CORRECTING THE SELECTOR: TRAINING-FREE PROPOSAL-FIELD CALIBRATION FOR OPEN-VOCABULARY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A frozen open-vocabulary detector can already contain a correct candidate and still fail to return it. Adding labels can evict an unchanged detection from a finite output list; within a class, confidence can favor a poorly localized box over a better alternative. We identify the selector as a distinct intervention point: contextual and geometric relations remain available in the dense proposal field even when scalar-score selection discards them. We introduce **VoPR**, a training-free correction of the proposal field for fixed-vocabulary accuracy before the unchanged selector. Proposal-Conditioned Vocabulary Routing (PVR) adjusts category admission, and Candidate Quality Calibration (CQC) combines semantic competition, frozen distributional box redecoding, and proposal agreement to refine candidate ordering. For class-separable scoring, we localize vocabulary coupling to shared finite caps. Under fixed-order selection, class-broadcast routing admits only a prefix of each class's NMS-survivor sequence; candidate-specific calibration supplies the complementary ability to change the sequence itself. VoPR improves WeDetect Tiny/Base/Large by $+0.35/+0.24/+0.20$ AP on COCO; frozen profiles also improve all nine LVIS and COCO-O transfer settings. With distractor labels kept active, VoPR gains $+0.344$–$+0.637$ target-label AP under held-out distractor content. These paired results connect class admission and candidate refinement to improved detection without additional training, parameter updates, or another visual forward pass.

1 INTRODUCTION

Open-vocabulary detection expands what a model can recognize by changing its text vocabulary. Yet a larger vocabulary need not expand what the model returns. A correct detection can disappear even when its score and box are unchanged. Likewise, at a fixed vocabulary, a well-localized proposal can already exist but lose to a more confident, less accurate alternative. These failures (Figure 1) expose a distinction between *producing useful evidence* and *selecting it*: improving the representation is not the only way to improve the detector.

Vocabulary-induced degradation is known from LOVD and vocabulary adaptation (Tang et al., 2024; Liu et al., 2025). Our question is more specific: *where can new labels suppress old detections when old scores do not change?* Under class-separable sigmoid scoring, classes do not share a softmax denominator. They do, however, share finite pre-NMS and final output lists. Newly appended high-score candidates can therefore consume the rank headroom of an existing candidate. We call this event *finite-cap eviction*. The relevant pressure is the number of added candidates crossing an admission boundary, rather than vocabulary size by itself.

The same scalar-score selector also discards relational evidence within a class. Confidence does not necessarily rank boxes by localization quality (Tychsen-Smith & Petersson, 2017; Jiang et al., 2018). However, the frozen proposal field contains more than scalar scores: image-wide class support, text-neighbor responses, box distributions, and overlapping peer proposals. These signals remain available before selection, including for candidates that final-output post-processing can no longer recover. This suggests a common intervention: *use the proposal field to inform its own admission and ordering*.

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.

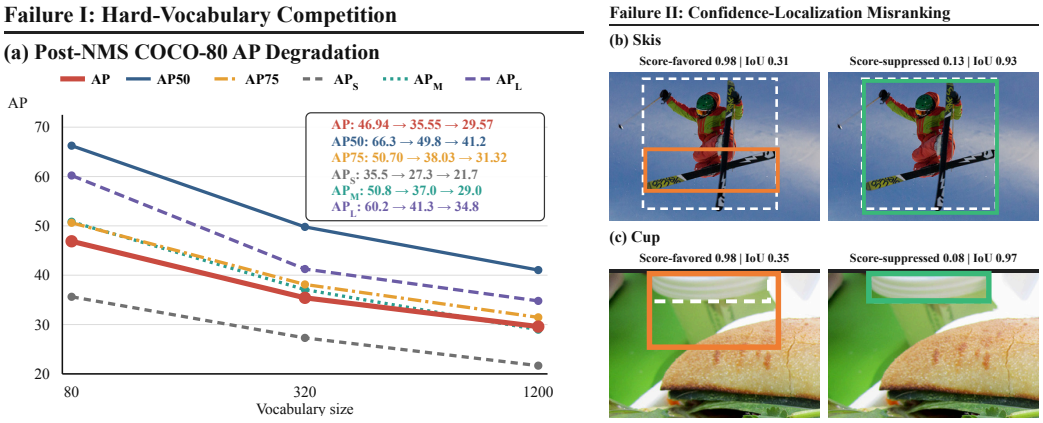

Figure 1: **Vocabulary degradation and confidence–localization inversion.** Left: a separate 500-image OWLv2 motivation diagnostic evaluates COCO-80 targets while nested hard vocabularies grow from 80 to 320 and 1200 classes. Right: frozen WeDetect-Large proposals illustrate confidence–localization inversions; white dashed boxes are ground truth, orange boxes are score-favored, and green boxes are better-localized alternatives. The right panel shows selected matched examples of the inversion. Section 4.2 isolates the finite-cap pathway on WeDetect and tests VoPR under controlled vocabularies.

The objective is to improve which candidates are returned at a fixed queried vocabulary and output budget. Holding capacity constant isolates the value of proposal-field evidence for selection.

VoPR (**Vo**cabulary-to-**P**roposal **R**ectification) implements this idea in two stages. Proposal-Conditioned Vocabulary Routing (PVR) broadcasts a bounded, image-conditioned class offset, preserving within-class proposal order. Candidate Quality Calibration (CQC) combines local semantic competition, top-L redecoding of native Distribution Focal Loss (DFL) outputs, and class-score-free proposal agreement. Both stages operate between one frozen forward pass and the unchanged selector. Under the class-wise selector analyzed below, PVR changes the admitted prefix of a fixed NMS-survivor sequence; CQC can change the sequence itself. The stages therefore control which classes receive capacity and which candidates occupy it.

We develop selector-local results that make the intervention testable. Under class-separable scoring, we locate the coupling operators and exact eviction boundary. A class-wise prefix constraint separates class admission from candidate replacement, while ordering and displacement bounds describe the implemented correction. Controlled component experiments then test the two-stage composition at the detection endpoint.

The evidence connects diagnosis to correction. On WeDetect-Base, relaxing only the final cap reduces vocabulary-induced AP loss from 0.1117 to 0.0177 (84% recovery). On WeDetect-Large, full VoPR exceeds either stage in ten controlled vocabulary–image conditions; held-out distractors yield $+0.344$ to $+0.637$ AP with positive paired 95% interval lower bounds. Tiny/Base/Large improve by $+0.35/+0.24/+0.20$ AP on COCO, and their frozen profiles remain positive across all nine LVIS and COCO-O transfer settings. These benchmarks test vocabulary and image-distribution changes without changing the COCO-developed configuration. Controlled mechanisms, component comparisons, and transfer thus support a reproducible correction to frozen inference.

Contributions. **(1) Locating recoverable selection error:** we distinguish finite-cap eviction from score dilution and characterize the class-wise prefix constraint that separates admission control from candidate replacement. **(2) Correcting admission and ordering without training:** PVR changes class-level admission, while CQC refines proposal scores and geometry, using the same frozen field before the unchanged selector. **(3) Testing the correction at three levels:** controlled vocabulary interventions identify the selection pathway, component comparisons test the two-stage design, and frozen-profile transfer tests its reach beyond COCO.

2 Related Work

Open-vocabulary representations. Open-vocabulary detectors transfer image–text representations (Radford et al., 2021) through distillation, regional alignment, grounding, and large-scale concept supervision (Zareian et al., 2021; Gu et al., 2021; Zhou et al., 2022; Zhong et al., 2021; Li et al., 2022; Liu et al., 2024b; Yao et al., 2024). OWL-ViT/OWLv2 scale dual-encoder detection, YOLO-World and YOLOE target efficient open-world prediction, and WeDetect exposes dense region–text retrieval across three model scales (Minderer et al., 2022; 2023; Cheng et al., 2024; Wang et al., 2025; Fu et al., 2025a). VoPR keeps these learned representations fixed and studies their conversion into a finite detection list.

Vocabulary adaptation and semantic context. LOVD investigates large distractor vocabularies using image-to-region filtering and cross-view verification; VocAda uses caption-derived vocabulary filtering (Tang et al., 2024; Liu et al., 2025). SIC-CADS supplies learned image-level context, SHiNe enriches text prototypes, and RALF retrieves semantic negatives (Fang et al., 2024; Liu et al., 2024a; Kim et al., 2024). Descriptions, cross-modal attention, and support sets provide other routes to contextualization (Pratt et al., 2023; Guo et al., 2022; Udandarao et al., 2023). PVR instead derives image-conditioned evidence from the detector's own proposals. It retains all queried classes and prototypes and adjusts their competition before shared caps. The distinguishing operation is a proposal-derived class adjustment at the admission interface, before candidates are discarded.

Calibration and localization-aware selection. Soft-NMS modifies suppression; Fitness NMS and IoU-Net address score–quality misalignment (Bodla et al., 2017; Tychsen-Smith & Petersson, 2017; Jiang et al., 2018). GFL, GFLv2, and VarifocalNet learn distributional or quality-aware heads (Li et al., 2020; 2021; Zhang et al., 2021); VoPR instead directly redecodes frozen distributions without learning a quality predictor. MoCaE calibrates experts; AggDet combines non-self overlap agreement with confidence before selection, and ProCal uses proposal priors (Oksuz et al., 2023; Zheng & Liu, 2024; Hong et al., 2026). FRACAL is also training-free: it uses class-frequency and spatial-distribution statistics for long-tailed logit calibration (Alexandridis et al., 2025). VoPR combines bounded class admission with candidate-specific redecoding and agreement before truncation, without AggDet's base-class instance prototypes or FRACAL's dataset-level statistics. Vocabulary/selector controls and component comparisons test this composition. Fixed AP changes evaluator-side coupling (Dave et al., 2021); VoPR instead changes the candidate field supplied to a fixed inference selector.

3 Method: Calibrating the Frozen Proposal Field

3.1 Interface and Two-Stage Design

VoPR is an inference operator on a frozen proposal field. Given class responses, text embeddings, and distributional boxes, PVR calibrates class admission and CQC redecodes and reranks individual proposals; the native selector then returns the detections (Figure 2). For one image, pyramid logits $\mathbf{Z}_\ell \in \mathbb{R}^{C\times H_\ell\times W_\ell}$ flatten to $\mathbf{Z} = [z_{r,c}] \in \mathbb{R}^{R\times C}$, with $R = \sum_\ell H_\ell W_\ell$. The head also supplies frozen text embeddings $\mathbf{T} \in \mathbb{R}^{C\times d_t}$, box-distribution logits $\mathbf{G} \in \mathbb{R}^{R\times 4\times(D+1)}$, and anchor/stride metadata. σ is the logistic sigmoid. VoPR maps $\mathbf{Z} \xrightarrow{\text{PVR}} \mathbf{Z}_{\text{in}} \xrightarrow{\text{CQC}} (\mathbf{Z}^\star, \mathbf{B}^\star)$ before applying the native selector

$$\mathcal{S} = \text{Cap}_{K_f} \circ \text{NMS}_{\text{cls}} \circ \text{Cap}_{K_p} \circ \text{Thresh}_\tau \,. \tag{1}$$

Thresholding is strict; global caps use fixed tie-breaking. Native coordinate restoration and class-wise NMS remain unchanged. Both stages reuse one frozen pass, with no parameter update, external encoder, or additional proposals.

3.2 Proposal-Conditioned Vocabulary Routing (PVR)

Category support. A single maximum cannot distinguish repeated category support from an isolated high response, while a mean over the whole dense field includes many background locations. Top-M aggregation instead summarizes a class's strongest local evidence with a fixed support budget, without requiring object-count estimates. For $p_{r,c} = \sigma(z_{r,c})$ and $M_I = \min(32, R)$, define

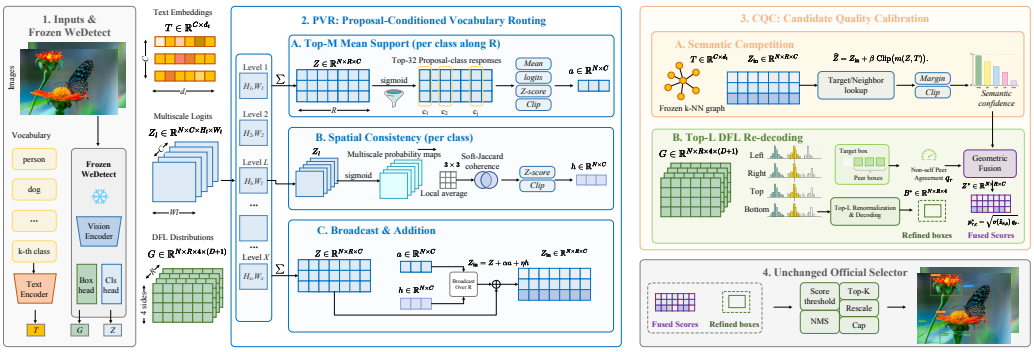

Figure 2: **VoPR's shared proposal-field inference path.** A frozen detector pass supplies logits $\mathbf{Z}$, DFL distributions $\mathbf{G}$, text embeddings $\mathbf{T}$, and anchors. **PVR** combines top-M support and centered 3×3 spatial consistency into a bounded class offset. **CQC** computes competition from $\mathbf{Z}$, adds it to $\mathbf{Z}_{\text{in}}$, redecodes top-L mass, and geometrically fuses confidence with class-score-free agreement (Eq. (6)). Output logits and boxes enter the unchanged selector. $N, R, C, d_t, D+1, \ell$ denote batch size, proposals, classes, text dimension, DFL bins, and pyramid level.

$\rho_c = M_I^{-1} \sum_{r \in \text{Top}_{M_I}(p_{\cdot,c})} p_{r,c}$. Mapping this support to logit space, standardizing across active classes, and clipping to $[-B_{\text{clip}}, B_{\text{clip}}]$ gives $a_c(I)$. Logit conversion places the correction in the same space as the native scores; standardization measures relative class support, and clipping prevents unusually concentrated evidence from dominating admission.

Spatial support and broadcast. Response strength alone omits spatial arrangement: equally strong peaks can be isolated or supported by their neighbors. Comparing the response map with its local average makes this distinction explicit. Soft-Jaccard normalization relates shared activation to their union instead of simply favoring a larger activated area; pyramid-wide accumulation combines the available scales. At level ℓ, let $A_{\ell,c} = \sigma(\mathbf{Z}_{\ell,c})$ and $\overline{A}_{\ell,c}$ be its centered 3×3 local average. The pyramid-wide soft Jaccard $\sum_{\ell,x} A_{\ell,c}(x)\overline{A}_{\ell,c}(x) / \sum_{\ell,x}[A_{\ell,c}(x) + \overline{A}_{\ell,c}(x) - A_{\ell,c}(x)\overline{A}_{\ell,c}(x)]$, standardized and clipped across classes, gives $h_c(I)$. The route is

$$z_{\text{in},r,c} = z_{r,c} + \xi_c(I), \qquad \xi_c(I) = \alpha a_c(I) + \eta h_c(I). \tag{2}$$

For a batch, the two $N \times C$ offsets broadcast over $\mathbf{Z} \in \mathbb{R}^{N \times R \times C}$. No category is deleted. Every proposal of a class receives the same offset, so routing preserves within-class proposal order while changing cross-class admission. Appendix C specifies normalization and numerical safeguards exactly.

