.3         |
| ViLD           | 22.1 | 36.1   | 39.9 | 56.6    | 43.2  | 34.9     | 24.5    | 30.1  | 35.9         |
| CORA           | 13.8 | 20.0   | 20.4 | 35.1    | 25.0  | 19.3     | 22.0    | 27.9  | 22.9         |
| OV-DINO        | 18.6 | 28.4   | 25.0 | 54.3    | 35.6  | 30.0     | 21.0    | 24.2  | 29.6         |
| Grounding DINO | 17.0 | 28.4   | 31.0 | 62.5    | 41.4  | 30.3     | 31.0    | 26.2  | 33.5         |
| + HA-FGOVD     | 19.2 | 32.3   | 34.0 | 62.2    | 41.5  | 33.0     | 32.1    | 29.2  | 35.4 (+1.9)  |
| + GUIDED       | 35.1 | 49.3   | 52.8 | 57.1    | 49.7  | 57.0     | 26.4    | 39.6  | 45.9 (+12.4) |
| + GUIDED++     | 36.7 | 49.7   | 52.7 | 56.5    | 50.3  | 58.6     | 31.4    | 42.3  | 47.3 (+13.8) |
| OWL-ViT(L/14)  | 26.6 | 39.8   | 44.5 | 67.0    | 44.0  | 45.0     | 36.2    | 29.2  | 41.5         |
| + HA-FGOVD     | 31.4 | 46.0   | 50.7 | 67.2    | 48.4  | 48.5     | 38.0    | 32.7  | 45.4 (+4.3)  |
| + GUIDED       | 46.8 | 59.4   | 64.1 | 66.2    | 60.4  | 58.9     | 44.7    | 54.5  | 56.9 (+15.4) |
| + GUIDED++     | 48.3 | 59.6   | 63.8 | 66.6    | 61.2  | 60.6     | 48.6    | 56.9  | 58.2 (+16.7) |
| LaMI-DETR      | 29.2 | 40.6   | 42.9 | 63.5    | 49.5  | 39.2     | 34.6    | 46.2  | 43.2         |
| + Finetune     | 39.5 | 50.7   | 54.2 | 66.0    | 51.9  | 53.7     | 42.1    | 49.1  | 50.9 (+7.7)  |
| + HA-FGOVD*    | 33.5 | 45.9   | 47.5 | 63.8    | 52.7  | 42.6     | 36.9    | 50.1  | 46.6 (+3.4)  |
| + GUIDED       | 57.5 | 69.5   | 73.3 | 72.6    | 64.8  | 68.5     | 62.0    | 63.4  | 66.4 (+23.2) |
| + GUIDED++     | 59.6 | 69.7   | 73.2 | 72.9    | 65.8  | 69.9     | 67.4    | 67.0  | 68.2(+25.0)  |

**Table 3** The ablation of main components in the GUIDED++ framework on the FG-OVD dataset. The results are shown in mAP (%). 'Baseline' denotes the LaMI-DETR. 'CMAV' denotes conjunctive multi-attribute verification. 'COC' denotes the coarse-grained OVD classifier

| Method                    | Hard | Medium | Easy | Trivial | Color | Material | Pattern | Transp | Average |
|---------------------------|------|--------|------|---------|-------|----------|---------|--------|---------|
| Baseline                  | 29.2 | 40.6   | 42.9 | 63.5    | 49.5  | 39.2     | 34.6    | 46.2   | 43.2    |
| & Train from scratch      | 31.7 | 39.6   | 43.4 | 42.6    | 32.4  | 33.0     | 30.8    | 20.4   | 34.2    |
| & Finetune                | 39.5 | 50.7   | 54.2 | 66.0    | 51.9  | 53.7     | 42.1    | 49.1   | 50.9    |
| & GUIDED++ w/o CMAV       | 53.0 | 65.2   | 69.5 | 72.0    | 64.1  | 63.3     | 49.0    | 63.4   | 62.4    |
| & GUIDED++ w/o COC        | 50.8 | 60.0   | 64.6 | 63.2    | 59.9  | 59.0     | 52.7    | 61.3   | 58.9    |
| & GUIDED++ w/o projection | 56.8 | 68.2   | 69.7 | 72.0    | 64.6  | 66.6     | 58.6    | 62.7   | 64.9    |
| & GUIDED++                | 59.6 | 69.7   | 73.2 | 72.9    | 65.8  | 69.9     | 67.4    | 67.0   | 68.2    |

