

---

# GUIDED: Granular Understanding via Identification, Detection, and Discrimination for Fine-Grained Open-Vocabulary Object Detection (Supplementary Document)

---

## 1 More Details

### 1.1 More Details about Datasets.

**FG-OVD** FG-OVD [1] augments the part-level annotations of PACO[6] with natural-language captions automatically generated by an LLM, creating a rigorous benchmark for attribute-aware object detection. Positive captions are produced by prompting OpenAssistant-LLAMA-30B[3] with PACO’s structured JSON for every object; the resulting sentences are manually verified to ensure grammatical fluency and semantic fidelity. Negative captions are derived from the same templates but replace one to three attribute tokens, keeping the syntax intact while maximizing semantic ambiguity. This design yields a challenging open-set vocabulary without additional manual labeling.

FG-OVD offers eight complementary test tracks. Four are difficulty-oriented—Trivial, Easy, Medium, and Hard—in which the number of substituted attributes in the negatives decreases from three to one, progressively tightening the visual–textual gap. The other four tracks are attribute-centric, including Color, Material, Pattern, and Transparency. Each isolating a single attribute category inherited directly from PACO.

The training split contains 27,684 annotated objects, shared across all tracks. The test split provides separate evaluation subsets with the following numbers of annotations: 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.

**3FOVD** The 3F-OVD [4] benchmark 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 collected by the authors, it comprises 145825 images, 676471 bounding boxes, and 719 fine-grained classes drawn from two domains: NEU-171K-C (89363 street-view photographs spanning 598 passenger-car models) and NEU-171K-RP (56462 warehouse images covering 121 retail products). Each class is represented by a single, class-level caption reused across all images that contain that class. The combination of a pronounced long-tailed distribution and minimal inter-class visual differences—e.g. the head-lamp shape that separates a Ford Focus from a Ford Fiesta—makes 3F-OVD considerably more challenging than existing open-vocabulary benchmarks.

In the official evaluation protocol, captions are supplied to the detector one at a time; the caption is tokenised, and every box returned for any token is consolidated via Clip-NMS and an area-based filter to yield the final prediction set. To better reflect the single-shot inference paradigm prevalent in real-world detectors, we adopt a one-pass variant in which all captions are processed concurrently. For NEU-171K-RP, we employ a large-language model to identify the head noun of each caption, whereas for NEU-171K-C we fix the subject to “car” for every caption, mirroring the dataset’s automotive focus.

