RETHINKING OVER-SEGMENTATION IN FEW-SHOT OBJECT DETECTION WITH VISION FOUNDATION MODELS: A TRAINING-FREE PROTOTYPE REFINEMENT AND OCCLUSION-AWARE DISCRIMINATION APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot object detectors built on frozen vision foundation models (VFMs) achieve strong cross-domain performance without task-specific training, but they suffer from *over-segmentation*: a single object is fragmented into multiple proposals, and non-maximum suppression (NMS) can delete the correct whole-object proposal in favour of a higher-confidence fragment. Our diagnostic study shows that this failure differs qualitatively from the classical one: the largest proposal almost always covers the object, and few-shot classification confidence is the bottleneck instead. We address the two sides of this failure with two training-free modules. *Training-free Prototype Refinement* (TPR) repairs the confidence signal: it retrieves an adjacency set of hard negatives from a frozen memory bank and moves the class prototype away from them, within a trust region derived from the support set. *Occlusion-Aware Discrimination* (OAD) repairs the selection rule: it separates whole objects from fragments by their confidence decline under partial occlusion, and retains genuine small objects that NMS would otherwise suppress. Together, the two modules form Prototype Refinement and Occlusion-Aware Inference-Time Selection (POISE), which improves over state-of-the-art methods on standard FSOD benchmarks while adding no trainable parameters.

1 INTRODUCTION

Few-shot object detection (FSOD) aims to detect novel categories from a handful of annotated examples. Classical approaches either meta-learn a class-agnostic comparator (Yan et al., 2019; Xiongwei Wu, 2020; Han et al., 2022b) or fine-tune a base detector on the novel examples (Wang et al., 2020); both require task-specific training on base classes and degrade under base–novel domain shift. Frozen vision foundation models (VFMs) offer an alternative: DINOv2 features separate object categories without supervision (Oquab et al., 2024; Darcet et al., 2024), the Universal Proposal Network (UPN) generates class-agnostic proposals across domains (Jiang et al., 2024), and SAM provides promptable masks (Kirillov et al., 2023; Ravi et al., 2024). Pipelines such as FSOD-VFM (Feng et al., 2026) combine these frozen components into strong training-free or near-training-free detectors that deploy to new domains without gradient-based adaptation.

These pipelines, however, exhibit a recurring failure: *over-segmentation*. A single object is split into several overlapping proposals, and NMS may keep a fragment while deleting the correct whole-object proposal, because the fragment receives higher classification confidence. Classical work on duplicate suppression (Hosang et al., 2017; Bodla et al., 2017) rests on the assumption that the highest-confidence box is the correct one. Whether this assumption survives in the few-shot regime, where the classifier is a cosine similarity to a prototype estimated from as little as one support example, has not been systematically examined.

Our diagnostic study (Section 3.2) reaches three conclusions, and Figure 2 illustrates the underlying failure chain. First, coverage is not the bottleneck: on Pascal-5_i, the largest proposal in an over-segmentation component achieves mean IoU 0.91 with the ground truth (Finding F1). Second, the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

failure is a *confidence* failure: at 1-shot, the largest proposal is the most confident one in only 39–45% of cases on three cross-domain datasets, rising to 71–75% at 10-shot (F3); with strong backbones and abundant support the phenomenon largely disappears (F2). Third, the small member of a nested pair is occasionally a genuine, distinct object that NMS deletes as a fragment (F4). In short, biased support sets shift the prototype towards part-views, so confidence ceases to be a reliable selection signal exactly where over-segmentation occurs.

We therefore intervene on both sides of this failure with two training-free modules, which together form POISE. *Training-free Prototype Refinement* (TPR, Section 3.3) repairs the confidence signal: it retrieves hard negatives from a frozen memory bank and moves the class prototype away from them, within a support-derived trust region. *Occlusion-Aware Discrimination* (OAD, Section 3.4) repairs the selection rule: for nested same-class pairs it compares confidence decline under partial occlusion, and retains both proposals when their scores are too close to adjudicate, protecting genuine small objects. Across eight benchmarks evaluated on the complete standard splits, POISE improves over FSOD-VFM and CD-ViTO on every benchmark.

Contributions.

- **A diagnostic study** of over-segmentation in VFM-based FSOD that reframes it as a *selection* failure rather than a coverage failure: coverage is nearly always sufficient (F1), few-shot classification confidence is unreliable exactly where over-segmentation occurs (F2–F3), and NMS additionally deletes occasional genuine small objects (F4).
- **Training-free Prototype Refinement (TPR)**, which repairs the confidence signal by moving each class prototype away from an adjacency set of retrieved hard negatives.
- **Occlusion-Aware Discrimination (OAD)**, which repairs the selection rule by adjudicating nested same-class proposal pairs through their confidence decline under partial occlusion, separating whole objects from fragments while retaining genuine small objects. Combined as POISE, the two modules improve over state-of-the-art methods on standard FSOD benchmarks.

2 Related Work

Few-shot object detection. Classical FSOD methods either meta-learn a class-agnostic comparator (Yan et al., 2019; Xiongwei Wu, 2020; Han et al., 2022b) or fine-tune a base detector on novel examples (Wang et al., 2020; Han et al., 2022a; Sun et al., 2021); all require task-specific training on base classes and degrade under base–novel domain shift. A parallel route replaces visual support with text prompts: open-vocabulary detectors (Zareian et al., 2021; Li et al., 2022; Minderer et al., 2022; Liu et al., 2024; Cheng et al., 2024; Zhou et al., 2022) transfer through language. Appendix F compares the two routes empirically.

Detection with vision foundation models. Self-supervised encoders (Caron et al., 2021; Oquab et al., 2024; Darcet et al., 2024), class-agnostic proposal generators (Jiang et al., 2024), and promptable segmentation models (Kirillov et al., 2023; Ravi et al., 2024) enable detection pipelines with little or no training. FSOD-VFM (Feng et al., 2026) integrates these components with graph diffusion; CD-ViTO (Wang et al., 2024) uses SAM masks and a lightweight trained adapter. POISE replaces the graph-diffusion post-processing of FSOD-VFM with two training-free modules.

Duplicate suppression and prototype classification. NMS greedily suppresses overlapping boxes by confidence; Soft-NMS (Bodla et al., 2017) decays confidence instead of deleting boxes, and learned alternatives (Hosang et al., 2017) predict suppression directly. All assume that the highest-confidence box is the correct one—the assumption our diagnostic study shows fails in the few-shot regime. On the classification side, prototype networks (Snell et al., 2017) classify by distance to the support mean, and refinements repel prototypes from confusable regions. TPR shares this negative-repulsion intuition but instantiates it without training: negatives are retrieved from a frozen memory bank and the update is bounded by a support-derived trust region (Section 3.3). In FSOD, prototypes are built from RoI-pooled support features (Feng et al., 2026), and the CD-ViTO adapter (Wang et al., 2024) can be read as a trained prototype refinement.

Occlusion as evidence. Occlusion-sensitivity analysis probes which image regions drive a decision (Zeiler & Fergus, 2014); erasing-based regularisers remove regions during training (DeVries

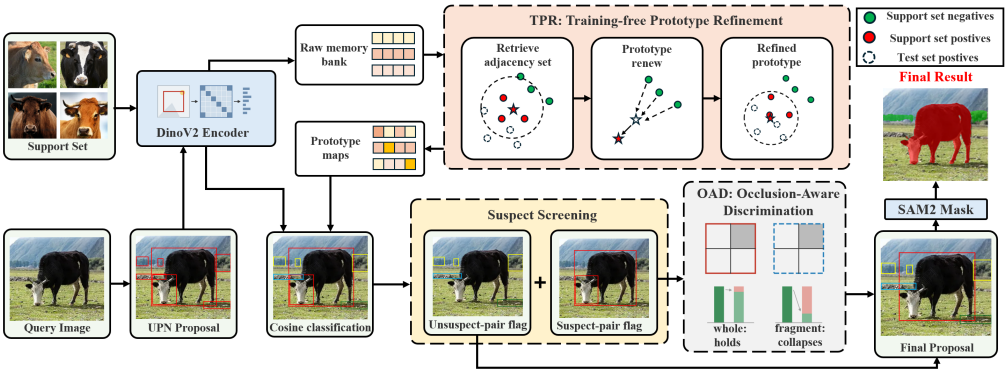


Figure 1: POISE overview. The support set is encoded by a frozen DINOv2 encoder into a raw memory bank and per-class prototype maps. *TPR* retrieves an adjacency set of hard negatives, renews each prototype by a bounded one-step update, and writes the refined prototype back. UPN proposals are scored by cosine classification against the refined prototypes; *suspect screening* flags nested same-class pairs and passes only those to *OAD*, which separates whole objects from fragments.

& Taylor, 2017; Zhong et al., 2020); adversarial occlusion hardens trained detectors (Wang et al., 2017); and test-time augmentation ensembles predictions over transformed inputs (Shanmugam et al., 2021). OAD differs in purpose: it applies structured occlusion at inference as a *comparative probe* of object completeness within a nested proposal pair, using the whole/fragment asymmetry of the decline rate (Table 9e) as a selection signal rather than an optimisation target.

3 Method

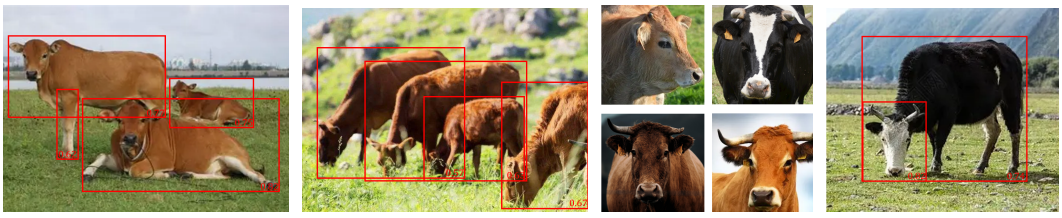
Section 3.1 gives an overview of the POISE pipeline, Section 3.2 summarises the diagnostic evidence that motivates the design, and Sections 3.3–3.4 present the two modules. Sampling protocols, derivations, boundary cases, and pseudocode are deferred to Appendices A–C; the UPN-free extension is evaluated in Section 4.3.

