DECIDING WHICH DETECTIONS TO RELEASE: BLOCK CALIBRATION FOR OPEN-VOCABULARY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary detectors let users locate objects through text, but a confident prediction for an absent category can become a false alert or an incorrect training label. Reliable use therefore requires deciding which detections to release while retaining useful objects. This decision is complicated by the detector itself: each query produces a dependent set of candidates whose size changes with the image, query, and score filtering. Pooling these candidates for calibration gives queries with more proposals greater weight and can fail to control the average error across queries. We propose block calibration, a postprocessing framework that aligns calibration with this query-level objective. The method keeps each image and its queries together, averages false-detection evidence within queries and across the image, and uses the resulting reference with the e-value Benjamini–Hochberg rule to select detections. Under exchangeability of complete images and their queries, it controls the expected false fraction averaged across queries, allowing score-dependent candidate counts and arbitrary dependence within an image. A maximum-reference variant also controls the expected false fraction in an image's combined output. We further characterize the exact release condition, explaining how evidence strength, candidate count, and rank determine which objects survive. Experiments with GroundingDINO and YOLO-World on COCO, LVIS, and Open Images examine both high-precision release and its cost in object recovery. On COCO/GroundingDINO-tiny, the learned-evidence rule recovers 16.51% of objects at 98.12% pooled precision, more than doubling the recovery of binary evidence at the same nominal error level. Filtering and prompt-bank studies further show how candidate generation shapes the usefulness of error-controlled release.

1 INTRODUCTION

Open-vocabulary detection (OVD) allows users to search for objects simply by naming them, extending object detection to flexible visual search and automatic annotation (Gu et al., 2021; Liu et al., 2024; Cheng et al., 2024). In these applications, however, a requested category need not appear in the image: a system asked to find people and donuts in a street scene must also decide when there is no donut to report. A confident prediction for an absent object can become a false alert or an incorrect training label, and recent evaluations show that hallucinated and semantically incorrect detections persist in capable OVD models (Ruis et al., 2026; Bianchi et al., 2024). Raising a score threshold can suppress these errors, but it also discards genuine objects and does not by itself establish an error guarantee. Figure 1 illustrates the resulting tension: filters remove the same false donut detection while retaining different numbers of people. Reliable use of a frozen detector therefore requires a principled *release* decision that limits false detections while preserving useful output.

Statistical calibration offers a route to this goal through conformal risk control, false-positive-controlled prediction sets, and conformal multiple testing (Angelopoulos et al., 2024; Fisch et al., 2022; Wang et al., 2024; Bashari et al., 2023). The challenge in OVD is to align the calibration unit with the error that matters to the user. Each image–query pair produces a *family* of overlapping candidates whose scores are dependent and whose number changes with the image, query, and score filtering. Pooling these candidates across calibration images gives a query with a hundred boxes far more weight than one with two. A system answering search requests, however, may need the fraction of incorrect results in each answer to carry equal weight. Our counterexample shows that

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.

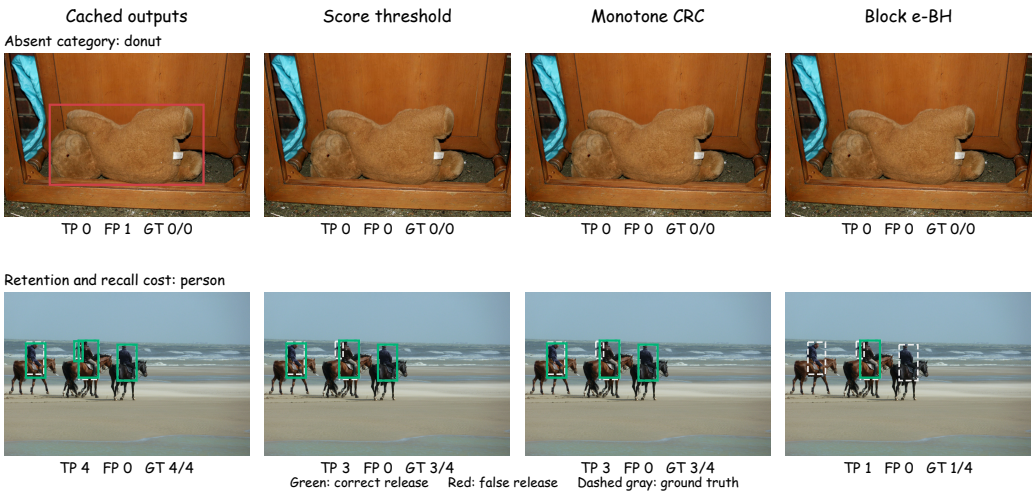

Figure 1: **Release decisions on COCO at $\alpha = 0.05$.** All three filters remove the false donut detection but retain different numbers of people. TP/FP count correct/false releases; GT shows covered/total objects.

this mismatch can invalidate query-average error control even when calibration and test images follow the same distribution (Example 1). This points to a calibration construction that preserves each image and its queries as one block, while accounting for the variable families inside it.

Following this observation, we propose *block calibration*, a postprocessing framework that turns detector scores into release decisions without retraining the detector. A score map learned on separate fitting images assigns evidence to candidates; calibration then averages false-detection evidence within each query family and across its image's queries. This construction gives every query equal weight and supplies a common reference for the e-value Benjamini–Hochberg rule (e-BH) (Wang & Ramdas, 2022). Under exchangeability of complete image blocks, we prove control of the expected false fraction averaged across queries, allowing score-dependent family sizes and arbitrary dependence within an image. A maximum-reference variant also controls the expected false fraction in an image's combined output. To understand when this guarantee leaves useful detections, we derive an exact release condition linking evidence strength, family size, and candidate rank, and extend it to groups of overlapping boxes. The condition exposes a practical bottleneck: weak proposals can increase the evidence needed for release, so removing them can help more surviving objects pass selection.

We examine this connection between error control and object recovery using GroundingDINO and YOLO-World on COCO, LVIS, and Open Images. At nominal level $\alpha = 0.025$, learned evidence on COCO/GroundingDINO-tiny recovers 16.51% of objects at 98.12% pooled precision, compared with 8.07% recall at 99.03% precision for binary evidence. At $\alpha = 0.05$, filtering weak proposals increases recall from 18.62% to 27.55%, consistent with the release condition. Comparisons across calibrated baselines and larger prompt banks identify both attainable trade-offs and settings where little useful output survives. The block construction aligns calibration with the error target, and the release condition explains how evidence and candidate generation determine object recovery. Code and data: `https://anonymous.4open.science/r/OVD-Hallucination-FDR/`.

2 Problem Setup

We split complete images into a fitting set $\mathcal{D}_{\text{fit}}$ that fixes the score map and preprocessing, n calibration images that determine the evidence normalization, and test images whose annotations measure outcomes.

Let Z_i denote an image block: an image x_i, its annotations for the queried categories, and Q queries. A query sampling rule, fixed before calibration and testing, assigns these queries to each image. For query index $q \in \{1, \dots, Q\}$, let K_{iq} be the candidate count, b_{iqk} a candidate box, and S_{iqk} its detector score. The frozen detector and fixed preprocessing return $G(x_i, q) = \{(b_{iqk}, S_{iqk})\}_{k=1}^{K_{iq}}$.

Preprocessing includes a score threshold and non-maximum suppression (NMS), which removes overlapping proposals.

Intersection over union (IoU) measures box overlap as intersection area divided by union area. A candidate is correct if its IoU with a non-crowd ground-truth box of the queried category is at least 0.5. All other candidates are *null*; their indices form $H_{0,iq}$. Multiple candidates may match one object. Object recall counts each matched object once.

Write $a \vee b = \max(a, b)$. For a released set $\mathcal{R}_{iq}$, the false-discovery proportion (FDP) is $\mathrm{FDP}_{iq} = |\mathcal{R}_{iq} \cap H_{0,iq}|/(|\mathcal{R}_{iq}| \vee 1)$, which assigns zero to an empty release. Index 0 denotes a future test image. We define the mean-family false discovery rate (FDR) as

$$\mathrm{mFDR} = \mathbb{E}\left[\frac{1}{Q}\sum_{q=1}^{Q}\mathrm{FDP}_{0q}\,\middle|\,\mathcal{D}_{\mathrm{fit}}\right]. \tag{1}$$

For a target $\alpha \in (0, 1)$ fixed before testing, the goal is $\mathrm{mFDR} \le \alpha$. The expectation averages over calibration and test images and their queries, conditional on the fitting set. We report the observed query average as mean-family FDP (mFDP). Pooled precision counts correct predictions among all releases and gives each release equal weight.

Each query carries equal weight in this target, matching the family and image averages used for calibration below.

Calibration annotations identify correct and null detections for the sampled queries. Calibration and test images follow the same image and query sampling rule, which may assign different query strings to different images. Section 4 specifies the benchmark queries; Appendix C.1 describes verified absence in partially annotated datasets.

3 Method

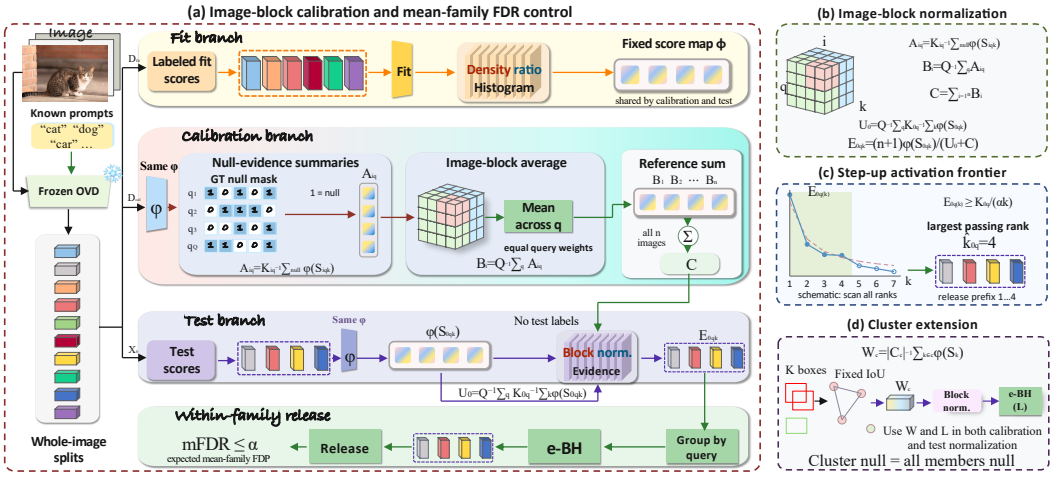

Figure 2: **Block calibration.** Image-level summaries normalize candidate evidence; e-BH selects boxes or groups of overlapping boxes within each query.

Figure 2 shows how fitting, calibration, and selection turn detector scores into release decisions.

3.1 Block calibration and control

Using labeled fit candidates, we learn a finite nonnegative score map $\phi : \mathbb{R} \to [0, \infty)$. Our histogram estimate assigns larger evidence to score ranges more common among correct than null detections (Appendix C). Control holds for any finite nonnegative map fixed on fit data; its separation of correct and null scores affects recall.

Average null and total evidence within each family, then across queries, to obtain the image summaries and calibration reference:

$$A_{iq} = \frac{1}{K_{iq}} \sum_{k \in H_{0,iq}} \phi(S_{iqk}), \quad T_{iq} = \frac{1}{K_{iq}} \sum_{k=1}^{K_{iq}} \phi(S_{iqk}),$$
$$B_i = \frac{1}{Q} \sum_q A_{iq}, \qquad U_i = \frac{1}{Q} \sum_q T_{iq}, \qquad C = \sum_{i=1}^{n} B_i. \tag{2}$$

Both family summaries divide by the total candidate count, including correct candidates. Set $A_{iq} = T_{iq} = 0$ for an empty family; n counts every calibration image, including images with zero summaries. Calibration annotations supply $B_1, \ldots, B_n$; test-time computation uses U_0, which needs no test labels. The common denominator gives the block-normalized evidence scores

$$E_{0qk} = \frac{(n+1)\phi(S_{0qk})}{U_0 + C}. \tag{3}$$

If $U_0 + C = 0$, all scores are set to zero.

For a nonempty family, order the scores as $E_{0q(1)} \geq \cdots \geq E_{0q(K_{0q})}$ and apply the e-BH step-up threshold (Wang & Ramdas, 2022):

$$\widehat{k}_{0q} = \max\left(\{k : E_{0q(k)} \geq K_{0q}/(\alpha k)\} \cup \{0\}\right) \tag{4}$$

and release the first $\widehat{k}_{0q}$ candidates. Empty families release nothing and have FDP zero. Sorting uses a fixed tie-breaking rule.

We assume *exchangeability* of complete calibration and test blocks: reordering them leaves their joint distribution unchanged, conditional on fit. Images sampled independently from a common distribution with the same query rule satisfy this condition.

Example 1 (Candidate pooling gives the wrong weight)**.** *Suppose each image has one query and only false candidates. Two image types are equally likely. Type A produces 100 boxes with raw evidence* $1/100$ *each; type B produces two boxes with evidence* 1 *each. With one calibration image and one independent test image, a candidate-pooled rule at* $\alpha = 0.10$ *releases both boxes when calibration sees A and testing sees B. This occurs with probability* $1/4$*; the other pairings release nothing. Its false-discovery rate is* $1/4$*, above the target. Appendix G.1.2 gives the four calculations. Block calibration uses the image as the exchangeable unit and controls the query-average target under this same image law.*

Proposition 1 (Image-block mean-family FDR control)**.** *Conditional on* $\mathcal{D}_{\text{fit}}$*, assume that the* n *calibration blocks and a test block are jointly exchangeable under the pre-specified image and query law. Apply the same frozen detector, preprocessing, matching, and finite nonnegative* ϕ *to every block. The evidence scores in* (3) *followed by* (4) *satisfy* $\mathbb{E}[Q^{-1}\sum_q \mathrm{FDP}_{0q} \mid \mathcal{D}_{\text{fit}}] \leq \alpha$*. Candidate counts may be random and score-dependent; dependence among candidates and among query families within an image is unrestricted.*

Proof sketch. For every realized nonempty family, the step-up threshold bounds FDP by $\alpha K_{0q}^{-1} \sum_{k \in H_{0,0q}} E_{0qk}$. Average over queries and use $U_i \geq B_i$ to obtain

$$\frac{1}{Q} \sum_q \mathrm{FDP}_{0q} \leq \frac{\alpha(n+1)B_0}{U_0 + C} \leq \frac{\alpha(n+1)B_0}{B_0 + C}. \tag{5}$$

Exchangeability of complete blocks makes the expected normalized share of B_0 at most $1/(n+1)$, which proves the claim. Appendix A covers zero denominators, grouped releases, and multiple test images.

The two averages are essential: dividing by each realized family size prevents a query from gaining weight merely by producing more boxes, and keeping the queries together preserves their dependence within an image. Exchangeability is required only across complete blocks. The realized-output FDP bound therefore accommodates detector-dependent counts without treating the selected boxes as independent calibration samples. The proof extends to general prediction families with the same block structure (Appendix A.1).

The error among released detections. The mean-family target gives each query equal weight. To measure the error in an image's released output, let r_{0q} and v_{0q} be its release and false-release counts for query q. The within-image pooled false-discovery proportion is

$$\mathrm{pFDP}_0 = \frac{\sum_q v_{0q}}{(\sum_q r_{0q}) \vee 1}. \tag{6}$$

This quantity weights each released detection equally. A block reference based on the largest query summary protects it. Set $M_i = \max_q A_{iq}$ for each calibration image and use $E^{\max}_{0qk} = (n+1)\phi(S_{0qk})/(T_{0q} + \sum_{i=1}^{n} M_i)$, with zero scores when the denominator vanishes. The same within-family e-BH rule then applies.

Lemma 1 (Pooled error is bounded by the largest family error)**.** *For any release sets in an image,* $\mathrm{pFDP}_0 \le \max_q \mathrm{FDP}_{0q}$.

Corollary 1 (Expected maximum-family and pooled control)**.** *Under the assumptions of Proposition 1, the maximum-reference rule satisfies* $\mathbb{E}[\max_q \mathrm{FDP}_{0q} \mid \mathcal{D}_{\mathrm{fit}}] \le \alpha$ *and* $\mathbb{E}[\mathrm{pFDP}_0 \mid \mathcal{D}_{\mathrm{fit}}] \le \alpha$.

Appendix B proves the corollary and compares the two references' observed errors.

3.2 When detections are released

Once the test block is observed, the denominator $D_0 = U_0 + C$ is shared by its candidates. Sorting E is equivalent to sorting ϕ. For $D_0 > 0$, define the rank-specific threshold

$$T^{\mathrm{block}}_{q,k} = \frac{K_{0q} D_0}{\alpha k (n+1)}. \tag{7}$$

Theorem 1 (Exact release condition)**.** *For a nonempty family and* $D_0 > 0$*, e-BH releases candidates up to the largest rank* k *satisfying* $\phi(S_{0q(k)}) \ge T^{\mathrm{block}}_{q,k}$*. In particular, at least* r *candidates are released if and only if some* $k \ge r$ *crosses its threshold.*

Substituting (3) into (4) gives this condition. A later crossing releases every candidate above it, including candidates whose own ranks fall below their thresholds. The rule checks the full ordered sequence.

Table 10 illustrates a seven-candidate family at $\alpha = 0.10$: only the leading rank crosses its threshold ($152.30 > 70.0$), releasing one correct box.

At fixed D_0, larger families require stronger evidence and later ranks have lower thresholds. This motivates filtering weak proposals with a *score floor*, the minimum detector score retained throughout fitting, calibration, and testing. We also group overlapping boxes for release.

