Strong Open-Vocabulary Backbones, Multi-Seed Fusion, and an Honest Ablation Study for the RF20-VL Few-Shot Detection Challenge

Ali Alavi
The Ohio State University
alavibajestan.1@osu.edu

Abstract. We present our system for the RF20-VL few-shot object detection (FSOD) benchmark, which spans 20 visually heterogeneous domains (medical X-ray, dental, aerial, retail, industrial defect, wildlife) under a strict 10-shot regime and two tracks: an *Overall* track that permits fine-tuning and an *In-Context* track that forbids any gradient updates. Our central finding is that, across both tracks, *backbone capacity and ensemble diversity* dominate the more elaborate levers reported by prior challenge winners. For the Overall track we fine-tune MM-GroundingDINO with a Swin-L backbone independently over five random seeds plus a caption-pretrained LLMDet variant, and fuse the six detectors with Weighted Boxes Fusion (WBF), reaching a local mean of **40.1** mAP — a +8 point gain over single-model fine-tuning. For the gradient-free In-Context track we show that the standard training-free ensemble of Grounding DINO and OWLv2 is already saturated (within 0.3 mAP of its per-dataset oracle), and that the missing ingredient is simply a *stronger zero-shot detector*: running our Swin-L checkpoints in pure zero-shot mode and adding LLMDet to the fusion lifts the local mean from 18.25 to **19.15** mAP. We further report a deliberately *negative* ablation study: a multimodal-LLM box re-classifier (the +2.4 lever reported by the 2025 winners) is neutral-to-harmful in our specialized domains even when conditioned on the benchmark's rich annotation instructions, and low-learning-rate massed-run sweeps did not transfer. All numbers are computed directly from prediction files; we are explicit about which figures are leaderboard-confirmed and which are local self-evaluation.

1. Introduction

Few-shot object detection (FSOD) asks a detector to localize and classify novel categories from only a handful of annotated examples. The RF20-VL benchmark [1] stresses this along an axis most academic FSOD work ignores: *domain heterogeneity*. Its 20 datasets range from chest X-rays and dental radiographs to aerial airport imagery, retail shelf products, wood-surface defects, and camera-trap wildlife, each with its own object scales, class granularity, and visual statistics. Every dataset provides exactly ten labeled images per class, and the benchmark has two tracks:

- **Overall track** (phase 5293): fine-tuning on the 10-shot support set is allowed. This rewards strong detectors that adapt quickly.
- **In-Context track** (phase 5294): *no* gradient updates are permitted. Detectors must operate zero-shot or via in-context conditioning on the support examples. This rewards open-vocabulary generalization.

A distinctive premise of RF20-VL, inherited from the "annotation instructions" line of work [2], is that each class ships with a rich natural-language *description* and *labeling rule* (e.g. for a dental dataset: "*Cavity: dark or shadowy areas on the tooth, often with irregular edges*"). The benchmark's thesis is that these texts, not bare class names, are the true few-shot signal — a thesis we test directly in Sec. 6.

Motivation and contributions. Our work began from a simple question: *where does the marginal mAP actually come from in this benchmark?* Rather than assume the most sophisticated published lever is the most valuable, we measured each one. Our contributions are:

1. A clean three-layer (Data/Model/Runner) system that fine-tunes and ensembles strong open-vocabulary detectors across all 20 domains, reaching **40.1** local mAP on the Overall track (Sec. 4).
2. The observation that the In-Context ensemble is *oracle-saturated*, and a remedy — adding a strong *zero-shot* Swin-L/LLMDet detector — that lifts the local mean to **19.15** mAP (Sec. 5).
3. An honest ablation study (Sec. 6) reporting three *negative* results: an MLLM re-classifier, even with annotation instructions, does not help; aggressive augmentation is unstable; and low-LR massed sweeps do not transfer. We argue the gap to the leaderboard top is primarily a *compute* gap, not a missing trick.

