Hybrid Fine-Tuned Open-Vocabulary Detection for RF20-VL FSOD

Abu Noman Md Sakib
University of Texas at San Antonio, San Antonio, TX, USA
abunomanmd.sakib@gmail.com

Abstract

This report describes our best Overall Track submission for the CVPR 2026 Roboflow-20VL Foundational Few-Shot Object Detection Challenge. The method starts from open-vocabulary detector checkpoints and adapts them with few-shot training for each RF20-VL dataset. We then run checkpoint inference, score the candidate prediction files locally, keep the best file for each dataset, and export a top-300 EvalAI package after a zero-based category-ID audit. This fine-tuned system is used only for the Overall Track, where gradient-based adaptation is allowed. The selected public submission is EvalAI ID `573896`, with official leaderboard mAP of **47.356**. Code is available here: https://github.com/anmspro/cvpr-rf20vl

1 Introduction

RF20-VL evaluates few-shot object detection across twenty datasets. The domains include aerial imagery, medical images, industrial defects, sports, consumer products, smoke, water meters, trail-camera data, and fine-grained objects. This setting is hard because many target domains are far from the data used to pretrain common open-vocabulary detectors. Each category also has only a small number of labeled examples.

Our final Overall submission uses one selected pipeline. We fine-tune detector variants, evaluate their prediction files per dataset, keep the strongest candidate for each dataset, and package the twenty selected files into the final EvalAI zip.

2 Method

2.1 Base detector family

The submitted solution uses open-vocabulary detector checkpoints from the GroundingDINO, MM-GroundingDINO, and LLMDet family. These models condition detection on text prompts, which makes them useful for unseen or rare categories. Few-shot training then adapts the detector to the visual style of each RF20-VL dataset.

2.2 Data preparation

Each RF20-VL dataset is prepared in COCO-style format for training and inference. We also follow the organizer guidance on category indexing. The final predictions use real categories starting at `0`; they are not shifted by a dummy `none` class.

Table 1: Overall Track leaderboard status at initial report time.

Team	Submission ID	mAP
`anmspro`	573896	47.356

The submitted zip contains twenty `.pkl` files, one for each dataset.

2.3 Few-shot fine-tuning

For the submitted solution, each dataset uses the strongest available fine-tuned checkpoint prediction selected by the local scoring pipeline. The final package was submitted publicly as EvalAI ID `573896`.

2.4 Augmentation and checkpoint selection

The fine-tuning pipeline uses standard low-shot detection augmentations:

- multi-scale resizing and scale jitter,
- horizontal or directional flips when domain-appropriate,
- photometric perturbations,
- crop/resize variation,
- conservative handling of dense or small objects.

When validation annotations are available, each candidate checkpoint is scored locally. The final zip is then assembled by choosing the strongest prediction source for each dataset. This per-dataset selector is the central part of the submitted method.

2.5 Inference-time post-processing

Inference uses confidence threshold tuning, NMS threshold tuning, top-300 prediction export, and per-dataset candidate replacement.

3 Results

Table 1 reports the official public EvalAI score for our best Overall Track submission at the time of writing. The local estimate for the same zip was 47.894 mAP. The official EvalAI score is the number used for ranking.

The system works best when open-vocabulary pretraining and the few-shot examples describe the target concept clearly. The weakest cases are specialized medical or fine-grained

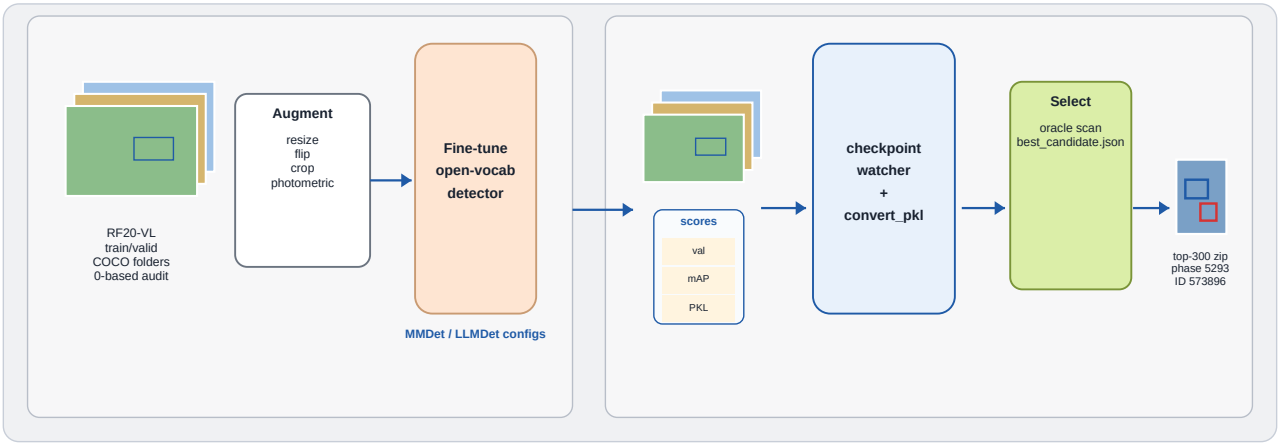

Figure 1: Best Overall Track solution: fine-tuned detector checkpoints are converted, scored per dataset, selected into `best_candidate.json`, and submitted as EvalAI ID `573896`.

datasets, along with small and dense objects. In those settings, class ambiguity and localization errors still hurt performance.

4 Reproducibility

The best solution can be reproduced with the following high-level steps:

1. Download RF20-VL using the organizer-recommended data path.
2. Prepare each dataset in COCO-style format.
3. Generate the selected MMDetection detector configurations.
4. Fine-tune the selected open-vocabulary detector variants per dataset.
5. Run `rf20vl_checkpoint_watcher.py` to infer checkpoints and convert them to challenge PKLs.
6. Run `rf20vl_oracle_scan.py` to choose the best per-dataset PKL and update `best_candidate.json`.
7. Export the top-300 zip and verify zero-based category IDs.
8. Submit the zip to phase 5293 with `evalai_submit_watcher_overall_only.sh`.

The open-source release contains the selected reproduction scripts, final submitted zip, audit script, and candidate-selection state for this Overall Track method.

5 Conclusion

Our Overall Track solution combines fine-tuned open-vocabulary detectors with dataset-specific candidate selection. It follows the Overall Track rules and reaches 47.356 official mAP with submission `573896`. The main path for improvement is stronger handling of weak domains, especially medical images, fine-grained categories, and dense small-object scenes.

References

[1] S. Liu et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. ECCV, 2024.

[2] X. Zhao et al. An open and comprehensive pipeline for unified object grounding and detection. arXiv, 2024.

[3] S. Fu et al. LLMDet: Learning strong open-vocabulary object detectors under the supervision of large language models. arXiv, 2025.

[4] R. Solovyev, W. Wang, and T. Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing, 2021.

[5] P. Robicheaux et al. Roboflow100-VL: A multi-domain object detection benchmark for vision-language models. arXiv, 2025.