3.1 Method Overview

Figure 1 summarises the POISE pipeline. On the support side, the frozen DINOv2 encoder turns the few-shot support examples into a raw memory bank of features and a set of per-class prototype maps; TPR (dashed pink panel) then retrieves an adjacency set of hard negatives from the memory bank, renews each support-mean prototype by a bounded one-step update, and writes the refined prototype back to the prototype maps. On the query side, UPN proposals from the query image are scored by cosine classification against the refined prototypes, and a *suspect-screening* stage (dashed yellow panel) splits nested same-class pairs into unsuspect pairs, which are kept as-is, and suspect pairs, which are passed to OAD. OAD probes each member of a suspect pair with partial occlusions: the confidence of a whole object holds, whereas that of a fragment collapses, so the pair is adjudicated by their joint score, with both retained when the scores are too close to separate. The surviving final proposals are finally converted to masks by SAM2 to produce the output. An optional UPN-free branch replaces proposals with SAM's automatic mask generation (Section 4.3).

3.2 Problem Setup and Diagnostic Findings

Problem setup. For each novel class c, the support set $\mathcal{S}_c$ contains K annotated examples. UPN produces proposals $\{b_j\}$ and DINOv2 provides proposal features $\{F_j^q\}$; after selection, SAM2 converts the retained proposals into masks $\{M_j\}$ for the final output. The predicted class is $\hat{c}_j = \arg\max_c \cos(F_j^q, \hat{p}_c)$, where $\hat{p}_c$ is the ℓ_2-normalised support-feature mean. A same-class pair is an *over-segmentation pair* when one box is largely contained in the other (containment IoU ≥ 0.75); the largest box in a connected component is the candidate whole-object proposal.



(a) (b) (c) (d)

Figure 2: Why over-segmentation defeats confidence-based suppression in the few-shot regime. (a) Real scenes contain multiple overlapping same-class instances, so class-agnostic proposal generators fragment objects into nested, mutually overlapping proposals that simple merging cannot consolidate. (b) The smaller member of a nested pair is occasionally a genuine, distinct object, so the covered box cannot simply be discarded. (c) Few-shot support sets are frequently biased: a handful of examples may emphasise part-views and fail to represent the class as a whole. (d) Under such bias, the prototype shifts towards part-views, and an over-segmentation fragment can attain *higher* classification confidence than the correct whole-object proposal.

F1: coverage is rarely the bottleneck. On Pascal-5_i (DINOv2-B, 5-shot, all four splits), the largest proposal has mean IoU 0.91 with the ground truth, and 96.9% of the smaller proposals are contained within it at containment IoU $\geq$ 0.95 (Table 9a). Over-segmentation in this pipeline is therefore a *selection* failure, not a proposal-generation failure.

F2: strong backbones and abundant support rarely over-segment. With DINOv2-L and 10-shot support, the Pascal-5_i over-segmentation error rate falls below 1% (Table 9b): the phenomenon concentrates in low-shot settings and with weaker backbones.

F3: few-shot confidence is unreliable. On samples of 2,000 over-segmentation components per dataset and shot, the largest proposal is ranked first by classification confidence at 1-shot in only 39.2% (DIOR), 44.6% (ArTaxOr), and 41.3% (NEU-DET) of cases, rising to 71.4–74.8% at 10-shot (Table 9c, with Wilson 95% confidence intervals). Confidence-based NMS therefore deletes the whole-object candidate in a majority of 1-shot cases; a smaller pooled audit on Pascal-5_i is directionally consistent (Appendix A). Figures 2c–2d show a concrete case of this mechanism: a part-biased support set lets a fragment outscore the whole object.

F4: the small member of a pair is not always a fragment. It can be a genuine, distinct same-class object whose confidence is lower simply because it aggregates fewer and weaker features; NMS then deletes it as a fragment. Our statistics provide only an upper bound on how often this happens (Table 9d); four unambiguous real-world cases confirm that it occurs (Appendix D); Figure 2b shows a representative case.

Design implication. TPR addresses the cause identified by F2–F3—the biased confidence signal; OAD addresses the selection problem identified by F3–F4 with a signal that does not rely on raw confidence alone (validated in Table 9e).

3.3 Training-free Prototype Refinement (TPR)

Prototype-based FSOD represents each class by the mean of support features (Snell et al., 2017; Feng et al., 2026)—training-free, but sensitive to biased support sets (F3). TPR instead retrieves an *adjacency set* $\mathcal{A}_c$ of hard negatives for class c, using either cosine top-k retrieval or an encoder-confusion criterion (Appendix B.3). Let $\hat{a}_c$ denote the ℓ_2-normalised mean of the adjacency features, so that both terms of the update live on the same unit sphere as $\hat{p}_c$. The bounded one-step update is

$$\tilde{p}_c = \hat{p}_c + \eta\,(\hat{p}_c - \hat{a}_c)\,, \tag{1}$$

followed by trust-region projection and ℓ_2 normalisation:

$$\hat{p}_c^{\text{new}} = \ell_2\text{-norm}\left(\hat{p}_c + \min\left(1, \frac{r_c}{\|\tilde{p}_c - \hat{p}_c\|_2}\right)(\tilde{p}_c - \hat{p}_c)\right). \tag{2}$$

r_c is the mean distance from support features to $\hat{p}_c$. At 1-shot this radius is degenerate, so we set a fixed threshold 0.3, matched to the measured support-feature dispersion on unit sphere (Table 16).

Table 1: Main results on CD-FSOD subsets (nAP50, 5-shot, mean±std over three support-set seeds). Rows marked ‡ are fine-tuning-based baselines reproduced once under our unified protocol with their standard training recipes; their absolute values are included as order-of-magnitude references only.

Method	ArTaxOr	Clipart1K	DIOR	DeepFish	NEU-DET	UODD	Avg
TFA‡ (Wang et al., 2020)	22.4	31.2	28.6	19.8	18.4	15.2	22.6
Meta Faster R-CNN‡ (Han et al., 2022b)	28.6	36.8	33.4	24.1	22.3	18.7	27.3
HeteroGraph‡ (Han et al., 2022a)	31.2	38.4	35.8	26.7	24.1	20.4	29.4
FSOD-VFM (Feng et al., 2026)	38.7±0.6	44.5±0.3	42.1±0.6	36.2±1.0	31.2±0.6	26.2±0.7	36.5±0.6
CD-ViTO (Wang et al., 2024)	40.1±0.5	45.8±0.5	43.5±0.8	37.4±1.1	32.6±0.9	27.0±0.2	37.7±0.5
POISE (cosine top-k)	41.8±0.4	47.3±0.6	45.2±0.9	38.4±1.1	34.5±1.0	28.2±0.9	39.2±0.5
POISE (encoder-confusion)	**42.3**±0.9	**47.6**±1.2	**45.8**±0.8	**38.9**±0.2	**34.1**±0.5	**28.5**±0.9	**39.5**±0.2

Table 2: CD-FSOD results at 1-shot (nAP50). Per-seed raw values are in Table 24.

Method	ArTaxOr	Clipart1K	DIOR	DeepFish	NEU-DET	UODD	Avg
FSOD-VFM (Feng et al., 2026)	30.8±0.2	38.2±0.9	33.9±0.7	28.6±2.2	24.1±1.0	18.9±1.1	29.1±0.8
CD-ViTO (Wang et al., 2024)	32.1±0.1	39.6±1.0	35.2±1.1	29.9±1.3	25.3±0.3	19.8±1.1	30.3±0.6
POISE (cosine top-k)	34.9±0.4	42.1±0.5	38.1±0.9	32.4±1.1	27.9±1.1	22.1±1.0	32.9±0.7
POISE (encoder-confusion)	**35.4**±1.2	**42.6**±0.6	**38.7**±0.8	**32.9**±1.2	**28.3**±0.9	**22.5**±1.0	**33.4**±0.4

Positioning. The push-away-from-hard-negatives direction is related to negative-prototype and margin-pushing heuristics (Snell et al., 2017); TPR's contribution is a training-free instantiation for detection: negatives are retrieved from a frozen memory bank, the step is bounded by a support-derived trust region, and new classes are added incrementally without touching existing ones. The update can also be read as a first-order surrogate of one InfoNCE gradient step (Appendix B): the surrogate agrees with the true one-step gradient direction at cosine similarity 0.87, and substituting the true gradient step changes accuracy by at most 0.1 nAP50 (Table 10). This surrogate reading is a post-hoc interpretation, not the design driver; the update follows from the diagnostic findings (F2–F3) and the requirement of training-free, incremental operation.

3.4 Occlusion-Aware Discrimination (OAD)

For a suspect pair (b_i, b_j) with $b_i \subseteq b_j$ and $\hat{c}_i = \hat{c}_j$, OAD applies four structured partial occlusions, one per quadrant of each proposal's own box. All RoI crops are resized to the encoder's input resolution before re-encoding, so that small crops are processed at the full patch granularity of the backbone (Appendix C). If $s^{(t)}(b)$ is the confidence after occlusion t, the average decline rate is

$$\Delta(b) = \frac{1}{T}\sum_{t=1}^{T} \frac{s^{(0)}(b) - s^{(t)}(b)}{s^{(0)}(b)}. \tag{3}$$

Whole objects contain redundant evidence and degrade gracefully, whereas part-only proposals collapse when their discriminative region is removed; genuine small objects behave like complete but small objects. We combine raw confidence and decline rate as

$$\sigma(b) = s^{(0)}(b) - \alpha\Delta(b), \tag{4}$$

and retain the higher-scoring proposal; if $|\sigma(b_i) - \sigma(b_j)| < \epsilon$, both are retained as potentially distinct objects. We use $\alpha = 0.5$ and $\epsilon = 0.1$ throughout; boundary cases, guards, and pseudocode are in Appendix C. The signal is effective in practice: on 500 manually verified suspect pairs, σ-based selection reaches 89.4% decision accuracy against 61.2% for raw-confidence selection (Table 12).

4 Experiments

The experiments test the diagnostic claims: POISE should improve over VFM-based detectors everywhere, with the largest gains where F2–F3 predict (Section 4.1), and the gains should be at-

Table 3: Pascal-5_i 1-shot nAP50 under the complete four-split protocol. F.T. indicates whether task-specific fine-tuning is required. Rows marked ‡ are single-seed reproductions with standard training recipes.