3.3 Candidate Quality Calibration (CQC)

Semantic competition. Image-wide support cannot determine which of two related labels best explains a specific region. CQC therefore measures local dominance against nearby text classes. Using the strongest neighbor tests the closest competing explanation; restricting the comparison to semantic neighbors avoids letting unrelated labels define that region's discrimination margin. Let $\mathcal{N}_k(c)$ contain the $\min(k, C-1)$ other classes with highest cosine similarity to frozen text embedding $\mathbf{t}_c$, with class-index tie-breaking. Define $m_{r,c} = z_{r,c} - \max_{d \in \mathcal{N}_k(c)} z_{r,d}$, or zero for $C = 1$. The calibrated logit is

$$\widehat{z}_{r,c} = z_{\text{in},r,c} + \beta \, \text{clip}_{[-B_m, B_m]}(m_{r,c}). \tag{3}$$

The margin is anchored to original logits: routing does not redefine native neighbor dominance. It concentrates the correction on plausible semantic confusions instead of uniformly increasing confidence.

Reading the native box distribution. Low-probability bins can shift the full expectation away from dominant mass. Top-L redecoding concentrates on that mass while retaining a weighted

estimate across several bins, rather than collapsing to a single argmax. The retained mass also gives an explicit bound on the resulting displacement (Lemma 1). A Distribution Focal Loss (DFL) head (Li et al., 2020) predicts logits $g_{r,s,b}$ over $D+1$ ordered distance bins for each side $s \in \mathcal{S}_{\text{side}} = \{\text{L}, \text{T}, \text{R}, \text{B}\}$. VoPR reuses this representation and replaces the full expectation with a top-L truncated expectation:

$$\widetilde{d}_{r,s} = \frac{\sum_{b \in \mathcal{B}_L(r,s)} b\, e^{g_{r,s,b}}}{\sum_{b \in \mathcal{B}_L(r,s)} e^{g_{r,s,b}}}, \quad 1 \le L \le D+1, \tag{4}$$

where $\mathcal{B}_L(r,s)$ contains the largest-logit bins with index tie-breaking. Native anchor geometry converts the distances to $\widetilde{\mathbf{b}}_r$. No regressor is added or fitted.

Exact proposal agreement and fusion. Following overlap-based confidence aggregation (Zheng & Liu, 2024), we measure non-self agreement and fuse it geometrically with confidence. VoPR evaluates this cue on its redecoded boxes after class routing and semantic calibration. Excluding self-overlap prevents trivial support; highest-IoU peers concentrate aggregation on overlapping predictions rather than the whole image, independently of the proposal's class score. For $R > 1$, let $\mathcal{N}_r^{\text{box}}$ contain the $K_{\text{peer}} = \min(K_q, R-1)$ highest-IoU non-self redecoded peers. Define

$$q_r = \frac{1}{K_{\text{peer}}} \sum_{j \in \mathcal{N}_r^{\text{box}}} \text{IoU}(\widetilde{\mathbf{b}}_r, \widetilde{\mathbf{b}}_j), \qquad q_r = 0 \text{ if } R \le 1. \tag{5}$$

This class-score-free quantity measures peer support, not ground-truth IoU. The fixed equal-weight geometric mean combines it with semantic confidence:

$$p^\star_{r,c} = \sqrt{\max(\sigma(\widehat{z}_{r,c}), \epsilon_f)\, \max(q_r, \epsilon_f)}. \tag{6}$$

The geometric mean treats confidence and agreement symmetrically in log space: both must support a candidate, so high confidence cannot by itself remove the influence of weak geometric support. Because q_r is shared across classes, fusion cannot reverse a proposal's class order, although flooring may create ties. Tiny additionally uses the frozen DFL-concentration residual $z^\star_{r,c} = \text{logit}(p^\star_{r,c}) + \delta u_r$, where u_r is the image-standardized logit of normalized top-J four-side mass; Base and Large set $\delta = 0$. This residual distinguishes the concentration of a proposal's own boundary distribution from agreement among different proposals, providing Tiny with a small within-head ordering signal while leaving its class preferences unchanged. Appendix D defines this residual. The resulting logits and redecoded boxes enter the unchanged selector.

3.4 Selector-Level Properties

The analysis answers three questions: where selection can lose preserved evidence, what each calibration stage can change, and how large its perturbation can be. PVR controls class admission; CQC can change the selected representative. Full proofs are in Appendix D.

Premise A1 (class-separable frozen scoring). With preprocessing, preserved prompts, checkpoint, and selector fixed, appending labels leaves preserved logits and box distributions unchanged. A1 concerns native scores; PVR and semantic competition deliberately introduce vocabulary-conditioned corrections before selection.

Proposition 1 (selector-coupling locus). Under A1 and a common deterministic order, adding labels cannot change a preserved candidate's threshold membership or its same-class NMS outcome conditional on reaching NMS. A retention change must pass through the shared pre-NMS cap K_p, final cap K_f, or both. Thus a degradation with invariant scores can be a selection failure rather than score dilution.

Proposition 2 (exact stage-local eviction). At stage $t \in \{p, f\}$, condition on unchanged upstream membership and order. A baseline-admitted candidate x has rank headroom $H_t(x) = K_t - \text{rank}_t(x) + 1$. If $\Pi^{(t)}_{\mathcal{D}}(x)$ counts appended-class candidates ordered before x, then eviction occurs exactly when $\Pi^{(t)}_{\mathcal{D}}(x) \ge H_t(x)$. This links correction to a binding rank boundary, not class count alone.

Proposition 3 (bounded routing). Since $a_c, h_c \in [-B_{\text{clip}}, B_{\text{clip}}]$, $|\xi_c| \le \Delta = B_{\text{clip}}(|\alpha| + |\eta|)$. A pairwise margin changes by at most 2Δ, and same-class offsets cancel. For a fixed stage input, sufficiently separated threshold and rank decisions are preserved; routing acts near admission boundaries.

Theorem 1 (class-wise prefix constraint). For one image, fix candidate identities, boxes, thresholds, capacities, and the within-class total order. Assume fixed, candidate-wise, score-independent validity filtering before greedy class-wise NMS; only the final global cap removes survivors. If $V_c = (v_{c,1}, \dots, v_{c,m_c})$ is the complete native NMS-survivor sequence of eligible class-c candidates, every class-broadcast route returns $S_c(\xi) = (v_{c,1}, \dots, v_{c,\ell_c(\xi)})$, where $0 \le \ell_c(\xi) \le m_c$. Thus routed outputs are nested within each class; equal counts imply equal candidate identity sets. *Proof sketch.* Thresholding, pre-NMS truncation, and fixed validity filtering admit an eligible class prefix. Each candidate then encounters the same preceding NMS survivors as in the complete run. The final cap truncates their prefix again (Appendix D.4, Theorem 1).

An eligible candidate outside V_c cannot be returned by a class offset (Appendix D.7, Corollary 2). PVR controls how much of a survivor sequence is admitted; CQC can change the sequence through candidate-specific priorities and geometry. Matched PVR/CQC/Full comparisons test their combined accuracy.

Proposition 4 (conditional ordering reversal). Write $\bar{p}_r = \max(\sigma(\widehat{z}_{r,c}), \epsilon_f)$ and $\bar{q}_r = \max(q_r, \epsilon_f)$. Before endpoint clipping and Tiny's residual, a lower-confidence proposal r_l has a strictly larger fused score than r_h exactly when

$$\frac{\bar{q}_{r_l}}{\bar{q}_{r_h}} > \frac{\bar{p}_{r_h}}{\bar{p}_{r_l}}. \tag{7}$$

Agreement must outweigh the opposing confidence advantage. The condition links the fusion rule to a measurable change in proposal ordering, whose end-to-end contribution is evaluated by the matched component controls.

Lemma 1 (redecoding perturbation). For native bin probabilities $\pi_{r,s,b}$, retained top-L mass $\omega_{L,r,s}$, and full expectation $d_{r,s}$, $|\widetilde{d}_{r,s} - d_{r,s}| \le D(1 - \omega_{L,r,s})$. Thus retained mass controls the maximum change in native distance units.

3.5 Frozen Configuration

The shared settings are $M = 32, B_{\text{clip}} = 3, \alpha = 0.25, \eta = 0.10, k = 5, B_m = 8, L = 4, K_q = 3$. Tiny/Base/Large use $\beta = 0.025/0.05/0.025$ and $\delta = 0.05/0/0$, with $J = 4$ for Tiny. These profiles were developed on COCO and frozen across transfer vocabularies and domains at each model scale. Image-dependent corrections are computed from the current frozen field without test-image parameter optimization. Appendices C and J give the complete implementation and deployment details.

4 Experiments

We first isolate selection-induced error, then test whether correcting the frozen field improves detection across scales and transfer settings. Component and sensitivity studies connect these results to the two-stage design.

4.1 Evaluation Protocol

Full VoPR uses exact proposal agreement and is evaluated on WeDetect Tiny/Base/Large. Each same-process pair fixes checkpoint, images, resolution, vocabulary, prompts, selector, and evaluator. The controlled vocabulary tests retain WeDetect's native $K_p = 30{,}000, K_f = 300$: matching these caps separates better selection from increased output capacity. The relaxed-cap controls in Appendix F vary capacity separately. We use COCO val2017 (Lin et al., 2014), LVIS minival/full validation with Fixed AP (Gupta et al., 2019; Dave et al., 2021) and six-domain COCO-O (Mao et al., 2023). Additional ODinW13/35 paired diagnostics use a frozen substitute prompt (Li et al.,

Table 1: **Correction under held-out distractor content.** WeDetect-Large, one fixed 1,000-image slice, unchanged selector, and original COCO-80 evaluation targets. Bank B uses distractor ranks 1121–2240. Intervals use 1,000 paired image-bootstrap replicates and compare Full with Baseline; the PVR/CQC comparisons are point estimates. All values reproduce the Bank-B comparison in Appendix F.

Bank B	C	Baseline	PVR	CQC	Full	Full–Baseline (95% CI)
Far	320	46.795	47.048	47.156	**47.432**	+0.637 [+.367,+.955]
Far	1200	46.558	46.767	46.937	**47.147**	+0.588 [+.430,+.865]
Hard	320	46.732	46.994	46.968	**47.265**	+0.533 [+.283,+.795]
Hard	1200	46.156	46.323	46.348	**46.500**	+0.344 [+.110,+.553]

2022; Fu et al., 2025a); their protocol and published-baseline alignment are detailed in Appendices B and E. Published baselines provide context; all gains use the reproduced checkpoint. The published comparison follows WeDetect and the corresponding detector reports (Fu et al., 2025a; Jiang et al., 2024; Wang et al., 2024; Zhao et al., 2024; Fu et al., 2025b; Ren et al., 2024).

Development and confirmation. We developed scale-specific inference profiles on COCO val2017 while keeping all model parameters frozen. Full-COCO bootstrap intervals condition on those profiles; a five-fold outer audit separately evaluates selection within the fixed component family. Vocabulary comparisons use mutually image-disjoint subsets of that development set. Bank B holds out distractor content; dataset transfer keeps the COCO-developed profiles frozen. All AP differences are percentage points computed from unrounded outputs.

4.2 From Finite-Cap Failure to Controlled Correction

On 1,000 WeDetect-Base images, nested $C = 80 \to 320$ vocabularies leave preserved scores unchanged within 1.97×10^{-6}, with identical boxes and best original labels. Retained COCO-80 AP nevertheless falls by 0.1117 points. Changing only the final cap from 300 to 1200 reduces the loss to 0.0177, recovering 84%. On WeDetect-Large, the hard 1,200-class vocabulary lowers target AP by 0.726 under native budgets but by 0.068 when both candidate budgets are relaxed (Appendix F). These interventions isolate the finite-selection pathway on the evaluated detector family.

Routing improves selection at both tested capacities. The WeDetect-Large stage audit supplies the relevant control: PVR remains positive in all five target/far/hard vocabulary conditions under both official and relaxed budgets (Appendix F). Its gain ranges from $+0.196$ to $+0.284$ AP with official budgets and $+0.190$ to $+0.275$ with relaxed budgets. Image-conditioned routing can change threshold admission, cross-class ranking within an image, and same-class ranking across images. These results establish PVR's value at both tested capacities.

Two-stage correction with distractors active. Across five vocabularies and two image-disjoint 1,000-image slices, PVR and CQC improve the baseline, and Full is best in all ten conditions (Table 11). Full gains $+0.381$ to $+0.802$ AP and exceeds the stronger stage by $+0.126$ to $+0.317$. A further held-out distractor rank block, Bank B, tests new far/hard content at $C = 320/1200$: Full gains $+0.637/+0.588$ and $+0.533/+0.344$, respectively, with all four paired 95% intervals above zero (Table 1; full details in Appendix F). All distractors remain active through selection while evaluation retains the COCO-80 targets. The cap intervention diagnoses selection-induced loss; paired VoPR comparisons establish higher AP at each fixed vocabulary. Appendix F distinguishes this accuracy endpoint from cross-vocabulary sensitivity.

4.3 Family-Wide Accuracy and Frozen Transfer

VoPR improves COCO AP by $+0.35/+0.24/+0.20$ for Tiny/Base/Large and AP_{75} by $+0.44/+0.33/+0.39$. The corresponding paired 95% AP intervals are $[+0.2136, +0.4570]$, $[+0.1198, +0.3277]$, and $[+0.1121, +0.3004]$ (Appendix G). AP_{50} changes by at most 0.06 AP, whereas the larger AP_{75} gains are consistent with the candidate-ordering and bounded-redecoding mechanisms in Section 3.4. These paired gains recover detection accuracy from frozen representations.

Table 2: **Unified family comparison.** Bold marks paired gains over reproduced WeDetect; published values provide context and are not ranked by typography. Additional ODinW diagnostics and reported/reproduced alignment are in Appendix E.

Method	LVIS mini	LVIS val	COCO	COCO O
Tiny				
YOLO-World-L	35.4	26.8	44.9	32.5
YOLOE-v8-L	35.9	–	–	–
WeDetect reproduced	37.42	31.21	44.90	38.59
VoPR (ours)	37.65	31.35	45.25	39.16
Δ paired	$+0.23$	$+0.14$	$+0.35$	$+0.57$
Base				
GLIP	26.0	17.2	46.1	29.0
Grounding DINO	27.4	20.1	48.4	37.6
DetCLIP	35.9	28.4	–	43.3
DetCLIPv2	40.4	32.8	–	–
DetCLIPv3	47.0	38.9	47.2	38.5
T-Rex2	42.8	34.8	45.8	–
OV-DINO	40.1	32.9	50.2	–
MM-GDINO	41.4	31.9	50.4	34.0
LLMDet	44.7	34.9	55.6	36.1
DINO-X Edge	44.5	38.4	48.7	–
WeDetect reproduced	47.33	41.19	52.12	44.15
VoPR (ours)	47.51	41.25	52.36	44.69
Δ paired	$+0.18$	$+0.06$	$+0.24$	$+0.54$
Large				
GLIP-L	37.3	26.9	49.8	–
DetCLIPv3-L	48.8	41.4	48.5	48.8
WeDetect reproduced	54.88	49.12	54.49	46.59
VoPR (ours)	54.96	49.18	54.69	47.02
Δ paired	$+0.08$	$+0.06$	$+0.20$	$+0.42$

Table 3: **Full paired COCO components on WeDetect-Large.** VoPR is best in AP, AP_{75}, all size-stratified AP, and all displayed AR.