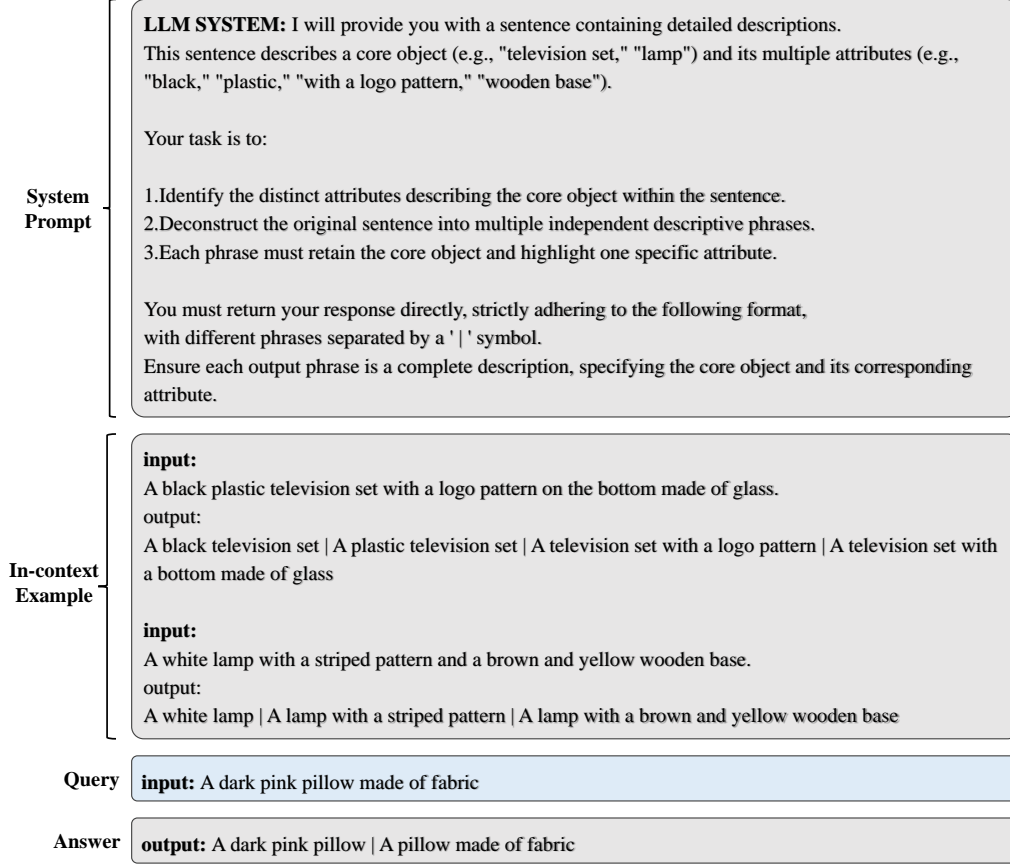


Figure 1: Illustration of the LLM prompt for attribute identification.

## 1.2 More Details about Prompts used in Subject Identification.

We present the prompt used to generate the attribute embeddings with LLM in subject identification, which is illustrated in Figure 1.

## 2 More ablation and analysis

**More analysis of refined CLIP used in GUIDED.** To further validate the necessity of the refined CLIP design in our GUIDED, we illustrate the embeddings generated by the refined CLIP in Figure 5, which reveals that captions sharing the same subject are clustered within the model’s latent space. It shows that the embeddings in refined CLIP are capable of recognizing the subject of the fine-grained class name, achieving better alignment with fine-grained text for FG-OVD.

**More analysis of our attribute embedding fusion module.** we also present the ablation about the attribute embedding fusion module in Table 1. As shown in the table, our attribute embedding fusion leads to a significant improvement on FG-OVD, showcasing the effectiveness of our proposed AEF on integrating helpful attribute information. When only conditioning on the fine-grained class embedding without the subtraction, the AEF drops by 1.3%, validating the subtraction operation on key states highlights the attribute for better attention estimation. Furthermore, our proposed AEF can also be applied to a traditional open vocabulary task by extending the class names with LLM to generate the fine-grained class embeddings. Applying AEF to the LaMI-DETR leads to a performance of 1.0% on  $AP_r$ , demonstrating the generalization of AEF.

Furthermore, we compare the mAP of subtraction-based attention with additive fusion and concatenation in FGOVD in Table 2. For concatenation, we apply a linear projection layer to transform the

Table 1: The ablation of attribute embedding fusion on the FG-OVD and the LVIS dataset. The results in the FGOVD dataset are shown in average mAP across the eight FG-OVD subsets. 'AEF' denoted attribute embedding fusion. 'w/o' Subtract denotes we do not subtract the subject embedding for the key states estimation.

Method	FG-OVD Average	LVIS			
		$AP_r$	$AP_c$	$AP_f$	AP
GUIDED w/o AEF	62.5	43.2	39.3	43.5	41.6
GUIDED w/o Subtract	65.1	43.4	38.4	43.6	41.4
GUIDED	66.4	43.9	39.2	43.7	41.8

dimensions back to the original dimensions. The results show that our AEF outperforms both additive fusion and concatenation by a clear margin, demonstrating the effectiveness of the subtraction-based attention on attribute integrations.

Table 2: The ablation study of the proposed AEF module with simpler structures on FG-OVD.

Arch	mAP
Concatenation	61.1
Addition	63.2
AEF	66.4

Table 3: The ablation of different fusion strategies on FG-OVD.