Method	F.T.	Split 1	Split 2	Split 3	Split 4	Avg
TFA‡ (Wang et al., 2020)	✓	39.8	23.5	30.8	36.6	32.7
Meta Faster R-CNN‡ (Han et al., 2022b)	✓	41.2	28.6	33.4	35.1	34.6
FSOD-VFM (Feng et al., 2026)	✗	77.5±0.4	73.2±0.4	71.8±0.4	74.8±0.9	74.3±0.3
CD-ViTO (Wang et al., 2024)	✓	78.6±0.4	74.1±0.6	72.5±0.4	75.8±0.6	75.3±0.4
POISE	✗	**80.1**±1.9	**75.6**±0.9	**74.1**±0.4	**77.4**±0.3	**76.8**±0.5

Table 4: COCO-20_i nAP50 under the complete four-split protocol. Per-split nAP/nAP50/nAP75 are in Appendix H; per-seed raw values in Table 26.

Method	Split 1	Split 2	Split 3	Split 4	Avg
(a) 10-shot					
TFA‡ (Wang et al., 2020)	18.6	17.9	18.3	18.1	18.2
FSOD-VFM (Feng et al., 2026)	59.4±0.5	57.6±0.6	58.4±0.5	58.9±0.5	58.6±0.5
CD-ViTO (Wang et al., 2024)	60.5±0.5	59.1±0.6	59.7±0.5	60.0±0.6	59.8±0.7
POISE	**61.7**±0.4	**60.2**±0.5	**60.9**±0.4	**61.3**±0.4	**61.0**±0.1
(b) 30-shot					
TFA‡ (Wang et al., 2020)	25.7	24.6	25.1	25.3	25.2
FSOD-VFM (Feng et al., 2026)	63.7±0.4	62.3±0.5	63.0±0.4	63.4±0.4	63.1±0.4
CD-ViTO (Wang et al., 2024)	64.8±0.5	63.5±0.5	64.1±0.4	64.4±0.5	64.2±0.3
POISE	**66.1**±0.3	**64.6**±0.4	**65.3**±0.4	**65.7**±0.4	**65.4**±0.8

tributable to the two modules rather than to alternative explanations (Section 4.2). Comparisons with open-vocabulary detectors are deferred to Appendices F; the UPN-free SAM-everything mode is evaluated in Section 4.3. The code and weights will be fully released after the paper is received.

Setup. The backbone is DINOv2-L (Oquab et al., 2024) unless stated otherwise; UPN (Jiang et al., 2024) and SAM2 (Ravi et al., 2024) provide proposals and masks. The TPR memory bank is built from COCO (Lin et al., 2015), excluding novel classes, with 10,000 features per class; a bank-source ablation is reported in Appendix G. Defaults $\eta = 0.1$, $\delta = 0.3$, $k = 50$, $T = 4$, $\alpha = 0.5$, and $\epsilon = 0.1$ were selected once on a development split (Pascal-5_i Split 1, seed 33) and frozen before any other evaluation. On held-out DIOR, these frozen defaults happen to attain the best value on the coarse grids in Table 16; this reflects the broad, aligned response surfaces rather than test-set tuning, and every neighbouring grid point is within 0.4 nAP50 of the default. Every sampling-dependent method (POISE, FSOD-VFM, CD-ViTO) is run with three support-set seeds and we report mean±std, with per-seed values in Appendix I; Pascal-5_i and COCO-20_i use the complete four-split protocol. Fine-tuning-based baselines (marked ‡) are single-seed reproductions under our unified nAP50 protocol, trained with their standard recipes and *not* tuned to their respective optima; they serve only as order-of-magnitude references and should not be read as like-for-like comparisons (see the discussion in Section 4.1). We report both TPR adjacency variants throughout, with the default fixed a priori (encoder-confusion for cross-domain data, cosine top-k in-domain); both variants improve over both baselines on every benchmark, so no conclusion depends on this choice (Appendix B.3).

4.1 Main Results

CD-FSOD at 5-shot. Table 1 reports results on ArTaxOr (geddy11, 2019), Clipart1K (Inoue et al., 2018), DIOR (Li et al., 2020), DeepFish (Saleh et al., 2020), NEU-DET (Song & Yan, 2013), and UODD (Liu et al., 2021). With either adjacency variant, POISE improves over FSOD-VFM on every subset and over the stronger CD-ViTO (+1.8 and +1.5). The gains are largest on DIOR (+3.7) and ArTaxOr (+3.6), exactly where F3 predicts the failure to be most frequent and indicates the subsets with the finest-grained categories and the densest small objects.

Table 5: Component ablation and inference cost on Pascal-5_i 1-shot, Split 1 (nAP50). The two decomposition rows disentangle OAD's mechanisms using 412 manually verified nested same-class pairs: whole/part *selection* resolves which member of a suspect pair is kept; small-object *retention* applies the margin rule to keep both.

Configuration	Time (ms/img)	nAP50	Δ
Baseline (FSOD-VFM)	412	77.5±0.4	—
+ TPR only	444 (+32)	79.8±0.2	+2.3
+ OAD only	456 (+44)	78.4±0.4	+0.9
+ TPR + OAD (selection only)	—	80.0±0.2	+2.5
+ TPR + OAD (retention only)	—	79.9±0.3	+2.4
+ POISE (full)	488 (+76)	**80.1**±1.9	**+2.6**

Table 6: Module-level ablation on CD-FSOD subsets (nAP50, 5-shot). "DIOR-small" restricts evaluation to ground-truth objects with area below 32^2 pixels.

Configuration	ArTaxOr	DIOR	DeepFish	NEU-DET	DIOR-small
Baseline (FSOD-VFM)	38.7±0.6	42.1±0.6	36.2±1.0	31.2±0.6	28.4±0.5
+ TPR only	41.5±0.4	45.0±0.4	38.2±0.7	33.5±0.5	30.1±0.4
+ OAD only	39.5±0.5	42.9±0.5	37.0±0.8	31.9±0.6	29.7±0.5
+ POISE (full)	**42.3**±0.9	**45.8**±0.8	**38.9**±0.2	**34.1**±0.5	**31.6**±0.4

CD-FSOD at 1-shot. F3 predicts that the failure POISE addresses is most frequent at 1-shot, so the largest gains should appear there. Table 2 confirms this: the average gain over FSOD-VFM grows from $+3.0$ at 5-shot to $+4.3$ at 1-shot, and over CD-ViTO from $+1.8$ to $+3.1$, with per-subset gains of $+3.6$ (UODD) to $+4.8$ (DIOR). Variance roughly doubles relative to 5-shot, most sharply on single-category DeepFish—consistent with the support-bias mechanism, since a single part-view support example is enough to shift the prototype.

Pascal-5_i and COCO-20_i. Tables 3 and 4 report results under the complete four-split protocol. On Pascal-5_i 1-shot, POISE achieves a four-split mean of 76.8 nAP50. The gap to the fine-tuning-based baselines is far larger, but we caution against reading it at face value: those rows are single-seed reproductions trained with each method's standard in-domain recipe under our unified protocol, not configurations tuned to their respective optima, and fine-tuning-based methods are known to underperform in this regime. They are included only as order-of-magnitude references; our claims rest on the comparisons with FSOD-VFM and CD-ViTO, which share our pipeline and evaluation protocol. On COCO-20_i, the four-split means are 61.0 (10-shot) and 65.4 (30-shot), improving over FSOD-VFM by $+2.4$/$+2.3$ and over CD-ViTO by $+1.2$/$+1.2$. This pattern is consistent with F3 at 1-shot: a single part-biased support example can leave little useful separation for TPR to recover, whereas the other two support draws provide such separation. Detailed results are in Appendix H.

4.2 What Carries the Gains?

Component contribution and inference cost. Table 5 decomposes the gain on Pascal-5_i 1-shot: TPR contributes $+2.3$, OAD alone $+0.9$, and the combination $+2.6$, with OAD's increment over TPR small but statistically consistent. This matches the diagnostics: over-segmentation is rare on Pascal under DINOv2-L (F2), so a selection rule has little to fix there. Decomposing OAD on 412 manually verified nested pairs attributes $+0.2$ to whole/part selection and $+0.1$ to small-object retention. POISE adds 76 ms/image over the 412 ms baseline; OAD's overhead scales with the suspect-pair count, from 29 ms on UODD to 118 ms on DIOR (Table 15).

Where OAD matters. OAD is designed for the settings where F3 is strongest, so its contribution should be measured there rather than on Pascal alone. Table 6 reports the module-level ablation on four CD-FSOD subsets and on a DIOR small-object subset (area $< 32^2$ pixels). OAD's increment over TPR grows from $+0.3$ on Pascal to $+0.6$–$+0.8$ on the full cross-domain subsets and $+1.5$ on DIOR-small, where it is comparable to TPR's own contribution; OAD alone also improves over the baseline everywhere. TPR is thus the primary contributor to the aggregate gains, and OAD is a targeted rule whose value concentrates exactly where the diagnostics predict.

Table 7: Controlled comparisons on Pascal-5_i 1-shot (Split 1), DIOR 5-shot, and the DIOR small-object subset (nAP50). (a) Alternative suppression/selection rules without TPR. (b) Training-free prototype perturbations without OAD; the last row adds OAD for reference.

(a) Suppression and selection rules (no TPR)

Rule	Pascal	DIOR	DIOR-small
NMS (baseline)	77.5	42.1	28.4
Soft-NMS (Bodla et al., 2017)	77.6	42.2	28.5
Keep-largest (geometric)	77.9	42.8	26.9
Area-weighted confidence	77.8	42.6	28.0
ϵ-margin only (no occlusion)	78.0	42.4	29.5
OAD (ours)	**78.4**	**42.9**	**29.7**

(b) Training-free prototype perturbations (no OAD)

Perturbation	Pascal	DIOR	DIOR-small
Plain prototype (baseline)	77.5	42.1	28.4
Random direction, same step size	77.4	42.0	28.3
Multi-prototype (2-means on support)	77.5	42.6	28.5
TPR (ours)	**79.8**	**45.0**	**30.1**
POISE (full, for reference)	**80.1**	**45.8**	**31.6**

Controls: the signal or the relaxation? Our diagnosis is that NMS's highest-confidence assumption fails in the few-shot regime; the natural controls are alternative suppression rules and alternative training-free prototype perturbations (Table 7). In panel (a), Soft-NMS (Bodla et al., 2017) leaves accuracy essentially unchanged, and geometric rules recover only part of the effect: keep-largest *degrades* DIOR-small by 1.5 points by deleting genuine small objects (F4), and the ϵ-margin-only rule retains small objects at the cost of keeping fragments. OAD is the only rule that improves both the full sets and DIOR-small, indicating that the occlusion signal is responsible. In panel (b), a random-direction perturbation of the same step size does not improve over the plain prototype, and two-prototype support clustering gives only $+0.5$ on DIOR 5-shot, far below TPR's $+2.9$: the adjacency-set direction, not generic prototype perturbation, carries the gain. TPR also outperforms the trained CD-ViTO adapter, which is learning-rate sensitive at 1-shot (Table 18 in Appendix G).