Cluster release. Within each image–query family, link boxes whose IoU is at least a threshold fixed before calibration and testing. Boxes connected directly or through other boxes form a *cluster*. We release its members together and count the cluster as correct when at least one member is correct. For cluster c with member set $\mathcal{C}_{iqc}$, let $W_{iqc} = |\mathcal{C}_{iqc}|^{-1} \sum_{k \in \mathcal{C}_{iqc}} \phi(S_{iqk})$. A cluster is null if every member is null. Replace the candidate evidence $\phi(S_{iqk})$ by W_{iqc} and the candidate count K_{iq} by the cluster count L_{iq} in (2) and (3). This defines $A^{\mathrm{cl}}_{iq}, T^{\mathrm{cl}}_{iq}, B^{\mathrm{cl}}_{i}, U^{\mathrm{cl}}_{i}$ and E^{cl}_{0qc}, with zero summaries for empty families.

Proposition 2 (Mean-family cluster FDR)**.** *Under the assumptions of Proposition 1, the cluster-normalized evidence from a fixed measurable clustering pipeline, followed by e-BH using* L_{0q}*, controls expected mean-family cluster FDP at* α.

The proof replaces candidate summaries by cluster summaries in (5). Membership can depend on detector scores and boxes: nonnegative evidence gives $U^{\mathrm{cl}} \ge B^{\mathrm{cl}}$ for every partition. The maximum-reference control in Corollary 1 also holds for clusters.

Cluster precision gives each released cluster equal weight. Object recall counts the distinct ground-truth objects covered by its members, counting each object once.

4 EXPERIMENTS

We evaluate frozen GroundingDINO-tiny/base (GD-T/B) (Liu et al., 2024) and YOLO-World (YOLO) (Cheng et al., 2024). The datasets are COCO (Lin et al., 2014), Large Vocabulary Instance Segmentation (LVIS) (Gupta et al., 2019), and Open Images V5 (OI) (Kuznetsova et al., 2020). Our main COCO/GroundingDINO-tiny experiment uses ten 1,000-image splits: 600 fit, 200 calibration, and 200 test images. Each image receives one annotated-present and two verified or annotated-absent category queries. We save detector proposals above score 0.05. The primary rule applies a score floor of 0.40 and clusters boxes at IoU 0.50.

Object recall divides distinct covered objects by all ground-truth objects in the tested families. Each box uses its best ground-truth match. Precision and FDP use clusters as the release unit. We pool recall and precision counts across splits and report split variation. Appendix C gives the full protocol.

The baselines cover expected-risk calibration and multiple testing. We compare monotone conformal risk control (CRC), finite-grid non-monotone CRC (CRC-NM), Learn then Test (LTT), and false-positive-controlled prediction sets (FP-CP). Empirical p-value baselines apply Benjamini–Hochberg (p-BH) or Benjamini–Yekutieli (p-BY) to pooled calibration scores. We also compare a calibrated score threshold and binary evidence, which gives equal weight to every score above a fit-selected threshold. Appendix C.3 states each method's target and assumptions.

4.1 HIGH-PRECISION RELEASE AND MATCHED COMPARISONS

At the high-precision settings in Table 1, block e-BH recovers 16.51% of objects at 98.12% pooled precision. Monotone CRC recovers 8.03% at 97.60% precision. Each row uses a method-specific nominal level from its measured precision–recall curve.

Table 1: High-precision settings on COCO/GroundingDINO-tiny, ten splits. Each method uses a nominal level chosen from its measured curve; rates are percentages and counts are released clusters. LTT: $\delta = 0.05$.

Method	Venue	α	Recall	Precision	mFDP	Total	TP	FP
Empirical p-BH	JRSS B'95	0.05	11.73	98.02	0.128	607	595	12
Empirical p-BY	Ann. Stat.'01	0.075	13.90	97.91	0.214	719	704	15
Monotone CRC	ICLR'24	0.0075	8.03	97.60	0.104	417	407	10
LTT	AoAS'25	0.075	7.53	97.69	0.096	390	381	9
Block e-BH	–	0.025	**16.51**	**98.12**	0.206	852	836	16

For paired outputs from the same setting and split with observed test precision within one percentage point, block e-BH gains 4.65, 6.88, and 4.05 percentage points of recall over empirical p-BH, p-BY, and binary evidence. Appendix Table 15 gives all comparators and confidence intervals; several other baselines recover more objects in this comparison. Both this pairing and Table 1 describe observed trade-offs using test outcomes. The calibration-only study below evaluates a separate decision rule that chooses a release setting before those outcomes are known.

At the common target $\alpha = 0.025$, block e-BH releases detections in all seven settings, while CRC-NM, LTT, and FP-CP release nothing (Table 2). With 200 calibration images, CRC-NM's finite-sample correction prevents any fitted threshold from passing below about 0.0662 (Appendix D.1).

Table 2: Release at $\alpha = 0.025$ (%). R/P: recall/precision. CRC-NM, LTT, FP-CP, and the oracle report maximum tested recall; † uses known present/absent query roles. Bold compares block e-BH and CRC precision.

		Block e-BH			Monotone CRC			CRC-NM	LTT	FP-CP		Oracle†
Detector	Data	R	P	mFDP	R	P	mFDP	R	R	Max	Sum	R
GD-T	COCO	16.51	**98.12**	0.206	39.40	89.21	1.724	0.00	0.00	0.00	0.00	16.80
GD-B	COCO	7.15	**95.15**	0.317	20.57	92.50	1.357	0.00	0.00	0.00	0.00	19.77
GD-B	LVIS	0.33	**79.17**	0.083	3.22	70.15	1.333	0.00	0.00	0.00	0.00	4.47
GD-B	OI	0.15	**100.00**	0.000	5.85	71.99	1.550	0.00	0.00	0.00	0.00	17.16
YOLO	COCO	21.34	**93.67**	0.796	38.79	89.12	1.763	0.00	0.00	0.00	0.00	12.36
YOLO	LVIS	13.62	**91.16**	0.910	20.89	88.72	1.547	0.00	0.00	0.00	0.00	7.13
YOLO	OI	2.10	**90.53**	0.150	6.00	74.39	1.242	0.00	0.00	0.00	0.00	8.44

Figure 3 traces the risk–recall trade-off. At $\alpha = 0.05$, block e-BH has lower mFDP than FP-CP Sum, which has higher recall and precision (Table 12). LTT uses failure probability $\delta = 0.01$; FP-CP outputs are capped at 20. Appendix C.3 gives the baseline target conversions.

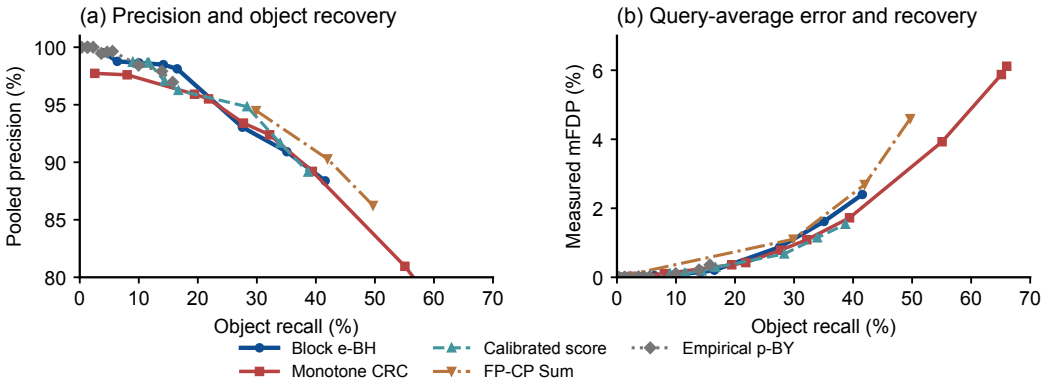

Figure 3: Precision and query-average error versus object recall on COCO/GroundingDINO-tiny. Markers denote evaluated nominal levels; full curves appear in Appendix I.

Choosing a release setting without test labels. A release policy pairs a fitted evidence map with a nominal level. Calibration data choose among these policies, with the score floor and cluster threshold fixed. We select the policy with the most calibration releases among those reaching 83% precision and at most 2.5% mean-family FDP. Across seven settings, pooled test precision exceeds the 80% target in every setting; mean object recall is 19.68%. Splits with no releases contribute zero recall. These are empirical results for an adaptive selection policy: Proposition 1 establishes control for a fixed nominal level and does not itself certify a level selected on the same calibration data. Appendix D.3 gives the selection rule and results; Appendix D.2 studies higher precision targets using fitting images alone.

With the evidence map, test images, and release unit fixed, nested calibration subsets of 25, 50, 100, and 200 images yield 97.53–98.68% pooled precision at $\alpha = 0.025$. Recall rises from 8.95% with 25 images to 16.51% with 200 (Appendix Figure 8).

4.2 Transfer across detectors and datasets

Across the tested detectors and datasets at the common nominal target, block e-BH has higher pooled precision and lower mean-family FDP than monotone CRC, while CRC recovers more objects (Figure 4). Split ranges show the variation behind these pooled values.

4.3 How filtering changes object recovery

Removing weak proposals can improve object recovery. With clustering IoU 0.50 and $\alpha = 0.05$, raising the score floor from 0.05 to 0.40 increases recall from 18.62% to 27.55% on COCO/GroundingDINO-tiny. Appendix D.4 separates changes in candidate coverage from changes in statistical selection. Candidate-pooled studies examine the same design questions across additional detectors and calibration choices (Appendix G).

Figure 5 holds image splits, matching, and the ground-truth recall denominator fixed. A higher floor removes some objects from the candidate set but lets more surviving objects pass selection. Appendix Figure 9(d) separates these losses and gains.

The released clusters also preserve high precision when their member boxes are counted individually. At floor 0.40 and $\alpha = 0.025$, 850 of 852 released clusters contain one box. Member-box precision is 98.13%, alongside cluster precision of 98.12%. At $\alpha = 0.05$, 1,490 of 1,496 released clusters contain one box, with corresponding precisions of 93.08% and 93.05%. These counts connect the cluster-level release rule to the boxes passed on to downstream use.

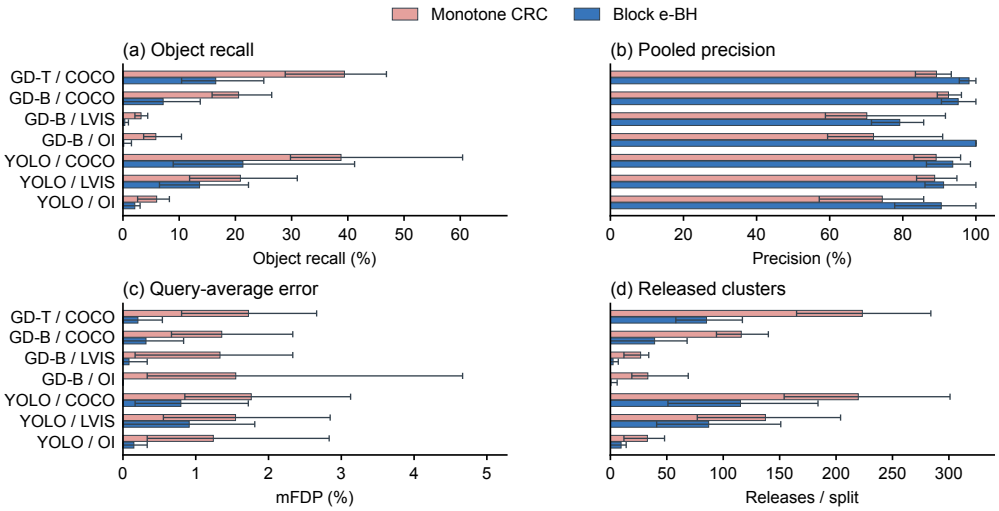

Figure 4: Cross-setting results at $\alpha = 0.025$, ten matched splits. Recall/precision pool counts; mFDP/releases are split means. Whiskers show split ranges; abbreviations match Table 2.

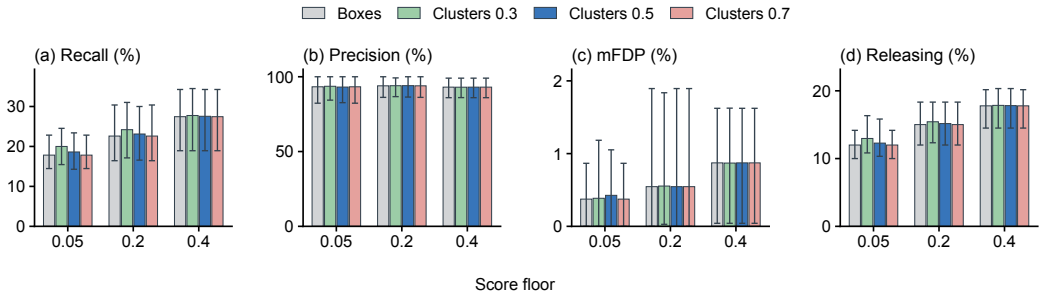

Figure 5: Filtering and release units at $\alpha = 0.05$, ten matched splits. Recall uses a common GT denominator; precision uses the indicated unit. Bars: pooled rates (a,b), split means (c,d). Whiskers: split ranges.

Learning the evidence map. This comparison isolates the score map. Set t_{fit} to the 90th percentile of fitting scores. The binary map $\phi(s) = \mathbf{1}\{s \geq t_{\text{fit}}\}$ gives every higher score equal evidence. The histogram map weights each range by its correct-to-false ratio.

Table 3: COCO evidence maps; rates in %.

Map	α	Recall	Precision	mFDP
Binary	0.025	8.07	99.03	0.067
Learned histogram	0.025	**16.51**	98.12	0.206
Binary	0.05	8.10	99.04	0.067
Learned histogram	0.05	**27.55**	93.05	0.873

At $\alpha = 0.025$, the learned map recovers about twice as many objects as binary evidence, with both precisions above 98% (Table 3). Raising the nominal target to 0.05 changes the binary rule's recall little and increases recall for the learned map. The evidence map supplies the score distinctions that the rank thresholds use to admit more objects. Appendix Figure 11 shows the full curves.

4.4 Prompt-bank scaling and detection outcomes

A prompt bank contains alternative text templates for the same categories. We expand this bank from 80 to 1,280 strings while keeping the vocabulary at 80 categories. Each of three splits divides 200 images into 120 fit, 40 calibration, and 40 test images. Outputs from synonymous prompts are grouped by image and category, using the same matching and clustering rules as the three-query study. Each image now contributes 80 families. Figure 6 varies the calibration-image count with fit and test fixed.

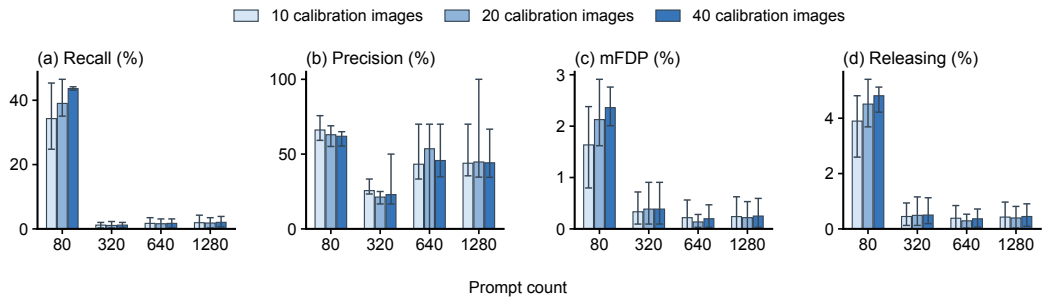

Figure 6: Prompt scaling over 80 categories at $\alpha = 0.10$, three splits. Bars: pooled recall/precision (a,b), mean mFDP/release frequency (c,d). Whiskers: split ranges. Colors indicate calibration-image budgets.

With 40 calibration images, expanding the bank from 80 to 1,280 prompts reduces recall from 43.62% to 2.04%, with fewer families releasing detections (Figure 6). Appendix E examines a separate one-category calibration design.

Figure 1 and Appendix Figure 19 show retained objects, suppressed detections, and remaining errors on identical detector outputs.

5 Discussion and Conclusion

Block calibration makes the image and its queries the unit of calibration, aligning the reference distribution with the error target used to evaluate release decisions. Under exchangeable blocks, the resulting guarantee allows candidate counts to depend on scores and permits arbitrary dependence within an image. The exact release condition then connects this guarantee to useful output: evidence strength, family size, and rank jointly determine which detections survive. The experiments show that learned evidence and removal of weak proposals can improve recovery, while larger prompt banks can leave very few objects to release.

The choice of error target matters when interpreting these results. Mean-family FDP weights each query equally, including queries that release nothing; pooled precision gives each released detection equal weight. A method can therefore have a low query-average error while placing more errors in families with many releases. This explains why the risk and precision curves need not rank methods in the same order. The maximum-reference construction protects the expected pooled error within an image by controlling its largest family FDP. Neither result is a deterministic precision guarantee for every query, nor a guarantee for precision pooled over an entire dataset. Applications should choose the reference according to the output whose errors they need to control and report recovery alongside the protected risk.

The guarantee also depends on preserving the image and query law between calibration and use. New prompt templates, a different mixture of present and absent categories, or a changed preprocessing rule can alter both the candidate families and their null evidence. Such changes call for a matching calibration design; robustness to arbitrary distribution shift does not follow from within-image dependence tolerance. Calibration further requires annotations that identify false candidates for the sampled queries. Verified absence is therefore relevant to the statistical construction, particularly on partially annotated datasets. Appendix Figures 15 and 16 examine annotation perturbations and query-law changes.

Finally, controlled release cannot recover an object absent from the detector's proposals. Filtering trades candidate coverage against the ability of the surviving evidence to cross the release threshold, and the fitted score map determines how well correct detections separate from false ones. Our calibration-budget and prompt-bank studies show why error control alone is insufficient to characterize usefulness. A practical evaluation must also measure how many objects survive, how often the rule releases nothing, and how those outcomes change with the intended query bank. Block calibration provides an error-controlled release layer for a frozen detector; its value ultimately depends on the detections available to that layer.