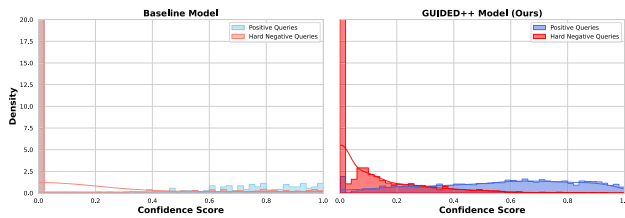
the full matching logits and the attribute matching logits is also crucial. Removing the attribute term  $\mathbf{I}^{\text{attr}}$  makes the model lose its fine-grained precision, causing a 2.0% drop. Conversely, removing the full matching logits  $\mathbf{I}^{\text{full}}$  forces the model to rely solely on the attribute match. This leads to overfitting on individual attributes, resulting in a performance decrease on the subset with multiple different attributes (Medium, Easy, Trivial). This principle is mirrored in the sampling of final logits in our training strategy. Removing the full matching logits ( $\mathbf{I}_{j,c} = \mathbf{I}_{j,c}^{\text{diff}}$ ) causes the training to overfit on attribute signals, neglecting the global object concept and leading to a 0.8% drop. Similarly, removing the attribute-specific logits ( $\mathbf{I}_{j,c} = \mathbf{I}_{j,c}^{\text{full}}$ ) from the subset of dis-

tinguishing attributes prevents the model from learning the conjunctive check in the first place, degrading performance by 1.2%. These ablations prove that CMAV's success in preventing the attribute under-representation.

**More analysis about our CMAV module.** Fig. 4 visualizes the confidence score distributions on the FG-OVD Hard benchmark to probe for attribute under-representation. The baseline model (left) reveals a critical failure in compositional reasoning. First, a vast majority of its predictions, for both positive and hard negative queries, are clustered near zero, indicating a general lack of discriminative confidence. More importantly, for the few instances that do elicit a high-confidence response, the distributions for positive and

**Table 4** The ablation of CMAV on the FG-OVD dataset. ‘Mean’ indicates using the mean operator instead of the min operator in Eq. (6)

| Method                              | Hard | Medium | Easy | Trivial | Color | Material | Pattern | Transp | Average |
|-------------------------------------|------|--------|------|---------|-------|----------|---------|--------|---------|
| GUIDED++                            | 59.6 | 69.7   | 73.2 | 72.9    | 65.8  | 69.9     | 67.4    | 67.0   | 68.2    |
| & w/o $I^{\text{attr}}$             | 56.8 | 68.9   | 72.5 | 72.6    | 64.3  | 67.2     | 63.9    | 63.1   | 66.2    |
| & w/o $I^{\text{full}}$             | 61.0 | 67.3   | 69.2 | 60.7    | 66.5  | 70.8     | 68.2    | 70.1   | 66.7    |
| & Mean                              | 55.9 | 69.7   | 73.3 | 71.6    | 63.5  | 66.7     | 62.7    | 63.8   | 65.9    |
| & $I_{j,c} = I_{j,c}^{\text{diff}}$ | 60.2 | 68.2   | 70.8 | 71.2    | 66.0  | 66.4     | 67.1    | 69.0   | 67.4    |
| & $I_{j,c} = I_{j,c}^{\text{full}}$ | 58.2 | 69.6   | 71.7 | 72.6    | 65.4  | 69.2     | 63.6    | 65.5   | 67.0    |

**Fig. 4** The score distributions of the baseline method (LaMI-DETR) and our GUIDED++ method

hard negative queries are highly overlapping and largely indistinguishable. These provide direct evidence of attribute under-representation: the model’s reliance on semantic averaging within a single holistic embedding prevents it from penalizing critical single-attribute mismatches. In stark contrast, GUIDED++ (right) demonstrates robust conjunctive verification. It achieves a decisive separation between the two distributions. Our conjunctive multi-attribute verification (CMAV) mechanism effectively suppresses scores for hard negatives by enforcing a strict logical ‘AND’ across all attributes, forcing the vast majority of them towards zero. This validates our approach’s ability to resolve the semantic ambiguity inherent in baseline models and perform true compositional reasoning.