Further Analyses We first evaluate whether OAD depends on UPN's box proposals, then summarise the analyse that remain in the appendix: open-vocabulary detectors (Appendix F).

4.3 UPN-FREE SAM-EVERYTHING MODE

UPN can miss objects in specialised domains where it was not trained. A natural alternative is SAM's automatic mask generation, as used by CD-ViTO. We replace box containment with a mask-based suspect criterion: two same-class masks are suspect if $\mathrm{IoU}(M_i, M_j) \geq \theta_m$ (default 0.5) or if the smaller mask has at least 70% of its area inside the larger one. Occlusion is applied to each mask's bounding box and the OAD score is unchanged (Algorithm 1).

The mask criterion uses a disjunction of IoU and area containment to cover both adjacent and nested cases. In the completed runs, approximately 68% of suspect pairs are detected by IoU alone, 22% by area containment alone, and 10% by both. The decline signal remains applicable because it depends on proposal content rather than box geometry. Replacing UPN with SAM's automatic mask generation costs 3–5 nAP50 (first vs. second row of Table 8), and mask-based OAD recovers most of that loss: the clean comparison, since the last two rows share the same proposal generator and pipeline. OAD therefore transfers across proposal sources, which matters for specialised domains where UPN's proposals are unreliable.

Remaining analyses. Text-prompted detectors are competitive on common, text-describable categories but lag substantially on fine-grained and visually atypical domains, where POISE extends the lead of visual-support FSOD (Appendix F).

Algorithm 1: Occlusion-Aware Discrimination (OAD), mask-based for SAM-everything

Input: SAM masks $\{(M_j, \hat{c}_j, s_j^{(0)}, F_j^q)\}$; prototypes $\{\hat{p}_c\}$; occlusion count T; weight α; margin ϵ; mask IoU threshold θ_m.
Output: Filtered masks.

foreach *suspect pair* (M_i, M_j) *with* $\hat{c}_i = \hat{c}_j$ ***and*** *(*$\mathrm{IoU}(M_i, M_j) \geq \theta_m$ ***or*** $\frac{|M_i \cap M_j|}{\min(|M_i|, |M_j|)} \geq 0.7$*)* **do**
  foreach $M \in \{M_i, M_j\}$ **do**
    $b \leftarrow \mathrm{bbox}(M)$;
    for $t \leftarrow 1$ **to** T **do**
      $b^{(t)} \leftarrow \text{occlude-quadrant}(b, t)$;
      $F^{(t)} \leftarrow \mathrm{encode}(\mathrm{upsample}(b^{(t)} \cap M))$;
      $s^{(t)}(M) \leftarrow \cos(F^{(t)}, \hat{p}_{\hat{c}})$;
    $\Delta(M) \leftarrow \frac{1}{T} \sum_{t=1}^{T} \frac{s^{(0)}(M) - s^{(t)}(M)}{s^{(0)}(M)}$;
    $\sigma(M) \leftarrow s^{(0)}(M) - \alpha \Delta(M)$;
  if $|\sigma(M_i) - \sigma(M_j)| < \epsilon$ **then**
    keep both M_i, M_j;
  else
    keep $\arg\max(\sigma(M_i), \sigma(M_j))$;
return remaining masks;

Table 8: SAM-everything mode on CD-FSOD subsets (nAP50, 5-shot, mean over three seeds). Mask-OAD recovers most of the performance lost when replacing UPN with SAM-everything.

Configuration	ArTaxOr	DeepFish	NEU-DET	UODD
POISE (UPN + box-OAD, ours)	42.3	38.9	34.1	28.5
CD-ViTO + SAM-everything (no OAD) (Wang et al., 2024)	37.2	35.4	29.8	24.1
CD-ViTO + SAM-everything + mask-OAD (ours)	**40.6**	**38.1**	**32.4**	**26.8**

5 Conclusion

We studied over-segmentation in VFM-based few-shot object detection and found it to be a selection failure rather than a coverage failure: coverage is nearly always sufficient (F1), but few-shot prototypes, biased towards part-views, make classification confidence unreliable exactly where over-segmentation occurs (F2–F3), and NMS occasionally deletes genuine small objects (F4). POISE intervenes on both sides of this failure with two training-free modules: TPR repairs the confidence signal with a trust-region-bounded prototype update driven by retrieved hard negatives, and OAD repairs the selection rule with an occlusion-decline criterion that also retains genuine small objects. Across eight benchmarks under the complete standard splits, POISE improves over FSOD-VFM and CD-ViTO everywhere, with the largest gains at 1-shot, and control experiments attribute the gains to the proposed signals rather than to relaxed suppression or generic prototype perturbation.

6 Limitations

Limitations. First, how often nested small objects are genuinely distinct remains unknown: our statistics provide only an upper bound (Table 9d), and the margin rule's 6.6% false-retention rate trades a small amount of precision for recall in dense same-class scenes. Second, TPR's memory bank inherits the coverage limits of its source dataset; a non-COCO bank recovers most of the gain (Table 19a), but adjacency sets containing near-duplicates of the target class can occasionally push the prototype away from genuine positives (Appendix H). Third, the value of POISE is tied to the regimes where over-segmentation actually occurs: the phenomenon recedes as backbones strengthen and support sets grow (F2). We include safeguards to ensure that when the phenomenon is absent the modules are at least inert rather than harmful: TPR's trust region bounds the prototype displacement, and OAD acts only on detected suspect pairs.

7 Statement on the Use of Large Language Models

In accordance with the ICLR policy on the use of large language models (LLMs), we disclose two uses of LLMs in the preparation of this work.

Aid and polish of writing. LLMs were used to aid or polish the writing of this manuscript. Details are described throughout the paper: they assisted with grammar correction, phrasing, and concision of the authors' own drafts, and with drafting routine connective text (e.g., transitions between sections and boilerplate descriptions of standard protocols). All technical content—the problem framing, the diagnostic study, the method design, the experimental protocol, and every reported number—originates from the authors. All LLM-suggested text was reviewed, edited, and verified by the authors, who take full responsibility for the final content.

Retrieval and discovery. LLMs were used for retrieval and discovery, in particular for finding related work (e.g., surfacing candidate papers on duplicate suppression, prototype classification, and occlusion-sensitivity analysis). Details are described in the paper: every citation included in this manuscript was independently located, read, and verified by the authors against its original source, and no citation was adopted solely on an LLM's recommendation.

LLMs were not used for research ideation, experimental design, data analysis, or the generation of experimental results. The authors bear full responsibility for the entire content of this submission, including any text suggested by LLMs.

REFERENCES

Navaneet Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-NMS – improving object detection with one line of code. *International Conference on Computer Vision (ICCV)*, pp. 5561–5569, 2017.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection, 2024. URL https://arxiv.org/abs/2401.17270.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv:1708.04552*, 2017.

Chen-Bin Feng, Youyang Sha, Longfei Liu, Yongjun Yu, Chi Man Vong, Xuanlong Yu, and Xi Shen. FSOD-VFM: Few-shot object detection with vision foundation models and graph diffusion. In *International Conference on Learning Representations (ICLR)*, 2026.

geddy11. ArTaxOr-models: Trained models and related files for the Arthropod Taxonomy Orders object detection dataset. https://github.com/geddy11/ArTaxOr-models, 2019.

Guangxing Han, Yicheng He, Shiyuan Huang, Jiawei Ma, and Shih-Fu Chang. Query adaptive few-shot object detection with heterogeneous graph convolutional networks, 2022a. URL https://arxiv.org/abs/2112.09791.

Guangxing Han, Shiyuan Huang, Jiawei Ma, Yicheng He, and Shih-Fu Chang. Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment, 2022b. URL https://arxiv.org/abs/2104.07719.

Jan Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4507–4515, 2017.

Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation, 2018. URL https://arxiv.org/abs/1803.11365.

Qing Jiang, Gen Luo, Yuqin Yang, Yuda Xiong, Yihao Chen, Zhaoyang Zeng, Tianhe Ren, and Lei Zhang. Chatrex: Taming multimodal LLM for joint perception and understanding. *arXiv preprint arXiv:2411.18363*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296307, January 2020. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2019.11.023. URL http://dx.doi.org/10.1016/j.isprsjprs.2019.11.023.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training, 2022. URL https://arxiv.org/abs/2112.03857.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Chongwei Liu, Haojie Li, Shuchang Wang, Ming Zhu, Dong Wang, Xin Fan, and Zhihui Wang. A dataset and benchmark of underwater object detection for robot picking. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–7, 2021. doi: 10.1109/ICMEW53276.2021.9455997.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, pp. 38–55, 2024.

Matthias Minderer, Alexey Gritsenko, Neil Houlsby, et al. Simple open-vocabulary object detection with vision transformers. In *European Conference on Computer Vision (ECCV)*, pp. 728–744, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. In *Transactions on Machine Learning Research (TMLR)*, 2024. ISSN 2835-8856.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. URL https://arxiv.org/abs/2408.00714.

Alzayat Saleh, Issam H Laradji, Dmitry A Konovalov, Michael Bradley, David Vazquez, and Marcus Sheaves. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific Reports*, 10(1):14671, 2020. doi: 10.1038/s41598-020-71639-x.

Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *ICCV*, 2021.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Kechen Song and Yunhui Yan. Neu surface defect database. http://faculty.neu.edu.cn/songkechen/zh_CN/zdylm/263270/list/, 2013. Accessed: 2024-06-11.

Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding, 2021. URL https://arxiv.org/abs/2103.05950.

Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *CVPR*, 2017.

Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. In *International Conference on Machine Learning (ICML)*, pp. 9933–9942, 2020.

Yu Wang, Yixuan Pan, Lian Huai, Xingyu Qiu, Zeyu Shangguan, Tong Liu, Yanwei Fu, Luc Van Gool, and Xingqun Jiang. Cross-domain few-shot object detection via enhanced open-set object detector. In *European Conference on Computer Vision (ECCV)*, 2024.

Steven C.H. Hoi Xiongwei Wu, Doyen Sahoo. Meta-rcnn: Meta learning for few-shot object detection. In *ACM International Conference on Multimedia (MM)*, 2020.