Ethics Statement

Filtering false detections can reduce incorrect training labels and unnecessary review. Withheld genuine objects also matter in monitoring and annotation tasks. Consequential uses should retain human review and account for both kinds of error.

AI Use Statement

ChatGPT was used only for literature search. The authors are responsible for the final content.

References

Tareq Aldirawi, Yun Li, and Wenge Guo. Conformal risk control under non-monotone losses: Theory and finite-sample guarantees. *arXiv preprint arXiv:2604.01502*, 2026. URL https://arxiv.org/abs/2604.01502.

Anastasios N. Angelopoulos. Conformal risk control for non-monotonic losses. *arXiv preprint arXiv:2602.20151*, 2026. URL https://arxiv.org/abs/2602.20151.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Tian Bai and Ying Jin. Conformal selective prediction with general risk control. *arXiv preprint arXiv:2603.24704*, 2026. URL https://arxiv.org/abs/2603.24704.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Meshi Bashari, Amir Epstein, Yaniv Romano, and Matteo Sesia. Derandomized novelty detection with fdr control via conformal e-values. *Advances in Neural Information Processing Systems*, 36: 65585–65596, 2023.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Yoav Benjamini, Abba M Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *biometrika*, pp. 491–507, 2006.

Lorenzo Bianchi, Fabio Carrara, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22520–22529, 2024.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16901–16911, 2024.

Jang Hyun Cho and Philipp Krähenbühl. Language-conditioned detection transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16593–16603, 2024.

Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14084–14093, 2022.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Conformal prediction sets with limited false positives. In *International Conference on Machine Learning*, pp. 6514–6532. PMLR, 2022.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1100–1126, 2025.

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.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5356–5364, 2019.

Nikolaos Ignatiadis, Bernd Klaus, Judith B Zaugg, and Wolfgang Huber. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, 13(7): 577–580, 2016.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

Lihua Lei and William Fithian. Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):649–679, 2018.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10965–10975, 2022.

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 *European conference on computer vision*, pp. 740–755. Springer, 2014.

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 *European conference on computer vision*, pp. 38–55. Springer, 2024.

Ariane Marandon, Lihua Lei, David Mary, and Etienne Roquain. Machine learning meets false discovery rate. *arXiv preprint arXiv:2208.06685*, 2, 2022.

Yannis Montreuil, Shu Heng Yeo, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. A two-stage learning-to-defer approach for multi-task learning. *arXiv preprint arXiv:2410.15729*, 2024.

Frank Ruis, Gertjan J. Burghouts, and Hugo Kuijf. Fantastic tractor-dogs and how not to find them with open-vocabulary detectors. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jUuXNrG7wh.

Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. Conformal prediction for zero-shot models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19931–19941, 2025.

John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):479–498, 2002.

Alexander Timans, Christoph-Nikolas Straehle, Kaspar Sakmann, and Eric Nalisnick. Adaptive bounding box uncertainties via two-step conformal prediction. In *European Conference on Computer Vision*, pp. 363–398. Springer, 2024.

Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852, 2022.

Xiaoning Wang, Yuyang Huo, Liuhua Peng, and Changliang Zou. Conformalized multiple testing after data-dependent selection. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1867.

Mengmeng Xu, Yancheng Bai, Bernard Ghanem, Boxiao Liu, Yan Gao, Nan Guo, Xiaochun Ye, Fang Wan, H You, D Fan, et al. Missing labels in object detection. In *CVPR workshops*, volume 3, 2019.

Yuewei Yang, Kevin J Liang, and Lawrence Carin. Object detection as a positive-unlabeled problem. *arXiv preprint arXiv:2002.04672*, 2020.

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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16793–16803, 2022.

SUPPLEMENTARY MATERIAL

The appendices follow the method's construction and evaluation. Appendices A–B give the proofs and stronger error target. Appendices C–D specify the protocol and extend the matched comparisons. Prompt-bank and annotation studies follow. Appendix G groups the candidate-pooled experiments by the design questions they address; Appendix H discusses related work.

A PROOFS OF MAIN RESULTS

A.1 AVERAGE NULL EVIDENCE AND ZERO-SIZE FAMILIES

The family and block summaries used in the main construction are

$$A_{iq} = \frac{1}{K_{iq}} \sum_{k \in H_{0,iq}} \phi(S_{iqk}), \qquad T_{iq} = \frac{1}{K_{iq}} \sum_{k=1}^{K_{iq}} \phi(S_{iqk}), \tag{8}$$

with both summaries zero when $K_{iq} = 0$, and

$$B_i = \frac{1}{Q} \sum_q A_{iq}, \qquad U_i = \frac{1}{Q} \sum_q T_{iq}, \qquad C = \sum_{i=1}^{n} B_i. \tag{9}$$

For $K_{iq} > 0$,

$$T_{iq} - A_{iq} = \frac{1}{K_{iq}} \sum_{k \notin H_{0,iq}} \phi(S_{iqk}) \geq 0.$$

For $K_{iq} = 0$ both summaries are zero. The inequality holds for every matching outcome, including detector-dependent null sets. It gives $U_i \geq B_i$ without conditioning on the realized candidates.

For a realized nonempty test family, a released null candidate has evidence at least $K_{0q}/(\alpha|\mathcal{R}_{0q}|)$. Summing over released null candidates gives

$$\mathrm{FDP}_{0q} \leq \frac{\alpha}{K_{0q}} \sum_{k \in H_{0,0q}} E_{0qk}. \tag{10}$$

Averaging over q and substituting (3) yields

$$\frac{1}{Q} \sum_q \mathrm{FDP}_{0q} \leq \frac{\alpha(n+1)B_0}{U_0 + C} \leq \frac{\alpha(n+1)B_0}{B_0 + C},$$

where the second inequality uses $U_0 \geq B_0$. Empty families contribute zero.

Let $S_B = \sum_{i=0}^{n} B_i$ and define $B_i/S_B = 0$ when $S_B = 0$. Conditional on fit, joint block exchangeability gives

$$\mathbb{E}\left[\frac{B_i}{S_B} \,\middle|\, \mathcal{D}_{\mathrm{fit}}\right] = \mathbb{E}\left[\frac{B_0}{S_B} \,\middle|\, \mathcal{D}_{\mathrm{fit}}\right].$$

Summing over i yields

$$(n+1)\mathbb{E}\left[\frac{B_0}{S_B} \,\middle|\, \mathcal{D}_{\mathrm{fit}}\right] = \Pr(S_B > 0 \mid \mathcal{D}_{\mathrm{fit}}) \leq 1.$$

If $U_0 + C = 0$, all evidence scores are zero and e-BH makes no releases. If $S_B = 0$ but $U_0 + C > 0$, then $B_0 = 0$ and all test-null evidence is zero. Again the FDP bound holds. This completes all cases in Proposition 1.

For multiple test images sharing calibration blocks, the proposition holds for each calibration–test block tuple with the same exchangeable law. Linearity gives

$$\mathbb{E}\left[\frac{1}{N} \sum_{j=1}^{N} \frac{1}{Q} \sum_{q=1}^{Q} \mathrm{FDP}_{jq} \,\middle|\, \mathcal{D}_{\mathrm{fit}}\right] \leq \alpha.$$

No independence among test images is needed. Algebraically the empirical mean is

$$\frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{Q}\sum_{q=1}^{Q}\mathrm{FDP}_{jq}\right)=\frac{1}{NQ}\sum_{j=1}^{N}\sum_{q=1}^{Q}\mathrm{FDP}_{jq}\,.$$

For unequal numbers of queries per image, the unmodified block mean weights images equally. A uniform mean over all families would require an explicitly aligned weighting law.

General prediction families. Let $Z_0,\dots,Z_n$ be complete data blocks that are jointly exchangeable conditional on fit. A common prediction rule, fixed conditional on fit, produces Q finite families and finite nonnegative evidence weights from each block's observed input. A common annotation rule identifies null members. Family sizes may depend on the input, and dependence within a block is unrestricted. Replace $\phi(S_{iqk})$ by these weights in (2) and (3), keeping the empty-family conventions. Within-family e-BH then satisfies $\mathbb{E}[Q^{-1}\sum_q \mathrm{FDP}_{0q} \mid \mathcal{D}_{\mathrm{fit}}] \le \alpha$. The proof above applies directly: the common rules preserve exchangeability of $B_0,\dots,B_n$, and nonnegative weights give $U_i \ge B_i$. In OVD, the detector supplies the families, box matching defines correctness, and clustering chooses the release unit.

A.2 Cluster proof

The cluster normalization is

$$E^{\mathrm{cl}}_{0qc}=\frac{(n+1)W_{0qc}}{U^{\mathrm{cl}}_0+\sum_{i=1}^{n}B^{\mathrm{cl}}_i}. \tag{11}$$

A cluster's evidence W_{iqc} is nonnegative. Its null set is a subset of all clusters, so $T^{\mathrm{cl}}_{iq}\ge A^{\mathrm{cl}}_{iq}$ for each realized partition. Set both summaries to zero for $L_{iq}=0$. The same deterministic mapping from a complete block to B^{cl}_i is applied to calibration and test blocks. Joint exchangeability is preserved by this mapping.

For a nonempty family,

$$\mathrm{FDP}^{\mathrm{cl}}_{0q}\le\frac{\alpha}{L_{0q}}\sum_{c:\,\mathcal{C}_{0qc}\ \text{null}}E^{\mathrm{cl}}_{0qc}.$$

Averaging over q gives the analogue of (5), and the normalized-share identity proves Proposition 2. This proof averages raw nonnegative evidence within each cluster before block calibration. Averaging marginal candidate e-values over an output-dependent cluster would require a separate validity argument.

A.3 Rank thresholds

The rank-specific threshold is

$$T^{\mathrm{block}}_{q,k}=\frac{K_{0q}D_0}{\alpha k(n+1)}. \tag{12}$$

For $D_0>0$, $E_{0qk}=(n+1)\phi(S_{0qk})/D_0$ preserves the ordering of ϕ. At rank k,

$$E_{0q(k)}\ge\frac{K_{0q}}{\alpha k}\quad\Longleftrightarrow\quad\phi(S_{0q(k)})\ge\frac{K_{0q}D_0}{\alpha k(n+1)}.$$

Taking the maximum crossing rank proves Theorem 1. If $D_0=0$, the zero-score convention produces an empty set. Replacing K by L and ϕ by W gives the cluster release condition.

A.4 Common and unequal template mixtures

Proposition 3 (Common-mixture evidence bound)**.** *For each template t, let p_t and $q_{0,t}$ be fit-reference and null-score densities with p_t absolutely continuous with respect to $q_{0,t}$. Use the same fixed weights $\pi_t\ge 0$, $\sum_t\pi_t=1$, in $p_{\mathrm{pool}}=\sum_t\pi_t p_t$ and $q_{0,\mathrm{pool}}=\sum_t\pi_t q_{0,t}$. Where $q_{0,\mathrm{pool}}(s)>0$,*

$$\phi_{\mathrm{pool}}(s)=\sum_t\mu_t(s)\phi_t(s),\quad \mu_t(s)=\frac{\pi_t q_{0,t}(s)}{\sum_\ell\pi_\ell q_{0,\ell}(s)},\quad \phi_t(s)=\frac{p_t(s)}{q_{0,t}(s)}. \tag{13}$$

The weights form a convex combination, so $\phi_{\mathrm{pool}}(s) \le \max_t \phi_t(s)$ *and* $\operatorname{ess\,sup} \phi_{\mathrm{pool}} \le \max_t \operatorname{ess\,sup} \phi_t$.

With the common-mixture assumptions in Proposition 3,

$$\frac{\sum_t \pi_t p_t(s)}{\sum_t \pi_t q_{0,t}(s)} = \sum_t \frac{\pi_t q_{0,t}(s)}{\sum_\ell \pi_\ell q_{0,\ell}(s)} \frac{p_t(s)}{q_{0,t}(s)}.$$

Terms with zero null density have zero weight and are assigned zero product; absolute continuity excludes unexplained numerator mass at those points. The weighted-average inequality gives the claimed bound. Equality may hold whenever positive-weight terms attain the same pointwise maximum; identical ratios on the entire support are unnecessary.

If numerator weights are ω_t and denominator weights are $\pi_t > 0$ for every $\omega_t > 0$, then

$$\phi_{\mathrm{pool}}(s) = \sum_t \frac{\omega_t q_{0,t}(s)}{\sum_\ell \pi_\ell q_{0,\ell}(s)} \phi_t(s) \le \left(\max_{t:\omega_t>0} \frac{\omega_t}{\pi_t} \right) \max_t \phi_t(s).$$

Compression affects release power through the evidence map. The block guarantee holds for a fixed nonnegative map under block exchangeability. Template-conditioned calibration uses exchangeability within each template and calibration samples for each group.

B STRONGER ERROR TARGETS

Proof of Lemma 1. When no query releases a detection, both sides are zero. Otherwise, $\mathrm{pFDP}_0 = \sum_q w_q \mathrm{FDP}_{0q}$ with nonnegative weights $w_q = r_{0q} / \sum_\ell r_{0\ell}$ summing to one. A weighted average cannot exceed its largest term. □

Proof of Corollary 1. The query collection is finite and pre-specified, with its sampling included in the complete-block law. Write $C_M = \sum_{i=1}^n M_i$. For each family with positive denominator,

$$\mathrm{FDP}_{0q} \le \frac{\alpha(n+1)A_{0q}}{T_{0q} + C_M} \le \frac{\alpha(n+1)M_0}{M_0 + C_M}.$$

For $C_M > 0$, this uses $T_{0q} \ge A_{0q}$, $A_{0q} \le M_0$, and the monotonicity of $a/(a + C_M)$. If $C_M = 0$ and $M_0 > 0$, the last ratio is one. If $M_0 = C_M = 0$, test-null evidence vanishes and every family FDP is zero. The bound holds for all queries at once. Exchangeability of $M_0, \ldots, M_n$ makes the expected normalized share of M_0 at most $1/(n+1)$. Lemma 1 gives the pooled bound. Replacing candidate summaries by cluster summaries preserves the same argument. □

For any score-measurable choice of one of the realized queries, its FDP is bounded by the family maximum. The corollary also bounds the mean within-image pooled FDP over several deployment images by linearity.

Table 4: Optional maximum-reference rule on ten COCO/GroundingDINO-tiny splits, fixed floor 0.40 and $\alpha = 0.10$. The box and cluster recall denominators differ as stated.

Release unit	Released	Correct / false	Precision	Recall
Boxes	1,099	1,058 / 41	96.27%	13.93% candidate
IoU clusters	1,102	1,061 / 41	96.28%	20.98% object

Both rows have 1.554% mean block-maximum FDP and nonzero releases in every split.

The two error summaries measure different views of the same output. Across seven settings and ten splits per setting at $\alpha = 0.025$, the mean-reference rule has 0.352% query-average FDP and 1.007% within-image pooled FDP. The maximum-reference rule has 0.046% within-image pooled FDP on those same settings and splits.

Table 5: Main block experiment. All preprocessing choices are shared by fit, calibration, and test.

Component	Specification
Detector and data	Frozen GroundingDINO-tiny, COCO val2017
Sampling	Ten splits, seeds 0–9; 1,000 images per split
Queries	One uniformly sampled annotated-present category and two sampled absent categories per image
Split	600/200/200 complete fit/calibration/test images; 1,800/600/600 families
Evidence	Up to 80 quantile bins, add-one class counts, density-ratio cap 10^6, learned from fit labels
Candidate floor	Detector outputs retained from 0.05; analysis floor 0.40 applied before fitting and calibration
Clusters	Within-family connected components of the IoU ≥ 0.50 graph; mean member evidence
Risk target	Expected arithmetic mean of query-family FDPs under the image-block and query law
Test annotations	Used for evaluation only

C Benchmark protocol and glossary

Family sizes and cluster counts include all surviving candidates or clusters. Empty families remain in the empirical average. Calibration blocks whose entire reference equals zero remain in n. Each individual split contains 600 test families; the ten-split mean weights all 6,000 family occurrences equally.

Candidate recall uses pre-floor correct candidates as its denominator. For object recall, each ground-truth instance is counted once within a tested image–category family if at least one released cluster covers it. Across repeated splits, counts refer to test occurrences. The cluster experiment has 5,071 ground-truth object occurrences and the no-floor cluster rule covers 1,234. A released cluster may cover more than one object; correct-cluster count and covered-object count are distinct.

C.1 Verified negative queries in LVIS and Open Images

The LVIS and Open Images rows in Table 2 use the same three-query block structure as COCO. For LVIS, each image contributes one annotated-present category and two verified-negative categories. Categories marked as incompletely annotated are excluded from the absent-query pool. For Open Images V5, each image contributes one boxed positive category and two categories with verified negative human image labels (confidence zero). The three queries carry equal weight and remain together in fit, calibration, and test splits. These metadata define the admitted families; matching uses the same non-crowd IoU protocol as the COCO block experiment.

C.2 Disjoint-image check

We formed a separate COCO check from images appearing in exactly one of the ten saved splits. Fitting and calibration use split 0; the holdout contains 228 test images from other splits, with no image overlap. At $\alpha = 0.025$, the frozen rule releases 69 clusters, of which 64 are correct. Pooled precision is 92.75%, and mean-family FDP is 0.731%. The measured mean-family FDP stays below its nominal level at all five tested levels; the smallest level makes no releases.

C.3 Control targets and calibration units

The table distinguishes expected-error control from high-probability bounds and count-based targets. Here i.i.d. means independently and identically distributed.

The stability-based construction of Angelopoulos (2026) bounds non-monotone risk through algorithmic stability; its correction differs from the finite-grid CRC-NM implementation evaluated here. The risk-adjusted e-value method SCoRE (Bai & Jin, 2026) controls risk among cases selected as trusted. Our calibrated unit is the complete image block, containing its detector-selected families.