Variant	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
WeDetect	54.49	72.86	59.72	42.17	59.94	66.49
PVR-only	54.58	**72.93**	59.87	42.38	59.98	66.61
CQC-only	54.59	72.88	59.94	42.26	60.06	66.58
VoPR	**54.69**	72.89	**60.11**	**42.45**	**60.09**	**66.65**

Variant	AR_{100}	AR_M	AR_L
WeDetect	74.32	79.53	83.58
PVR-only	74.49	79.79	83.77
CQC-only	74.83	79.94	83.96
VoPR	**74.85**	**80.00**	**84.04**

The frozen profiles improve all nine LVIS and COCO-O transfer settings in Table 2. LVIS gains are positive at every scale and split; COCO-O gains are $+0.57/+0.54/+0.42$ AP, with positive differences in all 18 scale–domain cells. Transfer spans both vocabulary and image-distribution changes. The selector account relates these transfer gains to available proposals and actual crossings of the candidate-admission boundaries.

4.4 Composition, Selection, and Sensitivity

Routing contributes more at AP_{50}, CQC at AP_{75} and recall; Full leads on AP, size-stratified AP, and displayed recall (Table 3). In the complete 2^4 comparison, every singleton improves AP. An AggDet-style overlap-agreement control gains $+0.044$ AP on the same frozen field; Full exceeds it by $+0.157$ (paired 95% interval $[+0.062, +0.259]$) and the strongest singleton, PVR, by $+0.108$ ($[+0.047, +0.157]$; Table 7).

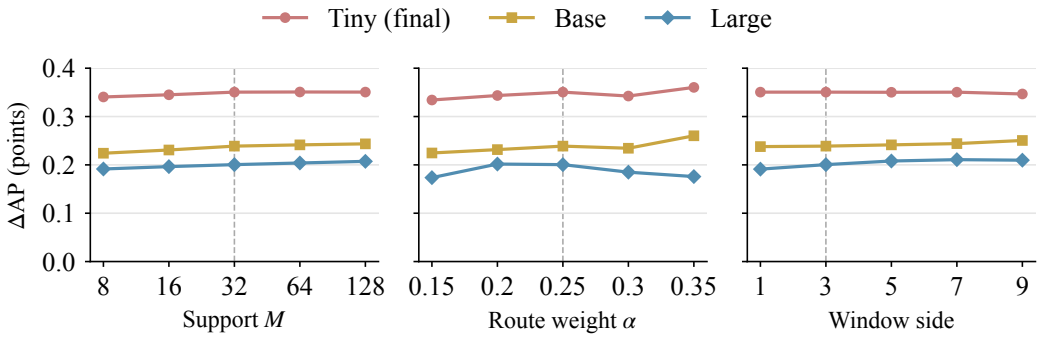

Figure 3: **Stable positive routing ranges.** Existing one-factor COCO sweeps of proposal support M, routing strength α, and spatial window size; dashed markers indicate the frozen settings. Full settings, other axes, and the exact Tiny profile are reported in Appendix H.

The five-fold outer audit chooses a composition on four folds and evaluates only on the remaining fold. Concatenated held-out predictions improve AP by $+0.1608$, with a paired 95% interval of $[+0.0797, +0.2737]$. The audit tests selection within the fixed component family while keeping the COCO-developed hyperparameters fixed. Full one-factor sweeps in Appendix H remain positive across all tested values on the reported axes. For Large, changing semantic-neighbor count from 1 to 20 spans only 0.0052 AP; the M sweep spans 0.0158. These sweeps and Figure 3 show stable positive routing ranges around the frozen settings across Tiny, Base, and Large.

4.5 Architecture-Native Illustrations

Table 4: **Cross-architecture proposal-field calibration.** All paired variants improve; details are in Appendix I.1.

Detector	Benchmark	Native operator	Baseline	Variant	ΔAP
WeDetect-Large	COCO	Full VoPR	54.49	**54.69**	$+0.20$
Grounding DINO-T	COCO	Routing + layer consensus	48.51	**48.88**	$+0.37$
YOLO-World-L	COCO	Routing + class competition	44.94	**45.23**	$+0.29$
YOLOE-v8-L	LVIS minival	Routing + proposal quality	35.89	**36.06**	$+0.17$

The three native ports in Table 4 improve AP by $+0.37/+0.29/+0.17$, respectively. Their operators are specified in Appendix I.1. These modes, selected on their evaluation splits, instantiate proposal-field calibration through architecture-native operators. The class-separable theory applies to fields satisfying A1; each port retains its native scoring path.

5 Conclusion

Stable preserved scores need not yield stable detections: shared finite caps can change which candidates survive even when their scores and boxes remain fixed. VoPR turns the frozen proposal field into evidence for its own selection. PVR calibrates category admission, while CQC combines semantic competition, native box redecoding, and proposal agreement to refine candidate ordering before the unchanged selector, without additional training or parameter updates.

Controlled cap interventions isolate a selection pathway; component comparisons establish the value of the two-stage correction, and frozen profiles yield positive paired results across all nine LVIS and COCO-O transfer settings, covering vocabulary and image-distribution changes. Together, these findings distinguish producing useful candidates from returning them. Selection-aware calibration offers an actionable complement to representation learning: improving a detector can begin with evidence already present in its frozen predictions.

REFERENCES

Konstantinos Panagiotis Alexandridis, Ismail Elezi, Jiankang Deng, Anh Nguyen, and Shan Luo. Fractal calibration for long-tailed object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15139–15150, 2025.

Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *ICCV*, pp. 5561–5569, 2017.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *CVPR*, pp. 16901–16911, 2024.

Achal Dave, Piotr Dollár, Deva Ramanan, Alexander Kirillov, and Ross Girshick. Evaluating large-vocabulary object detectors: The devil is in the details. *arXiv preprint arXiv:2102.01066*, 2021.

Ruohuan Fang, Guansong Pang, and Xiao Bai. Simple image-level classification improves open-vocabulary object detection. In *AAAI*, volume 38, pp. 1716–1725, 2024.

Shenghao Fu, Yukun Su, Fengyun Rao, Jing Lyu, Xiaohua Xie, and Wei-Shi Zheng. Wedetect: Fast open-vocabulary object detection as retrieval. *arXiv preprint arXiv:2512.12309*, 2025a.

Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *CVPR*, pp. 14987–14997, 2025b.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. *arXiv preprint arXiv:2209.14169*, 2022.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pp. 5356–5364, 2019.

Jae-Ryung Hong, Ho-Joong Kim, and Seong-Whan Lee. Procal: Inference-time proposal calibration for open-vocabulary object detection. *arXiv preprint arXiv:2607.01759*, 2026.

Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *ECCV*, pp. 784–799, 2018.

Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy. In *ECCV*, pp. 38–57. Springer, 2024.

Jooyeon Kim, Eulrang Cho, Sehyung Kim, and Hyunwoo J Kim. Retrieval-augmented open-vocabulary object detection. In *CVPR*, pp. 17427–17436, 2024.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, pp. 10965–10975, 2022.

Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *NeurIPS*, 33:21002–21012, 2020.

Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In *CVPR*, pp. 11632–11641, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755. Springer, 2014.

Mingxuan Liu, Tyler L Hayes, Elisa Ricci, Gabriela Csurka, and Riccardo Volpi. Shine: Semantic hierarchy nexus for open-vocabulary object detection. In *CVPR*, pp. 16634–16644, 2024a.

Mingxuan Liu, Tyler L Hayes, Massimiliano Mancini, Elisa Ricci, Riccardo Volpi, and Gabriela Csurka. Test-time vocabulary adaptation for language-driven object detection. In *2025 IEEE International Conference on Image Processing (ICIP)*, pp. 540–545. IEEE, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, pp. 38–55. Springer, 2024b.

Xiaofeng Mao, Yuefeng Chen, Yao Zhu, Da Chen, Hang Su, Rong Zhang, and Hui Xue. Coco-o: A benchmark for object detectors under natural distribution shifts. In *ICCV*, pp. 6339–6350, 2023.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *ECCV*, pp. 728–755. Springer, 2022.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *NeurIPS*, 36:72983–73007, 2023.

Kemal Oksuz, Selim Kuzucu, Tom Joy, and Puneet K Dokania. Mocae: Mixture of calibrated experts significantly improves object detection. *arXiv preprint arXiv:2309.14976*, 2023.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, pp. 15691–15701, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PmLR, 2021.

Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, et al. Dino-x: A unified vision model for open-world object detection and understanding. *arXiv preprint arXiv:2411.14347*, 2024.

Shiyu Tang, Zhaofan Luo, Yifan Wang, Lijun Wang, Huchuan Lu, Weibo Su, and Libo Liu. Lovd: Large-and-open vocabulary object detection. In *ACM MM*, pp. 9321–9329, 2024.

Lachlan Tychsen-Smith and Lars Petersson. Improving object localization with fitness nms and bounded iou loss. *arXiv preprint arXiv:1711.00164*, 2017.

Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *ICCV*, pp. 2725–2736, 2023.

Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yoloe: Real-time seeing anything. In *ICCV*, pp. 24591–24602, 2025.

Hao Wang, Pengzhen Ren, Zequn Jie, Xiao Dong, Chengjian Feng, Yinlong Qian, Lin Ma, Dongmei Jiang, Yaowei Wang, Xiangyuan Lan, et al. Ov-dino: Unified open-vocabulary detection with language-aware selective fusion. *arXiv preprint arXiv:2407.07844*, 2024.

Lewei Yao, Renjie Pi, Jianhua Han, Xiaodan Liang, Hang Xu, Wei Zhang, Zhenguo Li, and Dan Xu. Detclipv3: Towards versatile generative open-vocabulary object detection. In *CVPR*, pp. 27391–27401, 2024.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, pp. 14393–14402, 2021.

Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. Varifocalnet: An iou-aware dense object detector. In *CVPR*, pp. 8514–8523, 2021.

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

Yanhao Zheng and Kai Liu. Training-free boost for open-vocabulary object detection with confidence aggregation. *arXiv preprint arXiv:2404.08603*, 2024.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. *arXiv preprint arXiv:2112.09106*, 2021.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, pp. 350–368. Springer, 2022.

AI USE STATEMENT

AI tools assisted with manuscript formatting, language polishing, coding, experiment planning and interpretation, and developing and checking the selector-level theoretical analysis. The authors take full responsibility for the content of the paper.

A Supplementary Roadmap and Evidence Index

Reading the supplement. Appendices B–D give the evaluation protocol, complete operators, and proofs. Appendices E–G provide full comparisons, controlled vocabulary tests, and paired uncertainty. Sensitivity, architecture-native illustrations, deployment measurements, and qualitative examples follow in Appendices H–J and the final galleries.

B Evaluation and Reproducibility Protocol

Evaluation questions. The audit asks five evidence-level questions. **Family efficacy:** what paired differences arise for Tiny, Base, and Large under the same checkpoint and evaluator? **Component attribution:** what do PVR, CQC, and full VoPR contribute under matched component controls? **Transfer:** do COCO-developed frozen profiles retain their effect on LVIS, COCO-O, and the substitute-prompt ODinW paired diagnostic? **Scope:** which effects have paired-resampling support and which transfer across model scales and benchmarks? **Robustness:** does the restricted output Core (top-4 redecoding plus proposal agreement) survive a selection-disjoint image partition, and do frozen profiles survive one-factor controls?

Datasets and fixed protocol. The evaluations use COCO val2017, LVIS v1.0 minival and full validation with Fixed AP, the six-domain COCO-O release, and the released ODinW13/35 task selection and valid split with background-stripped annotations. Within each pair, checkpoint, backbone, resolution, vocabulary, images, prompts, preprocessing, selector, and evaluator are fixed. The exact author ODinW prompt remains unavailable; ODinW therefore remains a controlled same-process diagnostic.

Detector capacity versus evaluator truncation. The inference cap K_f bounds the score-ranked union of all active classes within an image, before predictions reach the evaluator. In category-aware COCO bounding-box evaluation (`useCats=1`), `maxDets=100` truncates the detections for each image–category pair. These are different operators: a target candidate removed from the global inference list cannot be restored by the evaluator's per-category truncation. Enlarging K_f can therefore return additional target detections and change COCO AP while leaving `maxDets` unchanged. Distractor classes remain active during inference; target-only evaluation does not filter them before the selector.

Uncertainty. For the fixed 2^4 COCO component family, a deterministic five-fold outer audit selects on four folds, concatenates held-out predictions, and applies a 1,000-replicate paired-image bootstrap. Frozen Tiny, Base, and Large selected-profile COCO comparisons also use 1,000 paired-image replicates. The selection-disjoint 15,000-image LVIS restricted-Core audit uses a 50-group paired image-jackknife and a complementary category bootstrap. The original COCO family and five-fold bootstraps use seed 20260727; the Full-versus-singleton composition bootstrap uses seed 20260729; and the held-out Bank-B bootstrap uses seed 20260730. LVIS/COCO-O deltas for which no interval is reported remain descriptive; ODinW is excluded from bootstrap claims.

Reproducibility files. Machine-readable files contain SHA-256 hashes for the three WeDetect checkpoints and configs, COCO/LVIS annotations, evaluator scripts, controllers, prediction shards, and every table source. Figure-selection protocols and rendering scripts accompany the assets.

C Complete Operator and Implementation Details

Algorithm 1 summarizes the complete PVR–CQC inference path.