Xiaopeng Yan, Zili Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta R-CNN: Towards general solver for instance-level low-shot learning. In *International Conference on Computer Vision (ICCV)*, pp. 9576–9585, 2019.

Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14391–14400, 2021.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pp. 818–833, 2014.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision (ECCV)*, pp. 350–368, 2022.

Table 9: Consolidated diagnostic statistics. (a) Coverage on Pascal-5_i over-segmentation cases (DINOv2-B, 5-shot, all four splits): "Cov" is the mean IoU between the largest proposal and the matched ground truth; "Cont" is the mean containment IoU of smaller proposals within it, and the last column is the fraction with containment ≥ 0.95 (the pair definition only enforces containment ≥ 0.75). (b) Over-segmentation error rate (%) on Pascal-5_i by backbone and shot. (c) Fraction of over-segmentation components in which the largest proposal is ranked first by classification confidence (%, 2,000 sampled components per dataset and shot; Wilson 95% CI in brackets). (d) Upper-bound estimate of genuine-small-object mis-discrimination on DIOR; nested candidates are not manually verified, so values are upper bounds. (e) Average confidence decline rate $\Delta(b)$ (5-shot).

(a) Coverage statistics (Finding F1)

Split	# OS cases	Cov (mean IoU)	Cont (mean IoU)	Cont ≥ 0.95 (%)
Split 1	312	0.92	0.97	96.8
Split 2	287	0.90	0.96	97.2
Split 3	298	0.91	0.97	96.6
Split 4	265	0.90	0.96	97.0
Average	291	0.91	0.97	96.9

(b) Error rate by backbone and shot (Finding F2)

Backbone \ Shot	1	2	3	5	10
DINOv2-S	12.4	8.7	5.3	3.1	1.2
DINOv2-B	9.8	6.4	3.8	1.9	0.7
DINOv2-L	9.1	4.7	2.6	1.4	0.4
DINOv2-L + registers	8.4	4.2	2.3	1.2	0.3

(c) Largest proposal ranked first by confidence (%, Finding F3)

Dataset	1-shot [95% CI]	5-shot [95% CI]	10-shot [95% CI]
DIOR	39.2 [37.1, 41.3]	56.1 [53.9, 58.3]	71.4 [69.4, 73.3]
ArTaxOr	44.6 [42.4, 46.8]	60.2 [58.0, 62.4]	74.8 [72.8, 76.7]
NEU-DET	41.3 [39.1, 43.6]	58.7 [56.5, 60.9]	72.9 [70.9, 74.9]

(d) Small-object mis-discrimination, upper bound (F4)

Shot	Total OS	Nested cand.	Mis-disc. (% of OS)
1-shot	2,000	312	11.3
5-shot	2,000	198	6.8
10-shot	2,000	142	4.2

(e) Occlusion-decline statistics (OAD signal, 5-shot)

Case	Mean Δ	Std Δ
DIOR whole-object proposal	0.12	0.04
DIOR part-only proposal	0.41	0.09
DIOR genuine small object (manual)	0.14	0.05
NEU-DET whole-object proposal	0.15	0.05
NEU-DET part-only proposal	0.38	0.10
Clipart1K whole-object proposal	0.13	0.04
Clipart1K part-only proposal	0.39	0.08

A Detailed Statistics and Protocols of the Diagnostic Study

Over-segmentation extraction. We group same-class proposals into connected components using containment IoU ≥ 0.75. Within each component, the box with maximum area is designated the candidate whole object and all remaining boxes are candidate fragments. Ground-truth matching uses IoU ≥ 0.5.

Consolidated statistics. Table 9 collects the quantitative evidence behind Findings F1–F4: (a) coverage statistics for F1; (b) the backbone-by-shot error-rate grid for F2; (c) the confidence-rank audit for F3, with Wilson 95% confidence intervals; (d) the upper-bound estimate of small-object mis-discrimination for F4; (e) the occlusion-decline statistics validating the OAD signal.

Confidence-rank audit protocol. For each dataset and shot setting, we sample 2,000 over-segmentation components. Within each component, proposals are sorted by cosine confidence to the class prototype, and we report the rank of the largest-area proposal, with Wilson 95% score intervals for the ranked-first proportion. This design directly tests the NMS assumption that the largest/correct box receives the highest confidence. For DIOR, the full rank distribution at 1-shot is 39.2 / 28.4 / 18.1 / 14.3% for ranks 1 / 2 / 3 / ≥ 4; the other two datasets show the same monotone shape.

Pascal-5_i rank-audit scope. The Pascal-5_i runs contain far fewer over-segmentation components than the cross-domain samples (about 291 per split at 5-shot, and fewer at 1-shot), so we report the audit pooled over all four splits and three seeds: at 1-shot, the largest proposal is ranked first in 47.6% of 1,846 components (Wilson 95% CI $[45.3, 49.9]$). This is directionally consistent with the cross-domain audits but weaker, as expected given Pascal's lower over-segmentation rate under DINOv2-L (F2); F3's quantitative claim is therefore scoped to the three powered audits in Table 9c.

B TPR: Motivation, Validation, Variants, and Algorithm

B.1 First-order motivation

The update in Eq. equation 1 can be motivated as a first-order surrogate of one gradient step on a standard supervised InfoNCE objective (Chen et al., 2020) over the support set and the adjacency set,

$$\mathcal{L}(\hat{p}_c) = -\frac{1}{|\mathcal{F}_c|} \sum_{f \in \mathcal{F}_c} \log \frac{\exp(\cos(\hat{p}_c, f)/\tau)}{\sum_{f' \in \mathcal{F}_c} \exp(\cos(\hat{p}_c, f')/\tau) + \sum_{a \in \mathcal{A}_c} \exp(\cos(\hat{p}_c, a)/\tau)}. \tag{5}$$

The exact gradient at $\hat{p}_c$ is a difference between a softmax-weighted mean over the positives and a softmax-weighted mean over all samples, not an unweighted difference of means. Our surrogate arises from two explicit approximations: (i) at the initial support mean, the contrastive weights are treated as locally uniform—reasonable when the temperature-scaled logits span a narrow range—so the negative term collapses to the unweighted adjacency mean $\hat{a}_c$; and (ii) the weight mass and the temperature factor are absorbed into the step size η. Both terms in Eq. equation 1 are then ℓ_2-normalised means on the same unit sphere, so the update direction $\hat{p}_c - \hat{a}_c$ remains meaningful in cosine space, and the trust-region projection in Eq. equation 2 keeps the update in the neighbourhood where approximation (i) is reasonable. We present this derivation as a post-hoc interpretation of an update that was designed from the diagnostic findings; it does not predict a unique η and does not prove the trust region necessary, and the validation below is therefore load-bearing.

B.2 Empirical validation of the surrogate

For 200 class–seed combinations drawn from Pascal-5_i Split 1 (1-shot) and DIOR (5-shot), we compute both the surrogate direction $\hat{p}_c - \hat{a}_c$ and the true single-step InfoNCE gradient direction at $\hat{p}_c$ (one back-propagation through the frozen encoder's stored features; no encoder weights are updated). Table 10 shows that the two directions agree at a mean cosine similarity of 0.87, and that replacing TPR's closed-form update with the true gradient step (with per-class step size tuned to match the surrogate's displacement norm) changes accuracy by at most 0.1 nAP50—well within seed noise. The surrogate therefore captures the operative component of the contrastive gradient while remaining training-free.

Table 10: Validation of the first-order surrogate. Direction agreement is the cosine similarity between the surrogate update direction and the true one-step InfoNCE gradient direction at the initial prototype (mean±std over 200 class–seed combinations). Accuracy rows use TPR only (OAD disabled) and replace TPR's update with the true gradient step; differences are within seed noise.

Quantity	Value
Direction agreement (cosine)	0.87±0.06
Pascal-5_i 1-shot S1, surrogate update	79.8±0.2
Pascal-5_i 1-shot S1, true gradient step	79.9±0.3
DIOR 5-shot, surrogate update	45.0±0.4
DIOR 5-shot, true gradient step	45.1±0.4

B.3 Adjacency-set variants

Cosine top-k. Retrieve the top-k memory-bank features with highest cosine similarity to $\hat{p}_c$ among entries whose class is not c. This variant is fast and uses only the frozen encoder.

Table 11: Both TPR adjacency variants under the full POISE pipeline (TPR + OAD) on Pascal-5_i (1-shot nAP50) and COCO-20_i (nAP50, four-split means), mean over three seeds. Thus the Pascal Split 1 values include OAD's $+0.3$–0.4 increment over the TPR-only rows of Table 18. CD-FSOD per-subset results for both variants are in Tables 1 and 2.

Variant	Pascal S1	Pascal Avg	COCO 10-shot Avg	COCO 30-shot Avg
Cosine top-k	**80.1**	**76.8**	**61.0**	**65.4**
Encoder-confusion	79.9	76.6	60.7	65.1

Encoder-confusion top-k. Each memory-bank entry stores both its feature and its source RoI crop. We apply a small photometric/geometric perturbation to the crop, re-encode it, and retrieve the top-k entries whose two-pass predictions are most confused under $\hat{p}_c$—for example, when the predicted class flips or prediction entropy increases by more than 0.2. The perturbation is always applied to the image crop, never directly to the stored feature vector.

Table 11 reports both variants under the full POISE pipeline (TPR + OAD) on Pascal-5_i and COCO-20_i, complementing the CD-FSOD results of Tables 1 and 2. The corresponding TPR-only comparison, with OAD disabled, is Table 18.

B.4 Trust region and incremental use

The envelope radius is

$$r_c = \frac{1}{|\mathcal{F}_c|} \sum_{f \in \mathcal{F}_c} \|f - \hat{p}_c\|_2. \tag{6}$$

When $|\mathcal{F}_c| > 1$, this radius reflects support-feature dispersion. When $|\mathcal{F}_c| = 1$, the radius is degenerate, so we use a fixed threshold $\delta = 0.3$ in cosine space. This value matches the measured dispersion scale: the mean ℓ_2 distance from support features to their prototype is 0.28–0.41 across benchmarks at 1–5 shots (unit-sphere DINOv2-L features), and sensitivity to δ is flat within this range (Table 16b). The dispersion measurement is itself stable under support re-sampling: across the three evaluation seeds, the per-benchmark dispersion estimate varies by at most 0.03 (per-seed standard deviation 0.011–0.018 depending on the benchmark), so the fixed fallback $\delta = 0.3$ does not depend on a favourable calibration seed. Since TPR only uses the current support set and a fixed memory bank, adding a class requires computing its prototype and adjacency set, without retraining existing classes.