Table 6: Control targets and assumptions. Three-query studies use 200 calibration images unless varied; maps and policy grids use fit data.

Method	Target	Cal. / release unit	Conditions
Block e-BH	Expected mean family FDP	Image block / query clusters	Exchangeable blocks; average null-evidence budget
Monotone CRC	Expected bounded risk	Image block / thresholded clusters	Monotone upper envelope of non-monotone FDP
CRC-NM, finite grid	Expected bounded risk	Image block / fixed policies	Fit-searched threshold or no release; finite-sample correction under i.i.d. blocks
LTT	High-probability risk bound converted to expectation	Image block / thresholded clusters	Fit-ordered fixed sequence; bounded i.i.d. losses
FP-CP Max/Sum	Expected false-positive count	Image block / nested cluster sets	Exchangeable blocks; count budget 3α bounds three-query mean FDP
Pooled empirical p-BH/BY	Empirical operating comparison	Null clusters / query clusters	Scores calibrated by pooling null candidates
Query-slot BY/Bonferroni	Fixed-cap family error	Slot-conditional blocks / capped clusters	Uses known present/absent query positions
Stability-based CRC (Angelopoulos, 2026)	Expected loss for non-monotone algorithms	Exchangeable data units	Requires a justified stability correction; discrete FDP needs an appropriate bound
SCoRE (Bai & Jin, 2026)	Marginal or selective deployment risk	Calibration cases / selected trusted cases	Risk-adjusted e-values; an OVD adaptation must specify the case and loss

Table 7: Terminology used in the method and experiments.

Term	Meaning
Family	Candidate boxes from one image–query invocation
Image block	One image and all queries assigned to it by the sampling rule
Null candidate	A box labeled unmatched by the fixed annotation and matching protocol
FDP	False discovery proportion: false releases divided by releases, zero with no releases
Mean-family FDR	Expected mean of within-family FDPs under the specified law
Pooled precision	Correct releases divided by all releases at the stated release unit
α	Pre-specified target level for the protected expectation
$\phi(s)$	Nonnegative evidence learned from a detector score on the fit split
E	Block-normalized evidence score used by e-BH
e-BH	E-value Benjamini–Hochberg step-up rule
NMS	Non-maximum suppression, detector postprocessing that suppresses overlapping proposals
IoU	Intersection over union of two boxes
Exchangeability	Invariance of the joint distribution under permutation of complete image blocks, conditional on fit
Validity and power	Respecting the expected error target, and the ability to release correct detections, respectively

For the worked example, $\sum_k \phi(S_{0qk}) = 344.05$ up to displayed rounding, so $T_{0q} \approx 49.15$. The other two families complete the block mean $U_0 \approx 17.04$. The multiplier uses unrounded saved evidence in evaluation; Table 10 rounds values for readability.

D Extended matched experiments

Table 2 reports the block-average cluster rule at $\alpha = 0.025$, floor 0.40. The GroundingDINO-tiny/COCO and YOLO-World/COCO studies use ten sampled 1,000-image pools with a union of 4,424 images. The other settings re-partition one 1,000-image pool ten times. These repetitions characterize sensitivity to splitting.

Table 8: Release-unit metrics on COCO/GroundingDINO-tiny at $\alpha = 0.10$, without a score floor. The rows use different recall denominators: candidate recall for boxes and object recall for IoU clusters.

Release unit	Recall metric	Recall	mFDP
Boxes	Candidate recall	15.84%	0.788%
IoU clusters	Object recall	24.33%	0.771%

Table 9: Image-block cluster operating curve with fixed score floor 0.40. Pooled precision and object recall describe utility; cluster mFDP estimates the protected expectation.

α	Released	Correct / false	Object recall	Pooled precision	Cluster mFDP
0.025	852	836 / 16	16.51%	98.12%	0.206%
0.050	1,496	1,392 / 104	27.55%	93.05%	0.873%
0.075	1,949	1,772 / 177	35.08%	90.92%	1.618%
0.100	2,376	2,100 / 276	41.57%	88.38%	2.399%

The comparisons include the two-policy CRC-NM rule (Aldirawi et al., 2026), LTT (Angelopoulos et al., 2025), and FP-CP Max and Sum (Fisch et al., 2022). LTT's high-probability bound is converted to an expectation target. FP-CP uses a false-positive count budget of 3α, which bounds the mean FDP of three families. Appendix C.3 states the calibration conditions. The query-slot baseline uses the known present/absent positions as an oracle comparison. Table 2 gives the shared-target results; Table 11 gives the monotone CRC operating curve.

D.1 Finite-sample baseline settings

For monotone-envelope CRC, fit evidence supplies at most 199 quantile thresholds plus a no-release policy. Increasing the included set gives a sequence of per-image mean-family FDP losses. The cumulative maximum along this sequence is a monotone upper envelope in $[0, 1]$. Calibration selects the largest policy whose corrected empirical envelope risk is at most α, with the usual $1/(n + 1)$ finite-sample term.

For CRC-NM, the fit stage searches distinct evidence thresholds and proposes one. Calibration compares this proposal with no release, giving two calibration policies. With $n = 200$, its correction is

$$D = \frac{1}{2}\sqrt{\frac{2\log(2m)}{n}} + \frac{\log(2m)}{3n}, \qquad m = 2.$$

The calibration cutoff is $((n+1)(\alpha-D)-1)/n$. Here $D = 0.06118$, so the cutoff can be nonnegative only when $\alpha \geq D + 1/(n+1) = 0.06616$. It is -0.04136 at $\alpha = 0.025$ and -0.01624 at 0.05. The searched threshold cannot pass at either level. On the ten-level grid, CRC-NM first releases at 0.075. Block e-BH has no fixed additive correction of this form and first releases at 0.0075. At $\alpha = 0.025$, it averages 48.5 releases per setting and split across the seven settings.

LTT uses at most 199 fit-evidence quantile thresholds plus no release, with a fixed-sequence testing order learned from fit risk and recall. For each fixed $\delta < \alpha$ in $\{0.005, 0.01, 0.02, 0.025, 0.03, 0.04, 0.05, 0.075\}$, the calibration target is $\tau = (\alpha - \delta)/(1 - \delta)$. A probability-$1 - \delta$ bound on selected risk at τ, together with losses bounded by one, gives expected risk at most $(1 - \delta)\tau + \delta = \alpha$. Reported comparisons consider each fixed (α, δ) configuration separately. Rows selected for best test performance summarize the observed range across configurations.

D.2 Selecting operating points before testing

We divide each split's 600 fitting images into 400 for learning the score map, 100 for preliminary calibration, and 100 for development. A fixed random seed determines this division. Every method uses the nominal grid $\{0.005, 0.0075, 0.01, 0.0125, 0.015, 0.02, 0.025, 0.05, 0.075, 0.10\}$. We select the highest-recall point reaching the requested development precision, 95% or 98%. Ties prefer higher precision, then smaller α; if no point qualifies, the rule releases nothing. We refit the maps and policy grids on all 600 fitting images, calibrate on 200 separate images, and evaluate on 200 test images. The study reuses the benchmark image pools with these separate roles within each split.

Table 10: Complete seven-candidate worked example, ordered by evidence. The detector and evidence rankings differ because ϕ need not be monotone.

Evidence rank	Detector rank	Score	$\phi(s)$	E	$7/(0.1k)$	Label	Release
1	1	0.8725	342.15	152.30	70.0	Correct	Yes
2	5	0.0617	0.47	0.21	35.0	Null	No
3	2	0.0833	0.37	0.16	23.3	Null	No
4	3	0.0700	0.34	0.15	17.5	Null	No
5	6	0.0548	0.27	0.12	14.0	Correct	No
6	7	0.0510	0.27	0.12	11.7	Null	No
7	4	0.0621	0.18	0.08	10.0	Null	No

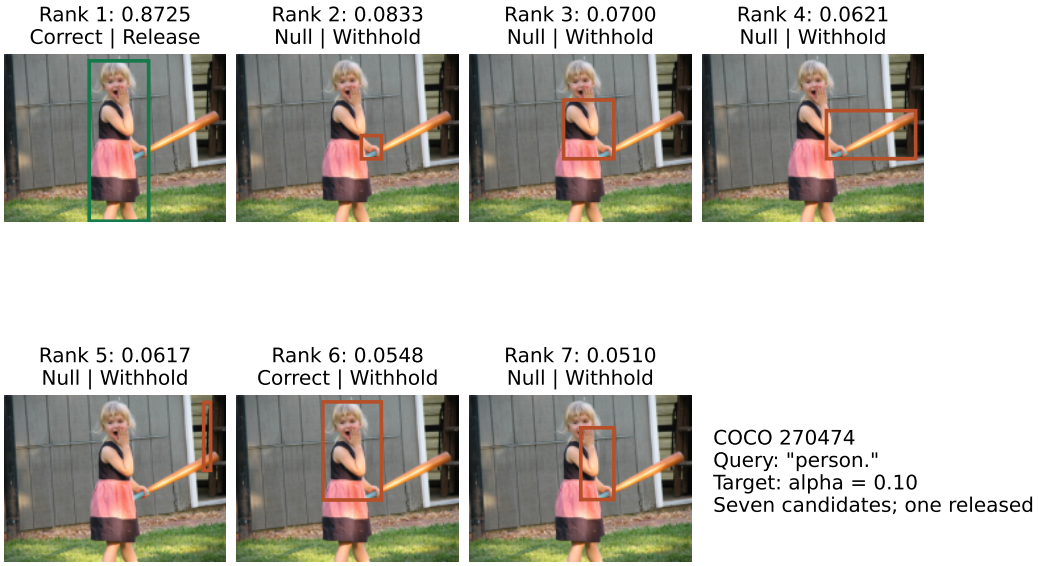

Figure 7: Seven proposals in the worked example: green is released; orange is withheld. Ranks 1 and 6 both match the same person at IoU ≥ 0.5, counted once for object recall.

LTT fixes $\delta = 0.01$ and FP-CP fixes cap 20. The score-threshold comparator chooses a fit-grid threshold attaining calibration precision $1 - \alpha$. The binary-evidence ablation replaces the fitted density ratio by $\phi(s) = \mathbf{1}\{s \geq t_{\text{fit}}\}$, where t_{fit} is the 90th percentile of fit scores. Both ablation and full method retain block normalization and e-BH.

Development selection assigns each method an operating point before test outcomes are measured. Tables 13 and 14 report the resulting recall and precision for every method and setting. The achieved precision varies across datasets and can fall below the development target. This experiment complements the full precision–recall curves by evaluating the selection rule itself.

D.3 Calibration-only selection

The score floor is fixed at 0.40 and the cluster IoU threshold at 0.50. We compare the histogram map with two models of candidate correctness: a monotone isotonic fit and a logistic model. All three maps use fitting images. For each setting and split, the rule searches nominal levels $\{0.025, 0.05, 0.075, 0.10, 0.15, 0.20, 0.30\}$. It keeps policies whose calibration precision is at least 83% and whose calibration mean-family FDP is at most 2.5%, then chooses the one with the most calibration releases. A split with no eligible policy releases nothing.

We reran level selection after changing test labels, with the histogram map, detector outputs, and calibration data fixed. For five splits in each of the seven settings, we inverted the labels, randomly flipped them, and marked every test detection false. The chosen level and calibration statistics stayed identical in all 35 cases.

Table 11: Monotone-envelope CRC on the same COCO/GroundingDINO-tiny block cluster data. Each row calibrates a fit-defined threshold grid at the listed target.

α	Object recall	Pooled precision	mFDP	Releases
0.0050	2.54%	97.73%	0.028%	132
0.0075	8.03%	97.60%	0.104%	417
0.0100	19.44%	95.91%	0.363%	1,028
0.0125	21.81%	95.51%	0.426%	1,158
0.0150	27.73%	93.40%	0.784%	1,501
0.0200	32.16%	92.39%	1.088%	1,761
0.0250	39.40%	89.21%	1.724%	2,233
0.0500	55.08%	80.95%	3.928%	3,439

Table 12: Risk allocation on COCO/GroundingDINO-tiny at $\alpha = 0.05$. Precision is undefined for zero releases.

Method	Object recall	Precision	mFDP
Block e-BH	27.55%	93.05%	0.873%
CRC (monotone envelope)	55.08%	80.95%	3.928%
LTT ($\delta = 0.01$)	16.72%	95.50%	0.360%
CRC-NM (two policies)	0.00%	–	0.000%
FP-CP Max (cap 20)	24.14%	94.58%	0.600%
FP-CP Sum (cap 20)	29.88%	94.49%	1.103%

D.4 Evidence, multiplicity, and object recovery

Figure 5 compares object recovery, precision, mFDP, and release frequency as the score floor and release unit change. The fitted evidence map can change detector ranking, and a later step-up crossing can release an entire prefix. In Figure 9(c), the evidence-to-threshold ratios are 0.61 and 1.14. The second rank crosses its threshold, so both clusters are released.

The floor-by-clustering study holds image splits, matching IoU, and GT denominators fixed. At $\alpha = 0.05$ and clustering IoU 0.50, increasing the floor from 0.05 to 0.40 raises object recall from 18.62% to 27.55%. Figure 9(d) explains this gain. The higher floor removes all candidates for some objects and lets more of the surviving objects pass statistical selection.

The saved release sets also show how cluster-level precision relates to the member boxes. At floor 0.40 and $\alpha = 0.025$, 850 of the 852 released clusters contain one box. Counting every released member gives box precision of 98.13%, alongside cluster precision of 98.12%. At $\alpha = 0.05$, 1,490 of 1,496 released clusters contain one box; the corresponding precisions are 93.08% and 93.05%. These counts use the same candidate matching as the main experiment.

Table 13: COCO/GroundingDINO-tiny at fixed levels and after development selection. Rates are percentages; R/P means object recall/pooled precision. *Below the development target; – denotes zero-release precision. Raw outputs include all post-floor clusters.

	$\alpha = 0.025$			$\alpha = 0.05$			Dev. target	
Method	R	P	mFDP	R	P	mFDP	95%: R/P	98%: R/P
All cached clusters	66.02	73.37	6.116	66.02	73.37	6.116	0.00/–	0.00/–
Calibration-tuned score	16.70	96.25	0.287	28.34	94.83	0.677	23.60/94.46*	13.23/94.64*
Empirical p-BH	6.31	99.38	0.022	11.73	98.02	0.128	15.70/96.02	7.97/97.36*
Empirical p-BY	5.54	99.65	0.017	9.96	98.44	0.106	12.27/97.80	6.37/98.78
Monotone CRC	39.40	89.21	1.724	55.08	80.95	3.928	17.12/95.37	3.55/99.44
CRC-NM (two policies)	0.00	–	0.000	0.00	–	0.000	4.38/77.27*	0.00/–
LTT	0.00	–	0.000	16.72	95.50	0.360	14.95/90.87*	0.00/–
FP-CP Max	0.00	–	0.000	24.14	94.58	0.600	22.24/91.74*	6.13/95.96*
FP-CP Sum	0.00	–	0.000	29.88	94.49	1.103	24.87/88.85*	0.00/–
Block e-BH, binary ϕ	8.07	99.03	0.067	8.10	99.04	0.067	7.83/99.25	5.62/98.96
Block e-BH	16.51	98.12	0.206	27.55	93.05	0.873	18.77/92.98*	6.03/98.08

Table 14: Test recall/pooled precision (%) after development selection, all ten splits including abstentions. *Precision below the development target.

Setting	CRC	LTT	FP-CP Max	FP-CP Sum	Block e-BH
Dev. precision: 95%					
GD-T / COCO	17.12/95.37	14.95/90.87*	22.24/91.74*	24.87/88.85*	18.77/92.98*
GD-B / COCO	13.84/93.51*	12.52/88.62*	12.37/87.14*	12.90/85.19*	7.91/90.75*
GD-B / LVIS	0.75/75.86*	0.84/65.33*	0.86/63.29*	0.96/58.95*	0.77/81.82*
GD-B / OI	2.08/86.73*	2.52/76.87*	2.01/67.21*	2.23/69.47*	1.08/83.02*
YW / COCO	11.26/91.95*	9.45/85.97*	28.81/87.02*	24.33/87.42*	24.57/89.94*
YW / LVIS	1.25/94.81*	2.90/89.42*	0.00/–	3.02/86.27*	4.08/94.07*
YW / OI	1.98/72.32*	1.98/73.64*	3.77/61.85*	1.98/68.07*	1.66/75.56*
Dev. precision: 98%					
GD-T / COCO	3.55/99.44	0.00/–	6.13/95.96*	0.00/–	6.03/98.08
GD-B / COCO	6.14/94.38*	2.24/90.70*	3.65/86.76*	3.71/82.83*	4.38/90.08*
GD-B / LVIS	0.75/75.86*	0.84/65.33*	0.86/63.29*	0.96/58.95*	0.77/81.82*
GD-B / OI	2.08/86.73*	2.52/76.87*	2.01/67.21*	2.23/69.47*	1.08/83.02*
YW / COCO	6.92/95.38*	3.39/88.21*	13.03/87.05*	12.84/88.30*	11.38/91.40*
YW / LVIS	0.00/–	0.00/–	0.00/–	0.00/–	3.03/94.15*
YW / OI	1.98/72.32*	1.98/73.64*	3.77/61.85*	1.98/68.07*	1.66/75.56*

Table 15: Block e-BH minus comparator recall (percentage points), paired within setting/split at test precision within one point. Intervals use a normal approximation; comparisons describe the observed grid.

Comparator	Pairs	ΔRecall (pp)	95% CI
Empirical p-BH	30	$+4.65$	$[+2.72, +6.58]$
Empirical p-BY	27	$+6.88$	$[+4.64, +9.12]$
Binary evidence	48	$+4.05$	$[+2.58, +5.52]$
FP-CP Sum	45	-0.94	$[-2.62, +0.75]$
LTT	42	-0.60	$[-1.96, +0.76]$
Calibrated score	79	-2.23	$[-3.47, -0.98]$
Monotone CRC	52	-5.79	$[-7.68, -3.90]$

Table 16: Test results after calibration-only selection, ten splits. Precision pools clusters; recall averages splits, with zero for abstentions. Selection targets 80% test precision using an 83% calibration threshold.