**Analysis of framework robustness to LLM choice.** We evaluate our framework’s robustness to the upstream LLM by analyzing the generation of the Subject query and the more complex Atomic queries. As detailed in Table 5, generating a set of Atomic queries is an inherently more challenging linguistic task than identifying a single Subject. Specifically, LLaMA–3.1-8B model struggles significantly more with Atomic query generation, producing 22 failures, compared to 9 failures for the simpler Subject task. This difficulty in parsing fine-grained semantics directly correlates with its suboptimal detection performance (65.9 mAP). Crucially, this challenge is effectively overcome by large-scale models. Both the proprietary GPT-4o and the open-source LLaMA–3.3-70B demonstrate exceptional stability, committing very few errors on either task. This high parsing fidelity translates directly to stable, state-of-the-art detection performance (68.4 and 68.2 mAP, respectively), confirming that our method’s success is not contingent on a specific

**Table 5** The number of failure cases of 300 samples in the FG-OVD dataset in category name parsing with different LLMs. We report the number of failure cases in the generation of the subject query (Noted as ‘Subject’) and the atomic queries (Noted as ‘Atomic’), respectively. The notation ‘ $x/y$ ’ denotes that there are  $y$  failure cases, of which  $x$  lead to detection errors. We also report the mAP(%) evaluation results on the FG-OVD benchmark with different LLMs

| LLM           | Subject | Atomic | mAP  |
|---------------|---------|--------|------|
| GPT4-o        | 1/2     | 1/3    | 68.4 |
| LLaMA–3.1-8B  | 9/9     | 10/22  | 65.9 |
| LLaMA–3.3-70B | 1/1     | 2/5    | 68.2 |

proprietary model. Furthermore, while subject identification errors are highly critical when they occur (e.g., 9/9 propagation for LLaMA-8B), the framework shows a notable tolerance for minor imperfections in Atomic queries (e.g., only 2/5 propagated for LLaMA-70B). Given that large-scale LLMs have virtually eliminated the critical subject errors and can reliably handle the complex atomic query generation, our framework stands as a robust and practical solution.

**Analysis of the coarse-grained detector.** Our two-stage framework relies on an initial coarse-grained detector to generate object proposals. To validate the effectiveness of this first stage, it is crucial to ensure that it does not become a performance bottleneck by failing to detect the ground-truth objects. Therefore, we conduct a recall analysis of the coarse-grained detector on the FG-OVD dataset. We measure the recall of the top- $k$  proposals at standard IoU thresholds of 0.5 and 0.75. The results are presented in Table 6. As the table shows, the detector’s performance is robust. While the single highest-scoring proposal (Top-1) captures approximately half of the ground-truth objects, the recall increases significantly when more proposals are considered. With 100 proposals per image, the detector successfully recalls 97.5% of ground-truth objects at an IoU of 0.5, and 91.4% even at the stricter IoU threshold of 0.75. This high recall rate confirms that the coarse-grained stage provides a comprehensive set of candidates to the subsequent fine-grained analysis stage, ensuring that the overall performance of our method is not constrained by missed detections in the initial phase.

**Table 6** Recall performance of the coarse-grained detector on the FG-OVD dataset

| Top-k Proposals | IoU=0.5 | IoU=0.75 |
|-----------------|---------|----------|
| Top-1           | 53.8%   | 49.1%    |
| Top-10          | 89.9%   | 82.9%    |
| Top-100         | 97.5%   | 91.4%    |

**Table 8** Ablation of  $\alpha$  on FG-OVD

| $\alpha$ | Average |
|----------|---------|
| 0.4      | 67.5    |
| 0.6      | 68.2    |
| 0.8      | 66.1    |

**Table 9** Ablation of  $p$  on FG-OVD

| $p$ | Average |
|-----|---------|
| 0.3 | 67.7    |
| 0.5 | 68.2    |
| 0.7 | 68.0    |

**Inference time analysis.** We analyze the computational overhead of GUIDED++ in Table 7. The results show that the significant accuracy gains from our new conjunctive multi-attribute verification (CMAV) module come at a negligible cost. The inference time for GUIDED++ is only 4.3 ms longer than our prior work, GUIDED, an overhead caused by processing a few extra atomic queries.

Crucially, the LLM-related latency is incurred only once per class name, not per image. For any given dataset or deployment scenario with a fixed vocabulary, the query parsing can be pre-computed and cached offline. Therefore, during image inference, the effective latency of GUIDED++ (with an 8B LLM) is only 216.4 ms (198.8 ms Detector + 17.6 ms CLIP), which is nearly identical to the baseline’s 211.7 ms, while delivering a massive +23.9% mAP improvement. For dynamic scenarios requiring on-the-fly parsing of novel classes, the choice of LLM presents a clear trade-off between performance and latency. As shown, using a more powerful LLaMA–3.3-70B model boosts mAP to 68.2% at the cost of higher latency. This flexibility allows practitioners to balance accuracy and speed based on their specific application needs.

#### 4.5 Analysis of Hyper-Parameters

We analyze the impact of the key hyperparameters  $\alpha$  and  $p$  that control the scoring mechanism in GUIDED++.