Fusion	mAP
Weighted Average	63.1
Ours	66.4

**Ablation of the score fusion strategy.** We conduct the ablation of different score fusion strategies used in Equation (5). The results are presented in Table 3. It shows that weighted multiplication achieves a better performance. Compared with the weighted average, the multiplication enforces a high score on both the coarse confidence score and the attribute similarity simultaneously. This prevents candidates with a poor attribute similarity score from being promoted simply because their confidence scores are high (and vice versa).

### 3 Analysis of hyper-parameters

In this section, we analyze the hyperparameters used in our method. For hyperparameters  $m_{\text{fine}}$  and  $m_{\text{coarse}}$ , we follow the conventional setting for scaling the similarities between the CLIP embeddings to set their value as 100. We analyze the choice of the other parameter  $\alpha$  used in our method.

**Choice of  $\alpha$ .**  $\alpha$  is a value controlling the effect of the detector's coarse confidence and attribute similarity score. To determine the value of  $\alpha$ , we compared model performance under different  $\alpha$  values for our method in Table 4. Setting  $\alpha$  to 0.4 only leads to a slight performance decline, which indicates that our proposed GUIDED is not overly sensitive to a smaller  $\alpha$ . When increasing  $\alpha$  to 0.8, the performance drops by 0.5%, demonstrating the effectiveness of the integration of attribute similarity scores.

### 4 More results on standard OVD benchmark.

We present the performance on the LVIS benchmark in Table 6. The results show fine-tuning on the FG-OVD dataset degrades the performance on the LVIS benchmark. However, co-training on LVIS and FG-OVD during the second stage not only alleviates this issue but surpasses the original LaMI-DETR baseline while mainly preserving FG-OVD performance. This enhancement stems from our AEF module's attribute integration and expanded training data from FG-OVD.

### 5 Failure Case in Limitations

We provide a qualitative example of a failure case depicted in our limitation section. For the query "suitcase on the conveyor belt", the conveyor belt (contextual attribute) may not be captured within the candidate box of any suitcase. Consequently, GUIDED may incorrectly associate the attribute with a suitcase spatially adjacent to the conveyor belt. This occurs because our method relies on the

Table 4: The ablation of  $\alpha$  on the FG-OVD dataset. The results are shown in average mAP across the eight FG-OVD subsets.

$\alpha$	Average
0.8	65.9
0.6	66.4
0.4	66.3
0.2	65.3

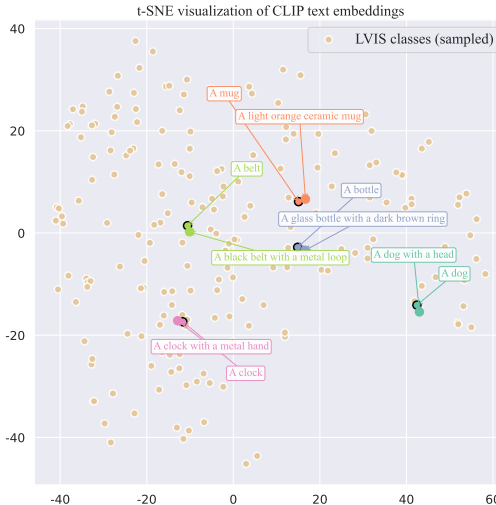
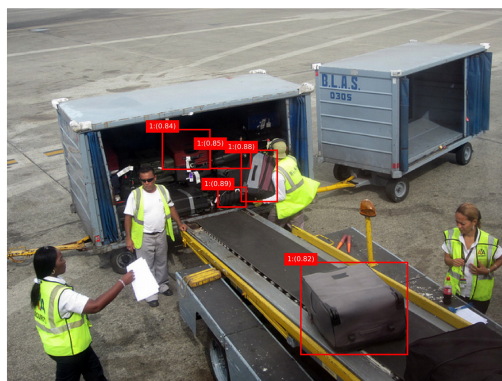


Table 5: The CLIP-embedding visualization for the refined CLIP used in GUIDED.