Setting	Selected	Recall (%)	Precision (%)
GroundingDINO-tiny / COCO	10/10	38.74	89.10
GroundingDINO-base / COCO	10/10	20.05	88.40
YOLO-World / COCO	10/10	45.20	87.81
GroundingDINO-base / LVIS	9/10	1.52	80.73
YOLO-World / LVIS	10/10	23.97	88.26
GroundingDINO-base / Open Images	10/10	2.34	88.79
YOLO-World / Open Images	10/10	5.97	84.10

D.5 Guide to additional experiments

The following table links each experimental question to its supporting figures.

Table 17: Guide to additional experiments.

Question	Figures
Full precision–recall curves	Figures 24–25
Target-dependent risk and utility	Figure 26
Calibration-image budget	Figure 8; 25/50/100/200 images, fixed fit and test
Evidence-map contribution	Figure 11; histogram versus fit-only binary evidence
Family and cluster structure	Figure 30 and Figure 9(b)
Floor/clustering interaction	Figure 12; common GT denominator
Complete step-up behavior	Figure 13; eight realized families
Stratified utility and support	Figure 14; size and category support
Prompt-bank size	Figure 6; 80 categories, 1/4/8/16 templates
Annotation perturbation	Figure 15; fixed map versus refitting
Query-law change	Figure 16; matched, shifted, recalibrated
Computational stages	Figure 17; postprocessing time on saved outputs
Score-selected random families	Figure 18; simulated calibration and test images
Extended detection cases	Figure 19; retention, suppression, misses, false release

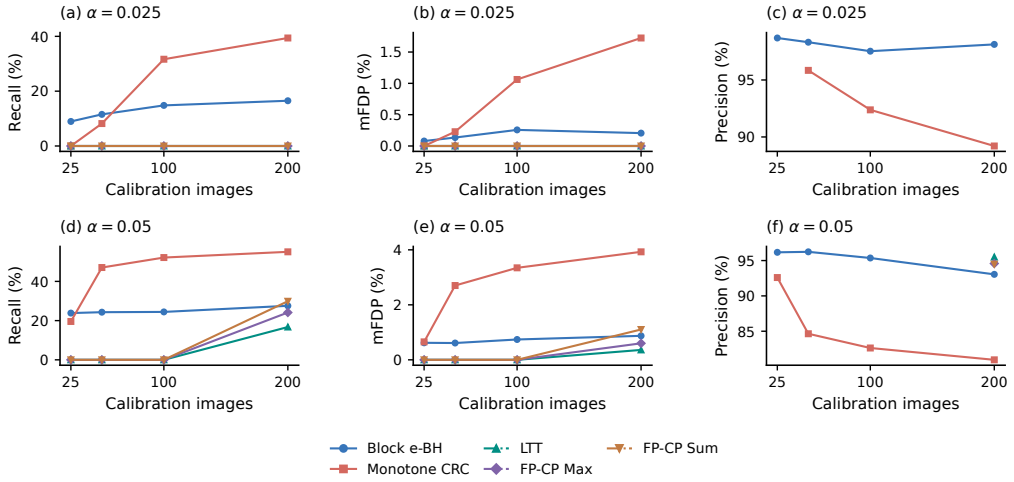

Figure 8: Calibration-image budgets using nested, label-independent subsets; fit and test images stay fixed. Rows: $\alpha = 0.025, 0.05$. Columns: pooled recall, mFDP, pooled precision.

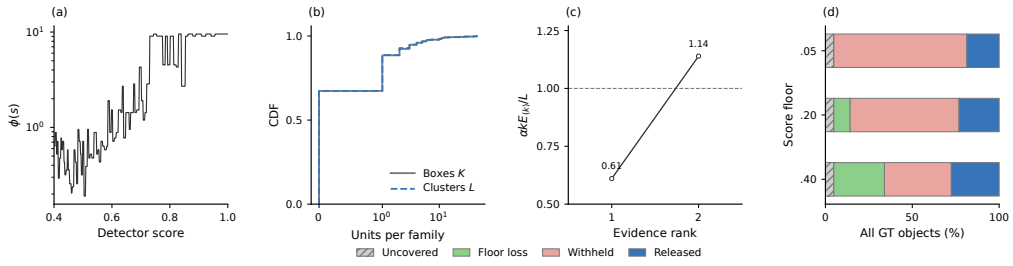

Figure 9: Evidence and filtering. (a,b) Fit map and family sizes in split 0, including empty families. (c) A second-rank crossing releases both clusters at $\alpha = 0.05$. (d) Object recovery and losses at clustering IoU 0.50.

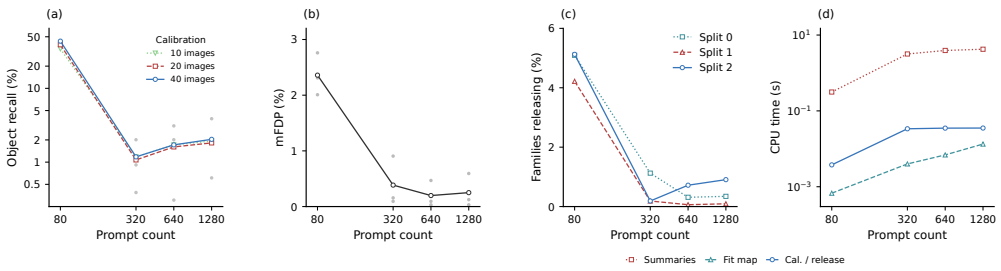

Figure 10: Prompt scaling over 80 categories. (a) Recall by calibration budget. (b,c) mFDP/release frequency with 40 calibration images. Gray points: three splits. (d) Median CPU postprocessing time, excluding detector inference. Prompt count, (a) recall, and (d) time use log axes.

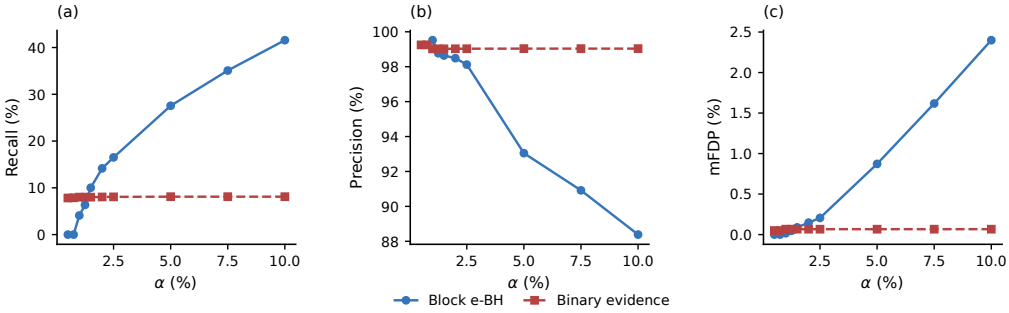

Figure 11: Learned versus binary evidence on COCO/GroundingDINO-tiny with identical splits, floor 0.40, clustering IoU 0.50, and block e-BH. The binary threshold is the fit-score 90th percentile.

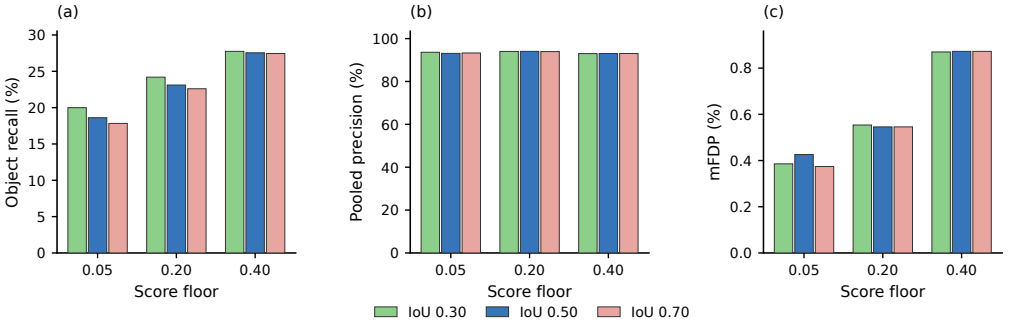

Figure 12: Floor/clustering interaction at $\alpha = 0.05$, ten matched splits. Matching IoU stays 0.50; each configuration refits the histogram after filtering. Recall uses a common pre-floor GT denominator.

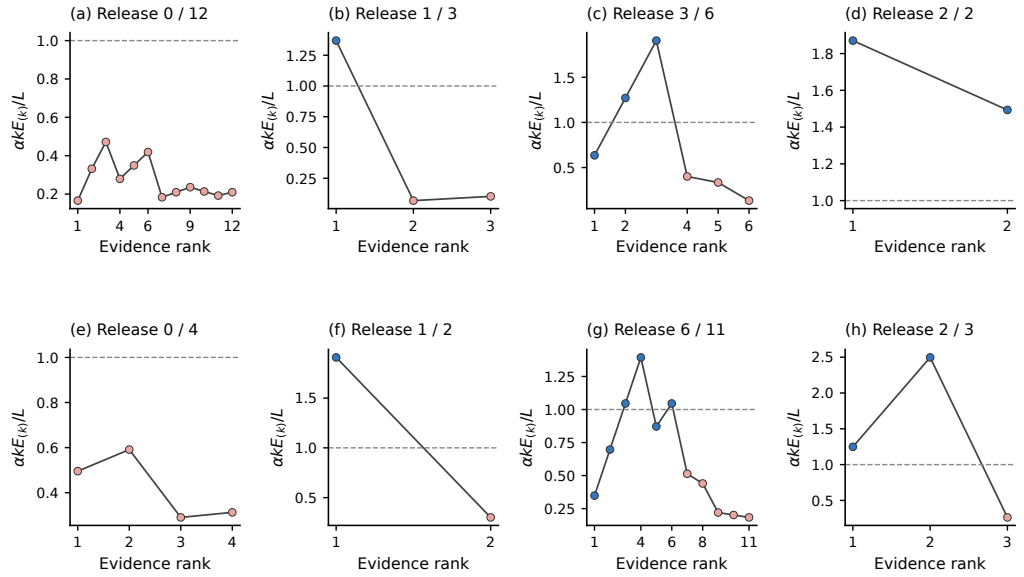

Figure 13: Eight families at $\alpha = 0.05$. Blue: released prefix ending at the last $\alpha k E_{(k)}/L \geq 1$ crossing. Pink: withheld. Columns show no, first-rank, later-rank, and multiple-object release, with two examples each.

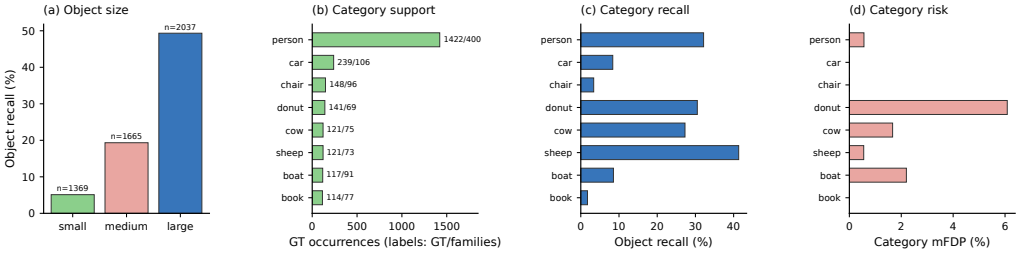

Figure 14: Size/category strata at $\alpha = 0.05$, ten splits. COCO size thresholds: $32^2, 96^2$; categories: eight with most GT support. Labels count GT/family occurrences. Category-wise risk is descriptive; control is query-averaged.

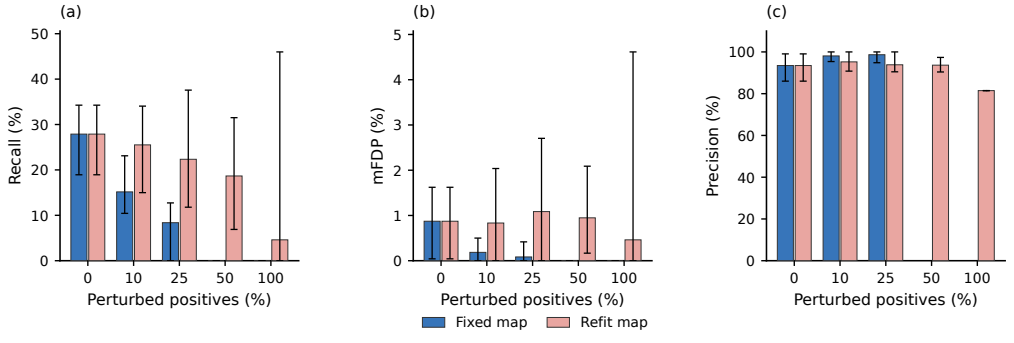

Figure 15: Annotation perturbations: blue relabels calibration positives as null with ϕ fixed; pink perturbs fit labels and refits ϕ. Other splits and candidates stay fixed. Bars/whiskers: split means/ranges; zero-release precision is omitted.

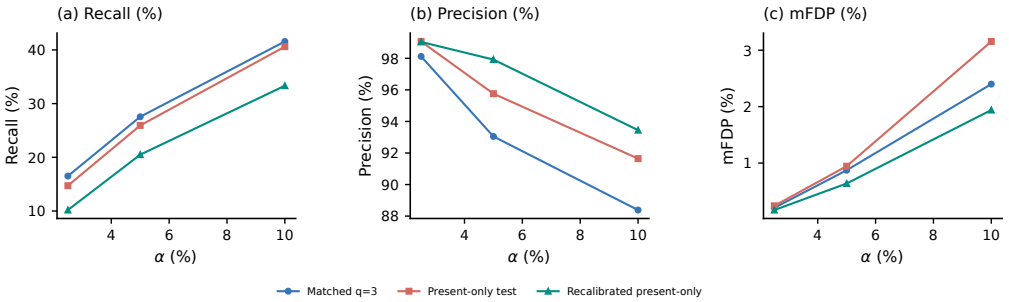

Figure 16: Query-law comparison on identical images. Matched: three queries in calibration/test. Shifted: three-query calibration, present-only test. Recalibrated: present-only calibration/test.

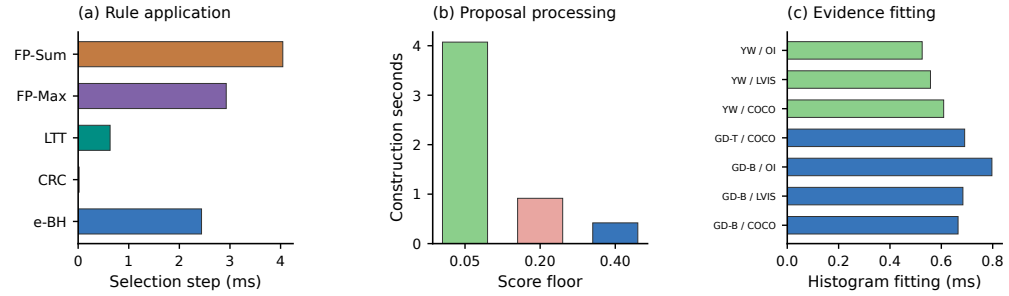

Figure 17: Median CPU postprocessing time across splits, excluding detector inference. Rule application also excludes evidence fitting, calibration-loss computation, and FP-CP probability-map preparation.

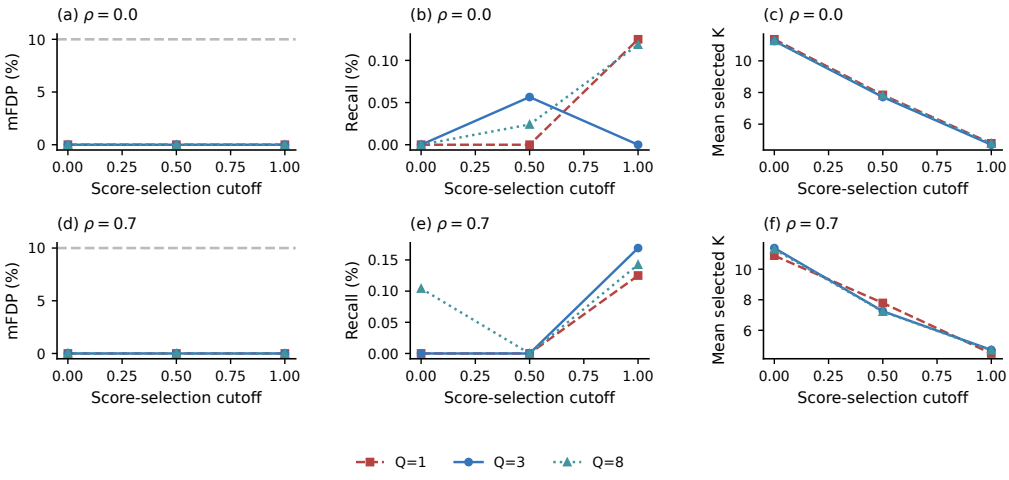

Figure 18: Score-selected families: 400 replicates, each with 200 calibration blocks and one test block, $\alpha = 0.10$, $Q = 1, 3, 8$. Each query starts with 20 candidates; Gaussian scores share image variance ρ and TP shift 1.5. Thresholding determines K; evidence is $\exp(S)$, clipped at $S = \pm 10$.

