No-Finetuning In-Context Detection Ensemble 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 valid In-Context Prompting Track submission for the CVPR 2026 Roboflow-20VL Foundational Few-Shot Object Detection Challenge. This track does not allow gradient-based fine-tuning. The submitted `v2` method therefore starts from an accepted no-finetuning base submission and replaces selected datasets with frozen FSOD-VFM predictions when local scoring improves. The final candidate is packaged as top-300 PKLs with zero-based category IDs and submitted publicly as EvalAI ID `573580`. It appears on the leaderboard under team `v2` with mAP **20.844**. Code is available here: https://github.com/anmspro/cvpr-rf20vl

1 Introduction

The In-Context Prompting Track measures how well pretrained detection systems can use limited visual and category context without updating model weights. This is stricter than the Overall Track. The model may use the few-shot examples and class names at inference time, but it cannot be trained or fine-tuned with gradient updates on RF20-VL.

Our final In-Context submission uses one valid method. An accepted no-finetuning root submission is used as the base. FSOD-VFM is run without gradient updates, then six locally improved datasets are replaced before packaging the final `v2` submission.

2 Method

2.1 Frozen pretrained detectors

The best valid submission uses FSOD-VFM as the frozen improvement source. FSOD-VFM is run with DINOv2 ViT-L/14 features and the RF20-VL support JSON/category list. The model weights are never updated on RF20-VL.

2.2 In-context category conditioning

For each dataset, we use the provided category vocabulary and few-shot context to condition detector inference. The category-conditioning process emphasizes consistency:

- preserve challenge category names,
- avoid category-ID shifts from raw Roboflow exports,
- keep one consistent label vocabulary per dataset,
- generate predictions automatically from model outputs.

Table 1: In-Context Track leaderboard status at initial report time.

Team	Track account	mAP
`v2`	573580	20.844

For the selected submission, category handling is implemented through `support.json`, `target_categories.json`, and a conversion script. The script maps FSOD-VFM's one-based predicted category IDs to the zero-based IDs expected by EvalAI.

2.3 Inference-time filtering and fusion

Since model weights are frozen, the main optimization lever is replacement-based candidate selection. The final method uses an accepted legal `v2` root submission as the base. A dataset is replaced only when the FSOD-VFM output improves the local score. The replacement datasets in the submitted report are:

- `all-elements-fsod-mebv`,
- `defect-detection-yjplx-fxobh-fsod-amdi`,
- `dentalai-i4clz-fsod-fsuo`,
- `lacrosse-object-detection-fsod-uxkt`,
- `new-defects-in-wood-uewd1-fsod-tffp`,
- `orionproducts-vtl2z-fsod-puhv`.

The resulting zip contains twenty dataset prediction files. Each image keeps at most the top 300 instances sorted by confidence.

3 Results

Table 1 shows the public leaderboard score for our best valid no-finetuning In-Context submission. The score is lower than the Overall Track, as expected, because the model cannot adapt weights to RF20-VL's specialized domains.

4 Reproducibility

The best valid In-Context pipeline can be reproduced as follows:

1. Download RF20-VL using the organizer-recommended data path.
2. Prepare `support.json` and `target_categories.json` for FSOD-VFM.

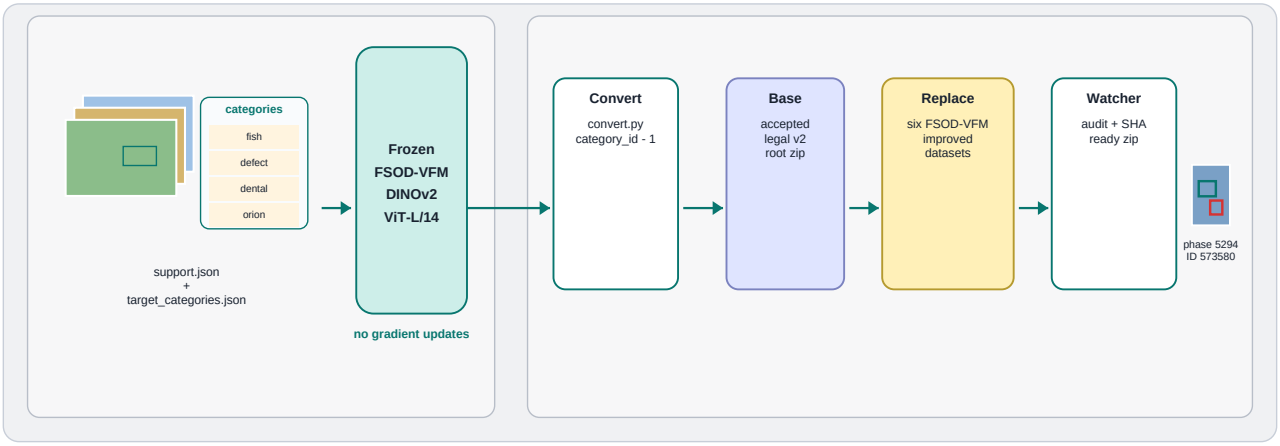

Figure 1: Best In-Context Track solution: accepted legal `v2` base plus FSOD-VFM gradient-free replacements, submitted as EvalAI ID `573580`.

3. Run `rf20vl_fsodvfm_queue.py` on the target datasets with frozen FSOD-VFM/DINOv2 features.
4. Convert FSOD-VFM JSON predictions with `rf20vl_fsodvfm_convert.py`.
5. Run `rf20vl_incontext_fsodvfm_watcher.py` to merge the accepted `v2` base with improved FSOD-VFM replacements.
6. Audit zero-based category IDs and package top-300 PKLs.
7. Submit the final zip to phase 5294 with team `v2`.

The open-source release contains the selected reproduction scripts, final submitted zip, audit script, accepted base artifact, and FSOD-VFM replacement report for this In-Context method.

5 Discussion

The In-Context Track exposes a gap between general open-vocabulary recognition and domain-specific few-shot detection. Without fine-tuning, the model has limited ability to handle specialized object appearances, subtle class differences, and dense small objects. Future work should improve visual in-context retrieval, automatic prompt descriptions, calibration across frozen detectors, and inference-time fusion while still avoiding gradient updates.

6 Conclusion

Our In-Context Track submission is a separate no-finetuning pipeline submitted under team `v2`. The best valid solution uses the accepted root-format `v2` base plus six FSOD-VFM gradient-free replacements. It achieves 20.844 public mAP with EvalAI submission `573580`.

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

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