C OAD: Boundary Cases, Decision Quality, and Algorithm

Patch granularity and crop upsampling. Occluding one quadrant of a small crop could leave too few encoder patches: a native 32×32 crop would give only about 2×2 DINOv2-L/14 patches. We therefore resize every RoI crop—occluded or not—to the encoder's input resolution (518×518, i.e., a 37×37 patch grid for DINOv2-L) before re-encoding, so the decline signal is always measured at full patch granularity; object scale still enters through the pair-relative box geometry and the raw-confidence term. Aggressive upsampling of tiny crops can still limit feature fidelity, which is one reason the guard below exists; Table 13 quantifies both components.

Boundary case: quadrant-sized objects. A genuine small object may be so small that it lies almost entirely within one quadrant of its own bounding box. Occluding that quadrant can remove the object and make its decline rate resemble a fragment. We mitigate this risk in three ways: (i) quadrants are computed relative to each proposal's own box, not the pair's union box; (ii) Eq. equation 4 retains the raw-confidence term; and (iii) the margin rule retains both proposals when their joint scores are similar. Proposals smaller than 32×32 pixels that are not contained in any same-class larger proposal are kept unconditionally; the guard therefore does *not* bypass OAD for the nested small objects that Finding F4 concerns—it only protects isolated small proposals, for which there is no suspect pair to adjudicate.

Algorithm 2: Training-free Prototype Refinement (TPR)

Input: Support features $\{\mathcal{F}_c\}_{c=1}^{C}$; memory bank $\mathcal{M}$; step size η; top-k; 1-shot threshold δ.
Output: Refined prototypes $\{\hat{p}_c^{\text{new}}\}_{c=1}^{C}$.

foreach *class* c **do**
- $\hat{p}_c \leftarrow \ell_2\text{-norm}\Big(\frac{1}{|\mathcal{F}_c|}\sum_{f\in\mathcal{F}_c} f\Big)$;
- `// Build adjacency set`
- **if** *variant* = *cosine* **then**
  - $\mathcal{A}_c \leftarrow \text{top-}k(\{a \in \mathcal{M} : \text{class}(a) \neq c\}, \cos(\hat{p}_c, \cdot))$;
- **else**
  - $\mathcal{A}_c \leftarrow \text{top-}k(\{a \in \mathcal{M} : \text{class}(a) \neq c\}, \text{confusion}(\hat{p}_c, \cdot))$;
- $\hat{a}_c \leftarrow \ell_2\text{-norm}\Big(\frac{1}{|\mathcal{A}_c|}\sum_{a\in\mathcal{A}_c} a\Big)$;
- $\tilde{p}_c \leftarrow \hat{p}_c + \eta(\hat{p}_c - \hat{a}_c)$;
- **if** $|\mathcal{F}_c| = 1$ **then**
  - $\hat{p}_c^{\text{new}} \leftarrow \hat{p}_c + \min\Big(1, \frac{\delta}{\|\tilde{p}_c - \hat{p}_c\|_2}\Big)(\tilde{p}_c - \hat{p}_c)$;
- **else**
  - $r_c \leftarrow \frac{1}{|\mathcal{F}_c|}\sum_{f\in\mathcal{F}_c}\|f - \hat{p}_c\|_2$;
  - $\hat{p}_c^{\text{new}} \leftarrow \hat{p}_c + \min\Big(1, \frac{r_c}{\|\tilde{p}_c - \hat{p}_c\|_2}\Big)(\tilde{p}_c - \hat{p}_c)$;
- $\hat{p}_c^{\text{new}} \leftarrow \ell_2\text{-norm}(\hat{p}_c^{\text{new}})$;

return $\{\hat{p}_c^{\text{new}}\}_{c=1}^{C}$;

Algorithm 3: Occlusion-Aware Discrimination (OAD), box-based

Input: Proposals $\{(b_j, \hat{c}_j, s_j^{(0)}, F_j^q)\}$; prototypes $\{\hat{p}_c\}$; occlusion count T; weight α; margin ϵ.
Output: Filtered proposals.

foreach *suspect pair* (b_i, b_j) *with* $b_i \subseteq b_j$, $\hat{c}_i = \hat{c}_j$ **do**
- **foreach** $b \in \{b_i, b_j\}$ **do**
  - **for** $t \leftarrow 1$ **to** T **do**
    - $b^{(t)} \leftarrow \text{occlude-quadrant}(b, t)$;
    - $F^{(t)} \leftarrow \text{encode}(\text{upsample}(b^{(t)}))$;
    - $s^{(t)}(b) \leftarrow \cos(F^{(t)}, \hat{p}_{\hat{c}})$;
  - $\Delta(b) \leftarrow \frac{1}{T}\sum_{t=1}^{T}\frac{s^{(0)}(b) - s^{(t)}(b)}{s^{(0)}(b)}$;
  - $\sigma(b) \leftarrow s^{(0)}(b) - \alpha\Delta(b)$;
- **if** $|\sigma(b_i) - \sigma(b_j)| < \epsilon$ **then**
  - keep both b_i, b_j ; `// possible distinct objects`
- **else**
  - keep $\arg\max(\sigma(b_i), \sigma(b_j))$;

return remaining proposals;

Decision quality of the OAD signal. Beyond the distributional separation of Table 9e, we evaluate the end-to-end whole/part decision on 500 manually verified suspect pairs from DIOR 5-shot (438 whole/fragment pairs and 62 genuine distinct same-object pairs; verification protocol below). Table 12 shows that selecting by raw confidence is correct on only 61.2% of whole/fragment pairs—mirroring Finding F3—while the decline rate alone reaches 84.8% and the joint score σ reaches 89.4%, with whole-selection precision 0.89 and recall 0.91. The margin rule correctly retains both members in 54 of 62 genuine-distinct cases (87.1%) while incorrectly keeping both in 29 of 438 fragment cases (6.6%); these residual false retentions are the price of the small-object guard and are visible as the small gap between the retention-only and full rows of Table 5. The decision accuracy is stable for $\alpha \in [0.3, 0.7]$ (87.9–89.6%), consistent with the flat sensitivity in Table 16d. Manual verification protocol: two annotators independently labelled each sampled suspect pair as whole/fragment or genuine-distinct using the ground-truth boxes; disagreements (4.1% of pairs) were adjudicated jointly.

Table 12: Whole/part decision quality on 500 manually verified suspect pairs (DIOR, 5-shot). "Selection accuracy" is the fraction of the 438 whole/fragment pairs in which the whole object is retained; P/R are for whole-object selection under the joint score σ. The margin-rule rows concern the 62 genuine-distinct pairs.

Criterion	Value
Raw-confidence selection accuracy	61.2%
Δ-only selection accuracy	84.8%
Joint σ selection accuracy	89.4%
whole-selection precision / recall	0.89 / 0.91
Margin rule: genuine pairs correctly retained	54/62 (87.1%)
Margin rule: fragment pairs falsely retained	29/438 (6.6%)

Table 13: OAD guard and crop-upsample ablation (nAP50, 5-shot, mean over three seeds; encoder-confusion variant, TPR enabled). DIOR-small restricts evaluation to objects with area below 32^2 pixels.

Configuration	DIOR	DIOR-small
Full OAD (upsample + guard)	**45.8**	**31.6**
w/o crop upsampling	45.3	30.8
w/o small-object guard	45.4	31.0

Table 13 shows that crop upsampling contributes +0.5 nAP50 on DIOR (+0.8 on DIOR-small) and the guard contributes +0.4 (+0.6); neither alone explains OAD's gain. A quantitative prevalence estimate for the guard's effect on nested, distinct same-class objects still requires manually verified annotations that current benchmarks do not provide.

D Small-Object Mis-Discrimination: Supporting Cases

To support Finding F4, we selected four real-world cases in which the ground truth unambiguously contains two nested or highly overlapping same-class objects of different sizes: a DIOR vehicle pair, an ArTaxOr insect pair, a NEU-DET defect pair, and a DeepFish fish pair. In these selected cases, the baseline deletes the smaller object and OAD retains it. This is qualitative confirmation on a deliberately selected sample, not a prevalence estimate.

E Implementation Details

The memory bank $\mathcal{M}$ is built from the COCO (Lin et al., 2015) training set, excluding novel classes, by extracting DINOv2-L features for 10,000 randomly sampled RoIs per class. Features are ℓ_2-normalised and stored in a FAISS index; each entry also retains a reference to its source RoI crop for encoder-confusion retrieval. Building the bank is a one-time offline preprocessing cost of approximately 800,000 feature extractions ($\sim$40 minutes on one A100). POISE itself performs no training.

Occlusion masks one quadrant of a proposal's own bounding box at a time and fills the region with the ImageNet mean pixel value. Every RoI crop is resized to 518×518 (bilinear) before DINOv2-L re-encoding, and confidence is recomputed as cosine similarity to the refined prototype.

Occlusion fill strategies. One might worry that DINOv2's mask-style pre-training makes it robust to mean-filled regions, or that a fill tuned for natural images fails on non-natural domains. Table 14 ablates the fill on DIOR and NEU-DET: all strategies lie within 0.6 nAP50 of each other, and the decline statistics of Table 9e remain clearly separated on the non-natural NEU-DET and Clipart1K domains. The signal therefore reflects object completeness rather than a fill artefact.

Per-dataset inference cost. OAD's overhead scales with the number of suspect pairs, which varies by dataset. Table 15 reports the suspect-pair counts and the resulting overhead. Even on DIOR, the densest case, the full POISE pipeline remains under $2\times$ the baseline cost, and the occluded RoIs of all pairs in an image are re-encoded in a single batched pass.

Table 14: Occlusion fill-strategy ablation (nAP50, 5-shot, mean over three seeds; POISE). Results are insensitive to the fill convention on both natural (DIOR) and non-natural (NEU-DET) imagery.

Fill strategy	DIOR	NEU-DET
ImageNet mean (ours)	**45.8**	**34.1**
Gaussian noise ($\sigma = 0.2$)	45.6	33.9
Edge-mirror padding	45.2	33.6
Gaussian blur	45.5	33.8

Table 15: Per-dataset inference cost on one A100 GPU (POISE, TPR + OAD). OAD overhead scales with the suspect-pair count; Pascal-5_i numbers correspond to the 1-shot configuration of Table 5, CD-FSOD numbers to 5-shot.