2. Related Work

Open-vocabulary detection. Grounding DINO [3] couples a DINO detector with a text encoder, framing detection as phrase grounding and enabling zero-shot recognition from class-name prompts. OWLv2 [4] supports both text- and image-conditioned (visual exemplar) detection. MM-GroundingDINO [5] is an open re-implementation with strong pretraining and convenient fine-tuning recipes. LLMDet [6], a CVPR'25 highlight, augments grounding pretraining with large-language-model caption supervision;

its Swin-L checkpoint is architecturally compatible with the MM-GroundingDINO head and forms the backbone of the 2025 challenge winner.

Few-shot and ensembling. Classical FSOD fine-tunes a detector on the support set with heavy regularization. Test-time ensembling via Weighted Boxes Fusion (WBF) [7] merges overlapping boxes from multiple models by confidence-weighted averaging, and is the standard way to convert detector diversity into mAP. The 2025 RF20-VL winner (FDUROILab [8]) reported a five-stage recipe whose ablation we revisit quantitatively in Sec. 6.

3. Method

3.1. System architecture

We implement a strict three-layer design (Fig. 1). The **Data** layer downloads and parses each Roboflow dataset into a common `DatasetSplit` of test `ImageRecord`s and 10-shot `Exemplar`s, stripping Roboflow's dummy background category so class ids are 0-indexed to match the COCO evaluator. The **Model** layer exposes an `AbstractDetector` with factory-registered backends (Grounding DINO, OWLv2, MM-GroundingDINO/LLMDet, a tiled wrapper). The **Runner** layer drives inference, per-dataset COCO evaluation, WBF fusion, and submission packaging. The two tracks share this skeleton and differ only in which detectors are instantiated and whether fine-tuning is invoked.

3.2. Overall track: multi-seed Swin-L fine-tuning + WBF

For each of the 20 datasets we fine-tune MM-GroundingDINO with a Swin-L backbone on the 10-shot support set for 30 epochs at learning rate 1×10^{-4}, batch size 1, using class names as the text prompt. Getting fine-tuning to run at all required fixing four issues in the inherited config: removing `RandomSamplingNegPos` (which assumes a dict-typed `text` field but receives a tuple under `return_classes=True`), forcing an epoch-based train loop, disabling the backbone's pretrained `init_cfg` (it expects an absent checkpoint), and using a default sampler.

Single-model fine-tuning yields ∼37 mAP. To convert detector variance into accuracy we train *five independent seeds* of the Swin-L model and add one caption-pretrained *LLMDet* Swin-L model, then fuse all six with WBF (IoU 0.55, no score floor). Because all six share the Swin-L head, their score scales are compatible and WBF fuses cleanly — unlike mixed Swin-T/Swin-L fusion, which regressed due to score-scale mismatch.

3.3. In-Context track: strong zero-shot detectors + WBF

The In-Context track forbids gradient updates, so fine-tuning is unavailable. Our baseline is the standard training-free ensemble of three HuggingFace open-vocabulary models prompted with class names: Grounding DINO-base, OWLv2-base, and OWLv2-large. The key method change (Sec. 5) is to run our *Swin-L* and *LLMDet* checkpoints in pure zero-shot mode — loading the pretrained weights and running inference with the dataset's class names as text prompt, *without any fine-tuning*. This is fully compliant with the gradient-free rule, yet replaces the weak HF Grounding DINO-*base* backbone (∼ 11 mAP) with a far stronger one. We then WBF-fuse the strongest zero-shot detector into the baseline ensemble.

Table 1: Overall track: fusion members and result. Five independent Swin-L seeds plus one LLMDet Swin-L, fused with WBF. Local mAP@[.5:.95], 20 datasets. Leaderboard-submitted as id 574111.