Algorithm 1 VoPR Inference: PVR Followed by CQC

Input: dense logits $\mathbf{Z}$, box-distribution logits $\mathbf{G}$, frozen text embeddings $\mathbf{T}$, frozen geometric metadata $\mathcal{A}$
Output: official WeDetect post-processing of $(\mathbf{Z}^\star, \mathbf{B}^\star)$

Set $\mathbf{Z}_{\text{in}} \leftarrow \mathbf{Z}$
if PVR is enabled **then**
  Compute category evidence a_c and spatial consistency h_c
  Set $z_{\text{in},r,c} \leftarrow z_{r,c} + \alpha a_c + \eta h_c$
end if
CQC: build the text-neighbor graph $\mathcal{N}_k(c)$
Measure margins on original $\mathbf{Z}$ and add them to $\mathbf{Z}_{\text{in}}$, yielding $\widehat{\mathbf{Z}}$
Decode $\mathbf{B}^\star$ from the top-L distributional expectation
Estimate q_r from K_{peer} non-self overlap peers, or set $q_r = 0$ if $R \le 1$
Apply equal-weight confidence–agreement fusion; add the DFL-concentration residual for Tiny only to form $\mathbf{Z}^\star$
Apply strict $\sigma(z^\star_{r,c}) > \tau$ filtering and the global pre-NMS cap K_p
Rescale without pre-NMS clamping, run class-wise NMS, and apply K_f

Frozen tensor interface. Pyramid logits $\mathbf{Z}_\ell \in \mathbb{R}^{N\times C\times H_\ell\times W_\ell}$ are flattened into $\mathbf{Z} \in \mathbb{R}^{N\times R\times C}$, where $R = \sum_\ell H_\ell W_\ell$. The same frozen forward supplies $\mathbf{G} \in \mathbb{R}^{N\times R\times 4\times(D+1)}$ and $\mathbf{T} \in \mathbb{R}^{C\times d_t}$, together with frozen geometric metadata $\mathcal{A}$. The outputs have shapes $\mathbf{Z}^\star \in \mathbb{R}^{N\times R\times C}$ and $\mathbf{B}^\star \in \mathbb{R}^{N\times R\times 4}$, with $\mathbf{B}^\star_{n,r,:} := \widetilde{\mathbf{b}}_{n,r}$. Here N is batch size, C the active-vocabulary size, R the proposal count, and d_t the text-embedding dimension. When PVR is disabled, $\mathbf{Z}_{\text{in}} = \mathbf{Z}$. The intermediate tensor is $\widehat{\mathbf{Z}} = [\widehat{z}_{n,r,c}] \in \mathbb{R}^{N\times R\times C}$. In full VoPR, PVR precedes CQC; the component audit also applies CQC directly to the original field. Tiny uses the additional DFL-concentration residual, whereas Base and Large use the shared path.

Parameter domains. We assume $N, R, C, M, k, K_q, K_p, K_f \in \mathbb{N}_+$, $D \in \mathbb{N}_+$, $0 < \tau < 1$, $1 \le L \le D+1$, $1 \le J < D+1$, and $B_{\text{clip}}, B_m \ge 0$. The scale-specific coefficients $\alpha, \eta, \beta, \delta$ are finite real numbers. These domains close the top-M, neighbor, peer, threshold, and distributional reductions used below; the $C = 1$ and $R \le 1$ conventions are stated in the main method.

For one image, let $p_{r,c} = \sigma(z_{r,c})$ and $M_I = \min(M, R)$. We use fixed safeguards $\epsilon_{\log} = \epsilon_{\text{std}} = 10^{-6}$, $\epsilon_{\text{union}} = 10^{-8}$, and $\epsilon_f = \text{eps}(\text{float32})$. The exact category-evidence reduction is

$$\begin{aligned} \rho_c(I) &= \frac{1}{M_I} \sum_{r \in \text{Top}_{M_I}(p_{\cdot,c})} p_{r,c}, \\ v_c(I) &= \text{logit}\left(\text{clip}_{[\epsilon_{\log}, 1-\epsilon_{\log}]}(\rho_c(I))\right), \\ a_c(I) &= \text{clip}_{[-B_{\text{clip}}, B_{\text{clip}}]}\left(\frac{v_c(I) - \mu_v(I)}{\max(\text{sd}_v(I), \epsilon_{\text{std}})}\right). \end{aligned} \tag{8}$$

Here $\mu_v(I)$ and $\text{sd}_v(I)$ are the population mean and standard deviation across the active classes; top-M_I uses proposal-index tie-breaking. Batched top-M_I reduction yields $\boldsymbol{\rho} = [\rho_c(I_n)]_{n,c}$ and $\mathbf{a} = [a_c(I_n)]_{n,c}$ in $\mathbb{R}^{N\times C}$, broadcast as $[N, 1, C]$. Spatial consistency gives $\mathbf{h} = [h_c(I_n)]_{n,c} \in \mathbb{R}^{N\times C}$ with the same broadcast layout.

Pyramid-wide spatial consistency. At pyramid level ℓ, let $z_{\ell,c}(x)$ denote the scalar logit at spatial position x, set $A_{\ell,c}(x) = \sigma(z_{\ell,c}(x))$, and let $\overline{A}_{\ell,c}(x)$ be its stride-one 3×3 local average, excluding padded values from boundary averages. Define $U_{\ell,c}(x) = A_{\ell,c}(x) + \overline{A}_{\ell,c}(x) - A_{\ell,c}(x)\overline{A}_{\ell,c}(x)$. Intersections and unions are accumulated over the complete pyramid before normalization:

$$\begin{aligned} \text{Int}_c(I) &= \sum_{\ell,x} A_{\ell,c}(x)\overline{A}_{\ell,c}(x), \qquad \text{Uni}_c(I) = \sum_{\ell,x} U_{\ell,c}(x), \\ h_c^{\text{raw}}(I) &= \frac{\text{Int}_c(I)}{\max(\text{Uni}_c(I), \epsilon_{\text{union}})}, \qquad h_c(I) = \text{clip}_{[-B_{\text{clip}}, B_{\text{clip}}]}\left(\frac{h_c^{\text{raw}}(I) - \mu_h(I)}{\max(\text{sd}_h(I), \epsilon_{\text{std}})}\right). \end{aligned} \tag{9}$$

Here $\mu_h(I)$ and $\mathrm{sd}_h(I)$ are the population mean and standard deviation across active classes. This is a pyramid-wide soft Jaccard rather than an average of independently normalized level scores. The odd-window sensitivity results are in Appendix H.

Finite-precision fusion. Before converting the fused probability to a logit, the frozen implementation clips it to $[\epsilon_f, 1-\epsilon_f]$, with $\epsilon_f = \mathrm{eps}(\mathrm{float32})$. Tiny adds its residual after this conversion. The ordering condition in Proposition 4 describes the fusion score before endpoint clipping; monotone clipping can create ties but cannot reverse this order before the residual.

Selector replay. The final $z^\star_{r,c}$ and redecoded $\widetilde{\mathbf{b}}_r$ enter WeDetect's unmodified multi-label path: strict $\sigma(z^\star_{r,c}) > \tau$ filtering, global pre-NMS top-K_p over flattened proposal–class pairs, padding removal and coordinate restoration without pre-NMS boundary clamping, optional undersized-box removal, class-wise NMS, the final top-K_f cap, and only then image-boundary clamping. No proposal or selector rule is added.

Computational cost and batch size. The frozen scale-specific settings are specified in the main method. The text-neighbor graph costs $O(C^2 d_t)$ per image in the audited implementation; an output-equivalent implementation caches it once per vocabulary. PVR costs $O(RC)$, semantic competition costs $O(RCk)$, and exact overlap agreement costs $O(R^2)$. Agreement is processed in chunks of 512 proposals to bound peak memory. No gradients, inference-time ground-truth annotations, learned adapters, additional visual encoders, or second forward pass are required. All image statistics are independent and the audited runs use batch size one.

D Proofs for the Theoretical Analysis

D.1 Proof of Proposition 1

Premise A1 fixes every preserved score and box. Hence comparison with the strict threshold is invariant. The global pre-NMS cap is the first operator that mixes classes. Conditional on a preserved candidate x surviving that cap, every same-class candidate ordered before x also survives, because it precedes x in the same fixed global order. Class-wise NMS then receives exactly the same possible suppressors of x, with unchanged scores, boxes, and tie-breaking; candidates of newly appended classes cannot enter this same-class suppression chain. Thus x's NMS outcome is unchanged conditional on reaching NMS. The only remaining cross-class operator is the final cap. Consequently, any change in final retention must pass through the pre-NMS cap, the final cap, or both.

D.2 Proof of Proposition 2

Condition on the same stage-local input specified in Proposition 2: preserved candidates reach stage t in unchanged order, and vocabulary expansion only appends candidates from the added-label set $\mathcal{D}$. At that stage, the target has $\mathrm{rank}_t(x) \le K_t$ and therefore is baseline-admitted; it has $\mathrm{rank}_t(x) - 1$ baseline candidates ahead of it. After vocabulary expansion its rank is

$$\mathrm{rank}_t(x) + \Pi^{(t)}_{\mathcal{D}}(x).$$

It remains within capacity exactly when this quantity is at most K_t. Therefore it is evicted exactly when

$$\Pi^{(t)}_{\mathcal{D}}(x) \ge K_t - \mathrm{rank}_t(x) + 1 = H_t(x), \tag{10}$$

which proves Eq. (10). Enlarging the added candidate set cannot decrease the indicator sum, establishing monotonicity at the same fixed stage.

D.3 Proof of Proposition 3

Fix image I and suppress its argument. Write the broadcast class offset as $\xi_c = \alpha a_c + \eta h_c$. Clipping gives

$$|\xi_c| \le B_{\mathrm{clip}}(|\alpha| + |\eta|) = \Delta.$$

For arbitrary candidates $x = (r, c)$ and $x' = (r', c')$,

$$(z_{\text{in},r,c} - z_{\text{in},r',c'}) - (z_{r,c} - z_{r',c'}) = \xi_c - \xi_{c'}, \qquad |\xi_c - \xi_{c'}| \le 2\Delta. \tag{11}$$

The triangle inequality yields $|\xi_c - \xi_{c'}| \le 2\Delta$, proving Eq. (11). If $c = c'$, the difference is exactly zero; PVR therefore preserves the complete within-class proposal order. Since Δ is a logit offset, candidates satisfying $|z_{r,c} - \text{logit}(\tau)| > \Delta$ cannot cross the threshold. Conditional on a fixed input candidate set, a top-K_t boundary gap exceeding 2Δ cannot reverse, proving selector-membership locality conditional on the fixed stage input.

For a target $x = (r, y)$ and distractor $x' = (r', d)$, a strict PVR score reversal means

$$z_{r,y} + \xi_y > z_{r',d} + \xi_d, \quad \iff \quad \xi_y - \xi_d > z_{r',d} - z_{r,y}. \tag{12}$$

Thus the sign of the endogenous target–distractor support gap determines the direction of margin motion, but admission changes only after a zero crossing. An initially lower distractor overtakes the target exactly when $\xi_d - \xi_y > z_{r,y} - z_{r',d}$.

Per-class filtering and a shared final budget. Per-class top-k filtering followed by a global cap still contains a cross-class selection operator. As a deterministic illustration, take one preserved-class candidate of score 0.6, one appended-class candidate of score 0.9, a per-class capacity of one, and a global capacity of one. Both pass the per-class filters, but the appended candidate evicts the preserved candidate at the global cap. This is a counterexample to guaranteed decoupling, not an empirical comparison. Removing all binding shared caps does decouple classes under A1, but changes the shared-output-budget constraint. Proposition 1 identifies this distinction; the unchanged-budget VoPR comparisons ask how to improve the candidates selected under that constraint. The relaxed-budget PVR controls test the separate question of whether routing contributes beyond increasing capacity.

D.4 Class-Broadcast Routing Preserves NMS-Survivor Prefixes

Vocabulary conditioning and within-class order. Native score invariance and calibrated vocabulary dependence concern different objects. Under A1, the preserved entries of $\mathbf{Z}$ stay fixed when classes are appended. PVR recomputes support normalization over the active classes, and semantic competition recomputes their text-neighbor sets. These are the method's vocabulary-conditioned interventions. At each fixed vocabulary, however, PVR adds the same ξ_c to every proposal of class c, so all within-class logit differences remain unchanged. The following result turns this algebraic property into a constraint on candidate identity: a class offset controls how much of a survivor sequence is admitted, whereas candidate-specific scores or geometry can change which sequence is available.

The pairwise cancellation in Proposition 3 implies an output-level constraint on routing. Fix one image, its candidate identities and boxes, the score threshold, both global capacities, and greedy class-wise NMS. Routing changes only $z_{r,c}$ to $z_{r,c} + \xi_c$. Assume a common deterministic within-class total order throughout; exact real-valued class offsets preserve this order, while an implementation that introduces new numerical ties must preserve it explicitly for this statement to apply. Any validity filtering is fixed, independent of scores and other candidates, and applied before greedy NMS (possibly after the pre-NMS cap). After NMS, the only candidate removal is the final global score cap.

For class c, let Q_c be the complete native score-ordered sequence of eligible candidates. Let $V_c = (v_{c,1}, \ldots, v_{c,m_c})$ be the survivors of greedy NMS on all of Q_c, before threshold or capacity truncation. This is a conceptual reference sequence, not an additional inference pass; it is independent of ξ.

Theorem 1 (class-wise prefix constraint). For every routing vector ξ, the final output of class c is

$$S_c(\xi) = (v_{c,1}, \ldots, v_{c,\ell_c(\xi)}), \qquad 0 \le \ell_c(\xi) \le m_c.$$

Thus any two routed outputs of the same class are nested, and equal output counts imply equal candidate identity sets.

Proof. Class-broadcast offsets leave all within-class logit differences unchanged. Strict thresholding therefore retains a prefix of the native class order. The global pre-NMS cap also retains a class prefix: if a candidate survives, every earlier same-class candidate precedes it globally and survives too. Removing the same fixed invalid candidates preserves the prefix relation in the eligible sequence Q_c.

Greedy NMS is prefix-consistent. Induct over a prefix of Q_c: the first candidate has the same decision as in the complete run. If the preceding decisions agree, the next candidate encounters exactly the same earlier survivors with the same boxes, so its decision agrees as well. Later candidates cannot change an earlier decision. Consequently the NMS survivors of the admitted prefix form a prefix of V_c. Finally, the global output cap cannot retain a survivor while discarding an earlier same-class survivor under the common order. It therefore truncates this prefix once more, proving the statement. The nesting and equal-count claims follow immediately.

Why the two intervention levels differ. PVR adjusts how far a class's survivor sequence is admitted, but cannot substitute a later survivor for an earlier one or recover a candidate suppressed in the complete native sequence. For example, if two same-class boxes a, b have scores $2, 1$ and overlap above the NMS threshold, the complete native sequence is $[a]$. Class-broadcast routing can return $\varnothing$ or $[a]$, never $[b]$. Reversing their candidate-specific priority or changing their suppression geometry can change this sequence. CQC provides precisely these intervention capabilities; the component experiments assess their instantiated combination. This structural result concerns candidate identities within one image and class. Image-conditioned routing can still change admission counts and same-class score ordering across images, both of which enter dataset AP. It requires the native class-wise greedy selector; cross-class suppression or score-reordering suppression rules are different operators.