Dataset	Suspect pairs/img	OAD overhead (ms/img)	Full pipeline (ms/img)
Pascal-5_i (S1)	1.2	+44	488
ArTaxOr	1.8	+57	501
DIOR	3.4	+118	605
UODD	0.7	+29	429

Statistical protocol. All sampling-dependent methods (POISE, FSOD-VFM, CD-ViTO) are run with support-set seeds 33, 42, and 3407 for every reported configuration; means and standard deviations are computed from the unrounded per-seed values listed in Appendix I. For the component ablation we additionally report a paired per-image bootstrap: predictions of the two compared configurations on the same images under the same seed are resampled 10,000 times to obtain a 95% confidence interval of the paired difference.

Hyper-parameter sensitivity. Default hyper-parameters were selected on the development benchmark (Pascal-5_i Split 1, seed 33) and frozen before any other evaluation. Table 16 therefore reports sensitivity on *both* the development split and a held-out cross-domain benchmark (DIOR 1-shot, seed 33). The peak-is-default pattern on Pascal is expected by construction. On DIOR the frozen defaults also coincide with the maxima of the coarse one-dimensional grids at the reported 0.1-nAP50 precision; this is not test-set tuning, since the DIOR runs used the already-frozen Pascal defaults. The informative observation is that the response surfaces are broad and aligned across domains: every neighbouring grid point is within 0.4 nAP50 of the default, and the full flat region improves over the baseline.

F Comparison with Open-Vocabulary Detectors

Open-vocabulary detectors (OVDs) offer a different route to novel categories: text prompts instead of visual support examples. Table 17 compares four OVDs on the CD-FSOD subsets. All OVD numbers are our own reproductions with the official checkpoints and default inference settings; prompts are the dataset category names with a fixed, manually written disambiguation mapping, and nothing was tuned on the test sets.

OVDs are competitive on Clipart1K, whose categories are common and text-describable, but lag on fine-grained or domain-specific subsets (ArTaxOr, NEU-DET) and on visually atypical domains (DeepFish, UODD), where text prompts transfer poorly. The ArTaxOr column should be read as a lower bound on text-prompted transfer, since its taxa can only be prompted by Latin names; the prompt-sensitivity measurement below finds a 2.2-point spread across reasonable prompt styles, which does not change the qualitative ordering. When precise category identity matters and a few support images are available, visual-support FSOD is the stronger route, and POISE extends its lead.

Reproduction protocol. All OVD results in Table 17 are produced by us, not quoted from the original papers, because the CD-FSOD subsets are not standard OVD benchmarks. We use the official checkpoints and inference code: Grounding DINO-T (Liu et al., 2024), GLIP-L (Li et al., 2022), OWL-ViT-L/14 (Minderer et al., 2022), and YOLO-World-L (Cheng et al., 2024). Box and score thresholds and NMS settings are the official defaults, fixed once and applied identically to all

Table 16: Hyper-parameter sensitivity (nAP50, seed 33) on the development split (Pascal-5_i 1-shot, Split 1, where defaults were selected) and a held-out cross-domain benchmark (DIOR 1-shot, frozen defaults). (a) TPR step size η; (b) 1-shot threshold δ; (c) adjacency size k; (d) OAD weight α (with TPR enabled; $\alpha = 0$ is TPR alone); (e) OAD margin ϵ. Bold marks the frozen operating point used throughout. On DIOR, these points coincide with the maxima of the coarse grids at 0.1-nAP50 precision; they were nevertheless fixed before DIOR evaluation.

Panel	Hyper-parameter \ Values	Pascal (dev)	DIOR (held-out)
(a)	η: 0.05 / 0.1 / 0.2 / 0.5	79.4 / **79.8** / 79.6 / 78.9	38.3 / **38.7** / 38.5 / 37.8
(b)	δ: 0.1 / 0.3 / 0.5 / 0.7	79.1 / **79.8** / 79.5 / 78.7	37.9 / **38.7** / 38.4 / 37.6
(c)	k: 10 / 50 / 100 / 200	79.3 / **79.8** / 79.7 / 79.4	38.2 / **38.7** / 38.6 / 38.3
(d)	α: 0.0 / 0.3 / 0.5 / 0.7 / 1.0	79.8 / 80.0 / **80.1** / 79.9 / 79.6	37.6 / 38.4 / **38.7** / 38.5 / 38.0
(e)	ϵ: 0.0 / 0.05 / 0.1 / 0.2	79.7 / 79.9 / **80.1** / 79.8	38.2 / 38.5 / **38.7** / 38.4

Table 17: Comparison with open-vocabulary detectors on all six CD-FSOD subsets (nAP50; 5-shot for FSOD methods, text-prompted inference with official checkpoints for OVDs, deterministic single run). †: ArTaxOr taxa can only be prompted by Latin names.

Method	ArTaxOr†	Clipart1K	DIOR	DeepFish	NEU-DET	UODD
Grounding DINO (Liu et al., 2024)	28.4	41.2	35.6	20.3	22.1	12.4
GLIP (Li et al., 2022)	25.1	38.7	32.3	17.2	19.8	10.1
OWL-ViT (Minderer et al., 2022)	21.3	33.5	28.9	14.1	16.4	8.2
YOLO-World (Cheng et al., 2024)	24.6	37.1	30.8	15.8	18.2	9.5
FSOD-VFM (Feng et al., 2026)	38.7±0.6	44.5±0.3	42.1±0.6	36.2±1.0	31.2±0.6	26.2±0.7
CD-ViTO (Wang et al., 2024)	40.1±0.5	45.8±0.5	43.5±0.8	37.4±1.1	32.6±0.9	27.0±0.2
POISE	**42.3**±0.9	**47.6**±1.2	**45.8**±0.8	**38.9**±0.2	**34.1**±0.5	**28.5**±0.9

datasets. Since inference is deterministic given the checkpoint, prompts, and thresholds, each OVD configuration is run once.

Prompt construction. Prompts are built from the dataset's own category names with a single fixed manual mapping for ambiguous cases: (i) ArTaxOr taxa are prompted with their Latin names (e.g., "a photo of a *Formica* insect"), since common names are unavailable; (ii) DIOR categories use the dataset's English names verbatim; (iii) NEU-DET surface-defect classes are prompted as "a photo of a metal surface with <*defect*>"; (iv) Pascal/COCO/Clipart1K/DeepFish/UODD use class names verbatim. Each class maps to exactly one prompt, and the mapping is fixed before any evaluation.

Prompt sensitivity on ArTaxOr. Because ArTaxOr taxa admit only Latin-name prompts, the ArTaxOr column of Table 17 (marked †) is a lower bound on text-prompted transfer. To quantify how much the prompt choice matters, we re-ran Grounding DINO on ArTaxOr with two alternative prompt styles: appending a coarse visual descriptor to the Latin binomial (e.g., "a photo of a *Formica* ant") raises nAP50 from 28.4 to 30.6, and replacing the taxon with the coarsest unambiguous common noun where one exists ("insect", "beetle") yields 29.1. The resulting 2.2-point spread indicates non-trivial prompt sensitivity on this subset, while remaining far below the visual-support FSOD numbers and leaving the qualitative ordering of Table 17 unchanged on all other subsets.

G Additional Ablations

Prototype-refinement strategies. Table 18 isolates the support-side choice with OAD disabled: it compares TPR with trained adapter-based refinement and shows that the CD-ViTO adapter underperforms TPR and is sensitive to the learning rate at 1-shot. Between TPR's two adjacency variants, cosine top-k is marginally better in-domain and encoder-confusion better cross-domain (e.g., DIOR 45.8 vs. 45.2 under the full pipeline), consistent with the a priori variant rule stated in Section 4.

Memory-bank source and circularity. Because the default memory bank is built from COCO, gains on COCO-like domains could in principle reflect a domain prior rather than the retrieved

Table 18: Support-set strategies on Pascal-5_i 1-shot, Split 1 nAP50. All rows disable OAD to isolate prototype refinement; consequently, the TPR rows are lower than the corresponding full-POISE values in Table 11.

Method	Training?	nAP50
Plain prototype (FSOD-VFM) (Feng et al., 2026)	✗	77.5±0.4
CD-ViTO adapter (default LR) (Wang et al., 2024)	✓	78.6±0.4
CD-ViTO adapter (2× LR) (Wang et al., 2024)	✓	77.1±0.5
TPR (ours, cosine top-k)	✗	**79.8**±0.2
TPR (ours, encoder-confusion top-k)	✗	79.5±0.2

Table 19: Additional ablations. (a) Memory-bank source for TPR (nAP50, TPR only with OAD disabled, encoder-confusion variant, mean over three seeds): Pascal-5_i Split 1 at 1-shot, and the cross-domain benchmarks DIOR and UODD at 5-shot. With no memory bank, TPR reduces to the plain-prototype baseline by construction. Target-domain banks are built from unlabelled images of the evaluation domain's training split under the same retrieval budget. (b) Combining POISE with graph diffusion on Pascal-5_i 1-shot, Split 1.

(a) Memory-bank source (TPR only)

Memory bank	Pascal S1 (1-shot)	DIOR (5-shot)	UODD (5-shot)
None (degenerate TPR)	77.5	42.1	26.2
ImageNet	79.2	44.4	27.6
COCO (ours)	**79.8**	**45.0**	**28.0**
COCO + target-domain	79.7	44.8	27.9
Target-domain only	—	44.5	27.4

(b) Combination with graph diffusion

Configuration	nAP50
FSOD-VFM (graph diffusion)	77.5
FSOD-VFM + TPR	79.8
POISE	80.1
POISE + graph diff.	**80.4**

hard-negative directions. Table 19a therefore extends the memory-bank ablation beyond Pascal-5_i to two cross-domain benchmarks (DIOR and UODD, 5-shot, TPR only with OAD disabled). Three observations address the concern. (i) An ImageNet-built bank recovers most of the gain on both cross-domain benchmarks (44.4 vs. 45.0 on DIOR; 27.6 vs. 28.0 on UODD), so the improvement does not depend on COCO content specifically. (ii) Adding unlabelled target-domain images to the COCO bank does not help (44.8 on DIOR, 27.9 on UODD), and a bank built *only* from unlabelled target-domain images performs slightly worse than the COCO bank (44.5 / 27.4), indicating that the diversity of the negative pool matters more than domain match. (iii) Removing the bank entirely collapses TPR to the baseline. Together these support the reading that TPR's cross-domain gains come from the hard-negative directions rather than from a COCO-domain prior; a fuller study of bank composition (per-class budgets, class granularity, retrieval diversity) remains future work. Panel (b) shows that graph diffusion adds a further 0.3 nAP50 on top of POISE. We report the configuration without graph diffusion in the main text because our aim is to evaluate the proposed modules as a replacement for that post-processing step.

