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

Lingyi Hong[1] Mingxi Cheng [1] Keliang Yin [1] Runze Li[2] Xingdong Sheng[2] Wenqiang Zhang[1,3]*

[1] Shanghai Key Lab of Intelligent Information Processing,
College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China
[2] Lenovo Research
[3] College of Intelligent Robotics and Advanced Manufacturing,
Fudan University, Shanghai, China

{lyhong22, mxchen24, klyin25}@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 [3].

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
- Rest of the team members: Mingxi Cheng, Keliang Yin, Runze Li, Xingdong Sheng, Wenqiang Zhang
- Affiliation: Fudan University, Lenovo Research
- Team name on the Roboflow-20VL Few-Shot Object Detection Challenge competitions: FDUROILab_Lenovo

3. Method

In this section, we firstly introduce our approach to aligning foundation models with target concepts through few-shot multimodal fine-tuning. Then, we describe a post-processing strategy utilizing a MLLM.

3.1. Efficient Tuning

To enhance the model's alignment with target domain concepts in few-shot multi-modal detection, we propose an efficient fine-tuning strategy. Our approach leverages data augmentation techniques to expand the training set and improve the model's ability to recognize objects in the target domain, using merely 10 annotated samples per category.

Given a setting where each category has only ten labeled samples, we employ a structured fine-tuning pipeline, which is shown in Figure 1. (1) **Data Augmentation.** We apply various data augmentation techniques directly to the original training images, including:

- *CachedMosaic*: Combines four images into one composite image, allowing the model to encounter multiple contexts and object scales within a single training sample.
- *YOLOXHSVRandomAug*: Applies photometric variations through hue, saturation, and value (HSV) adjust-

*Corresponding Author

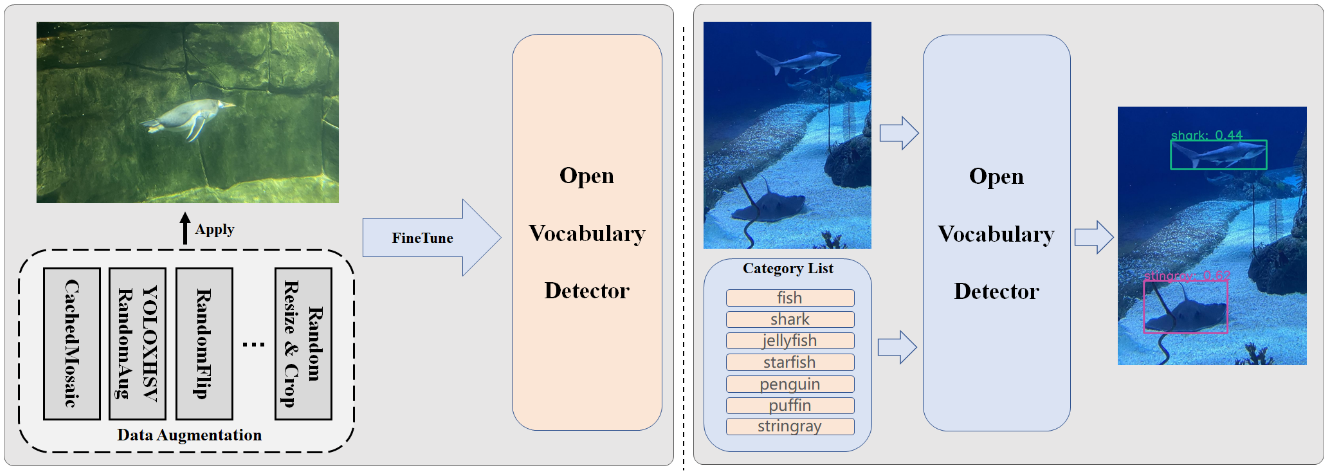

Figure 1. Overview of our efficient tuning and inference.

ments to simulate diverse lighting or weather conditions.