D.5 Proof of Proposition 4

For analysis only, consider the generalized unequal-weight family

$$p_{r,c}^{(\gamma)} = \bar{p}_{r,c}^{\gamma} \bar{q}_r^{1-\gamma}, \qquad 0 < \gamma < 1.$$

This family supports sensitivity analysis; the reported method fixes the equal-weight case. For the score-favored proposal r_h and the agreement-favored proposal r_l, the logarithm of the generalized second-to-first score ratio is

$$\log \frac{p_{r_l,c}^{(\gamma)}}{p_{r_h,c}^{(\gamma)}} = \gamma \log \frac{\bar{p}_{r_l,c}}{\bar{p}_{r_h,c}} + (1-\gamma) \log \frac{\bar{q}_{r_l}}{\bar{q}_{r_h}}.$$

Under the confidence inversion, the first term is non-positive (strictly negative unless flooring ties the two scores). If $\bar{q}_{r_l} \le \bar{q}_{r_h}$, the second is non-positive, so the generalized fusion cannot reverse the ordering. If $\bar{q}_{r_l} > \bar{q}_{r_h}$, reversal is equivalent to

$$(1-\gamma) \log \frac{\bar{q}_{r_l}}{\bar{q}_{r_h}} > \gamma \log \frac{\bar{p}_{r_h,c}}{\bar{p}_{r_l,c}}, \qquad \gamma < \frac{\log(\bar{q}_{r_l}/\bar{q}_{r_h})}{\log(\bar{q}_{r_l}/\bar{q}_{r_h}) + \log(\bar{p}_{r_h,c}/\bar{p}_{r_l,c})}. \tag{13}$$

Setting $\gamma = 1/2$ cancels the common factor and recovers Eq. (7), which is the reported equal-weight method. Equality produces tied fused scores and is handled by the fixed selector tie-breaker. With Tiny's residual enabled, the unclipped comparison becomes

$$\operatorname{logit}(p_{r_l,c}^{\star}) - \operatorname{logit}(p_{r_h,c}^{\star}) + \delta(u_{r_l} - u_{r_h}) > 0. \tag{14}$$

The implemented comparison substitutes $\operatorname{clip}_{[\epsilon_f, 1-\epsilon_f]}(p^{\star})$ for each $p^{\star}$ inside the logits. At $\delta = 0$, the shared equal-weight condition is recovered when endpoint clipping is inactive; saturation can otherwise create ties.

D.6 Proof of Lemma 1 and Exact Tiny Residual

Distributional quality estimation versus direct redecoding. GFLv2 (Li et al., 2021) feeds top-bin probability statistics into a learned quality predictor. Equation (4) instead retains the distance indices and computes a truncated conditional expectation, directly changing box coordinates without

fitting a predictor. Tiny's top-J residual is closer in its statistical input, but uses a fixed concentration formula and image-wise standardization rather than a learned mapping to IoU. Distribution statistics are established evidence; VoPR's use of them is a frozen decoding and ordering intervention, combined with class admission and peer agreement.

Fix a proposal–side pair (r, s) and suppress these two indices. Let $\mathcal{B}_L = \mathcal{B}_L(r, s)$ retain mass $\omega_L = \omega_{L,r,s}$. For $0 < \omega_L < 1$, write the full expectation as

$$d = \omega_L \widetilde{d} + (1 - \omega_L) d_{\text{tail}},$$

where $d_{\text{tail}} \in [0, D]$ is the conditional expectation outside $\mathcal{B}_L$. Since both conditional expectations lie in $[0, D]$,

$$|\widetilde{d} - d| = (1 - \omega_L)|\widetilde{d} - d_{\text{tail}}| \le D(1 - \omega_L).$$

The case $\omega_L = 1$ is immediate. The constant D is uniformly sharp over distributions and retained sizes: for $L = 1$, mass $\omega > 1/2$ at bin zero and $1 - \omega$ at bin D gives $\widetilde{d} = 0$ and $d = D(1 - \omega)$. For $D > 1$, finite softmax logits approach this equality as intermediate-bin mass tends to zero. Sharpness here concerns distance perturbation, not an IoU or AP improvement guarantee.

For completeness, let $\pi_{r,s,b} = \mathrm{softmax}_b(g_{r,s,\cdot})$, and for $1 \le J < D + 1$, let $\mathcal{B}_J(r, s)$ contain its J largest-probability bins. Tiny's side-wise top-J concentration is

$$\begin{aligned} \kappa_{r,s} &= \sum_{b \in \mathcal{B}_J(r,s)} \pi_{r,s,b}, \\ \widetilde{\kappa}_{r,s} &= \frac{\kappa_{r,s} - J/(D+1)}{1 - J/(D+1)}. \end{aligned} \tag{15}$$

The sum of the largest J probabilities is at least the uniform reference $J/(D+1)$, hence $\widetilde{\kappa}_{r,s} \in [0, 1]$. We average the four sides, clip to $[\epsilon_f, 1 - \epsilon_f]$, map to logit space, and standardize across the image's R proposals:

$$\begin{aligned} \bar{\kappa}_r &= \frac{1}{4} \sum_{s \in \mathcal{S}_{\text{side}}} \widetilde{\kappa}_{r,s}, \qquad \lambda_r = \mathrm{logit}\big(\mathrm{clip}_{[\epsilon_f, 1-\epsilon_f]}(\bar{\kappa}_r)\big), \\ u_r &= \frac{\lambda_r - \mu_\lambda(I)}{\max(\mathrm{sd}_\lambda(I), \epsilon_{\text{std}})}. \end{aligned} \tag{16}$$

Here $\mu_\lambda(I)$ and $\mathrm{sd}_\lambda(I)$ are the population mean and standard deviation over $r = 1, \ldots, R$, with $\epsilon_{\text{std}} = 10^{-6}$ and $\epsilon_f = \mathrm{eps}(\mathrm{float32})$. This proposal-level residual is shared across classes and therefore cannot alter the class order within a proposal.

Scope and strength of the results. Proposition 1 exactly localizes cross-category coupling for the audited class-separable selector. Proposition 2 gives an exact stage-local eviction law; Proposition 3 supplies sufficient stability bounds and exact pairwise crossing conditions; Proposition 4 gives the exact conditional fusion boundary; Theorem 1 characterizes the complete class-wise survivor prefix under routing; and Lemma 1 bounds box displacement. Together, these results separate cross-class admission, within-class survivor identity, and geometric displacement. Matched evaluations quantify their end-to-end effect on dataset AP, including matching, duplicates, and false positives.

D.7 Consequences for Admission and Survivor Identity

Corollary 1 (cap-free invariance). Assume A1, fixed candidate validity, and the common deterministic order of Proposition 1. If the expanded vocabulary has at most K_p threshold survivors and at most K_f class-wise NMS survivors, every preserved class returns exactly its original detections. *Proof.* Neither cap binds for the expanded vocabulary. Preserved threshold survivors and their class-wise NMS inputs are unchanged by A1; appended classes add only their own candidates and NMS survivors. Hence neither cap binds at the original vocabulary either. Each preserved class therefore returns the same class-wise NMS result. This states an exact output-invariance regime rather than an accuracy guarantee for a calibrated detector.

Corollary 2 (access to suppressed candidates). Under Theorem 1, let V_c be the complete native NMS-survivor sequence of class c, and let $b \in Q_c \setminus V_c$ be an eligible candidate. No class-broadcast offset can return b. Candidate-specific rescoring can make b survive NMS if it ranks b ahead of every same-class candidate whose overlap with b exceeds the suppression threshold. It then returns b provided b passes the score threshold, pre-NMS cap and final cap. *Proof.* Theorem 1 restricts every routed output to a prefix of V_c, which excludes b. Under the stated rescoring, no earlier candidate can suppress b, so greedy NMS retains it. Passing the remaining selector stages gives its final admission. Thus replacing a suppressed candidate requires changing candidate-specific order or suppression geometry; this is the degree of freedom supplied by CQC, not a uniqueness claim about its particular operators.

Operator-local bounds and preserved orders. The semantic update alone moves each logit by at most $|\beta|B_m$, so it changes any pairwise logit margin by at most $2|\beta|B_m$; it leaves boxes unchanged. Top-L redecoding alone leaves classification logits unchanged and obeys Lemma 1 in native distance-bin units. Native stride scaling converts that distance bound to image-coordinate displacement. The subsequent agreement computation can change because it reads the redecoded boxes. For a fixed proposal, confidence–agreement fusion and Tiny's class-shared residual preserve the ordering of the input classes up to numerical ties. Across proposals, Eq. (7) applies before endpoint clipping and Tiny's residual; with that residual, the correct comparison is Eq. (14). These local properties distinguish admission, geometry and ordering without conflating their composed effects.

E Complete Metrics and Baseline Context

This section expands the family and component comparisons in Section 4: baseline alignment establishes the paired reference, and the complete component results resolve the contribution of each retained operator. Figure 4 summarizes the paired gains across scales and benchmarks.

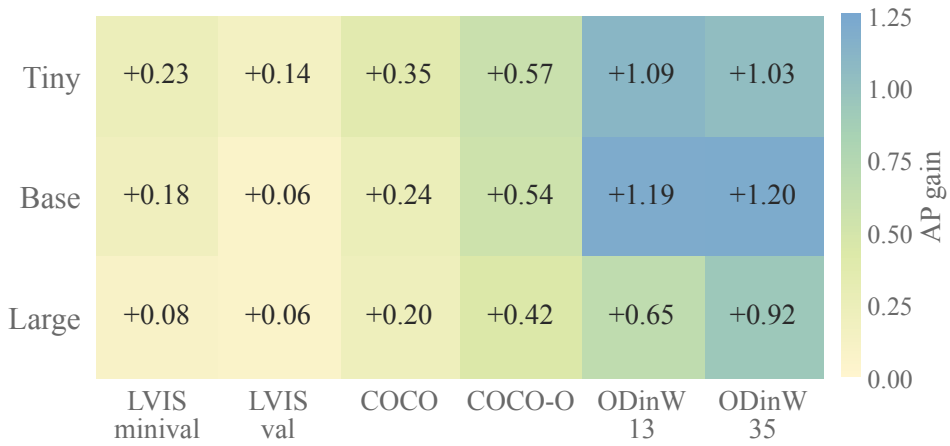

Figure 4: **Family-wide paired AP gains.** Each cell compares frozen-profile VoPR with reproduced WeDetect. The ODinW columns follow the paired diagnostic protocol in Section 4.1.

Table 5: **WeDetect reported/reproduced alignment (reported/reproduced AP).** Across the four prompt-compatible benchmarks, every absolute discrepancy is at most 0.41 AP, supporting the fidelity of the same-process baselines used for VoPR comparisons. ODinW is reported separately in the historical tables because the exact author prompt is unavailable.

Scale	LVIS minival	LVIS val	COCO	COCO-O
Tiny	37.4/37.42	31.4/31.21	44.9/44.90	38.6/38.59
Base	47.3/47.33	41.4/41.19	52.1/52.12	44.1/44.15
Large	55.0/54.88	49.4/49.12	54.5/54.49	47.0/46.59

The at-most 0.41-AP discrepancy in Table 5 supports the reproduced rows as paired anchors for VoPR; ODinW remains a separately qualified paired diagnostic.

Component names and stage mapping. In the factorial comparison, PVR includes both Top-M category support and spatial consistency. SC denotes semantic competition, DBD denotes Top-L distributional box redecoding, and OQF denotes overlap-based confidence–agreement fusion. CQC combines SC, DBD, and OQF; the full method is PVR followed by CQC. The factorial table uses Large, whose concentration residual is zero. The separate Tiny profile includes its stated residual. The restricted "Core" in the validity study contains DBD and OQF, not full CQC or full VoPR.

Table 6: **Complete outcome-independent 2^4 component comparison on COCO val2017.** Every combination of proposal-conditioned vocabulary routing (PVR), semantic competition (SC), top-L DFL box redecoding (DBD), and overlap-based proposal-agreement fusion (OQF) is evaluated on the same 5,000 images, frozen WeDetect-Large outputs, vocabulary, selector, and evaluator. ΔAP is relative to identity and uses unrounded evaluator outputs. Bold marks the best AP, AP50, and AP75 point estimate independently; the complete configuration maximizes AP, whereas the metric-specific AP50/AP75 maxima differ slightly.

Configuration	PVR	SC	DBD	OQF	AP	ΔAP	AP50	AP75
Identity	–	–	–	–	54.4884	$+0.0000$	72.8618	59.7233
OQF	–	–	–	Yes	54.5323	$+0.0439$	72.8253	59.8266
DBD	–	–	Yes	–	54.5080	$+0.0196$	72.8645	59.7981
DBD+OQF	–	–	Yes	Yes	54.5474	$+0.0590$	72.8223	59.8875
SC	–	Yes	–	–	54.5095	$+0.0211$	72.8969	59.7271
SC+OQF	–	Yes	–	Yes	54.5662	$+0.0778$	72.8772	59.8722
SC+DBD	–	Yes	Yes	–	54.5261	$+0.0377$	72.8946	59.8047
SC+DBD+OQF	–	Yes	Yes	Yes	54.5932	$+0.1048$	72.8784	59.9443
PVR	Yes	–	–	–	54.5812	$+0.0928$	72.9262	59.8679
PVR+OQF	Yes	–	–	Yes	54.6578	$+0.1695$	72.8804	60.0452
PVR+DBD	Yes	–	Yes	–	54.6013	$+0.1129$	72.9247	59.9430
PVR+DBD+OQF	Yes	–	Yes	Yes	54.6836	$+0.1952$	72.8934	**60.1099**
PVR+SC	Yes	Yes	–	–	54.5861	$+0.0977$	**72.9404**	59.8565
PVR+SC+OQF	Yes	Yes	–	Yes	54.6638	$+0.1754$	72.8854	60.0524
PVR+SC+DBD	Yes	Yes	Yes	–	54.6045	$+0.1161$	72.9366	59.9374
PVR+SC+DBD+OQF	Yes	Yes	Yes	Yes	**54.6888**	$\mathbf{+0.2005}$	72.8931	60.1090

Mechanistic reading. Every singleton improves AP over identity: PVR by $+0.0928$, semantic competition by $+0.0211$, top-L redecoding by $+0.0196$, and proposal-agreement fusion by $+0.0439$. PVR is the strongest singleton, consistent with category evidence acting directly on finite-cap admission. The complete configuration gives the highest overall AP, while the AP50 and AP75 maxima occur at nearby but different combinations; the operators thus do not behave as one uniform score offset. The held-out analysis below confirms that selecting within this fixed component family retains a positive effect on unseen folds.

Table 7: **Direct composition control on COCO val2017.** All rows reuse the frozen 5,000-image COCO development 2^4 component prediction set. The method is Full VoPR (PVR+SC+DBD+OQF); each reference retains exactly one component. Intervals condition on the fixed development-set configurations and are two-sided percentile intervals from 1,000 paired image-bootstrap replicates using the fixed seeds specified in Appendix B. The singleton comparator set is fixed before resampling; "strongest" denotes the largest full-sample singleton AP.

Singleton control	Control AP	Full AP	ΔAP	Paired 95% CI
PVR (strongest singleton)	54.581	54.689	$+0.108$	$[+0.047, +0.157]$
Semantic competition (SC)	54.510	54.689	$+0.179$	$[+0.093, +0.275]$
top-L DFL redecoding (DBD)	54.508	54.689	$+0.181$	$[+0.105, +0.285]$
Proposal-agreement fusion (OQF)	54.532	54.689	$+0.157$	$[+0.062, +0.259]$

Direct evidence for composition. Full VoPR (PVR+SC+DBD+OQF) exceeds every singleton on the primary AP metric. The closest comparison is the strongest singleton, PVR, for which the paired gain is $+0.108$ AP with a 95% interval of $[+0.047, +0.157]$ and 99.9% positive bootstrap mass. The lower endpoint is also positive against semantic calibration, top-L redecoding, and proposal agreement individually. The frozen composition therefore improves over each retained operator on the same images and evaluator. The five-fold audit below evaluates composition selection on held-out folds.

