Roboflow-20VL Few-Shot Object Detection Challenge Report -FDUROILab Lenovo-

Lingyi Hong[1] Mingxi Chen [1] Xingqi He [1] Runze Li[2] Xingdong Sheng[2] Wenqiang Zhang[1,3*]
[1] Shanghai Key Lab of Intelligent Information Processing,
School of Computer Science, Fudan University, Shanghai, China
[2] Lenovo Research
[3] College of Intelligent Robotics and Advanced Manufacturing, Fudan University, Shanghai, China
{lyhong22, mxchen24, xqhe25}@m.fudan.edu.cn, {lirz7, shengxd1}@lenovo.com, wqzhang@fudan.edu.cn

1. Introduction

Vision-Language Models have showcased exceptional detection capabilities on general images. Nevertheless, these foundational models may still fall short of optimal performance in specialized target applications, particularly in domains such as medical imaging analysis and aerial imagery interpretation. However, due to the lack of large amounts of annotated data in downstream domains, the challenge lies in how to efficiently transfer models to downstream scenarios with only a small number of annotated samples [2].

To address these challenges, we propose an efficient fine-tuning approach based on open-vocabulary detection. By applying transformations to the given samples for data augmentation, we enhance the adaptation capability of the model to the new domain. Furthermore, we introduce a novel post-processing method leveraging multimodal large language models (MLLMs) to achieve more precise classification. Our approach aims to improve detection performance in cross-domain scenarios with minimal supervision, ensuring better adaptability to unseen domains.

2. Team Details

- Team name: FDUROILab_Lenovo
- Team leader name: Lingyi Hong
- Team leader email: lyhong22@m.fudan.edu.cn
- Team members: Mingxi Chen, Xingqi He, Runze Li, Xingdong Sheng, Wenqiang Zhang
- Affiliation: Fudan University, Lenovo Research

3. Method

We build our method on an open-vocabulary detection framework and adapt it to cross-domain few-shot object detection through three key components: object-centric data augmentation, efficient target-domain fine-tuning, and MLLM-assisted prediction refinement. The overall pipeline first expands the limited annotated samples with instance-level augmentation, then fine-tunes the detector on each target domain, and finally uses a multimodal large language model to improve category-level reliability during inference.

3.1. Efficient Tuning

To improve the detector's adaptability under limited supervision, we adopt an efficient fine-tuning strategy for each target domain. The key idea is to increase the diversity of the few-shot training set while preserving the original category semantics. Instead of relying only on standard image-level augmentation, we construct additional training samples through an object-centric augmentation pipeline.

Given the few-shot annotations of each target dataset, we first crop target instances from the original training images according to the provided bounding boxes. These object crops are then augmented with geometric and photometric transformations, including random flipping, rotation, grayscale conversion, color jittering, and scale perturbation. Such instance-level transformations help the model observe more diverse object appearances while keeping the category identity unchanged.

After obtaining the augmented object crops, we randomly rescale them and paste them back into images from the same target domain. The pasted locations and object scales are randomly sampled to synthesize new object layouts and contextual combinations. This process increases the effective number of training samples and improves robustness to variations in object size, position, and background context. Since all pasted instances are derived from the provided few-shot annotations, the augmented data remains consistent with the few-shot setting.