**Choice of  $\alpha$ .** The hyperparameter  $\alpha$  balances the influence of the coarse-grained score ( $s^{\text{coarse}}$ ) and the fine-grained score ( $s^{\text{fine}}$ ) in the final prediction. As shown in Table 8, we tested values of 0.4, 0.6, and 0.8. Performance peaks at  $\alpha = 0.6$  (68.2% mAP). This suggests an optimal balance

where the robust localization from the coarse detector provides a strong foundation, while the fine-grained verification effectively refines the final decision. A lower value over-relied on the strict fine-grained check, while a higher value dilutes its impact.

**Choice of  $p$ .** We investigate the model’s sensitivity to the stochastic sampling probability  $p$ , which balances training between compositional and holistic representations. As presented in Table 9, our method demonstrates consistent performance across the tested range of  $p$  values. The performance remains stable, with only a minor variation between 67.7 and 68.2 Average, indicating that our framework is not highly sensitive to the value of  $p$ .

## 5 Conclusion

In this paper, we addressed the challenge of compositional reasoning in Fine-grained open-vocabulary detection (FG-OVD). We began by identifying that the prevalent strategy of compressing complex descriptions into a single vector creates a fundamental semantic ambiguity, which manifests as two distinct failure modes: attribute over-representation and the more insidious attribute under-representation. This core issue prevents models from performing the strict, compositional alignment that human language demands. To resolve this, we proposed GUIDED++, a framework built on a dual strategy that systematically dismantles both facets of semantic ambiguity. First, by employing architectural decoupling, we separate coarse-grained localization from fine-grained discrimination, effectively mitigating the mislocalizations caused by attribute over-representation. Second, and more critically, we introduced the novel conjunctive multi-attribute verification (CMAV) mechanism to directly combat attribute under-representation. By decomposing queries into their atomic conditions and enforcing a logical ‘AND’ through a minimal aggregation function, CMAV eradicates the semantic averaging effect and ensures that an object satisfies all specified attributes. Our extensive experiments demonstrate that GUIDED++ sets a new state-of-the-art, validating that true progress in compositional visual understanding requires moving beyond monolithic, holistic representations. The principles of decomposition and conjunctive verification presented here offer a robust and generalizable paradigm, paving the way for vision-language models that can interpret the rich, compositional structure of the world with greater fidelity and trustworthiness.

## Appendix A Dataset Details

**FG-OVD.** The FG-OVD benchmark (Bianchi et al., 2024a) is a rigorous testbed for attribute-aware object detection,

**Table 7** The comparison of inference time(ms) between Baseline(LaMI-DETR) and GUIDED++ for detecting one class in an image. We also report the average mAP(%) of each method in FGOVD

| Method                      | LLM   | Detector | CLIP | Overall | mAP  |
|-----------------------------|-------|----------|------|---------|------|
| Baseline                    | —     | 198.5    | 13.2 | 211.7   | 43.2 |
| GUIDED with LLaMA—3.1-8B    | 72.1  | 198.8    | 13.3 | 284.2   | 65.9 |
| GUIDED++ with LLaMA—3.1-8B  | 72.1  | 198.8    | 17.6 | 288.5   | 67.1 |
| GUIDED++ with LLaMA—3.3-70B | 193.8 | 198.8    | 17.6 | 410.2   | 68.2 |

built upon the part-level annotations of the PACO dataset (Ramanathan et al., 2023). Its captions are semi-automatically generated. Positive captions are produced by prompting OpenAssistant-LLaMA-30B (Kopf et al., 2023) with PACO's structured JSON data for each object, followed by manual verification to ensure grammatical fluency and semantic fidelity. The benchmark's primary challenge stems from its hard negatives, which are derived from the same templates but with one to three attribute tokens replaced. This design preserves the syntax while maximizing semantic ambiguity, forcing models to perform precise verification. This allows for the creation of a challenging open-set vocabulary without additional manual labeling.

FG-OVD features eight complementary test tracks. Four are difficulty-oriented—Trivial, Easy, Medium, and Hard—where the number of substituted attributes in the negative captions decreases from three to one, progressively tightening the visual-textual gap. The other four tracks are attribute-centric, isolating performance on a single attribute category inherited from PACO: Color, Material, Pattern, and Transparency. The training split contains 27,684 annotated objects, shared across all tracks. The test split provides separate evaluation subsets with the following annotation counts: 3,545 for Hard, 2,968 for Medium, 1,299 for Easy, 3,119 for Trivial, 3,193 for Color, 467 for Material, 3,545 for Pattern, and 409 for Transparency.