Selection-disjoint composition audit. The strongest singleton in Table 6 is PVR at 54.5812 AP, while the complete configuration reaches 54.6888 AP, a $+0.1077$-point descriptive difference. To prevent the same images from both choosing and testing a component composition, a deterministic five-fold outer audit selects one of all 16 rows on the other four folds. It selects the complete configuration on folds 0, 2, 3, and 4 and PVR+DBD+OQF on fold 1. Concatenating only held-out predictions gives $+0.1608$ AP and $+0.3456$ AP75 over identity; 1,000 paired-image bootstrap replicates give 95% intervals $[+0.0797, +0.2737]$ and $[+0.2208, +0.5074]$, respectively. Both intervals remain strictly positive, establishing held-out improvement over identity for the fixed component family.

Table 8: Full Tiny/Base historical context. Published rows are cross-paper context; primary gains use the same-process reproduced baseline. Daggered ODinW values use the substitute frozen prompt and are not official cross-paper comparisons. Published rows are context only; bold marks VoPR when it exceeds its paired reproduced value.

Method	LVIS minival	LVIS val	COCO	COCO-O	ODinW13	ODinW35
Tiny						
YOLO-World-L	35.4	26.8	44.9	32.5	38.4	17.1
YOLOE-v8-L	35.9	–	–	–	–	–
WeDetect reported	37.4	31.4	44.9	38.6	46.4	21.1
WeDetect reproduced	37.42	31.21	44.90	38.59	46.02†	20.58†
VoPR (ours)	**37.65**	**31.35**	**45.25**	**39.16**	**47.11**†	**21.61**†
Base						
GLIP	26.0	17.2	46.1	29.0	46.5	19.6
Grounding DINO	27.4	20.1	48.4	37.6	51.4	22.7
DetCLIP	35.9	28.4	–	43.3	–	–
DetCLIPv2	40.4	32.8	–	–	–	–
DetCLIPv3	47.0	38.9	47.2	38.5	–	–
T-Rex2	42.8	34.8	45.8	–	–	18.0
OV-DINO	40.1	32.9	50.2	–	–	–
MM-GDINO	41.4	31.9	50.4	34.0	52.5	23.1
LLMDet	44.7	34.9	55.6	36.1	52.1	23.8
DINO-X Edge	44.5	38.4	48.7	–	–	–

Table 8: *(continued)*

Method	LVIS minival	LVIS val	COCO	COCO-O	ODinW13	ODinW35
WeDetect reported	47.3	41.4	52.1	44.1	53.1	24.6
WeDetect reproduced	47.33	41.19	52.12	44.15	52.37†	23.95†
VoPR (ours)	**47.51**	**41.25**	**52.36**	**44.69**	**53.56**†	**25.15**†

Paired Tiny/Base transfer analysis. Relative to the same-process reproduction, VoPR improves all 12 paired Tiny/Base benchmark settings (Table 8). Tiny gains $+0.23/+0.14$ on LVIS minival/val, $+0.35$ on COCO, $+0.57$ on COCO-O, and $+1.09/+1.03$ on the substitute-prompt ODinW13/35 diagnostic; Base gains $+0.18/+0.06$, $+0.24$, $+0.54$, and $+1.19/+1.20$, respectively. The consistent sign across all six benchmarks shows that the frozen profiles transfer beyond COCO; the smaller LVIS gains are consistent with inference-only reranking acting on evidence already present in the proposal field.

Table 9: Full WeDetect-Large historical context. Published rows are cross-paper context; primary gains use the same-process reproduced baseline. Published rows do not enter boldface ranking; bold marks VoPR when it exceeds its paired reproduced value. Daggered ODinW values use the substitute frozen prompt and are not official cross-paper comparisons.

Method	LVIS minival	LVIS val	COCO	COCO-O	ODinW13	ODinW35
GLIP-L	37.3	26.9	49.8	–	–	–
DetCLIPv3-L	48.8	41.4	48.5	48.8	–	–
WeDetect-Large reported	55.0	49.4	54.5	47.0	53.4	25.8
WeDetect-Large reproduced	54.88	49.12	54.49	46.59	46.47†	20.74†
VoPR-Large (ours)	**54.96**	**49.18**	**54.69**	**47.02**	**47.12**†	**21.67**†

Paired Large-scale analysis. VoPR-Large improves its paired reproduction in all six columns of Table 9: $+0.08$ on LVIS minival, $+0.06$ on LVIS val, $+0.20$ on COCO, $+0.42$ on COCO-O, and $+0.65/+0.92$ on the qualified ODinW13/35 diagnostic. These paired gains extend selector-level correction to the strongest frozen scale. Published rows provide reference context; daggered ODinW values follow the specified substitute-prompt protocol.

F Mechanism Audit Details

Three comparisons support Section 4.2: varying the cap isolates the selection pathway, the routing-stage controls distinguish PVR from a budget increase, and the composition comparisons test correction with distractors active. Their protocols and results follow in that order.

The controlled WeDetect-Base cap intervention uses 1,000 images and a strictly nested $C = 80 \rightarrow 320$ vocabulary while retaining the original 80 labels for evaluation. Across preserved original anchor–class pairs, the maximum score change is 1.97×10^{-6}, boxes are identical, and the best original label never changes. Retained COCO-80 AP loses 0.1117 under the official final cap of 300 and 0.0177 when only that cap is relaxed to 1200. On the 0–100 AP scale, baseline/expanded AP is 53.9511/53.8394 at the 300 cap and 53.9763/53.9586 at the 1200 cap. The relative reduction in loss is $1 - 0.0176536/0.1117198 = 84.2\%$, rounded to 84% in the main text. This intervention tests the final-cap pathway without assuming that every upstream candidate set is identical.

Five additional signed routing-stage audits use WeDetect-Large on the same 1,000 images and fixed $C = 80$, far, and hard vocabularies. Here *hard* and *far* denote the pre-frozen semantically near and semantically distant distractor banks, respectively; their class lists and construction rules remain fixed across every paired run.

Table 10: **Controlled WeDetect-Large vocabulary and selector-stage audit.** Each row uses the same 1,000 COCO images and evaluates the original 80 target labels. The official budget is $nms_{\text{pre}} = 30{,}000$, $K_{\text{image}} = 300$; the relaxed budget uses 60,000 and 1,200. AP columns report target-label COCO AP. Thr./Pre/NMS/Final report the paired difference in covered non-crowd ground-truth instances after thresholding, pre-cap, class-wise NMS, and the official final cap; Final-R uses the relaxed final cap. All vocabularies and budgets are fixed before evaluation.

		Official budget			Relaxed budget			PVR net GT recovery				
Vocabulary	C	Baseline	PVR	ΔAP	Baseline	PVR	ΔAP	Thr.	Pre	NMS	Final	Final-R
COCO target	80	47.529	47.799	+0.270	47.546	47.821	+0.275	+86	+86	+93	+66	+90
Far	320	47.495	47.699	+0.204	47.547	47.746	+0.199	+96	+96	+102	+63	+94
Hard	320	47.388	47.643	+0.255	47.539	47.792	+0.253	+89	+88	+94	+44	+76
Far	1200	47.314	47.510	+0.196	47.524	47.714	+0.190	+98	+97	+103	+77	+85
Hard	1200	46.803	47.087	+0.284	47.478	47.730	+0.251	+94	+90	+98	+76	+60

PVR has a positive target-label AP point difference under both budgets in all five fixed conditions (Table 10): $+0.196$ to $+0.284$ under the official budget and $+0.190$ to $+0.275$ under the relaxed budget. For these five routing comparisons, the original-score parity check is exact. Under the official budget, the baseline drops 0.726 AP from $C = 80$ to hard $C = 1200$; relaxing only the two candidate budgets reduces the corresponding gap to 0.068 AP, consistent with the finite-cap eviction path localized by Proposition 1 and Proposition 2. Positive net recovery occurs at every audited selector stage, including $+44$ to $+77$ ground-truth instances at the official final cap. This fixed-population design isolates PVR under one bank per condition, complementing the full-VoPR family results. PVR's positive paired effect persists under both tested capacity settings.

Table 11: **Controlled composition under fixed vocabulary pressure.** Within each row, Baseline, PVR, CQC, and Full reuse identical frozen detector tensors, image slice, vocabulary bank, unchanged selector, and evaluator. The four configurations are fixed before evaluation. The two 1,000-image sets are image-disjoint and reuse the same five banks. AP is evaluated against the original COCO-80 labels; for $C = 320/1200$, every distractor remains active through the unchanged selector. "Confirm." denotes the image-disjoint confirmation set.

Split	Vocabulary	C	Baseline	PVR	CQC	Full
Primary	COCO target	80	47.529	47.799	47.841	**48.095**
	Far	320	47.495	47.699	47.785	**47.969**
	Hard	320	47.388	47.643	47.524	**47.769**
	Far	1200	47.314	47.510	47.575	**47.742**
	Hard	1200	46.803	47.087	46.905	**47.221**
Confirm.	COCO target	80	46.824	47.308	47.202	**47.625**
	Far	320	46.797	47.033	47.173	**47.413**
	Hard	320	46.696	47.051	46.851	**47.224**
	Far	1200	46.602	46.742	46.949	**47.109**
	Hard	1200	45.978	46.265	46.120	**46.420**

Accuracy endpoint. Fixed-vocabulary AP is the method's accuracy endpoint; Full improves it in every condition in Table 11. Cross-vocabulary sensitivity is a separate quantity: the baseline/Full hard-bank gaps are $0.726/0.874$ on the primary slice and $0.846/1.205$ on the confirmation slice. The cap intervention separately identifies the selection-induced loss pathway.

Vocabulary construction and split identities. A bank is a fixed set of queried class names. The distractor pool contains 30,679 canonicalized names assembled from OpenImages V6, V3Det, and Objects365 metadata. Each name is ranked by its maximum frozen OWLv2 text cosine similarity to a COCO-80 class, using the prompt "a photo of a {name}." Hard banks take the descending order and far banks the ascending order; ties use normalized UTF-8 name order. The original bank uses

ranks 1–1120 of each order. Bank B uses the disjoint ranks 1121–2240; its first 240 or all 1120 distractors give $C = 320$ or $C = 1200$ after adding the 80 target names. Bank B varies distractor content with the detection checkpoint and target categories fixed. The two 1,000-image slices are image-disjoint subsets of the COCO val2017 development dataset; their comparison evaluates frozen-profile composition across images.

Controlled composition analysis. PVR and CQC are individually positive in all ten paired conditions, while Full achieves the highest AP in every condition. Across the two image-disjoint populations, Full improves the frozen baseline by +0.381 to +0.802 AP and exceeds the stronger single stage by +0.126 to +0.317 AP. The fixed-bank comparison supports complementary class-level and candidate-level correction; Bank B evaluates the same composition under disjoint distractor content.

Table 12: **Full-versus-Baseline paired uncertainty under held-out Bank-B content.** AP is evaluated on the original COCO-80 labels while all distractors remain active through the unchanged selector. Bank B uses far/hard distractor ranks 1121–2240 on one designated 1,000-image slice. Panel (b) reports Full−Baseline AP points and two-sided 95% paired image-bootstrap intervals from 1,000 replicates; fixed seeds are specified in Appendix B.

(a) AP point estimates

Bank B	C	Baseline	PVR	CQC	Full
Far	320	46.795	47.048	47.156	**47.432**
Far	1200	46.558	46.767	46.937	**47.147**
Hard	320	46.732	46.994	46.968	**47.265**
Hard	1200	46.156	46.323	46.348	**46.500**

(b) Full-versus-Baseline paired uncertainty

Bank B	C	Full−Baseline	two-sided 95% CI
Far	320	**+0.637**	**[+0.367,+0.955]**
Far	1200	**+0.588**	**[+0.430,+0.865]**
Hard	320	**+0.533**	**[+0.283,+0.795]**
Hard	1200	**+0.344**	**[+0.110,+0.553]**

Bank-B uncertainty analysis. All four Full–Baseline paired 95% intervals in Table 12 are positive. PVR/CQC contrasts are point estimates; the 5,000-image singleton intervals in Table 7 concern a separate population.

G Paired Uncertainty and Selection-Disjoint Evaluation

The main accuracy comparisons report paired point estimates. This section adds their uncertainty and distinguishes fixed-profile resampling from selection-disjoint evaluation, so each interval is read against the comparison it actually tests.

Table 13: **Convergent evidence beyond the main tables.** Positive paired resampling, held-out composition, a selection-disjoint Core check, sign-consistent domain transfer, and the controlled cap intervention test complementary parts of the proposal-field account. Resampling units are stated per row.

Evidence	Paired protocol		Result	Uncertainty or qualification
Tiny selected profile	COCO val2017, images	5,000	$\Delta\mathrm{AP} = +0.3505$	1,000-replicate paired-image bootstrap 95% interval $[+0.2136, +0.4570]$
Base selected profile	COCO val2017, images	5,000	$\Delta\mathrm{AP} = +0.2388$	1,000-replicate paired-image bootstrap 95% interval $[+0.1198, +0.3277]$

Table 13: *(continued)*

Evidence	Paired protocol	Result	Uncertainty or qualification
Large selected profile	COCO val2017, 5,000 images	$\Delta\mathrm{AP} = +0.2005$	1,000-replicate paired-image bootstrap 95% interval $[+0.1121, +0.3004]$
Fixed 2^4 component-family audit	Five-fold COCO held-out concatenation	$\Delta\mathrm{AP} = +0.1608$; $\Delta\mathrm{AP}_{75} = +0.3456$	Paired-image 95% intervals $[+0.0797, +0.2737]$ and $[+0.2208, +0.5074]$
Restricted Core (top-4 redecoding + agreement; a restricted variant of VoPR)	15,000 LVIS val images excluding minival	$\Delta\mathrm{AP} = +0.0980$	50-group image-jackknife 95% interval $[+0.00057, +0.19539]$
Frozen domain transfer	COCO-O, Tiny/Base/Large, six domains	$+0.57/ + 0.54/ + 0.42$ AP	All 18 scale-by-domain point differences are positive; no bootstrap interval is imputed
Finite-cap intervention	WeDetect-Base, 1,000 images, $C = 80 \rightarrow 320$	Retained AP loss $0.1117 \rightarrow 0.0177$	Relaxing only the final cap removes about 84% of the loss; mechanism evidence, not a VoPR gain

Convergent-evidence reading. Paired uncertainty, held-out composition, selection-disjoint Core, and frozen domain transfer test complementary aspects of improvement (Table 13). The cap intervention measures eviction recovery. Each row retains its own evaluation population, resampling unit, and uncertainty estimate.