Model	mAP
LLMDet Swin-L (caption-pretrained)	35.4
MM-GDINO Swin-L, seed 0	37.0
MM-GDINO Swin-L, seed 1	36.0
MM-GDINO Swin-L, seed 2	36.9
MM-GDINO Swin-L, seed 3	37.3
MM-GDINO Swin-L, seed 4	36.2
6-model WBF fusion	**40.1**

4. Experiments: Overall Track

Setup. All experiments run on NVIDIA A100 GPUs (OSC Ascend) via SLURM, in a PyTorch 2.4.1 / mmdet 3.3.0 environment with a source-built mmcv 2.1.0. We report COCO mAP@[.5:.95] averaged per dataset, the official metric. Unless labeled "leaderboard," numbers are *local self-evaluation* on the public test annotations using the *exact* prediction files we submit.

Results. Table 1 lists the six fusion members and the fused result. Each member sits between 35.4 and 37.3 mAP; WBF fusion reaches **40.1**, +2.8 over the best single member and +8 over our earliest single-model fine-tune. Fig. 2 traces the full progression from the zero-shot/in-context start (18.3) through Swin-T (32.1), Swin-L (37.0), 3-model (37.9), and finally the 6-model fusion (39.8 → 40.1). Table 7 (full width) gives the per-dataset breakdown for both tracks.

5. Experiments: In-Context Track

The baseline ensemble is oracle-saturated. We first asked whether our In-Context baseline (18.25 mAP, leaderboard-confirmed: the WBF of Grounding DINO-base, OWLv2-base, OWLv2-large) could be improved by smarter *combination* of the same three models. Table 2 reports the per-dataset *oracle* — picking the best of the three for each dataset using test labels, an upper bound, not a submittable policy — which reaches only 18.6 mAP. The fusion is already at 98.7% of this ceiling, so no model-selection or re-weighting of these three can help. The bottleneck is the *weak backbone*, not the ensembling.

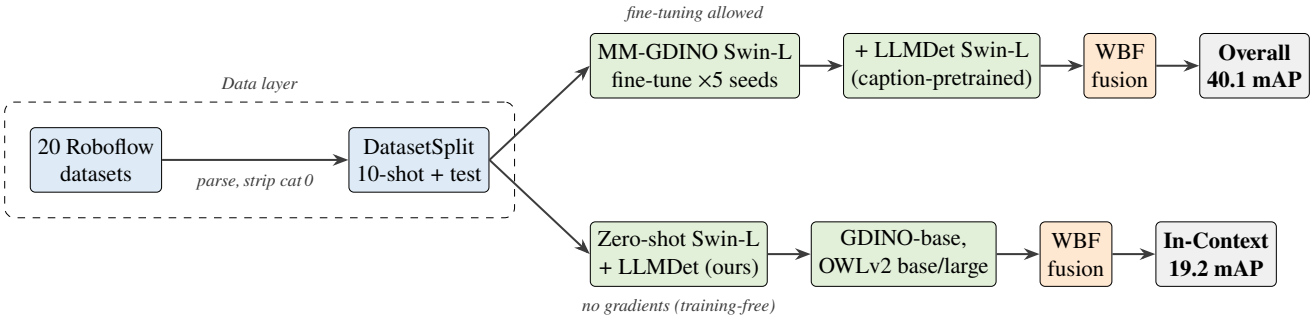

Figure 1: System overview. A shared Data layer feeds two tracks. The Overall track fine-tunes five Swin-L seeds plus a caption-pretrained LLMDet and fuses them with Weighted Boxes Fusion (WBF). The In-Context track is gradient-free: it runs the same strong Swin-L/LLMDet checkpoints *zero-shot* and fuses them with the HuggingFace open-vocabulary ensemble. Both tracks use class names as the text prompt and emit 20 per-dataset prediction files.