Table 6: The comparison on the standard open-vocabulary benchmark LVIS dataset and the FGOVD dataset.

Method	Stage1	Stage2	FG-OVD	LVIS			
			Average	$AP_r$	$AP_c$	$AP_f$	AP
Baseline	LVIS	-	43.2	43.2	39.3	43.5	41.6
GUIDED	LVIS	-	46.8	43.9	39.2	43.7	41.8
GUIDED	LVIS	FG-OVD	66.4	39.3	32.6	38.7	36.2
GUIDED	LVIS	FG-OVD & LVIS	65.5	44.4	39.4	43.7	41.9

detector’s candidate boxes, which may not encompass all required contextual element. This could be mitigated by using expanded region proposals or incorporating context-aware reasoning beyond bounding boxes.



Label 0: the suitcase in the container  
Label 1: the suitcase on the conveyor belt

Figure 2: A failure case of GUIDED for our limitations.

## 6 Qualitative Comparisons against Other Methods

We present several examples of qualitative comparisons between our method and current methods for FG-OVD. Figure 3 displays the qualitative comparisons on FG-OVD. Our baseline LaMI-DETR[2] often mis-localizes objects because prominent attribute tokens such as “orange” or “head” divert attention away from the object centers, whereas HA-FGOVD[5] partly corrects this drift but still suf-



Figure 3: Qualitative comparisons against other methods in FG-OVD. Each row juxtaposes four columns: (a) ground-truth boxes, (b) the baseline LaMI-DETR, noted as LaMI (c) LaMI-DETR + HA-FGOVD, and (d) our GUIDED framework. Green text represents positive labels, and red text represents negative labels. Green boxes represent correct classifications, and red boxes represent incorrect classifications.

fers from low-confidence outputs and occasional attribute confusions. By first anchoring localization with a coarse subject phrase and then refining labels with fine-grained attributes, GUIDED eliminates the drift, produces tighter boxes, and assigns materially and chromatically precise labels across varied instances—ladders, spoons, dogs, computer mice, clocks, and mugs—thereby delivering the clearest qualitative improvement over both baselines.

## 7 Broader impacts

The proposed GUIDED framework holds significant potential to advance critical societal applications by enabling precise, attribute-aware object detection without requiring exhaustive annotation efforts. The framework’s ability to recognize nuanced visual attributes enables transformative applications in scientific discovery (e.g., rare specimen analysis), sustainable development (e.g., precision resource management), and inclusive technology design (e.g., high-fidelity assistive systems). Its compositional reasoning supports cross-domain adaptation, empowering non-experts to deploy fine-

grained recognition in under-resourced scenarios like biodiversity monitoring or cultural heritage preservation.

## References

- [1] Lorenzo Bianchi, Fabio Carrara, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. 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*, pages 22520–22529, 2024.
- [2] Penghui Du, Yu Wang, Yifan Sun, Luting Wang, Yue Liao, Gang Zhang, Errui Ding, Yan Wang, Jingdong Wang, and Si Liu. Lami-detr: Open-vocabulary detection with language model instruction. In *Proceedings of the European conference on computer vision (ECCV)*, 2024.
- [3] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. corr, abs/2304.07327, 2023. doi: 10.48550. *arXiv preprint arXiv:2304.07327*.
- [4] Ying Liu, Yijing Hua, Haojiang Chai, Yanbo Wang, and TengQi Ye. Fine-grained open-vocabulary object detection with fined-grained prompts: Task, dataset and benchmark. *arXiv preprint arXiv:2503.14862*, 2025.
- [5] Yuqi Ma, Mengyin Liu, Chao Zhu, and Xu-Cheng Yin. Ha-fgovd: Highlighting fine-grained attributes via explicit linear composition for open-vocabulary object detection. *IEEE Transactions on Multimedia*, 2025.
- [6] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023.