G.1 Selection-Disjoint Validity Robustness

We reuse frozen dense outputs on 15,000 LVIS val images that exclude all 4,809 minival selection images. All comparisons use Fixed AP, 1,001 jointly valid categories, and 10,000 paired category-bootstrap replicates. Here *Core* is the restricted top-4 redecoding plus proposal-agreement operator, a restricted variant of VoPR. A sanitized implementation maps non-finite agreement values to zero and reproduces the frozen Core aggregate exactly; the validity-only official decoder matches reconstructed identity within 3×10^{-7} AP. In a validity-only control, a decoded box must have four finite coordinates, $x_2 > x_1$, and $y_2 > y_1$. Invalid boxes receive zero class probability before the native selector; valid boxes retain native classification scores, without routing or agreement fusion. "Official" uses the full DFL expectation, whereas "top-4" uses the truncated decoder.

Table 14: **Selection-disjoint validity robustness on LVIS val.** Every retained contrast has a positive paired category-bootstrap interval. Frozen and sanitized Core give identical aggregate metrics, and validity-only controls preserve the improvement.

Paired contrast	ΔAP	95% interval	Supported reading
Frozen Core – reconstructed identity	$+0.0980$	$[+0.0381, +0.1640]$	Positive selection-disjoint effect.
Sanitized Core – reconstructed identity	$+0.0980$	$[+0.0381, +0.1640]$	Sanitization preserves the result.
Sanitized Core – validity-only official decoder	$+0.0980$	$[+0.0381, +0.1640]$	Official validity handling preserves the gain.
Sanitized Core – validity-only top-4 decoder	$+0.0948$	$[+0.0421, +0.1566]$	The top-4 validity handling preserves the gain.
Agreement only – reconstructed identity	$+0.0917$	$[+0.0379, +0.1540]$	Agreement-only is positive.

The unchanged $+0.0980$ AP across validity paths (Table 14) places the gain beyond numerical cleanup. Positive agreement-only transfer independently supports proposal evidence on these selection-disjoint images.

H Frozen Hyperparameter Stability

These one-factor sweeps expand the compact sensitivity results in Section 4. Each sweep varies one choice on the COCO development protocol while holding the rest of the final profile fixed; the profiles used for cross-dataset transfer remain unchanged. Figure 3 summarizes the routing axes. Table 15 reports all tested values for each retained axis. The Large rows share the frozen full VoPR profile (PVR, spatial consistency, semantic competition, distribution-aware box decoding, and proposal-agreement fusion); the Tiny rows vary its family-specific DFL-concentration residual. Two further tables repeat every shared axis on Base and on the exact final Tiny profile, including its frozen DFL-concentration residual.

The reported sensitivity audit covers the validated retained axes: $M, B_{\text{clip}}, \alpha$, the spatial-consistency window and η, k, B_m, β, K_q, the generalized fusion weight, and Tiny's J, δ. The top-L redecoding uses the shared frozen $L = 4$; the sensitivity claim is limited to the axes listed here. Numerical safeguards, the native head and selector, and overlap chunking remain fixed.

Fusion-weight sensitivity analysis. The tables vary the generalized fusion exponent defined in Appendix D around the frozen equal-weight setting. The reported method retains its frozen equal-weight fusion rule.

Table 15: Reported Large-profile and Tiny-specific one-factor-at-a-time COCO val2017 sweeps. Each row uses 5,000 images, the same checkpoint and official evaluator, and changes only the named quantity. Values are descriptive development differences from the paired reproduced WeDetect baseline. Frozen settings are marked for comparison with the surrounding sensitivity range.

Profile	Axis	Swept values (in order)	Ordered ΔAP (points)	Frozen
Large	PVR support M	$8, 16, 32, 64, 128$	$+0.1915, +0.1964, +0.2005, +0.2039, +0.2073$	32
Large	route clip B_{clip}	$1, 2, 3, 4, 5$	$+0.1084, +0.1491, +0.2005, +0.1411, +0.1136$	3
Large	PVR weight α	$0.15, 0.20, 0.25, 0.30, 0.35$	$+0.1736, +0.2017, +0.2005, +0.1849, +0.1758$	0.25

Table 15: *(continued)*

Profile	Axis	Swept values (in order)	Ordered ΔAP (points)	Frozen
Large	Spatial weight η	$0.05, 0.075, 0.10, 0.125, 0.15$	$+0.1809, +0.2035, +0.2005, +0.1969, +0.1816$	0.10
Large	Semantic weight β	$0, 0.0125, 0.025, 0.0375, 0.05$	$+0.1952, +0.1975, +0.2005, +0.1989, +0.1938$	0.025
Large	Spatial window	$1, 3, 5, 7, 9$	$+0.1912, +0.2005, +0.2080, +0.2108, +0.2097$	3
Large	Semantic neighbors k	$1, 3, 5, 10, 20$	$+0.1970, +0.2022, +0.2005, +0.2011, +0.1996$	5
Large	Semantic clip B_m	$2, 4, 8, 12, 16$	$+0.1951, +0.1953, +0.2005, +0.2007, +0.2002$	8
Large	agreement K_q	$1, 2, 3, 5, 7$	$+0.1871, +0.1815, +0.2005, +0.2085, +0.1991$	3
Large	Generalized fusion weight γ	$0.40, 0.45, 0.50, 0.55, 0.60$	$+0.1728, +0.1942, +0.2005, +0.1991, +0.1828$	0.50
Tiny	Distribution mass J	$1, 2, 4, 6, 8$	$+0.3419, +0.3501, +0.3505, +0.3347, +0.3334$	4
Tiny	residual δ	$0, 0.025, 0.05, 0.10, 0.20$	$+0.3342, +0.3441, +0.3505, +0.3389, +0.2179$	0.05

Large and Tiny-specific analysis. For Large, every tested value remains above the paired baseline and most semantic, neighborhood, and fusion axes form shallow positive plateaus. Routing clipping matters more: overly tight or permissive bounds weaken the gain, consistent with bounding image evidence without letting it dominate frozen logits. Spatial windows from 1×1 through 9×9 all retain positive gains. Tiny also remains positive for every tested J and δ, including the zero-residual endpoint. Together, these development-set sweeps show positive neighborhoods around the reported settings.

Table 16: Full-COCO Base sweeps for the reported shared VoPR sensitivity axes. Each row uses 5,000 val2017 images, the paired official evaluator, and the frozen Base profile, varying one axis while retaining all other settings.

Axis	Swept values (in order)	Base ordered ΔAP	Frozen
PVR support M	$8, 16, 32, 64, 128$	$+0.2241, +0.2308, +0.2388, +0.2414, +0.2435$	32
Route clip B_{clip}	$1, 2, 3, 4, 5$	$+0.1844, +0.2106, +0.2388, +0.1793, +0.1523$	3
PVR weight α	$0.15, 0.20, 0.25, 0.30, 0.35$	$+0.2246, +0.2316, +0.2388, +0.2344, +0.2602$	0.25
Spatial weight η	$0.05, 0.075, 0.10, 0.125, 0.15$	$+0.2444, +0.2448, +0.2388, +0.2197, +0.2110$	0.10
Semantic weight β	$0, 0.025, 0.05, 0.075, 0.10$	$+0.2165, +0.2343, +0.2388, +0.2278, +0.2104$	0.05
Spatial window	$1, 3, 5, 7, 9$	$+0.2378, +0.2388, +0.2414, +0.2440, +0.2505$	3
Semantic neighbors k	$1, 3, 5, 10, 20$	$+0.2469, +0.2387, +0.2388, +0.2418, +0.2422$	5
Semantic clip B_m	$2, 4, 8, 12, 16$	$+0.2289, +0.2353, +0.2388, +0.2260, +0.2277$	8

Table 16: *(continued)*

Axis	Swept values (in order)	Base ordered ΔAP	Frozen
Agreement K_q	$1, 2, 3, 5, 7$	$+0.2393, +0.2373, +0.2388, +0.2036, +0.1633$	3
Generalized fusion weight γ	$0.40, 0.45, 0.50, 0.55, 0.60$	$+0.2041, +0.2320, +0.2388, +0.2461, +0.2366$	0.50

Base analysis. Base remains positive at every tested setting. Routing weights, spatial windows, and generalized fusion all form positive ranges around the frozen development point. The decline for large K_q is compatible with diffuse peer aggregation mixing less-local support, while the stable M, k, and B_m ranges show that no exact support or neighbor count is required.

Table 17: Full-COCO Tiny sweeps for the reported shared VoPR sensitivity axes. Each row uses 5,000 val2017 images, the paired official evaluator, and the exact final Tiny profile with $J = 4$ and $\delta = 0.05$ retained in every branch. The dedicated J and δ sweeps are in Table 15.

Axis	Swept values (in order)	Tiny ordered ΔAP	Frozen
PVR support M	$8, 16, 32, 64, 128$	$+0.3405, +0.3450, +0.3505, +0.3508, +0.3506$	32
Route clip B_{clip}	$1, 2, 3, 4, 5$	$+0.1889, +0.3092, +0.3505, +0.3159, +0.3031$	3
PVR weight α	$0.15, 0.20, 0.25, 0.30, 0.35$	$+0.3342, +0.3435, +0.3505, +0.3424, +0.3603$	0.25
Spatial weight η	$0.05, 0.075, 0.10, 0.125, 0.15$	$+0.3412, +0.3469, +0.3505, +0.3424, +0.3348$	0.10
Semantic weight β	$0, 0.0125, 0.025, 0.0375, 0.05$	$+0.3199, +0.3321, +0.3505, +0.3553, +0.3587$	0.025
Spatial window	$1, 3, 5, 7, 9$	$+0.3505, +0.3505, +0.3502, +0.3504, +0.3467$	3
Semantic neighbors k	$1, 3, 5, 10, 20$	$+0.3413, +0.3513, +0.3505, +0.3480, +0.3513$	5
Semantic clip B_m	$2, 4, 8, 12, 16$	$+0.3364, +0.3406, +0.3505, +0.3462, +0.3462$	8
Agreement K_q	$1, 2, 3, 5, 7$	$+0.3531, +0.3407, +0.3505, +0.3224, +0.2939$	3
Generalized fusion weight γ	$0.40, 0.45, 0.50, 0.55, 0.60$	$+0.3247, +0.3308, +0.3505, +0.3475, +0.3425$	0.50

Tiny analysis. Tiny has the flattest spatial-window response: all tested windows remain within 0.0038 AP. Larger semantic weights can improve the development point, whereas broad agreement neighborhoods reduce it, suggesting that the smaller head benefits from bounded semantic correction but is more sensitive to diffuse peer support. The profile stays above its paired baseline throughout the tested development ranges.

Across Tiny, Base, and Large, the full profile remains above its matched baseline throughout the one-factor sweeps of proposal support M, PVR strength α, spatial strength η, window size, semantic neighbors, margin, agreement, generalized fusion weight, and clipping. We retain the verified values in Tables 15–17.

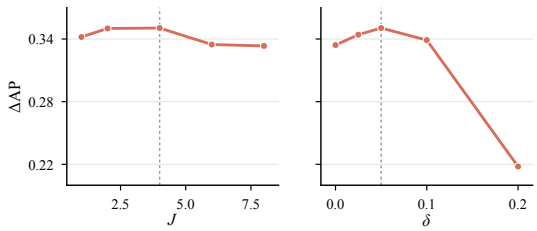

Figure 5: Tiny-specific sensitivity for the DFL-concentration bin count J and residual weight δ. The frozen $J = 4, \delta = 0.05$ setting lies inside the positive region rather than at a singular operating point.

Figure 5 shows that the complete Tiny profile remains above its paired baseline for every tested J and δ, including $\delta = 0$; the frozen $\delta = 0.05$ increases its development gain from $+0.3342$ to $+0.3505$ AP. Both the shared zero-residual path and the Tiny-specific profile therefore retain positive development gains.

All displayed points on the retained COCO axes remain positive. In particular, the M sweep spans only 0.0158 AP points, semantic-neighbor count spans 0.0052, semantic clipping spans 0.0056, and the Tiny mass sweep spans 0.0171. Tables 16 and 17 provide the corresponding Base and Tiny audits.

I Cross-Architecture Portability

This section explains how the pre-selector intervention adapts to the different native tensors exposed by each detector. Table 18 expands the compact main-paper comparison with reported-reference alignment; the implementation descriptions then specify which proposal-field signals each port uses.

Table 18: **Exploratory cross-architecture variants under paired, same-protocol evaluation.** Reported AP is an alignment check and does not enter the paired difference. AP50/AP75 differences for the four rows are $+0.03/+0.39$, $+0.37/+0.41$, $+0.39/+0.33$, and $+0.09/+0.26$. Grounding DINO-T and YOLO-World-L have 1,000-replicate paired-image one-sided 95% AP lower bounds of $+0.18$ and $+0.22$. YOLOE's 10,000-replicate paired-category bootstrap has a $+0.01$ AP one-sided 95% lower bound and a 95.85% positive-difference rate. Resampling uses images for Grounding DINO and YOLO-World and categories for YOLOE.

Detector	Benchmark	Added frozen operators	Reported	Reproduced	Variant	ΔAP
WeDetect-Large	COCO	Full VoPR	54.5	54.49	**54.69**	+0.20
Grounding DINO-T	COCO	Routing + layer consensus	48.4	48.51	**48.88**	+0.37
YOLO-World-L	COCO	Routing + competition	44.9	44.94	**45.23**	+0.29
YOLOE-v8-L	LVIS minival	Routing + proposal quality	35.9	35.89	**36.06**	+0.17

The positive paired differences on Grounding DINO-T, YOLO-World-L, and YOLOE-v8-L show that frozen proposal-field evidence can be instantiated beyond WeDetect. Each detector exposes a different frozen field, so the retained operator is architecture-native rather than one universal parameterization. AP75 improves at least as much as AP50 for WeDetect, Grounding DINO, and YOLOE: the observed gains extend to stricter localization criteria on these architectures.

I.1 Architecture-Specific Portability Implementations

All three adapters are deterministic, run under `no_grad`, and act on the last dense representation available before each detector's original candidate selection. They share only PVR's category-evidence primitive. Given dense class probabilities $p_{r,c}$, they apply the category-evidence reduction

in Eq. (8) with $M = 32$ and $B_{\text{clip}} = 3$. The offset $0.25a_c$ is broadcast to all proposals of class c. Unlike WeDetect PVR, these ports do not use the feature-pyramid spatial-consistency term because the three released heads do not expose the same spatial interface.

Grounding DINO-T. We intercept the Grounding DINO prediction head after its official token-positive map has converted token sigmoid responses into class probabilities. Let $z_{r,c}^{(j)}$ be their logits at decoder layer j, and let L_{dec} be the last decoder layer. The reported `VOPR_PORT_MODE=router_layer3` routing-plus-layer-consensus variant uses

$$z_{r,c}^{\text{port}} = \frac{1}{3} \sum_{j=L_{\text{dec}}-2}^{L_{\text{dec}}} z_{r,c}^{(j)} + 0.25a_c, \tag{17}$$

with a_c computed from the final decoder layer. The final-layer (c_x, c_y, w, h) box of every retained query is taken unchanged; score reranking can nevertheless change which queries are retained. The adjusted probabilities $\sigma(z_{r,c}^{\text{port}})$ enter the original flattened query–class top-K_{GD}, where K_{GD} is the native cap fixed by the official Grounding DINO configuration, followed by coordinate conversion, clipping, and rescaling. This port substitutes decoder-layer consensus for WeDetect's box-distribution/overlap branch; it neither adds distributional box redecoding nor introduces NMS into Grounding DINO.