**3F-OVD.** The 3F-OVD benchmark (Liu et al., 2025) is a recently released, fine-grained open-vocabulary detection dataset that we use to probe our model's capacity to distinguish subtle appearance cues under open-set conditions. Built on the NEU-171K corpus, it comprises 145,825 images, 676,471 bounding boxes, and 719 fine-grained classes across two domains: NEU-171K-C (89,363 street-view photographs of 598 passenger-car models) and NEU-171K-RP (56,462 warehouse images of 121 retail products). The dataset is characterized by a pronounced long-tailed distribution and minimal inter-class visual differences—for example, distinguishing a Ford Focus from a Ford Fiesta based on headlamp shape. Each class is represented by a single, class-level caption.

The official evaluation protocol supplies captions one at a time, tokenizes them, and consolidates box predictions. To better reflect the single-shot inference paradigm of real-world detectors, we adopt a one-pass variant where all captions are

processed concurrently. To apply our method, we identify the coarse-grained subject for each caption: for the automotive-focused NEU-171K-C, we fix the subject to “car” for all captions; for the diverse NEU-171K-RP domain, we employ a large-language model to identify the head noun of each caption.

## Appendix B Prompts Used in Category Name Phrasing

We present the prompt used to generate the attribute embeddings with LLM in subject identification, which is illustrated in Fig. 7.

## Appendix C Qualitative Comparisons

We conduct the qualitative comparisons for GUIDED++ against other methods in FG-OVD, which is illustrated in Fig. 5. As shown in row 1-2, GUIDED++ correctly predicts the bounding boxes while existing methods, which showcases better capability of GUIDED++ in tackling attribute under-representation and discriminating the subtle difference between different class names. As shown in line 3 and line 6, the existing methods LaMI-DETR and HA-FGOVD suffer from mislocalization, objects, because prominent attribute tokens such as “red” or “grey” divert attention away from the object centers, while our GUIDED++ correctly locates the objects. It validates the effectiveness of GUIDED++ in addressing the attribute over-representation.

Furthermore, Fig. 6 provides a detailed view of the internal two-stage mechanism of GUIDED++, illustrating the effectiveness of our components. This figure provides a step-by-step visualization of our model's detection process, organized to show the Ground Truth (GT), the output after our coarse-grained detector (CGOD), and the final result after our fine-grained module (CMAV). As demonstrated in the visualization, CGOD effectively generates accurate proposals for coarse-grained classes. Subsequently, the CMAV module then performs fine-grained verification. For a given fine-grained category, it evaluates these proposals and assigns a significantly higher score to the specific detection box that satisfies all compositional attributes. This process enables the