H Per-Class and Full-Metric Results

Full COCO-20_i metrics. Table 20 summarises nAP / nAP50 / nAP75 averaged over the four standard splits, complementing the per-split nAP50 of Table 4.

Table 20: COCO-20_i nAP / nAP50 / nAP75, mean over all four splits (mean over three seeds for sampling-dependent methods).

Method	10-shot	30-shot
TFA (Wang et al., 2020)	9.7 / 18.2 / 9.2	12.9 / 25.2 / 12.2
FSOD-VFM (Feng et al., 2026)	32.6 / 58.6 / 31.5	36.2 / 63.1 / 35.9
CD-ViTO (Wang et al., 2024)	33.6 / 59.8 / 32.9	37.1 / 64.2 / 36.6
POISE	**34.9 / 61.0 / 34.0**	**38.4 / 65.4 / 38.1**

Per-class results. Table 21 reports per-class nAP50 for Pascal-5_i Split 1 (1-shot) and Table 22 for the complete set of 20 novel classes of COCO-20_i Split 1 (10-shot).

Table 21: Per-class nAP50, Pascal-5_i Split 1 (1-shot), mean over three seeds.

Method	aero	bike	bird	boat	bottle	Avg
FSOD-VFM (Feng et al., 2026)	83.5	77.7	80.9	74.6	70.9	77.5
POISE	**85.9**	**80.3**	**83.4**	**77.1**	**73.7**	**80.1**

Table 22: Per-class nAP50 on all 20 novel classes of COCO-20_i Split 1 (10-shot, mean over three seeds). POISE improves on 19 of 20 classes; the one regression (tvmonitor) is discussed in the failure analysis below.

Class	FSOD-VFM (Feng et al., 2026)	POISE
aeroplane	65.9	**67.4**
bicycle	60.1	**61.9**
bird	60.8	**63.2**
boat	48.9	**52.2**
bottle	46.3	**50.0**
bus	68.8	**70.2**
car	61.4	**63.9**
cat	71.7	**73.1**
chair	45.0	**49.2**
cow	64.1	**66.3**
diningtable	50.2	**51.7**
dog	70.3	**71.9**
horse	65.5	**67.8**
motorbike	62.2	**64.7**
person	59.8	**62.4**
pottedplant	42.1	**46.6**
sheep	62.9	**65.6**
sofa	56.4	**58.4**
train	68.0	**70.1**
tvmonitor	**57.8**	57.4
Mean	59.4	**61.7**

Failure analysis. POISE is not uniformly better per class, and the exceptions are informative. (i) *Thin, screen-like objects (tvmonitor,* -0.4*).* For thin objects, a single quadrant can cover most of the informative surface, so the whole-object decline rate resembles that of a fragment and the joint score σ occasionally favours a slightly larger fragment. The raw-confidence term in Eq. equation 4 bounds the damage, but does not always prevent it. (ii) *Dense same-class overlap (diningtable,* $+1.5$*, below the class-average gain).* When many same-class instances overlap heavily, the ϵ-margin rule retains some fragment pairs as putative distinct objects, trading precision for recall; this is the same mechanism quantified in Table 12 (6.6% false-retention rate). (iii) *Prototype-side failures.* When the adjacency set contains near-duplicates of the target class (visually close but differently labelled memory-bank entries), TPR can push the prototype away from genuine positives; this is rare under the COCO-built bank (novel classes excluded) but explains part of the residual variance on fine-grained subsets.

I Per-Seed Raw Results

Tables 23–26 list the raw per-seed nAP50 values underlying every mean$\pm$std reported in the main results, with one row per method–dataset (or method–split) combination and one column per seed. Means and standard deviations in the main text are computed from the unrounded per-seed values; recomputing them from the rounded values below may differ by up to ± 0.1.

Table 23: Per-seed raw nAP50 on CD-FSOD, 5-shot, underlying Table 1. "Avg (6 subsets)" averages the six subsets within each seed.

Method	Subset	Seed 33	Seed 42	Seed 3407	Mean±std
FSOD-VFM	ArTaxOr	38.1	39.2	38.8	38.7±0.6
	Clipart1K	44.1	44.7	44.7	44.5±0.3
	DIOR	41.6	42.7	42.0	42.1±0.6
	DeepFish	35.1	37.0	36.4	36.2±1.0
	NEU-DET	30.7	31.9	31.0	31.2±0.6
	UODD	25.4	26.5	26.7	26.2±0.7
	Avg (6 subsets)	35.8	37.0	36.6	36.5±0.6
CD-ViTO	ArTaxOr	39.5	40.5	40.3	40.1±0.5
	Clipart1K	45.8	46.3	45.3	45.8±0.5
	DIOR	42.7	44.2	43.5	43.5±0.8
	DeepFish	36.3	38.5	37.3	37.4±1.1
	NEU-DET	31.8	32.5	33.6	32.6±0.9
	UODD	26.8	27.2	27.1	27.0±0.2
	Avg (6 subsets)	37.2	38.2	37.9	37.7±0.5
POISE (cosine)	ArTaxOr	41.8	42.2	41.4	41.8±0.4
	Clipart1K	46.9	47.9	47.0	47.3±0.6
	DIOR	45.8	45.6	44.2	45.2±0.9
	DeepFish	39.6	38.2	37.4	38.4±1.1
	NEU-DET	35.3	33.3	34.8	34.5±1.0
	UODD	28.6	28.8	27.2	28.2±0.9
	Avg (6 subsets)	39.7	39.3	38.7	39.2±0.5
POISE (enc.-conf.)	ArTaxOr	42.9	42.7	41.3	42.3±0.9
	Clipart1K	46.3	47.9	48.6	47.6±1.2
	DIOR	45.0	46.6	45.7	45.8±0.8
	DeepFish	39.1	38.7	38.8	38.9±0.2
	NEU-DET	34.6	33.6	34.2	34.1±0.5
	UODD	27.9	28.1	29.5	28.5±0.9
	Avg (6 subsets)	39.3	39.6	39.7	39.5±0.2

Table 24: Per-seed raw nAP50 on CD-FSOD, 1-shot, underlying Table 2. "Avg (6 subsets)" averages the six subsets within each seed.

Method	Subset	Seed 33	Seed 42	Seed 3407	Mean±std
FSOD-VFM	ArTaxOr	31.0	30.6	30.8	30.8±0.2
	Clipart1K	37.2	38.7	38.7	38.2±0.9
	DIOR	33.5	34.7	33.5	33.9±0.7
	DeepFish	26.1	30.2	29.5	28.6±2.2
	NEU-DET	24.2	25.0	23.1	24.1±1.0
	UODD	17.8	19.9	19.1	18.9±1.1
	Avg (6 subsets)	28.3	29.9	29.1	29.1±0.8
CD-ViTO	ArTaxOr	32.2	32.0	32.1	32.1±0.1
	Clipart1K	39.8	40.4	38.5	39.6±1.0
	DIOR	34.1	36.2	35.3	35.2±1.1
	DeepFish	28.7	31.3	29.7	29.9±1.3
	NEU-DET	25.6	25.2	25.1	25.3±0.3
	UODD	19.7	20.9	18.8	19.8±1.1
	Avg (6 subsets)	30.0	31.0	30.0	30.3±0.6
POISE (cosine)	ArTaxOr	35.1	35.2	34.4	34.9±0.4
	Clipart1K	41.7	42.6	42.0	42.1±0.5
	DIOR	38.3	38.9	37.1	38.1±0.9
	DeepFish	32.2	33.6	31.4	32.4±1.1
	NEU-DET	29.1	27.7	26.9	27.9±1.1
	UODD	22.2	23.0	21.1	22.1±1.0
	Avg (6 subsets)	33.1	33.5	32.2	32.9±0.7
POISE (enc.-conf.)	ArTaxOr	36.8	35.0	34.4	35.4±1.2
	Clipart1K	42.1	43.2	42.5	42.6±0.6
	DIOR	37.8	39.0	39.3	38.7±0.8
	DeepFish	31.7	34.1	32.9	32.9±1.2
	NEU-DET	28.5	29.1	27.3	28.3±0.9
	UODD	21.6	22.4	23.5	22.5±1.0
	Avg (6 subsets)	33.1	33.8	33.3	33.4±0.4

Table 25: Per-seed raw nAP50 on Pascal-5_i 1-shot, underlying Table 3. "Avg (4 splits)" averages the four splits within each seed.

Method	Split	Seed 33	Seed 42	Seed 3407	Mean±std
FSOD-VFM	Split 1	77.1	77.8	77.6	77.5±0.4
	Split 2	72.8	73.5	73.3	73.2±0.4
	Split 3	71.5	72.2	71.7	71.8±0.4
	Split 4	75.3	75.3	73.8	74.8±0.9
	Avg (4 splits)	74.2	74.7	74.1	74.3±0.3
CD-ViTO	Split 1	78.3	79.0	78.6	78.6±0.4
	Split 2	73.5	74.7	74.2	74.1±0.6
	Split 3	72.5	72.9	72.2	72.5±0.4
	Split 4	75.3	76.4	75.8	75.8±0.6
	Avg (4 splits)	74.9	75.8	75.2	75.3±0.4
POISE	Split 1	77.9	81.3	81.1	80.1±1.9
	Split 2	76.3	75.9	74.6	75.6±0.9
	Split 3	73.7	74.3	74.4	74.1±0.4
	Split 4	77.2	77.7	77.3	77.4±0.3
	Avg (4 splits)	76.3	77.3	76.9	76.8±0.5

Table 26: Per-seed raw nAP50 (four-split mean) on COCO-20_i, underlying the "Avg" columns of Table 4.

Method	Shot	Seed 33	Seed 42	Seed 3407	Mean±std
FSOD-VFM	10-shot	58.2	59.1	58.6	58.6±0.5
	30-shot	62.8	63.5	63.0	63.1±0.4
CD-ViTO	10-shot	60.3	60.1	59.0	59.8±0.7
	30-shot	63.9	64.3	64.4	64.2±0.3
POISE	10-shot	60.9	61.1	61.1	61.0±0.1
	30-shot	66.1	64.6	65.6	65.4±0.8