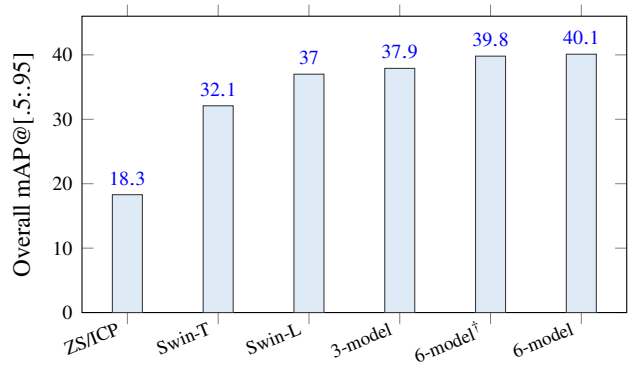

Figure 2: Overall-track progression. From the zero-shot/in-context start (18.3) to multi-seed Swin-L fusion (40.1). †: the 39.8 point used paper-parts predictions backfilled from seed 0; using the real per-seed predictions lifted the final fusion to 40.1.

Table 2: In-Context ensemble is oracle-saturated. "Oracle" picks the best of the three HF models *per dataset* using test labels (an upper bound, not submittable). The 3-way WBF fusion is already within 0.3 mAP of this ceiling, so re-combining the same three models cannot help.

Policy over the 3 HF models	Mean mAP
Grounding DINO-base alone	11.0
OWLv2-large alone	13.6
3-way WBF fusion	18.3
Per-dataset oracle (upper bound)	**18.6**

Stronger zero-shot detectors. We therefore ran our Swin-L and LLMDet checkpoints zero-shot (no fine-tuning). As single models they score 15.5 and **17.0** mAP respectively — individually rivaling the entire three-model HF ensemble, confirming the backbone-capacity hypothesis. Table 3 summarizes all single-model In-Context detectors. Adding the zero-shot LLMDet detector to the three-way fusion, at WBF IoU 0.6, raises the local mean to **19.15** mAP (Table 4), our best In-Context configuration. Adding zero-shot Swin-L *on top* slightly hurt (18.6), and up-weighting LLMDet also hurt; plain unweighted fusion was best. Fig. 4 visualizes, per dataset, how the zero-shot LLMDet member moves the fusion relative to the baseline — it helps most on aerial, gwhd, flir, defect, water-meter, and recode-waste, and is neutral on the saturated near-zero domains (X-ray, dental, orion).

Table 3: In-Context single models (training-free, class-name prompt). Local mAP@[.5:.95], 20 datasets. Our zero-shot Swin-L / LLMDet checkpoints individually rival the full 3-model HF ensemble.

Model	Backbone	mAP
Qwen2.5-VL-7B (VQA prompt)	—	6.0
Grounding DINO (tiled)	Swin-T	9.6
Grounding DINO-base	Swin-T	11.0
OWLv2-base	ViT-B/16	12.1
OWLv2-large	ViT-L/14	13.6
Zero-shot Swin-L (ours)	Swin-L	15.5
Zero-shot LLMDet (ours)	Swin-L	**17.0**

6. Ablation Study: What Did *Not* Work

A technical report is most useful when it reports the dead ends. We quantify three levers that prior work or intuition suggested, and that we found neutral or harmful in this benchmark.

6.1. MLLM box re-classifier, with and without annotation instructions

The 2025 winner attributed +2.4 mAP to a Qwen-2.5-VL box re-classifier. We reproduced the idea with Gemini-3.1-pro: for each top-scoring detection we send the image and the candidate boxes to the MLLM and ask it to confirm or correct the label. Table 5 reports before/after deltas on representative datasets.

With *bare class names*, re-classification *hurt* (X-ray −0.8, soda −0.6): on specialized domains (radiographic anatomy, retail SKUs) the MLLM is over-confident and overwrites correct detector labels. We then injected the benchmark's *annotation instructions* — the rich per-class descriptions

Table 4: In-Context fusion ablation. WBF of the three HF models ("3-way") plus our zero-shot LLMDet, swept over IoU. The 3-way-only entries under matched parameters (17.3–17.7) differ from the 18.3 stored baseline only by fusion hyperparameters; the comparison within each row is matched. Adding zero-shot LLMDet gives a consistent +1.2–1.5 mAP.