We then fine-tune the open-vocabulary detector on the

*Corresponding Author

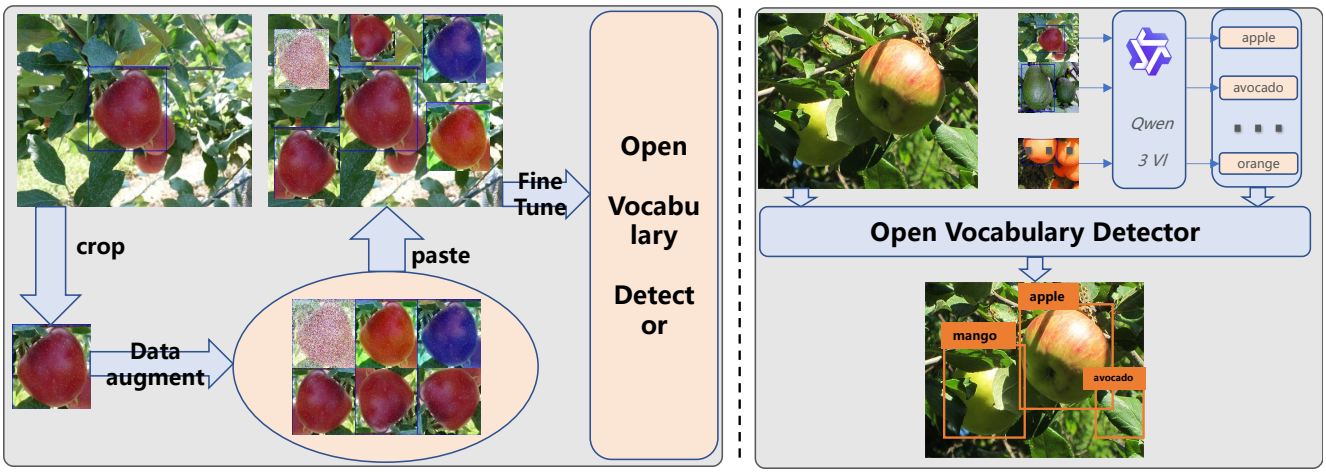

Figure 1. Overview of our efficient tuning and inference.

augmented training set. During fine-tuning, the detector learns target-domain visual patterns while preserving its pretrained vision-language alignment ability. For each target dataset, we train an independent model and select the final checkpoint according to validation mAP. This validation-guided checkpoint selection is important because few-shot fine-tuning is sensitive to optimization noise and may overfit when trained for too many iterations.

3.2. MLLM-Assisted Category Prompt Construction

Open-vocabulary detectors rely on textual category prompts to perform language-guided detection. However, raw category names are sometimes insufficient for fine-grained or domain-specific classes. A short class name may not fully describe the object's visual appearance, and visually similar categories may become difficult to distinguish when only the original label names are used.

To obtain more informative category prompts, we use Qwen3-VL [1] to generate visual descriptions for the target categories. For each category, we provide the model with the category name and representative few-shot examples, and ask it to summarize discriminative visual cues such as shape, texture, color, part structure, and typical context. The generated description is then combined with the original class name to form an enhanced textual prompt.

Formally, for a category c_i, we construct its textual query as

$$q_i = \text{Concat}(n_i, d_i), \tag{1}$$

where n_i denotes the original category name and d_i denotes the MLLM-generated visual description. The final category prompt set $Q = \{q_i\}_{i=1}^{C}$ is used as the language input of the open-vocabulary detector. Compared with directly using raw category names, these enhanced prompts provide richer visual semantics and improve alignment between the detector's language branch and target-domain concepts.

3.3. Inference

During inference, we use the enhanced textual prompts as category queries for the open-vocabulary detector. Given a test image, the detector predicts candidate bounding boxes, category labels, and confidence scores. The raw detection results are then passed to the post-processing stage for category calibration, false-positive suppression, and duplicate removal.

This inference design keeps the detector as the main localization module. The MLLM is not used to generate bounding boxes directly; instead, it provides category-level guidance and prediction refinement. This separation allows the detector to preserve accurate localization ability while using the MLLM to correct classification errors that are common in cross-domain few-shot scenarios.

3.4. Post-Process

We observe that the detector can often localize target objects but may confuse visually similar categories, especially when the category names are fine-grained, domain-specific, or visually ambiguous. Therefore, the main purpose of post-processing is not to replace the detector, but to improve the reliability of its category predictions.

We introduce Qwen3-VL as an auxiliary visual classifier. For each test image, we construct a multimodal prompt containing the test image, the candidate category list, and representative few-shot examples from the corresponding target dataset. Qwen3-VL is asked to infer which categories are likely to appear in the image. For uncertain detections, we further crop the detected regions and ask Qwen3-VL to classify each crop among the valid category set.

Based on the detector output and the MLLM prediction,

we apply the following post-processing strategies.

Category Filtering. If a detected category is inconsistent with the categories predicted by Qwen3-VL for the whole image, we suppress the detection when its confidence score is low. This step removes obvious false positives caused by prompt mismatch, background confusion, or cross-category misclassification.