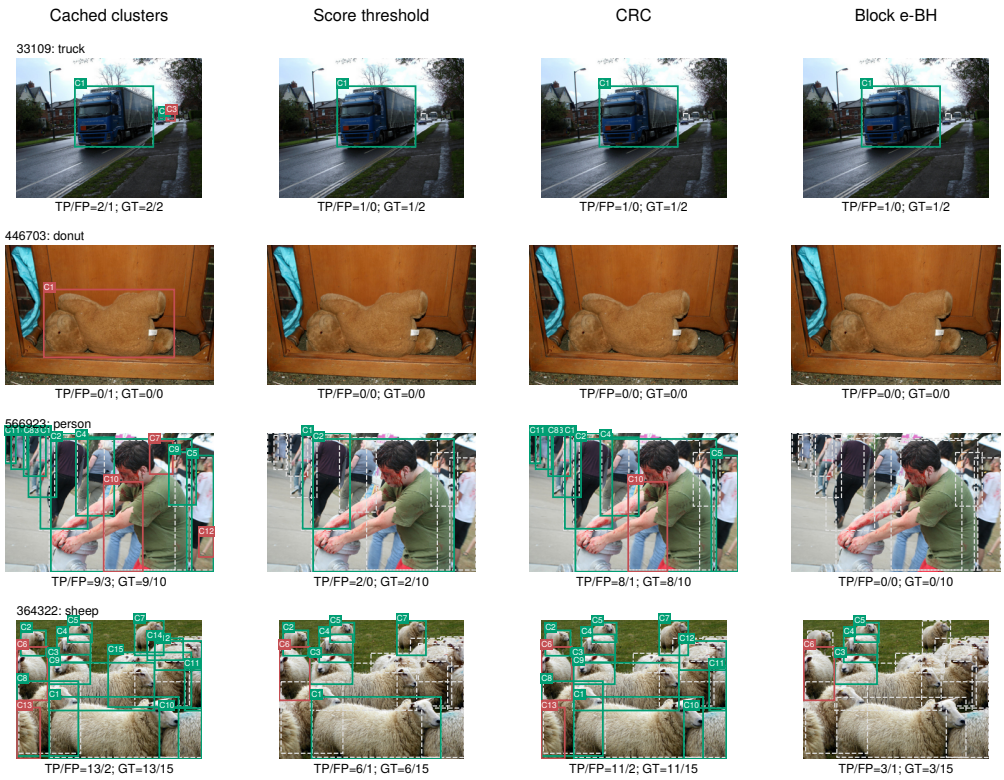

Figure 19: Additional releases under the protocol of Figure 1. C-labels identify the same clusters across columns. TP/FP count released clusters; GT counts distinct covered objects.

E PROMPT-BANK SCALING

This experiment calibrates a category sampled from each image. It uses greedy one-to-one matching after cross-template NMS, with detector-covered objects as the recall denominator. The main prompt-bank experiment in Figure 6 uses connected-component clusters, all annotated objects, and score floor 0.40. That experiment varies calibration support over 10, 20, and 40 images while keeping 120 fitting and 40 test images fixed.

The bank uses 80 categories and increasing numbers of synonymous templates, giving $V \in \{80, 320, 640, 1280\}$ prompts. Cross-template IoU clustering is applied within each image–category pair. The query rule draws a category uniformly from the fixed vocabulary. For selector repetition r, one category is drawn from each calibration image using randomization independent of labels and outputs. Its family null average is $A_i^{(r)}$, with value zero for an empty family. Each of the $n = 40$ calibration images contributes exactly one summary.

For a uniformly queried test image–category family f, form

$$E_{f,k}^{(r)} = \frac{(n+1)W_{f,k}}{T_f + \sum_{i=1}^{n} A_i^{(r)}}.$$

Under exchangeability of the image–category draws, including their selector randomization, the single-query block proof gives

$$\mathbb{E}\left[\frac{1}{L_f}\sum_{k \in H_{0,f}^{\mathrm{cl}}} E_{f,k}^{(r)} \,\middle|\, \mathcal{D}_{\mathrm{fit}}\right] \le 1.$$

The expression is zero for $L_f = 0$. The same statement holds for the average of $M = 200$ selector-specific evidence scores by linearity:

$$\overline{E}_{f,k} = \frac{1}{M}\sum_{r=1}^{M} E_{f,k}^{(r)}, \qquad \mathbb{E}\left[\frac{1}{L_f}\sum_{k \in H_{0,f}^{\mathrm{cl}}} \overline{E}_{f,k} \,\middle|\, \mathcal{D}_{\mathrm{fit}}\right] \le 1.$$

Dependence among selector repetitions is allowed. The expectation averages over the sampled calibration categories and the uniformly drawn test category. Evaluating all categories equally has the same mean as the uniform-category target.

Table 18: Block cluster prompt-bank study. Means and sample standard deviations are over three image splits. Recall and precision columns average the split-specific ratios.

Prompts	Released clusters	Correct clusters	Cluster mFDP	Object recall	Mean precision
80	121.33 ± 36.02	70.00 ± 19.16	$1.54 \pm 0.49\%$	$23.33 \pm 2.29\%$	$58.03 \pm 1.72\%$
320	49.00 ± 27.73	13.33 ± 6.43	$1.11 \pm 0.67\%$	$4.30 \pm 1.50\%$	$29.19 \pm 5.41\%$
640	38.33 ± 25.17	12.00 ± 7.55	$0.82 \pm 0.56\%$	$3.82 \pm 2.08\%$	$31.01 \pm 5.46\%$
1,280	35.00 ± 21.66	9.33 ± 3.51	$0.80 \pm 0.57\%$	$3.09 \pm 0.81\%$	$30.18 \pm 8.97\%$

At 1,280 prompts, 200 images produce 256,000 image–prompt families and 11.81 million candidate boxes. Collecting detector outputs takes 3.73 hours on five A10G GPUs, and cross-template post-processing takes 72 seconds for the three splits. The ratio of total correct clusters to total releases is approximately $28/105 = 26.67\%$; it differs from the mean split precision of 30.18%. Both differ from one minus mFDP because the latter gives empty-release families equal weight.

F ANNOTATION QUALITY AND CALIBRATION EVIDENCE

F.1 ONE-SIDED CALIBRATION ERRORS

Keep candidate sets, ϕ, and test blocks fixed. Suppose some correct calibration candidates are relabeled as null. Let H_{iq}^{added} contain their indices. Then

$$\widetilde{A}_{iq} = A_{iq} + \frac{1}{K_{iq}} \sum_{k \in H_{iq}^{\mathrm{added}}} \phi(S_{iqk}) \ge A_{iq},$$

with the empty-family convention. Hence $\widetilde{C} \geq C$ and $\widetilde{E}_{0qk} \leq E_{0qk}$ for all test candidates. Each crossing rank under the perturbed scores is also a crossing rank under the unperturbed scores. The release set is nested. For clusters, some correct clusters may become all-null; their nonnegative evidence likewise enlarges the null reference. This result holds with the fit map and candidate sets fixed.

Table 19 measures how added null labels change releases under candidate-pooled calibration. Figure 15 evaluates annotation changes under the block method.

Table 19: Candidate-pooled calibration-label sweep on ten splits. γ is the missing-positive perturbation rate, and the actual contamination fraction counts the affected entries in the null pool.

γ	Actual contamination	Released	Correct	False	mFDP	Candidate recall
0.00	0.00%	1,286	1,184	102	0.823%	15.67%
0.10	1.16%	1,063	992	71	0.490%	13.13%
0.30	3.62%	838	801	37	0.249%	10.60%

F.2 Population evidence bounds and formal corrections

Let p_0 and p_1 be the null and correct-detection score densities. A null reference with contamination fraction $\eta > 0$ has density $q_\eta = (1-\eta)p_0 + \eta p_1$. Its population evidence $\phi_\eta = p_1/q_\eta$ satisfies $\phi_\eta \leq 1/\eta$ wherever defined. For an estimated score map $\widehat{\phi}$, its peak $\widehat{\phi}_{\max}$ also depends on sampling and fitting. We use the reciprocal peak as a sensitivity diagnostic.

For the block release condition, if every observed ϕ is at most M and $M < D_0/[\alpha(n+1)]$, no rank can cross: $T^{\text{block}}_{q,k} \geq D_0/[\alpha(n+1)]$. This is a sufficient no-release condition conditional on the realized denominator. Evidence compression changes this power calculation without invalidating the exchangeability theorem.

A sufficient condition for a formal scaling correction can be stated directly on the full experiment. Let P be an exchangeable reference law conditional on fit, and suppose the actual joint law P' satisfies $dP'/dP \leq c$ for an independently justified constant $c \geq 1$. Running the block rule at α/c gives a nonnegative mean FDP F with

$$\mathbb{E}_{P'} F \leq c\,\mathbb{E}_P F \leq \alpha.$$

The required density bound concerns the full joint experiment. For a single score, $q_\eta \geq (1-\eta)p_0$ gives $c = 1/(1-\eta)$ if the remaining variables retain the same conditional law. Detector-dependent families require the corresponding joint bound. The one-sided fixed-fit result above applies directly with no contamination-rate estimate.

G Candidate-pooled calibration: comparisons and design analysis

Candidate-pooled calibration combines null candidates from different families into a shared reference. These experiments examine how that reference, the fitted score map, and the release unit affect detection quality. They complement the block experiments with additional detector comparisons and analyses of family size, query composition, and annotation quality. We first define the candidate-pooled construction and evaluation protocol, then group the results by these design questions. Appendix C specifies the complete-image protocol used in the main comparisons.

Table abbreviations are GroundingDINO (GDino), YOLO-World (YOLO), PASCAL Visual Object Classes (VOC), and released detections (Rej.). COCO-1k denotes a 1,000-image subset; COCO-tr denotes a subset of the COCO training split. A leave-one-out (LOO) check treats each calibration null in turn as the test score and averages its normalized evidence. The fitted evidence peak is $\widehat{\phi}_{\max} = \max_s \widehat{\phi}(s)$, abbreviated as $\phi_{\max}$ in the tables. A pooled pass rate counts splits whose pooled FDP is at most the nominal target.

G.1 Construction and evaluation protocol

G.1.1 Sampling, matching, and metrics

The following protocol fixes family formation, matching, and data splitting for the candidate-pooled comparisons. Individual studies state their changes to this protocol.

Task unit. The atomic unit is an *image-query family*: one image I paired with one text prompt q, yielding a set of candidate detections $\{(b_k, s_k)\}_{k=1}^{K}$ where b_k is a bounding box and s_k is the detector confidence score. A family is *present* if the prompt category has at least one ground-truth box in I, and *absent* otherwise.

Family construction. For each sampled image, draw $n_{\text{present}} = 1$ present prompt (uniformly from categories annotated in the image) and $n_{\text{absent}} = 2$ absent prompts (uniformly from the remaining dataset vocabulary). This yields $3N$ families from N images, with a fixed 1:2 present-to-absent benchmark ratio.

Ground-truth matching. A detection (b_k, s_k) is a true positive if $\text{IoU}(b_k, g) \geq \tau$ for some ground-truth box g of the queried category, where $\tau = 0.5$ by default. Greedy matching is used (each ground-truth box matches at most one detection, highest-IoU first). All unmatched detections are null (hallucinations).

Data splitting. Complete families are partitioned into *fit* (60%), *calibration* (20%), and *test* (20%) subsets. Splitting by family ensures that all candidates within a family stay together, preserving intra-family dependence structure.

Primary metrics.

- **Mean family FDP**: $\frac{1}{|\mathcal{F}|}\sum_f \text{FDP}_f$, where $\text{FDP}_f = \text{FP}_f / \max(R_f, 1)$.
- **Pooled FDP** (secondary, empirical diagnostic): $\sum_f \text{FP}_f \;/\; \sum_f R_f$, where FP_f and R_f are the number of false positives and total rejections in test family f. This summary weights families by their release counts.
- **Recall**: $\sum_f \text{TP}_f \;/\; \sum_f N_f^{\text{gt}}$, where N_f^{gt} is the number of ground-truth boxes in family f.
- **LOO mean**: leave-one-out mean e-value on calibration nulls, used to check the normalization calculation.
- **Pooled pass rate**: fraction of random splits where pooled FDP $\leq \alpha$.

Evaluation protocol. Run 20 independent random splits (seeds 0–19). Report mean $\pm$ standard deviation over splits for FDP and recall at $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$. A method *empirically passes the pooled-FDP criterion* if pooled FDP $\leq \alpha$ on all 20 splits; it *approximately passes* if this holds on ≥ 18 of 20.

Detector configuration. Detectors are run with a permissive score threshold (0.05) and standard NMS (IoU $= 0.7$, max 300 detections) to produce a large candidate pool. The release filter operates on fixed detector outputs.

Dataset requirements. Absent-category sampling uses annotations to identify categories missing from an image. COCO and VOC provide exhaustive annotations for their benchmark categories. LVIS (Gupta et al., 2019) and Open Images (Kuznetsova et al., 2020) provide partial annotations. Appendix G.2.6 evaluates subsets with verified category labels.

G.1.2 Evidence normalization and release thresholds

The candidate-pooled rule normalizes each score against a shared null reference. The identities below give its single-score expectation and exact release thresholds. These provide the analytical basis for the family-size and score-map experiments.

Lemma 2 (Single-score self-normalization)**.** *Conditional on the fit split, let $S_0, S_1, \ldots, S_m$ be jointly exchangeable null scores and let ϕ be a fixed nonnegative finite map. Define*

$$e(S_0) = \frac{(m+1)\phi(S_0)}{\phi(S_0) + C_m}, \qquad C_m = \sum_{i=1}^{m} \phi(S_i). \tag{14}$$

With zero assigned to a zero denominator, $\mathbb{E}[e(S_0) \mid \mathcal{D}_{\text{fit}}] \leq 1$.

Proof. The $m+1$ normalized shares have equal expectations by exchangeability. Their sum is the indicator that the denominator is positive. Multiplication by $m+1$ gives the claim. □

This lemma concerns one null score. The fixed-family testing condition is stated below.

Theorem 2 (Candidate-pooled release frontier)**.** *For a realized family of size K, let $C_m > 0$ and order the raw evidence as $\phi_{(1)} \geq \cdots \geq \phi_{(K)}$. For ranks with $\alpha k(m+1) > K$, set*

$$T_k(C_m) = \frac{KC_m}{\alpha k(m+1) - K}.$$

These rank thresholds form the activation frontier*. Candidate-pooled e-BH selects the largest crossing rank and releases that prefix; ranks with $\alpha k(m+1) \leq K$ cannot cross.*

Proof. The map $x \mapsto (m+1)x/(x+C_m)$ is increasing. Substituting it into the step-up comparison gives $\phi_{(k)}[\alpha k(m+1) - K] \geq KC_m$. The positive right-hand side KC_m requires a positive bracket. The crossing condition is then $\phi_{(k)} \geq T_k(C_m)$. Taking the largest crossing rank proves the result. □

Corollary 2 (Selected set)**.** *The selected set consists of the top $\widehat{k}$ candidates by ϕ, where $\widehat{k}$ is the largest crossing rank in Theorem 2. At least r discoveries require some crossing rank $k \geq r$, not necessarily a crossing at rank r.*

Proposition 4 (Fixed-family implication)**.** *For a fixed family of K hypotheses, if each null input satisfies $\mathbb{E}[e_k] \leq 1$, within-family e-BH controls FDR at α. This remains true conditional on a sigma-field fixing the family if the corresponding null expectations are bounded by one conditional on that field.*

Proof. The sample-wise bound $\text{FDP} \leq (\alpha/K)\sum_{k \in H_0} e_k$, followed by expectation, gives $\mathbb{E}[\text{FDP}] \leq \alpha|H_0|/K \leq \alpha$. The same calculation works conditionally under the stated conditional assumptions. □

Applying this proposition after detector selection requires the stated conditional expectations. The image-block construction handles that selection through the block-average bound.

A two-type image counterexample. Consider one calibration image and one independent test image, drawn from the same distribution. Each has one query and is equally likely to be type A or B. A type-A image produces 100 null candidates, each with raw evidence $\phi(S) = 1/100$. A type-B image produces two null candidates, each with $\phi(S) = 1$. The prediction rule and evidence map are fixed. Complete images are exchangeable, and all candidates within an image have the same evidence.

For the candidate-pooled rule (14), calibration on type A gives $(m, C_m) = (100, 1)$; calibration on type B gives $(m, C_m) = (2, 2)$. The normalized evidence of each test candidate is

	Test A	Test B
Calibration A	1	101/2
Calibration B	1/67	1

Set $\alpha = 0.10$. With equal evidence scores, e-BH releases the whole family exactly when their common value reaches $1/\alpha = 10$. Only the calibration-A/test-B pair crosses this threshold. It occurs with probability $1/4$ and releases two null candidates. Every other pair releases nothing, so the mean-family FDR is $1/4 > \alpha$.

The example shows why the calibrated unit matters even under a shared image law. Block calibration averages evidence within each family before comparing images. Its exchangeable image summaries give the bound in Proposition 1.

Prompt mixtures. The common-mixture premise and corrected bound are stated in Proposition 3, with the full unequal-mixture derivation in Appendix A.

Population contamination bound. For $q_\eta = (1-\eta)p_0 + \eta p_1$ and $\eta > 0$, $\phi_\eta = p_1/q_\eta \le 1/\eta$. For the candidate-pooled diagnostic, a sufficient no-release condition is $1/\eta < \min_k T_k(C_m)$ over eligible ranks, when the score map actually satisfies that bound. Estimated density-ratio maxima serve as sensitivity diagnostics. Appendix F states the fixed-fit one-sided result and the conditions needed for a formal scaling correction.

Lemma 3 (Single-score shift bound)**.** *For the candidate-pooled statistic in* (14)*, assume the calibration scores are independent draws from P_{pool}, independent of the test score drawn from P_g, conditional on fit. If $\mathrm{TV}(P_g, P_{\mathrm{pool}}) \le \delta$, then $\mathbb{E}[e(S_0)] \le 1 + (m+1)\delta$.*

Proof. The statistic lies in $[0, m+1]$. Replacing the test marginal changes the joint law from $P_{\mathrm{pool}}^{\otimes(m+1)}$ to $P_g \otimes P_{\mathrm{pool}}^{\otimes m}$ by total variation at most δ. For a function in $[0, m+1]$, the change in expectation is at most $(m+1)\delta$. Apply Lemma 2 under the exchangeable law. □

For a fixed family satisfying these input conditions, scaling by $1+(m+1)\delta$ restores the fixed-family bound in Proposition 4. For detector-dependent families, the stated input conditions must hold after selection.