YOLO-World-L. We concatenate the class logits from all prediction-pyramid levels into $\mathbf{Z} \in \mathbb{R}^{R\times C}$ before thresholding and NMS. Besides Eq. (8), the reported variant computes the top-one versus top-two class-logit margin $\chi_r^{\text{raw}} = z_{r,(1)} - z_{r,(2)}$ and sets

$$\chi_r = \text{clip}_{[-3,3]} \left(\frac{\chi_r^{\text{raw}} - \mu_\chi}{\max(\text{sd}_\chi, \epsilon_{\text{std}})} \right),$$

where μ_χ and sd_χ are computed across the image proposals. Its update is

$$z_{r,c}^{\text{port}} = z_{r,c} + 0.25a_c + 0.10\chi_r. \tag{18}$$

Because χ_r is shared by all classes of proposal r, it adjusts proposal priority without changing that proposal's internal class order. The adapter reshapes the logits back to their original pyramid levels and calls the unchanged YOLO-World prediction routine; box regression, objectness, rescaling, thresholding, and NMS are untouched. This margin cue is architecture-native and is not WeDetect's text-neighbor margin.

YOLOE-v8-L. The frozen YOLOE head exposes dense class probabilities $p_{r,c}$ and raw box-distribution logits. We define $z_{r,c} = \text{logit}(\text{clip}_{[\epsilon_f, 1-\epsilon_f]}(p_{r,c}))$. The reported routing–quality variant applies $z'_{r,c} = z_{r,c} + 0.25a_c$, then redecodes each side from the top four distribution bins and computes q_r by Eq. (5) with $K_q = 3$. This agreement is not ground-truth-IoU calibration. The adapter returns

$$p_{r,c}^{\text{port}} = \sqrt{\max(\sigma(z'_{r,c}), \epsilon_f) \max(q_r, \epsilon_f)}. \tag{19}$$

The redecoded boxes and fused scores then enter YOLOE's unchanged thresholding and multi-label, class-aware NMS; the reported LVIS minival run uses $\text{max_det} = 1000$ followed by the official Fixed-AP evaluator. Semantic competition and the Tiny-only concentration residual are disabled in this reported port. The YOLOE variant combines routing with proposal agreement; the YOLO-World variant combines routing with its proposal-margin cue.

These implementations test whether different frozen heads expose useful proposal-field evidence rather than enforcing one detector-independent operator. Each retained mode was chosen from an architecture-specific comparison on its evaluation split and then frozen; the reported resampling is therefore conditional development evidence for that retained mode.

J EFFICIENCY AND DEPLOYMENT EVIDENCE

Table 19: **Complete paired component metrics and timing.** All accuracy metrics use the same 5,000 COCO val2017 images and evaluator. Latency is the median increment on the matched V100 path with checkpoint, dense forward pass, and selector fixed.

Variant	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR_{100}	AR_M	AR_L	Δ latency (ms)
WeDetect	54.49	72.86	59.72	42.17	59.94	66.49	74.32	79.53	83.58	0.00
PVR-only	54.58	**72.93**	59.87	42.38	59.98	66.61	74.49	79.79	83.77	1.07
CQC-only	54.59	72.88	59.94	42.26	60.06	66.58	74.83	79.94	83.96	242.21
VoPR	**54.69**	72.89	**60.11**	**42.45**	**60.09**	**66.65**	**74.85**	**80.00**	**84.04**	242.73

Timing analysis. In Table 19, PVR adds only 1.07 ms, whereas CQC-only and full VoPR add 242.21 and 242.73 ms. The near-equality of the latter two isolates exact $O(R^2)$ peer agreement as the dominant measured overhead rather than linear routing; sparse or approximate peer search is the direct acceleration target.

Sparse-reference agreement. Let R remain the dense proposal count, and let R_s denote the reference budget. From the original logits, compute $s_r = \max_c \sigma(z_{r,c})$ and take the top $S = \min(R_s, R)$ proposals as references. All R redecoded boxes remain queries, including those not selected as references. For each query, compute IoUs against the S reference boxes; set its self-IoU to zero if it belongs to the reference set. The sparse agreement is the mean of the largest $\min(K_q, S-1)$ IoUs for every query. Set it to zero when $R \leq 1$ or $S \leq 1$. No query is removed, and score flooring, fusion, and the downstream selector are unchanged. Unlike exact agreement, sparse agreement uses class scores to select references, but the subsequent overlap average has no class-score weights. Work is $O(RS)$, with 512-query chunks and an $O(512S)$ overlap workspace, instead of the exact $O(R^2)$ work. The implementation uses the native tensor top-k operation for reference selection; equal salience values inherit its tie behavior.

The sparse accuracy slice consists of 1,000 COCO images at offset 1000 in ascending image-ID order, disjoint from the sparse-development slice within the COCO profile-development dataset. Table 20 compares full VoPR implementations: "Exact OQF" uses exact agreement, and the sparse rows use bounded reference sets.

Table 20: **WeDetect-Large selection-frozen accuracy–cost profile.** Accuracy uses the designated offset 1,000-image COCO confirmation slice. All non-baseline rows use the same Full VoPR operators; only proposal-agreement computation changes. Latency is synchronized batch-1 end-to-end inference over 200 images on one V100 verified idle immediately before measurement, with paired rotating method order and four CPU threads. Δt is the paired median increase over the official baseline. Confirmation budgets $R_s \in \{512, 1024, 2048\}$ are selected for positive development gains from the screened grid $\{256, 512, 1024, 2048\}$; $R_s = 256$ has development-only status, and the exploratory $R_s = 704$ candidate lies outside this grid.

Mode	R_s	AP	ΔAP	Gain kept	Median ms	Δt ms
Baseline	–	56.917	–	–	286.7	–
Sparse	512	57.073	+0.157	58.9%	374.9	+88.0
Sparse	**1024**	**57.174**	**+0.258**	**96.9%**	**376.2**	**+89.3**
Sparse	2048	57.168	+0.252	94.7%	378.9	+92.0
Exact OQF	–	57.183	+0.266	100.0%	530.5	+243.7

Accuracy–cost analysis. At $R_s = 1024$, the sparse implementation retains 96.9% of the exact OQF AP gain while reducing median end-to-end latency from 530.5 to 376.2 ms (29.1%). Relative to the official baseline, its paired median overhead is 89.3 ms (31.2%), versus 243.7 ms for exact OQF; this reduces the incremental overhead by 63.4%. These measurements characterize the specified single-V100 deployment.

Memory measurement. A separate synchronized replay of the last timing image, after resetting CUDA peak counters, records 3843.96 MiB peak allocated and 4064.00 MiB peak reserved for baseline, Exact, and each reported sparse mode. These totals include the resident model and active inference tensors and measure whole-run peak memory on that single image. The chunked implementation changes the overlap workspace without increasing this measured whole-run peak.

K VISUALIZATIONS AND ADDITIONAL QUALITATIVE EVIDENCE

Paired display protocol. Both galleries place reproduced WeDetect above VoPR for the same uncropped image and use identical eligible ground truth (white dashed boxes), stored final scores, a fixed score ≥ 0.35 cutoff, and stable descending top-eight display. Images are not reused across sets or scales. These selected pairs explain visible ranking changes; the 5,000-image evaluation establishes their population-level effect.

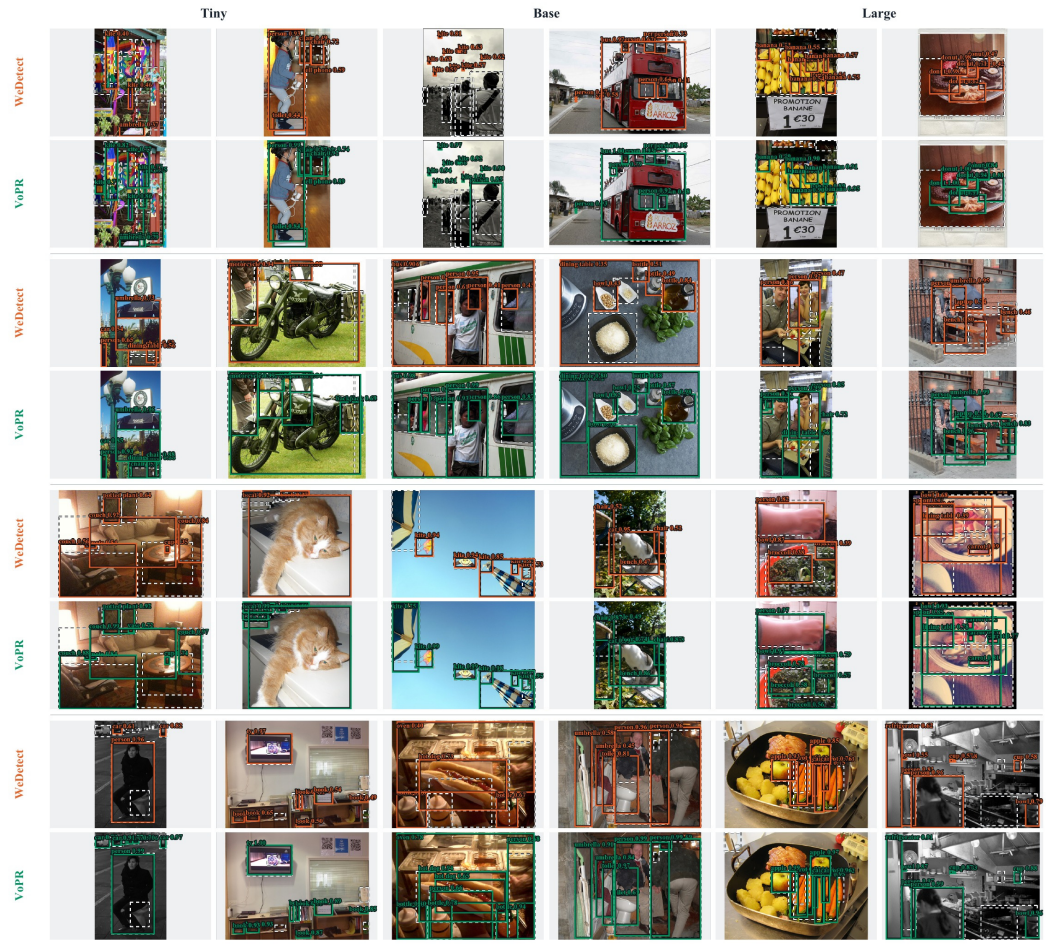

Figure 6: **Additional paired COCO diagnostics, set I.** Columns are grouped by Tiny, Base, and Large. The gallery contains 24 distinct images, eight per scale, under the shared protocol above; orange and green denote WeDetect and VoPR predictions.

Set-I analysis. Across Tiny, Base, and Large, the pairs in Figure 6 repeatedly show stronger retention of correct-category detections and reordering among competing boxes around the same object. This visual pattern is consistent with the two selector-level roles studied in the paper: PVR changes cross-class admission pressure, while CQC requires semantic confidence and proposal support to

agree. Because both branches are active, these examples illustrate their joint output without attributing a correction to either branch alone.

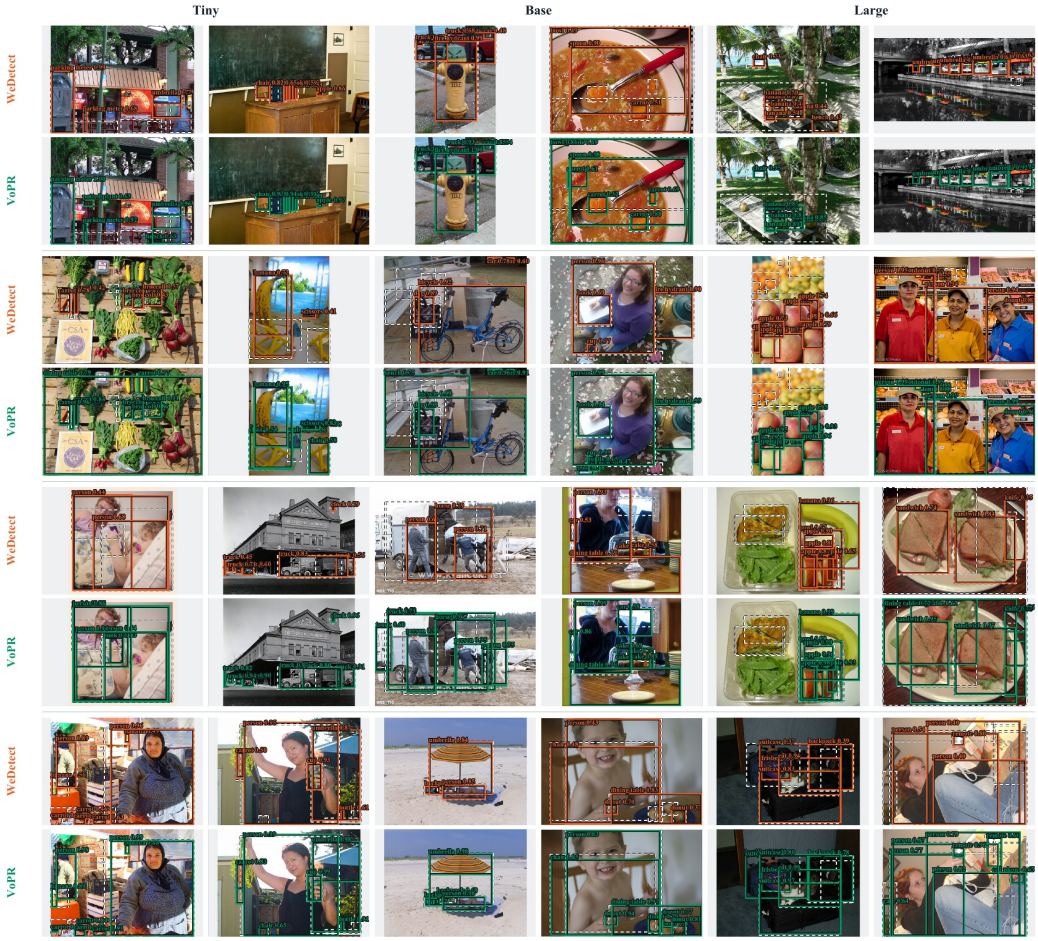

Figure 7: **Additional paired COCO diagnostics, set II.** This disjoint gallery adds 24 images, eight per scale, under the same paired protocol. Together, the two appendix galleries cover 48 images.

Set-II analysis. The disjoint second set (Figure 7) exhibits the same qualitative signature across new objects and scenes: supported detections move upward in the displayed ranking, and several competing boxes are replaced by candidates closer to the eligible ground-truth extent. Recurrence across all three frozen scales is compatible with a shared inference correction rather than checkpoint retraining. The two galleries illustrate the joint calibration of available candidates; full metrics and resampling quantify the paired effect.