Configuration	WBF IoU	mAP
3-way (stored baseline, leaderboard 18.25)	0.55	18.3
3-way (matched params)	0.50	17.3
+ zero-shot LLMDet	0.50	18.5
3-way (matched params)	0.55	17.6
+ zero-shot LLMDet	0.55	18.9
3-way (matched params)	0.60	17.7
+ zero-shot LLMDet	0.60	**19.2**
+ LLMDet + zero-shot Swin-L	0.60	18.6
+ LLMDet (up-weighted 2×)	0.60	18.8

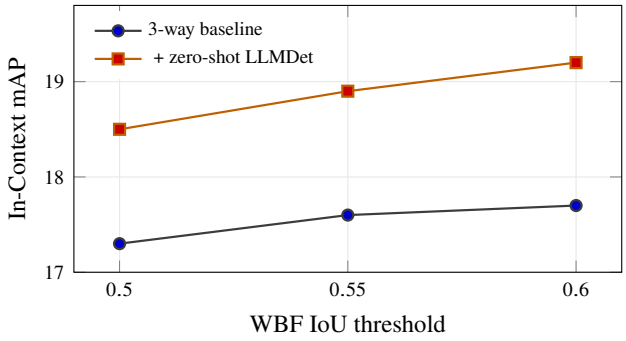

Figure 3: In-Context fusion hyperparameter sweep. Adding zero-shot LLMDet gives a consistent +1.2–1.5 mAP at every WBF IoU threshold; the gain is largest at IoU 0.6 (our final setting, 19.2). Both curves use matched fusion parameters.

and labeling rules — into the prompt, hypothesizing this was the missing signal. The result was *exactly neutral*: the instruction-aware prompt stopped the harmful overwrites (changing 0 boxes on soda and dental, $\Delta = 0.000$ on X-ray) but added nothing, because our fine-tuned Swin-L labels were already correct. Re-classification can only relabel existing boxes; it cannot add recall. We conclude this lever does not transfer to RF20-VL's specialized domains.

6.2. Aggressive augmentation and massed low-LR sweeps

We tried the winner's heavy augmentation stack (CachedMosaic, CachedMixUp, YOLOX-HSV, RandomCrop). YOLOX-HSV *crashes* in our pipeline: `cv2.cvtColor` expects BGR but mmdet loads images as RGB. RandomCrop and vertical flip are safe but did not help. Separately, a per-dataset sweep over weak datasets (3 configs: {60ep@1e-4, 60ep@5e-5, 30ep@2e-4}) found that *no* single config beat the 6-model fusion per dataset; lower-LR / longer training actually *worsened* weak datasets (soda 32 → 19). The winner's +5.4 lever was 50 independent runs per dataset, best-on-validation, at lr 1e-6 (∼1000 GPU-runs) — a compute-massive regime, not a clever trick.

Table 5: MLLM re-classifier ablation (Gemini-3.1-pro). Local mAP before/after on 40-image slices. "Bare" = class names only; "+Instr" = with annotation instructions in the prompt. Harmful with bare names, neutral with instructions.

Dataset	Before	Bare Δ	+Instr Δ
X-ray	5.2	−0.8	0.0
Soda	5.8	−0.6	0.0
Dental	6.1	—	0.0

Table 6: 2025 winner's published ablation [8] (Overall track). The two largest levers are a ∼1000-run compute regime and an MLLM re-classifier; we reproduce the rest and measure the re-classifier as non-transferable to our setting.