Hierarchical selection. The two-stage candidate-pooled argument requires valid family-level inputs at the first stage and conditional validity of the within-family inputs after the family selection. Independence of the first-stage data from those inputs suffices when the tested families and their null hypotheses are fixed. This requires control conditional on family selection. For selection among queries of the same image, Appendix B supplies the expected worst-family alternative.

G.1.3 Query-weighted and release-weighted errors

Mean-family FDP gives each query equal weight. Pooled FDP divides total false releases by total releases, giving more weight to families with larger outputs. The following table compares these empirical summaries in the candidate-pooled study.

Dataset	Detector	Pooled FDP	Per-family mean FDP
COCO	GDino	3.96%	0.65%
COCO	OWL-ViT	2.31%	0.34%
COCO	YOLO-World	3.08%	0.72%
VOC	GDino	5.62%	1.69%
VOC	OWL-ViT	2.54%	0.35%
VOC	YOLO-World	1.34%	0.90%

Table 20: Pooled and mean-family FDP at $\alpha = 0.10$ on 1,000-image subsets. Pooled FDP is substantially higher, showing that errors concentrate in families with more releases.

Error concentration. Pooled FDP is lower in small-family strata and less stable in the largest families (Table 28). This pattern is consistent with the release threshold's dependence on family size K. Table 20 records pooled FDP below the target in $19/20$ COCO/GDino splits and $20/20$ VOC/GDino splits at $\alpha = 0.10$.

Implication for score thresholding. A direct calibration-to-test concentration bound can be conservative when family sizes and release counts vary. The thresholding diagnostic used the margin $\varepsilon \ge K_{\max}\sqrt{\log(2/\delta)/(2N_{\mathrm{cal}})}/\bar{R}_{\mathrm{cal}} \approx 0.12$ at $\alpha = 0.10$. The monotone-envelope CRC comparison in Section 4.1 evaluates an alternative on complete image blocks.

G.2 Evaluation across detectors and image pools

G.2.1 Detector comparison on 1,000-image subsets

Table 21 reports per-family and pooled FDP, recall, $\phi_{\max}$, and the LOO self-normalization check for all six dataset-detector combinations on the 1,000-image subsets at $\alpha = 0.10$.

Dataset	Detector	Pooled FDP	Per-family FDP	Recall	$\phi_{\max}$	LOO mean
COCO-1k	GDino	$3.96 \pm 1.98\%$	0.65%	18.7%	357	1.000
COCO-1k	OWL-ViT	$2.31 \pm 1.84\%$	0.34%	11.2%	198	1.000
COCO-1k	YOLO-World	$3.08 \pm 2.30\%$	0.72%	15.3%	$>10^6$	1.000
VOC-1k	GDino	$5.62 \pm 1.96\%$	1.69%	32.6%	1146	1.000
VOC-1k	OWL-ViT	$2.54 \pm 2.76\%$	0.35%	9.7%	152	1.000
VOC-1k	YOLO-World	$1.34 \pm 1.12\%$	0.90%	40.9%	$>10^6$	1.000

Table 21: Candidate-pooled diagnostics on 1,000-image subsets, $\alpha = 0.10$, 20 splits. Pooled FDP: total FP/releases (mean $\pm$ s.d.); family FDP: mean within-family FDP. $\phi_{\max}$ is the median density-ratio peak; large peaks reflect sparse null-score bins. Random-family control requires additional conditions.

G.2.2 Stability across data splits

Table 22 reports pooled FDP, per-family mean FDP, recall, and pooled-FDP pass rate for all six dataset-detector combinations at all tested α levels.

At $\alpha \geq 0.10$, all six settings have pooled FDP below the target in all 20 splits. At $\alpha = 0.05$, the pass rate falls to 13/20 for VOC/GDino. These frequencies measure variation across the evaluated splits.

Table 22: Full 20-seed stability results. Pooled FDP = total FP / total rejections (mean $\pm$ s.d.); family FDP = per-family mean FDP; pass rate = fraction of seeds with pooled FDP $\leq \alpha$.

Dataset	Detector	α	Pooled FDP	Family FDP	Recall	Pass
COCO	GDino	0.05	$2.74 \pm 1.88\%$	0.42%	14.7%	19/20
		0.10	$3.96 \pm 1.98\%$	0.65%	18.7%	20/20
		0.15	$5.07 \pm 2.20\%$	0.95%	21.0%	20/20
		0.20	$5.98 \pm 2.12\%$	1.25%	22.8%	20/20
COCO	OWL-ViT	0.05	$1.51 \pm 1.71\%$	0.21%	8.2%	19/20
		0.10	$2.31 \pm 1.84\%$	0.34%	11.2%	20/20
		0.15	$3.53 \pm 3.02\%$	0.60%	13.4%	20/20
		0.20	$3.80 \pm 3.01\%$	0.81%	15.4%	20/20
COCO	YOLO	0.05	$2.92 \pm 2.20\%$	0.55%	13.1%	15/20
		0.10	$3.08 \pm 2.30\%$	0.72%	15.3%	20/20
		0.15	$3.42 \pm 2.92\%$	0.85%	16.5%	20/20
		0.20	$3.47 \pm 2.94\%$	0.92%	17.2%	20/20
VOC	GDino	0.05	$4.57 \pm 1.67\%$	1.16%	25.2%	13/20
		0.10	$5.62 \pm 1.96\%$	1.69%	32.6%	20/20
		0.15	$7.12 \pm 2.58\%$	2.15%	36.3%	20/20
		0.20	$8.66 \pm 2.83\%$	2.74%	38.8%	20/20
VOC	OWL-ViT	0.05	$1.81 \pm 2.54\%$	0.24%	7.8%	18/20
		0.10	$2.54 \pm 2.76\%$	0.35%	9.7%	20/20
		0.15	$2.39 \pm 2.65\%$	0.35%	11.1%	20/20
		0.20	$2.66 \pm 2.74\%$	0.39%	12.7%	20/20
VOC	YOLO	0.05	$1.07 \pm 0.86\%$	0.65%	36.8%	20/20
		0.10	$1.34 \pm 1.12\%$	0.90%	40.9%	20/20
		0.15	$1.36 \pm 1.11\%$	0.93%	42.3%	20/20
		0.20	$1.47 \pm 1.46\%$	0.99%	43.1%	20/20

G.2.3 FULL COCO VALIDATION SET

Table 23 reports the full COCO val (4,952 images) results across $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$ with GroundingDINO. At $\alpha = 0.10$, pooled FDP is $6.44 \pm 2.11\%$ and recall is 16.4%. Pooled FDP is below the target in 19 of 20 splits; seed 3 reaches 11.81%. The 1,000-image subset has lower mean pooled FDP, $3.96 \pm 1.98\%$. The two evaluations use different image pools and family counts.

	Betting e-BH				Baselines	
α	Rej.	Mean family FDP	Pooled FDP	Recall	p-BH pooled FDP	Binary recall
0.05	459.6	0.43%	4.9%	12.5%	17.9%	0.16%
0.10	615.1	0.78%	6.4%	16.4%	27.1%	0.18%
0.15	731.4	1.10%	8.6%	19.1%	35.0%	0.19%
0.20	823.0	1.45%	10.1%	21.1%	42.3%	0.83%

Table 23: Full-scale COCO val with GroundingDINO, 20 splits. Mean family FDP estimates the per-family FDR target; pooled FDP and recall are deployment diagnostics. This table reports the full-COCO candidate-pooled diagnostic values; the main result uses complete image-block calibration.

G.2.4 EVALUATION ON A SEPARATE COCO TRAINING SUBSET

As an independent validation on images unseen during any method development, we applied the full pipeline to a 1,000-image subset of COCO 2017 *train* with GroundingDINO. Table 24 reports the results over 20 random splits.

At $\alpha = 0.20$, pooled FDP is below the target in all 20 splits. At $\alpha = 0.10$, it is below the target in 15 splits, with a mean of $8.1 \pm 3.9\%$. The remaining five have FDP between 12.4% and 14.9%. COCO-train and COCO-val have similar median null scores (0.079 in both) and correct-detection scores (0.338 and 0.328). Their fitted evidence peaks differ (593 and 357). The LOO mean remains 1.000 across all splits despite the variation in test error.

Table 24: COCO-train (1,000 images) + GroundingDINO, 20 random splits.

α	Mean FDP	Std FDP	Recall	Pass rate
0.05	6.15%	3.75%	15.1%	10/20
0.10	8.13%	3.88%	20.0%	15/20
0.15	9.32%	4.12%	23.2%	17/20
0.20	10.08%	3.76%	25.4%	20/20

G.2.5 CALIBRATION CHOICES ON LVIS

We evaluate three targeted diagnostics on LVIS: (i) group-specific (Mondrian) calibration, separating absent and present queries; (ii) a minimum family-size filter ($k_{\min}$); (iii) a lower bound on the estimated null density to limit large density ratios. Observed error remains high under all three adjustments, with a fitted evidence peak of approximately 30. We next use category-annotation metadata to examine the role of absent-query labels.

G.2.6 AUDITED EVALUATION ON LVIS AND OPEN IMAGES

This study selects image–category families using annotation metadata. It compares GroundingDINO, OWLv2, and YOLO-World on LVIS, and GroundingDINO on Open Images version 7 (OI v7), using candidate-pooled calibration.

LVIS protocol. The curated LVIS pool combines native verified-negative categories with absences identified through an additional category-annotation audit. Categories with incomplete labels are excluded from the absent-query pool. It contains 536.6 families per split on average over 20 seeds and 1,000-image draws. Present families are retained separately.

Open Images version 7 protocol. Open Images provides positive and negative image-level verification flags. The audited construction admits absent-query families using verified negative labels

and retains annotated-present families separately. Its image-label-clean pool contains approximately 1,000 families per split.

Results ($\alpha = 0.10$, 20 seeds). All four detector–dataset rows have pooled FDP below the nominal target in every split:

Dataset	Detector	Rej.	Pooled FDP	Per-family FDP	Recall	Pool pass
LVIS (curated)	GDino	66.6	2.2%	0.4%	11.4%	20/20
	OWLv2	105.6	3.0%	1.2%	20.9%	20/20
	YOLO-World	31.2	3.7%	0.8%	12.6%	20/20
OI v7 (verified labels)	GDino	54.5	2.1%	0.7%	12.3%	20/20

Interpretation. Pooled FDP drops from LVIS-full's 17.7% (the baseline row of Table 30) to 2.2–3.7% under curated-pool restriction, at non-trivial recall (11–21%). Open Images shows the same pattern at 2.1% FDP with 12.3% recall. These numbers compare candidate-pooled evaluation pools: curated-pool restricts the category pairs, whereas the full LVIS pool includes assumed absences. These results quantify the effect of selecting queries with verified category labels. The block experiment uses the query pools specified in Appendix C.1.

G.3 Family Size and Candidate Filtering

G.3.1 Family-Size Distributions

Figure 20 shows the distribution of family sizes K (candidates per image-query family). On COCO/GDino, 58% of families have $K \leq 5$ and the median is 4; OWL-ViT produces fewer candidates per family (median 4, 59% ≤ 5). Theorem 2 implies that the rank-1 threshold has a finite denominator only when $\alpha(m+1) > K$; with typical $m \approx 500$–2000, this holds for all observed K. The displayed mean FDP is below nominal across the K bins; smaller families ($K \leq 5$) show more variable realized error fractions.

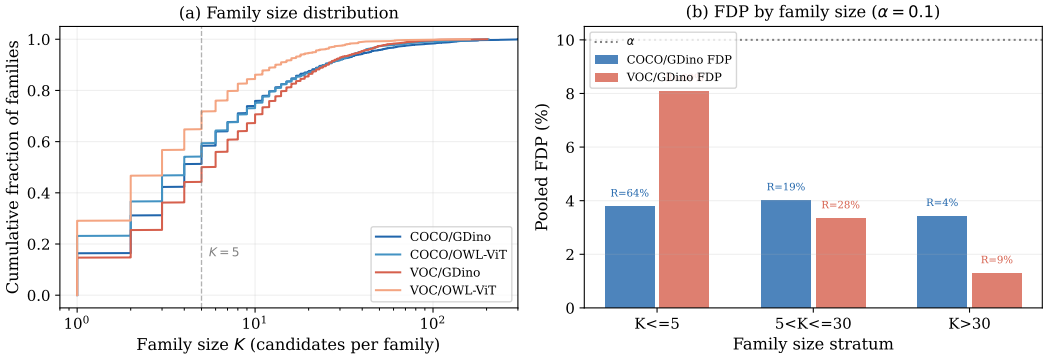

Figure 20: (a) Cumulative distribution of family sizes across four dataset-detector combinations. (b) Pooled FDP stratified by family size for COCO/GDino and VOC/GDino at $\alpha = 0.10$; all strata remain below the target. R = recall.

G.3.2 Evidence at the First Release Threshold

Table 25 reports the fraction of correct candidates whose evidence ϕ exceeds the first-rank threshold T_1 from Theorem 2. It covers all seven settings at $\alpha = 0.10$ over 20 seeds. The full e-BH rule also checks later ranks, where lower thresholds can release additional candidates.

Corollary 2 gives the full rank-dependent release rule. The following experiment examines whether removing low-score candidates improves recall.

Config	Recall	Rank-1 diag. ($s \geq 0$)	Rank-1 diag. ($s \geq 0.3$)	Rank-1 diag. ($s \geq 0.5$)
COCO-1k / GDino	18.7%	15.1%	28.8%	45.7%
COCO-full / GDino	16.4%	13.1%	24.3%	36.9%
VOC-1k / GDino	32.6%	25.6%	40.1%	48.2%
COCO-1k / OWL-ViT	11.2%	9.7%	19.8%	34.5%
VOC-1k / OWL-ViT	9.7%	8.9%	25.0%	51.4%
COCO-1k / YOLO	15.3%	14.1%	19.2%	22.8%
VOC-1k / YOLO	40.9%	37.8%	43.6%	47.0%

Table 25: Rank-1 diagnostic at $\alpha = 0.10$, 20 seeds: fraction of TPs above each score threshold satisfying $\phi(S) \geq T_1(K_g, C_m, m, \alpha)$. Recall is the observed candidate-pooled value; later-rank crossings can release additional detections.

G.3.3 Recall by detector score

Table 26 reports recall conditioned on true-positive detector score. The fitted map need not be monotone in score. Nearly all released true candidates in this diagnostic experiment have high detector confidence. Conditioning on score ≥ 0.5 raises recall from 16.4% to 46.5% on full COCO and from 18.7% to 56.3% on COCO-1k. This increase reflects the smaller evaluation denominator.

	COCO-1k (20 seeds)			COCO-full (20 seeds)		
Score $\geq$	#TP	#Rej TP	Recall	#TP	#Rej TP	Recall
0.0	692	129	18.7%	3500	575	16.4%
0.1	562	129	23.1%	2797	575	20.6%
0.2	434	129	29.7%	2247	575	25.6%
0.3	362	129	35.5%	1884	575	30.5%
0.4	290	129	44.3%	1558	575	36.9%
0.5	228	129	56.3%	1237	574	46.5%

Table 26: Score-conditional recall at $\alpha = 0.10$, betting e-BH/GDino, 20 seeds. #TP counts test TPs above the score threshold; #Rej TP counts selected TPs. Counts are seed averages.

G.3.4 Score-floor comparison across settings

Table 27 reports score-floor results across all seven dataset-detector settings at $\alpha = 0.10$ (20 seeds). The improvement is consistent: five of seven settings show 1.5–1.9$\times$ recall gains at $s_{\text{floor}} = 0.40$. GDino and OWL-ViT benefit most because their TP/null score distributions have substantial overlap in the $[0, 0.30]$ range, where floor pruning removes many nulls but few TPs. YOLO-World shows moderate gains on COCO (1.5$\times$) and no gain on VOC, where its baseline recall (40.9%) already reflects concentrated TP scores.

Per-family FDP remains below α in every configuration and at every score floor. The improvement saturates: beyond $s_{\text{floor}} = 0.40$, additional gains are marginal because most benefit comes from pruning the bulk of low-score nulls.

Figure 21 shows the score-floor frontier curves for COCO/GDino and VOC/GDino: both the rank-1 diagnostic (T_1 from Theorem 2) and the actual e-BH recall rise monotonically with s_{floor}, consistent with the threshold-reduction mechanism. The rank-1 diagnostic is conservative because multi-rejection families activate additional candidates at $T_2, T_3, \ldots$, which have lower thresholds; the gap is stable at $\sim$20% across all tested floors.

G.3.5 Observed error by family size

We apply within-family e-BH for each detector–dataset setting and data split, then group test families by size $K \in \{1, 2\text{–}5, 6\text{–}10, 11\text{–}20, >20\}$. Table 28 reports the mean across 20 seeds on COCO/GDino at $\alpha = 0.10$.