MLLM-Based Reclassification. For detections with reliable bounding boxes but uncertain labels, we use Qwen3-VL to reclassify the cropped detection region. If the MLLM prediction is consistent with the visual evidence and more plausible than the detector output, we replace the detector's predicted label with the MLLM-predicted category while keeping the original bounding box unchanged. This strategy is especially useful when the detector localizes the object correctly but assigns it to a visually similar class.

Geometry-Aware Filtering. Some target domains have strong object-size or aspect-ratio priors. We estimate the valid range of object area and aspect ratio from the few-shot training annotations. Detections with abnormal geometry, such as overly large boxes, extremely elongated boxes, or boxes covering irrelevant background regions, are filtered out when they violate the dataset-specific priors.

Non-Maximum Suppression. Finally, we apply Non-Maximum Suppression (NMS) to remove redundant overlapping detections. The confidence threshold and NMS threshold are selected according to validation performance for each target dataset.

The choice of post-processing strategy is dataset-specific. For datasets where the detector mainly suffers from false positives, category filtering is preferred. For datasets where localization is stable but classification is unreliable, MLLM-based reclassification is more effective. For datasets with clear object-scale priors, geometry-aware filtering provides additional gains.

3.5. Implementation Details

We use an open-vocabulary detection model as the baseline detector and Qwen3-VL-235B-A22B [1] as the auxiliary multimodal large language model. The MLLM is used in two stages: first, to generate category-level visual descriptions for prompt construction; second, to assist post-processing by providing image-level category prediction and crop-level reclassification.

All fine-tuning experiments are conducted on 8 NVIDIA RTX 3090 GPUs. We use a batch size of 8 and a base learning rate of 1×10^{-6}. During fine-tuning, the text branch of the detector is kept frozen, while the visual detection components are updated to adapt to the target-domain images. This design stabilizes training and avoids damaging the pre-trained language-alignment ability of the detector.

For each of the 20 target datasets, we train an independent model using the provided 10-shot examples per class. Since different datasets exhibit different domain shifts, object scales, and category ambiguity, we tune the number of fine-tuning iterations, confidence threshold, NMS threshold, and post-processing strategy separately on the validation set. The checkpoint with the best validation mAP is selected for final inference.

For post-processing, we choose among category filtering, MLLM-based reclassification, and geometry-aware filtering according to validation-set behavior. In domains with stable localization but frequent label confusion, we apply MLLM-based reclassification. In domains with many false positives, we apply category filtering. In domains with clear object-size priors, we additionally remove geometrically abnormal boxes. These dataset-specific choices improve the robustness of the final predictions across diverse target domains.

4. Results

Table 1. Ablation studies on Roboflow-20VL.

Method	Avg mAP	Δ
Baseline (Zero Shot)	18.3	-
+ *Finetune*	48.8	+32.7
+ *Post Process*	51.6	+33.5

Table 1 demonstrates the effectiveness of each component in our method. The zero-shot baseline achieves an average mAP of 18.3, showing that directly applying the open-vocabulary detector to target-domain images is insufficient under the few-shot setting. We conduct multiple training runs for each dataset and select the best-performing checkpoint according to validation performance, which further improves the Avg mAP to 48.8.

Finally, the MLLM-assisted post-processing strategy improves the Avg mAP from 48.8 to 51.6. This gain suggests that the detector's remaining errors are not only caused by localization failure, but also by category confusion among visually similar or domain-specific classes. By using the MLLM to refine category predictions and suppress unreliable detections, our final system achieves the best overall performance.

5. Conclusion

We propose an efficient few-shot adaptation method based on open-vocabulary detection. Our approach enhances target-domain adaptability through object-centric data augmentation and efficient fine-tuning, while Qwen3-VL-assisted prompt construction and post-processing improve category-level reliability. Experimental results demonstrate the effectiveness of our method across diverse target domains.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 2, 3

[2] Anish Madan, Neehar Peri, Shu Kong, and Deva Ramanan. Revisiting few-shot object detection with vision-language models. *arXiv preprint arXiv:2312.14494*, 2023. 1