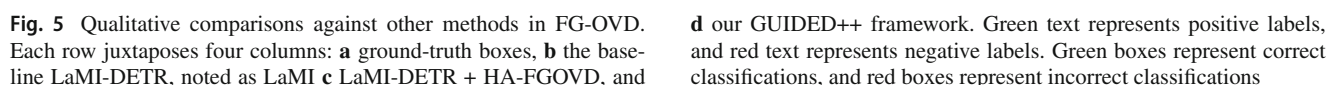
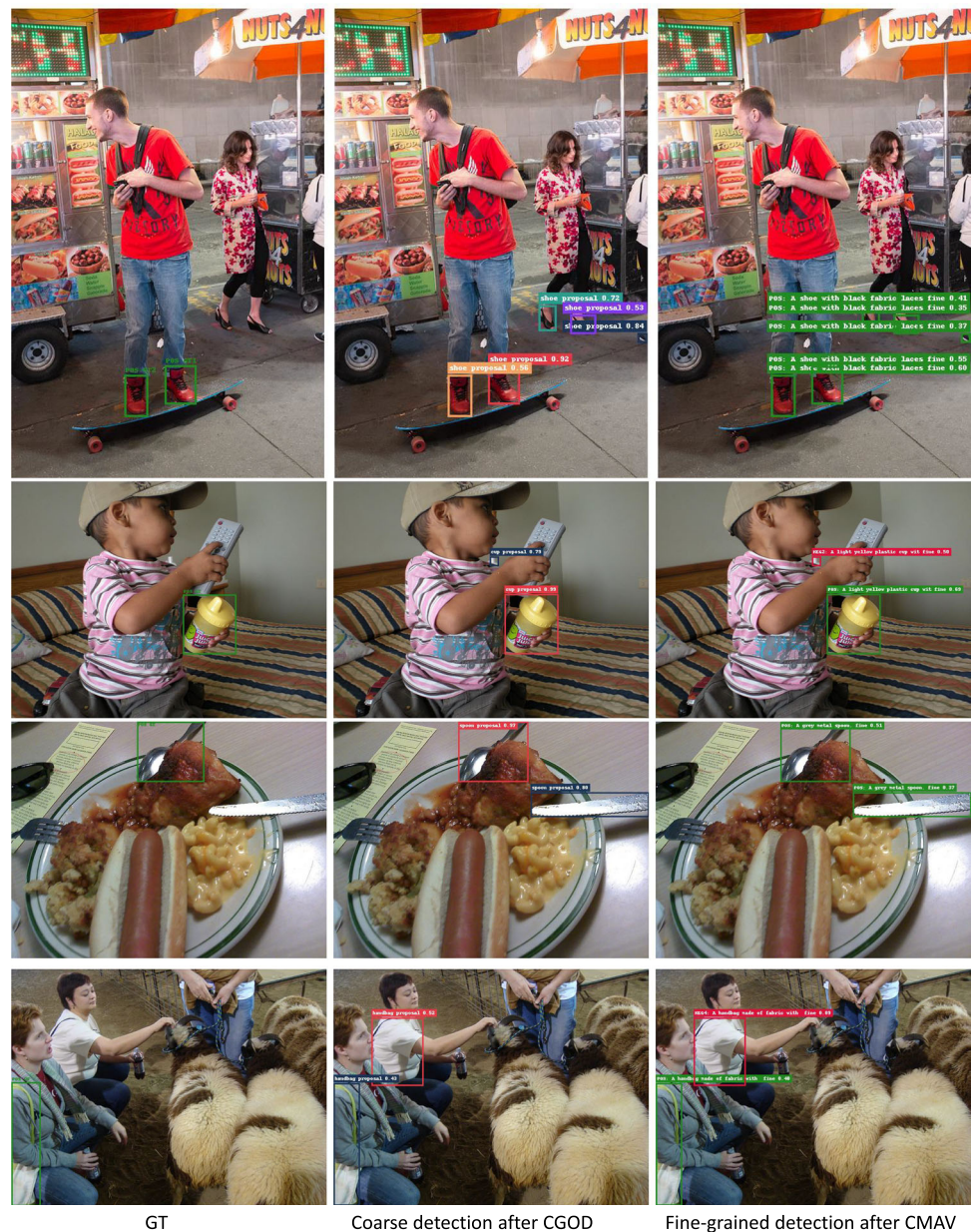


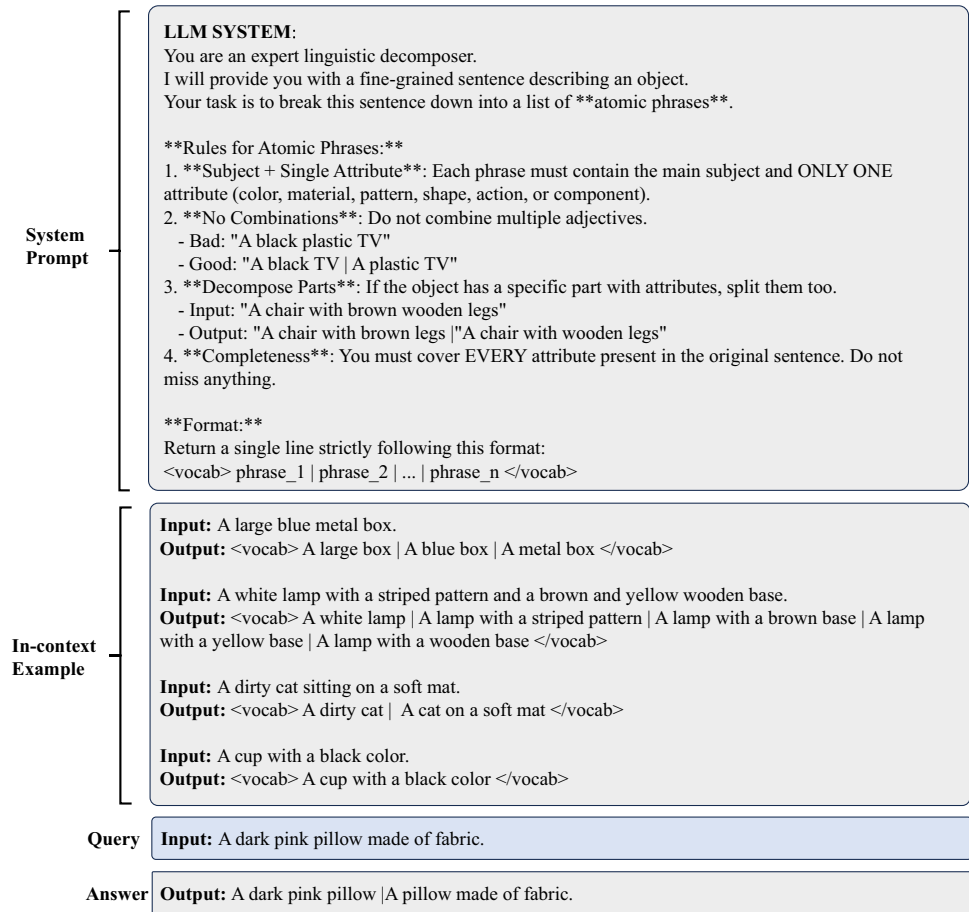
**Fig. 6** Qualitative visualization of our two-stage detection process. Each row presents a different example. The columns display: **a** the Ground-Truth (GT) bounding box; **b** the initial proposals generated by our coarse-grained detector (CGOD); and **c** the final fine-grained detection results from our framework. In the final results (c), green boxes represent correct detection