- *RandomFlip*: Randomly flips images horizontally or vertically to enhance training data diversity.
- *CachedMixUp*: Combines two images and their corresponding labels through linear interpolation, promoting the model's ability to generalize by moving beyond rigid decision boundaries.
- *RandomResize*: Performs dynamic image resizing during training to achieve scale invariance and enhance model robustness against objects of different sizes.
- *RandomCrop*: Extracts random image regions to simulate partial occlusions and diverse viewing angles, training the model to recognize objects from incomplete visual information.

This comprehensive augmentation strategy systematically addresses the challenges of limited training data while maintaining domain relevance.

(2) **Fine-Tuning with Augmented Data.** We finetune the open-vocabulary detection model with the augmented images, which enables the detector to better adapt to objects in the target domain, even with minimal labeled examples. Additionally, the augmented data effectively increases the number of training samples, mitigating the few-shot learning limitation and improving overall detection performance. Through this efficient fine-tuning approach, the finetuned model gains enhanced adaptability to new domains.

3.2. Post-Process

Although existing open-vocabulary detectors possess strong open-set detection capabilities, their performance on the challenge test set remains suboptimal. Upon further analysis, we found that while the detector can successfully identify most objects, its primary weakness lies in classification errors rather than detection failures. This indicates that the open-vocabulary detection model still struggles with accurate classification when adapting to objects in a new domain. To address this issue, we introduce Qwen2.5-VL as an auxiliary classifier to refine the final predictions. For each bounding box detected by the open-vocabulary detector in each test image, we use Qwen2.5-VL to determine whether the category predicted by the bounding box is correct. Additionally, we also ask Qwen2.5-VL to perform an extra category classification for this bounding box. According to the results of Qwen2.5-VL, if Qwen2.5-VL considers that the classification of this bounding box is incorrect or the classification result is inconsistent with that of open-vocabulary detector, we will modify the classification result of this bounding box to the classification result of Qwen2.5-VL. By leveraging Qwen2.5-VL as a post-processing step, we effectively correct classification errors and enhance the model's performance on unseen domains, leading to more accurate and reliable object detection results.

3.3. Implement Details

Our framework employs MM-GroundingDINO [4] as the foundational open-vocabulary detection model, utilizing Swin-Large [2] as its backbone architecture to ensure robust feature extraction. For multimodal comprehension and reasoning, we integrate Qwen2.5-VL-32B [1] as the core Multimodal Large Language Model (MLLM). All experiments are performed on a computational cluster equipped with 8 NVIDIA RTX 3090 GPUs, configured with a consistent batch size of 8 and an initial learning rate of 1e-6 to maintain training stability. To thoroughly evaluate model generalization across diverse data domains, we implement a rigorous training protocol: each of the 20 datasets undergoes

50 independent training runs, with every run executed for 20 full epochs to explore the optimization landscape comprehensively. The final model checkpoint for each dataset is determined by selecting the iteration that achieves peak performance on the respective validation set, ensuring a fair and reproducible evaluation process.

4. Results

Method	**Avg mAP**	**Δ**
Baseline (Zero Shot)	16.1	-
+ *Finetune*	38.3	+22.2
+ *More Aug*	41.8	+3.5
+ *More Runs*	47.2	+5.4
+ *Post Process*	49.6	+2.4

Table 1. Ablation studies on RoboFlow-20VL.

In the Table 1, we demonstrate the effectiveness of each component of our method. Fine-tuning on a small number of annotated images can effectively enhance adaptability to downstream scenarios. By adding more data augmentation methods, we can further improve the model's accuracy. Additionally, we find that repeated training on the same dataset leads to significant fluctuations in accuracy, so we conduct multiple training runs to obtain the best results. Finally, our post-processing also contributes to performance improvement.

5. Conclusion

we propose an efficient fine-tuning strategy based on open-vocabulary detection and a Qwen2.5-VL-assisted post-processing method. Our approach enhances model adaptability to new domains through data augmentation, while MLLM-based post-processing improves classification accuracy. Experimental results demonstrate the effectiveness of our method across different datasets.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[2] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[3] 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

[4] Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024. 2