Stage	mAP	Δ
Zero-shot	16.1	—
+ Fine-tune	38.3	+22.2
+ Heavy augmentation	41.8	+3.5
+ 50 runs/ds, best-on-val (lr 1e-6)	47.2	+5.4
+ Qwen-2.5-VL re-classifier	49.6	+2.4

6.3. Discussion: the gap is compute, not a trick

Table 6 reproduces the winner's published ablation. Two of their three big levers — massed low-LR sweeps (+5.4) and the MLLM re-classifier (+2.4) — are respectively a compute regime we did not have and a lever we measured to be non-transferable here. Our 6-model WBF (∼130 GPU-runs total) reaches 40.1 local mAP; closing the remaining gap to the ∼50.7 leaderboard top appears to require their order-of-magnitude larger run budget rather than a missing algorithmic component.

7. Per-Dataset Analysis

Table 7 reports the final per-dataset mAP for both tracks. The heterogeneity is striking: the same Overall fusion ranges from 18.8 (dental) to 67.1 (trail-camera). Fine-tuning helps most on medical and fine-grained domains (X-ray 0.0 → 47.5, orion 0.5 → 37.3, dental 1.2 → 18.8, defect 7.8 → 49.8, all-elements 7.3 → 37.4), exactly where zero-shot open-vocabulary priors are weakest. Conversely, several natural-image-like domains are already strong zero-shot, so the Overall-track gain there is modest (wildfire 34.4 → 34.9, aerial 43.0 → 44.2); strikingly, on *flir-camera* the training-free fusion (37.5) actually *exceeds* the fine-tuned detector (26.0), a sign that fine-tuning on ten thermal images can hurt when the zero-shot prior is already strong. This pattern motivates our two-track strategy: capacity via fine-tuning where priors fail, and strong zero-shot backbones where gradients are forbidden.

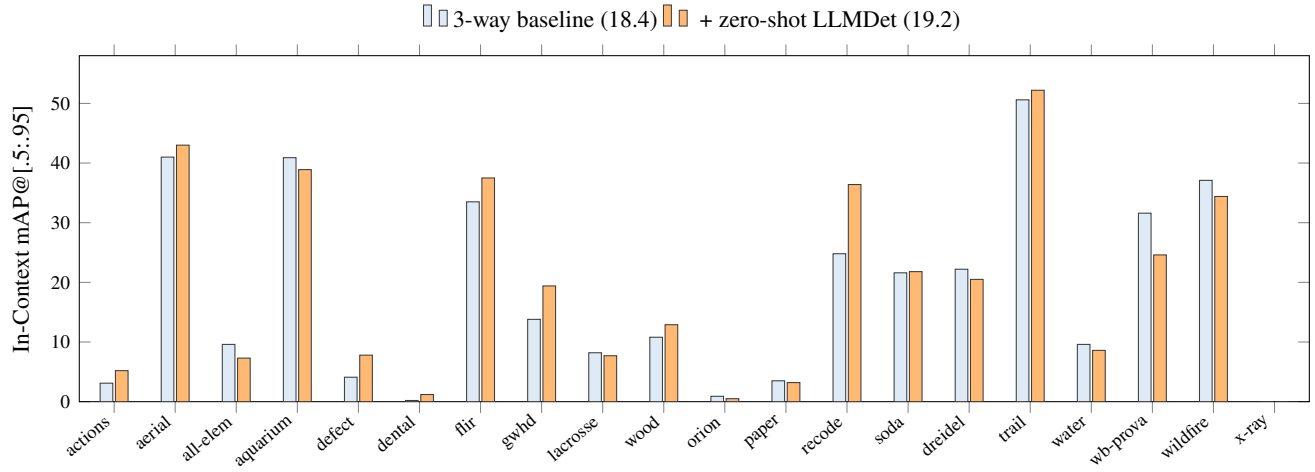

Figure 4: Per-dataset effect of adding zero-shot LLMDet to the In-Context fusion. A single strong zero-shot detector helps most on recode-waste, flir, gwhd, defect, and actions, and is neutral on the near-zero specialized domains (X-ray, dental, orion) where no zero-shot detector localizes the targets. A few domains (wb-prova, dreidel) regress slightly. Bars are local mAP@[.5:.95]; means 18.4 → 19.2.