model to successfully match the correct ground-truth class, as indicated by the green bounding boxes in the final column. This progression illustrates how a generic proposal is

refined into a specific, attribute-rich detection, validating the complementary roles of our two-stage design.

**Fig. 7** Illustration of the LLM prompt for constructing the atomic queries



**Acknowledgements** This work is supported in part by the National Key R&D Program of China (Nos.2024YFB3908503 and 2024YFB3908500), in part by the National Natural Science Foundation of China (No. 62322608) and in part by the PCL Major Key Project under Grant PCL2025A17-2 and Grant PCL2025A02-3. The authors would also like to thank the National Supercomputer Center in Guangzhou for providing high performance computational resources.

**Data Availability** All of the datasets used in this paper are available online. The [FG-OVD](#), [3F-OVD](#), [LVIS](#) dataset can be downloaded from their official website accordingly. Our source code will be available upon acceptance.

## References

- Bianchi, L., Carrara, F., Messina, N., Gennaro, C., & Falchi, F. (2024). The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 22520–22529).
- Bianchi, L., Carrara, F., Messina, N., & Falchi, F. (2024). Is clip the main roadblock for fine-grained open-world perception? In *2024 international conference on content-based multimedia indexing (CBMI)* (pp. 1–8). IEEE.
- Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., & Shan, Y. (2024). Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16901–16911).
- Dou, B., Wu, T., & Guo, Z. (2025). Fine-grained open-vocabulary object detection via attribute decomposition and aggregation. *AI*, 6(9), 201.
- Du, P., Wang, Y., Sun, Y., Wang, L., Liao, Y., Zhang, G., Ding, E., Wang, Y., Wang, J., & Liu, S. (2024). LaMI-DETR: Open-vocabulary detection with language model instruction. In *Proceedings of the European conference on computer vision (ECCV)*.
- Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., & Li, G. (2022). Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14084–14093).
- Fu, S., Yang, Q., Mo, Q., Yan, J., Wei, X., Meng, J., Xie, X., & Zheng, W.-S. (2025). LLMDet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *Proceedings of the computer vision and pattern recognition conference* (pp. 14987–14997).
- Gupta, A., Dollar, P., & Girshick, R. (2019). LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5356–5364).
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., & Vaughan, A. (2024). The llama 3 herd of models [arXiv:2407.21783](#) arXiv preprint.
- Gu, X., Lin, T.-Y., Kuo, W., & Cui, Y. (2021). Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint [arXiv:2104.13921](#).

- Ilharco, G., Wortsman, M., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., & Schmidt, L. OpenCLIP. <https://doi.org/10.5281/zenodo.5143773>.
- Kuo, W., Cui, Y., Gu, X., Piergiovanni, A., & Angelova, A. (2022). F-VLM: Open-vocabulary object detection upon frozen vision and language models. arXiv preprint [arXiv:2209.15639](https://arxiv.org/abs/2209.15639).
- Kim, J., Cho, E., Kim, S., & Kim, H.J. (2024). Retrieval-augmented open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17427–17436.
- Köpf, A., Kilcher, Y., Rütte, D., Anagnostidis, S., Tam, Z.-R., Stevens, K., Barhoum, A., Duc, N.M., Stanley, O., Nagyfi, R., et al. (2023). Openassistant conversations-democratizing large language model alignment. Corr. DOI: <https://doi.org/10.48550/arXiv.2304.07327>
- Liu, Y., Hua, Y., Chai, H., Wang, Y., & Ye, T. (2025). Fine-grained open-vocabulary object detection with fined-grained prompts: Task, dataset and benchmark. arXiv preprint [arXiv:2503.14862](https://arxiv.org/abs/2503.14862).
- Liu, M., Hayes, T. L., Ricci, E., Csurka, G., & Volpi, R. (2024). Shine: Semantic hierarchy nexus for open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16634–16644)
- Li, J., Liang, Z., Chen, W., Ma, L., & Li, G. (2025). Guided: Granular understanding via identification, detection, and discrimination for fine-grained open-vocabulary object detection. In *The thirty-ninth annual conference on neural information processing systems*.
- Li, W., Liu, X., Ma, J., & Yuan, Y. (2024). Cliff: Continual latent diffusion for open-vocabulary object detection. In *European conference on computer vision*, pp. 255–273. Springer.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C.L. (2014). Microsoft coco: Common objects in context. In: *Computer vision—ECCV 2014: 13th European conference, Zurich, Switzerland, September 6–12, 2014, proceedings, Part V 13*, (pp. 740–755). Springer.
- Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11976–11986)
- Li, J., Zhang, J., Li, J., Li, G., Liu, S., Lin, L., & Li, G. (2024). Learning background prompts to discover implicit knowledge for open vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (pp. 16678–16687).
- Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., & Zhang, L. (2023). Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. arXiv preprint [arXiv:2303.05499](https://arxiv.org/abs/2303.05499).
- Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., Chang, K.-W., & Gao, J. (2022). Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10965–10975.
- Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., & Houlsby, N. (2022). Simple open-vocabulary object detection. In *European conference on computer vision* (pp. 728–755). Springer.
- Ma, Y., Liu, M., Zhu, C., & Yin, X.-C. (2025). HA-FGOVD: Highlighting fine-grained attributes via explicit linear composition for open-vocabulary object detection. *IEEE Transactions on Multimedia*, 27, 3171–3183.
- Ma, S., Qian, D., Ye, K., & Zhang, S. (2025). Cake: Category aware knowledge extraction for open-vocabulary object detection. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 39, pp. 5982–5990)
- Ma, S., Wang, Y., Wei, Y., Zhang, E., Fan, J., Sun, X., & Chen, P. (2025). SKDF: A simple knowledge distillation framework for distilling open-vocabulary knowledge to open-world object detector. In: *IEEE Transactions on pattern analysis and machine intelligence*.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763) PMLR.
- Ramanathan, V., Kalia, A., Petrovic, V., Wen, Y., Zheng, B., Guo, B., Wang, R., Marquez, A., Kovvuri, R., Kadian, A., Mousavi, A., Song, Y., Dubey, A., Mahajan, D., & Paco. (2023). Parts and attributes of common objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7141–7151)
- Wang, K., Cheng, L., Chen, W., Zhang, P., Lin, L., Zhou, F., & Li, G. (2024). Marvelod: Marrying object recognition and vision-language models for robust open-vocabulary object detection. In *European conference on computer vision* (pp. 106–122). Springer.
- Wang, L., Liu, Y., Du, P., Ding, Z., Liao, Y., Qi, Q., Chen, B., & Liu, S. (2023). Object-aware distillation pyramid for open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11186–11196.
- Wang, H., Ren, P., Jie, Z., Dong, X., Feng, C., Qian, Y., Ma, L., Jiang, D., Wang, Y., Lan, X., & Liang, X. (2024). OV-DINO: Unified open-vocabulary detection with language-aware selective fusion. arXiv preprint [arXiv:2407.07844](https://arxiv.org/abs/2407.07844).
- Wu, S., Zhang, W., Jin, S., Liu, W., & Loy, C. C. (2023). Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15254–15264)
- Wu, X., Zhu, F., Zhao, R., & Li, H. (2023). 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* (pp. 7031–7040)
- Xi, X., Huang, Y., Luo, R., & Qiu, Y. (2025). Ow-ovd: Unified open world and open vocabulary object detection. In *Proceedings of the computer vision and pattern recognition conference* (pp. 25454–25464)
- Yao, L., Han, J., Liang, X., Xu, D., Zhang, W., Li, Z., & Xu, H. (2023). 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*, pp. 23497–23506.
- Yao, L., Pi, R., Han, J., Liang, X., Xu, H., Zhang, W., Li, Z., & Xu, D. (2024). 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*, (pp. 27391–27401).
- Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., & Misra, I. (2022). Detecting twenty-thousand classes using image-level supervision. In *European conference on computer vision* (pp. 350–368). Springer.
- Zhao, X., Liu, X., Wang, D., Gao, Y., & Liu, Z. (2024). Scene-adaptive and region-aware multi-modal prompt for open vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16741–16750.

- Zhao, S., Schuster, S., Zhao, L., Zhang, Z., Suh, Y., Chandraker, M., & Metaxas, D.N. (2024). Taming self-training for open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (pp. 13938–13947).
- Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., & Gao, J. (2022). RegionCLIP: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (pp. 16793–16803).

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.