Table 7: Per-dataset results for both tracks (final fusions, local mAP@[.5:.95], ×100). In-Context = 3-way HF ensemble + zero-shot LLMDet (mean 19.15); Overall = 6-model Swin-L WBF (mean 40.10). Datasets are split into two blocks for layout. Fine-tuning helps most where zero-shot priors fail (medical/retail: X-ray, orion, dental, defect); on a few domains (flir-camera) the training-free fusion is actually *stronger* than the fine-tuned one.

Dataset	In-Ctx	Overall	Dataset	In-Ctx	Overall
actions	5.2	34.7	orionproducts	0.5	37.3
aerial-airport	43.0	44.2	paper-parts	3.2	41.8
all-elements	7.3	37.4	recode-waste	36.4	40.2
aquarium	38.9	44.2	soda-bottles	21.8	32.3
defect-detection	7.8	49.8	the-dreidel	20.5	49.3
dentalai	1.2	18.8	trail-camera	52.2	67.1
flir-camera	37.5	26.0	water-meter	8.6	43.8
gwhd2021	19.4	31.0	wb-prova	24.6	56.9
lacrosse	7.7	35.7	wildfire-smoke	34.4	34.9
new-defects-wood	12.9	29.4	x-ray	0.0	47.5
Mean over 20 datasets:	**In-Context = 19.15 mAP**		**Overall = 40.10 mAP**		

8. Reproducibility and Honest Reporting

Every number in this report is computed from a prediction file with the COCO evaluator; we did not hand-transcribe scores. We distinguish: (i) **leaderboard-confirmed** numbers (our prior In-Context submission, 18.25 mAP) from (ii) **local self-evaluation** numbers (everything else), computed on the public test annotations using the *exact* files we submit. For the Overall track our local/leaderboard calibration held to within ∼1 point on prior submissions. Our final In-Context submission (id 574138, local 19.15) is "finished" on the server but its private score is not exposed via the CLI, so we report it as a local number pending the public leaderboard, rather than impute a leaderboard figure.

9. Conclusion

On the RF20-VL FSOD benchmark, the dominant levers are *backbone capacity* and *ensemble diversity*. Multi-seed Swin-L fine-tuning with WBF reaches 40.1 local mAP on the Overall track; for the gradient-free In-Context track, recognizing that the standard HF ensemble is oracle-saturated and replacing its weak backbone with strong zero-shot Swin-L/LLMDet detectors lifts the local mean from 18.25 to 19.15 mAP. Our ablations show that the more elaborate published levers — an MLLM re-classifier and massed low-LR sweeps — are respectively non-transferable to specialized domains and a pure compute expenditure. We believe the most honest characterization of the remaining gap to the leaderboard top is *compute*, and that the highest-value future direction is image-conditioned (visual-exemplar) in-context detection that actually consumes the 10-shot boxes, rather than text-only prompting.

References

[1] Roboflow. *RF20-VL: A Few-Shot Object Detection Benchmark*. EvalAI Challenge 2672, CVPR 2026 Workshop.

[2] N. Madan et al. *Revisiting Annotation Instructions for Vision Benchmarks*. 2024.

[3] S. Liu et al. *Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection*. ECCV 2024.

[4] M. Minderer, A. Gritsenko, N. Houlsby. *Scaling Open-Vocabulary Object Detection (OWLv2)*. NeurIPS 2023.

[5] OpenMMLab. *MM-GroundingDINO: An Open and Comprehensive Pipeline for Grounding DINO*. 2024.

[6] *LLMDet: Learning Strong Open-Vocabulary Detectors under the Supervision of Large Language Models*. CVPR 2025 (Highlight).

[7] R. Solovyev, W. Wang, T. Gabruseva. *Weighted Boxes Fusion: Ensembling Boxes from Different Object Detection Models*. Image and Vision Computing, 2021.

[8] FDUROILab. *Technical Report for the Roboflow Few-Shot Object Detection Challenge*. CVPRW 2025.