G.3.6 Score, object-size, and category groups

We group test candidates by score, object size, and category, complementing the family-size analysis in Appendix G.3.5. Each detector–dataset setting and target uses the same 20 data splits. Pooled

Config	s_{floor}	Per-family FDP	Recall	Gain
COCO/GDino	0.00	0.65%	18.7%	—
	0.30	1.93%	30.4%	1.62×
	0.40	2.59%	33.3%	1.78×
VOC/GDino	0.00	1.69%	32.6%	—
	0.30	4.25%	47.9%	1.47×
	0.40	4.42%	49.0%	1.50×
COCO/OWL-ViT	0.00	0.34%	11.2%	—
	0.30	1.55%	19.1%	1.71×
	0.40	1.77%	20.8%	1.86×
VOC/OWL-ViT	0.00	0.35%	9.7%	—
	0.30	0.48%	13.8%	1.42×
	0.40	0.63%	14.4%	1.48×
COCO/YOLO-World	0.00	0.72%	15.3%	—
	0.30	0.92%	22.5%	1.48×
	0.40	1.06%	23.2%	1.52×
VOC/YOLO-World	0.00	0.90%	40.9%	—
	0.30	0.82%	38.0%	0.93×
	0.40	0.83%	39.1%	0.96×
COCO-train/GDino	0.00	1.19%	20.0%	—
	0.30	2.69%	32.9%	1.65×
	0.40	3.12%	34.6%	1.73×

Table 27: Score-floor filtering at $\alpha = 0.10$, 1,000 images and 20 seeds per setting. VOC/YOLO-World is the only setting without a recall gain.

K bin	Family share	Mean K	Rejections	Pooled FDP	Recall	Pass rate
$K=1$	16.7%	1.0	17.5	3.04%	95.1%	0.95
2–5	41.9%	3.2	53.1	4.02%	58.0%	0.95
6–10	17.1%	7.5	26.6	2.32%	33.6%	1.00
11–20	11.8%	14.4	14.9	2.82%	12.7%	0.90
$K>20$	12.5%	55.5	22.1	7.51%	5.2%	0.70

Table 28: Family-size strata on COCO/GDino, $\alpha = 0.10$, 20 seeds. Pass rate is the fraction of seeds with pooled stratum FDP $\leq \alpha$; recall falls as families grow.

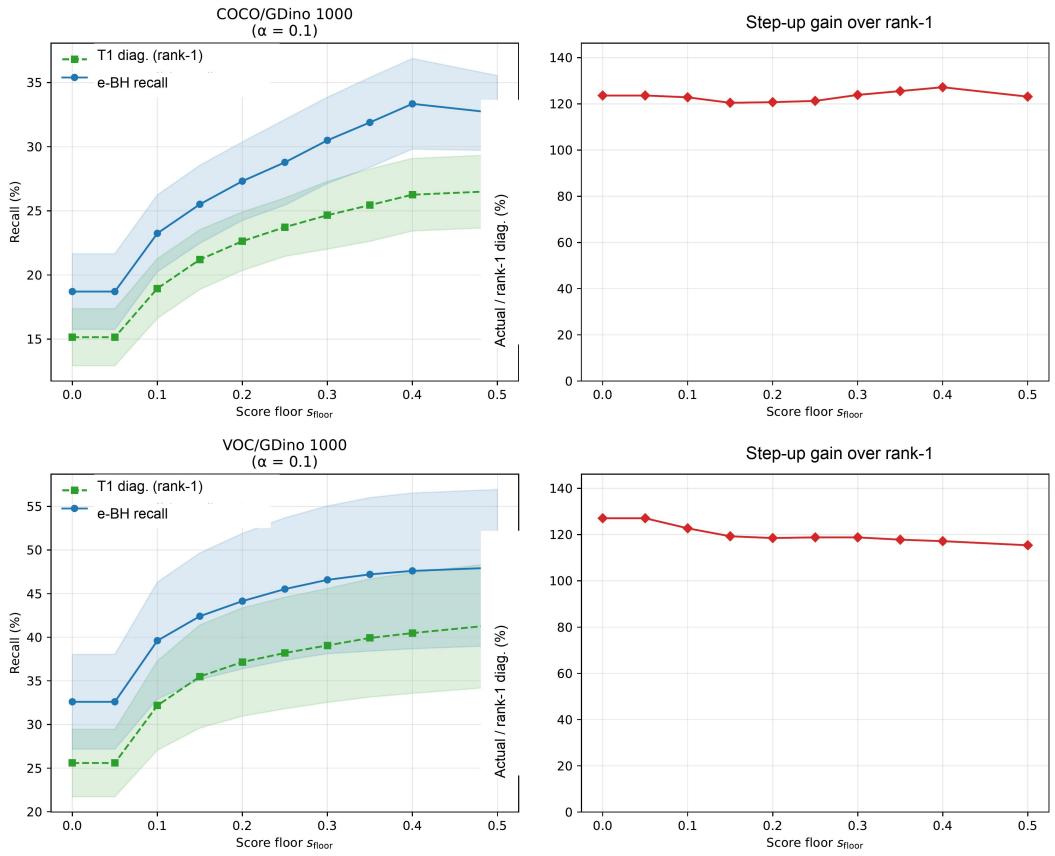

Figure 21: Score-floor frontiers at $\alpha = 0.10$, 20 seeds. Left: rank-1 diagnostic and e-BH recall. Right: their ratio, which can exceed 100% through later-rank crossings. Top/bottom: COCO/VOC with GDino.

FDP, recall, and LOO self-normalization means on nulls are reported per stratum; pass rate is the fraction of seeds with pooled stratum FDP $\leq \alpha$.

Score-quantile. Bins defined by per-seed null-score quantiles (5 equal-frequency bins). Table 29 (representative: COCO/GDino, $\alpha = 0.10$). Q1–Q4 receive zero rejections with LOO mean ≤ 0.54 (a conservative self-normalization check); Q5 (top quintile) holds all 134 rejections at pooled FDP 3.96% with LOO mean 3.29. The pattern appears in every setting (COCO/OWL-ViT, COCO/YOLO-World, VOC×3): rejections concentrate in the top quintile; pass rate 1.00 for every bin.

Bin	n_{cand}	Rej.	Pooled FDP	Recall	LOO mean	Pass
Q1 (low)	1131	0.0	0.00%	0.0%	0.26	1.00
Q2	1124	0.0	0.00%	0.0%	0.22	1.00
Q3	1145	0.0	0.00%	0.0%	0.37	1.00
Q4	1172	0.0	0.00%	0.0%	0.54	1.00
Q5 (high)	1592	134.0	3.96%	25.9%	3.29	1.00

Table 29: Score-quantile strata on COCO/GDino, $\alpha = 0.10$, 20 seeds. Releases concentrate in the highest-score quintile; LOO is a normalization diagnostic.

Object size (COCO area thresholds). Small ($\leq 32^2$), medium (32^2–96^2), large ($> 96^2$). At $\alpha = 0.10$, COCO/GDino has pooled FDP of 4.07%, 3.59%, and 3.97% for small, medium, and large objects. The small-object groups for OWL-ViT and YOLO-World reach pooled FDP as high as 16.7%. Errors can concentrate in small objects even when aggregate error is lower.

Per-category. Each category has only about 50 test families per seed, so its pooled FDP varies substantially across splits. Mean per-category pooled FDP is 2.28% on COCO/GDino and 12.6% on VOC/GDino at $\alpha = 0.10$. These summaries describe variation across categories; the main theorem averages over the specified query distribution.

G.4 Cluster construction and release comparisons

G.4.1 Candidate-pooled cluster calibration

The cluster ablations compare three candidate-pooled rules on the same detector output. The base rule tests boxes; E_{unif} averages candidate-pooled scores within each cluster; E_{cal} calibrates a cluster statistic, usually the maximum evidence, against null clusters. Let n_c count calibration null clusters. Specifically,

$$E_{\mathrm{unif}}(\mathcal{C}) = \frac{1}{|\mathcal{C}|}\sum_{k\in\mathcal{C}} e_k, \quad E_{\mathrm{cal}}(\mathcal{C}) = \frac{(n_c+1)R(\mathcal{C})}{R(\mathcal{C}) + \sum_{\ell=1}^{n_c} R(\mathcal{C}_\ell^{\mathrm{cal}})}.$$

The cluster statistic R is the maximum ($R_{\max}$), sum, or mean of the highest two member evidence values, as specified by each ablation.

Proposition 5 (Sufficient conditions for candidate-pooled cluster validity)**.** *For an all-null cluster, uniform averaging is valid if, conditional on the realized partition, every member input has conditional expectation at most one. The candidate-pooled calibrated statistic is valid if the test and calibration null-cluster statistics are jointly exchangeable and nonnegative.*

Proof. Under the first condition the member set and uniform weights are fixed after conditioning, so linearity gives a conditional expectation at most one. Under the second condition, the normalized-share argument in Lemma 2 applies to the cluster statistics. A zero denominator is assigned zero. □

Per-family FDR requires the fixed-family or conditional input conditions in Proposition 4. The block-normalized cluster rule and its proof in Section 3.2 and Appendix A.2 accommodate detector-dependent partitions directly. The tables below retain the candidate-pooled rules' measured comparisons.

G.4.1.1 Candidate-pooled cluster-aggregation diagnostic

Table 30 gives the six-configuration candidate-pooled comparison of baseline, E_{unif}, and E_{cal}. Its recall rows use different denominators. Figure 5 gives a block-protocol comparison with a common GT denominator.

G.4.1.2 Within-metric cluster-recall comparison

The comparison below gives every method the same denominator: the total number of correct clusters in the overlap graph. Recall counts the fraction selected. This isolates changes in release decisions under a common cluster-recall metric.

Config	Baseline (obj.)	E_{unif}	E_{cal}
COCO/GDino	29.7%	29.1% ($0.98\times$)	32.1% ($\mathbf{1.08}\times$)
VOC/GDino	51.6%	51.6% ($1.00\times$)	54.3% ($\mathbf{1.05}\times$)
COCO-tr/GDino	31.4%	31.5% ($1.00\times$)	33.8% ($\mathbf{1.08}\times$)
COCO/OWL-ViT	13.8%	13.7% ($0.99\times$)	15.6% ($\mathbf{1.13}\times$)
COCO/RegionCLIP	$21.1\%^{\ddagger}$	—	$\mathbf{37.4}\%^{\ddagger}$ ($\mathbf{1.77}\times^{\ddagger}$)

‡RegionCLIP uses no internal NMS. Its baseline and E_{cal} rows use clustering threshold $\tau = 0.10$, selected from the evaluated sweep (Table 31). The E_{unif} result uses $\tau = 0.50$ and is listed with that setting in Table 30. LVIS results are also reported there, alongside the available-label error rates. Under this common cluster-recall denominator, the measured E_{cal} ratios are 1.05–$1.13\times$ for internally NMS-filtered detectors and $1.77\times$ for the stated RegionCLIP setting.

Config	Method	Rej.	Pooled FDP	Mean FDP	Recall	Ratio‡
COCO/GDino	baseline	134.0	4.0%	0.7%	18.7%	—
	E_{unif}	129.7	4.4%	—	29.1%	1.56×
	E_{cal}	142.7	4.1%	0.7%	**32.1%**	**1.72×**
VOC/GDino	baseline	196.4	5.6%	1.7%	32.6%	—
	E_{unif}	193.5	6.0%	—	51.6%	1.58×
	E_{cal}	202.2	5.4%	1.6%	**54.3%**	**1.66×**
COCO-tr/GDino	baseline	151.6	8.1%	—	20.0%	—
	E_{unif}	151.7	8.4%	—	31.5%	1.58×
	E_{cal}	162.2	8.0%	1.2%	**33.8%**	**1.69×**
COCO/OWL-ViT	baseline	60.9	2.3%	—	11.2%	—
	E_{unif}	57.1	1.8%	—	13.7%	1.22×
	E_{cal}	65.0	1.7%	0.3%	**15.6%**	**1.39×**
COCO/RegionCLIP	baseline	52.9	0.1%	—	3.7%	—
	E_{unif}	29.1	0.2%	—	8.6%	2.30×
	E_{cal}	72.6	0.2%	0.0%	**21.6%**	**5.81×**
LVIS/GDino†	baseline	37.0	17.7%	—	3.9%	—
	E_{unif}	39.8	18.0%	—	6.4%	1.64×
	E_{cal}	54.0	18.7%	1.5%	8.6%	2.20×

Table 30: Cluster aggregation at $\tau = 0.50$, $\alpha = 0.10$: 20 seeds, except LVIS (5). Baseline recall counts candidates; $E_{\text{unif}}/E_{\text{cal}}$ count clusters. ‡: cluster/candidate recall ratio; Appendix G.4.1.2 uses a common denominator. †: LVIS uses available annotations; verified-query results appear in Appendix G.2.6.

G.4.1.3 IoU THRESHOLD SWEEP

Table 31 reports full 20-seed results at $\tau \in \{0.10, 0.20, 0.30, 0.50, 0.70\}$ for three detector–dataset combinations.

Config	τ	Rej.	Pooled FDP	Per-family FDP	Recall
COCO/GDino	0.30	143.8	3.1%	0.7%	39.2%
	0.50	142.7	4.1%	0.7%	32.1%
	0.70	134.0	4.0%	—	18.7%
VOC/GDino	0.30	196.3	5.4%	—	62.0%
	0.50	202.2	5.4%	1.6%	54.3%
	0.70	196.4	5.6%	—	32.6%
COCO/RegionCLIP	0.10	77.1	0.6%	0.1%	**37.4%**
	0.20	78.8	0.4%	—	35.3%
	0.30	77.8	0.4%	—	31.9%
	0.50	72.6	0.2%	0.0%	21.6%
	0.70	62.6	0.1%	—	8.5%

Table 31: Clustering IoU ablation at $\alpha = 0.10$, 20 seeds, using E_{cal} and $R_{\max}$. Lower τ produces coarser clusters.

G.4.1.4 CLUSTER STATISTICS AND GROUP-SPECIFIC CALIBRATION

Table 32 compares three choices of cluster statistic and a Mondrian (per-cluster-size) calibration pool. $R_{\max}$ has higher recall across three sources and all four α levels; the Mondrian variant underperforms the global pool because it shrinks the calibration sample available for large clusters.

G.4.2 CANDIDATE AND OBJECT RECALL

The candidate-level recall reported in Table 23 treats every GT-matched candidate as a separate TP. NMS-cluster aggregation groups overlapping candidates before testing. Table 33 compares candidate-level metrics from the candidate-pooled pipeline against object-level metrics from the calibrated NMS-aware variant ($\tau = 0.50$, E_{cal} with $R_{\max}$).

Source	R-stat / pool	Rej.	Pooled FDP	Per-family FDP	Recall
COCO/GDino	$R_{\max}$ (global)	142.7	4.1%	0.7%	**32.1%**
	R_{sum}	135.4	4.0%	0.7%	30.7%
	R_{top2}	131.2	4.0%	0.7%	29.8%
	$R_{\max}$ (Mondrian by cluster size)	131.1	3.9%	0.7%	29.6%
VOC/GDino	$R_{\max}$	202.2	5.4%	1.6%	**54.3%**
	R_{sum}	192.4	5.5%	1.6%	52.1%
	R_{top2}	188.6	5.5%	1.6%	51.1%
COCO-train/GDino	$R_{\max}$	162.2	8.0%	1.2%	**33.8%**
	R_{sum}	154.1	7.9%	1.2%	32.5%
	R_{top2}	149.5	7.9%	1.2%	31.7%

Table 32: Aggregation-statistic ablation at $\alpha = 0.10$, $\tau = 0.50$, 20 seeds. Mondrian calibration uses separate pools by family size.

Config	α	Cand. recall	Object recall	Cand. FDP	Object FDP	Gain (×)
COCO/GDino	0.05	14.7%	26.1%	2.74%	3.19%	1.77
COCO/GDino	0.10	18.7%	32.1%	3.96%	4.11%	1.71
COCO/GDino	0.15	21.0%	35.4%	5.07%	5.23%	1.68
COCO/GDino	0.20	22.8%	38.3%	5.98%	6.61%	1.68
COCO/OWL-ViT	0.10	11.2%	15.6%	2.31%	1.73%	1.39
VOC/GDino	0.10	32.6%	54.3%	5.62%	5.41%	1.67
VOC/GDino	0.20	38.8%	64.2%	8.66%	8.80%	1.65

Table 33: Cluster-level versus candidate-level recall and FDP at the same nominal level; the recall denominators differ.

G.4.3 Candidate-pooled baseline comparisons

The diagnostic baseline panel compares the unfiltered detector, calibration-tuned score thresholds, a CRC-style calibration, Storey-BH, Benjamini–Krieger–Yekutieli (BKY) two-stage BH, top-k release, and a family-size rule inspired by independent hypothesis weighting (IHW). They use the same candidate pools; p-value procedures share the pooled conformal p-values. Each baseline's implementation and measured behavior are described below. Block-based comparisons appear in Section 4.1.

Raw detector. With no post-processing, the released detection set has 88.7% pooled FDP on full-COCO/GDino; trivially candidate-level recall is 100%, but the precision penalty is severe. Comparing to candidate-pooled e-BH (pooled FDP 6.4%) gives an $8.3\times$ precision gain $(1 - 0.064)/(1 - 0.887)$ with the recall measured under each table's stated denominator. The same pattern holds across all six settings (Table 34); raw FDP ranges 39–90%.

Config	Raw rejections	Raw pooled FDP	Object-level recall proxy
COCO/GDino	6164	88.7%	1.00
COCO/OWL-ViT	2487	78.4%	1.00
COCO/YOLO-World	804	50.4%	0.94
VOC/GDino	5941	90.4%	1.00
VOC/OWL-ViT	1447	69.7%	1.00
VOC/YOLO-World	636	38.7%	1.00

Table 34: Raw-detector baseline (no e-BH, no thresholding) on the same families/splits as Table 23. Object-level recall proxy is $\min(1, |\text{TP cands}|/|\text{GT in present families}|)$.

Conformal Risk Control (Angelopoulos et al., 2024). The CRC-style row calibrates score thresholds using mean-family FDP and a finite-sample adjustment. FDP is not monotone in the threshold, so that adjustment alone does not establish the standard CRC guarantee. Its pooled FDP exceeds $\alpha = 0.10$ in four of six settings (COCO/GDino 33.9%, VOC/OWL-ViT 27.1%, COCO/YOLO 20.3%, VOC/GDino 20.2%). Pooled FDP and mean-family risk differ. The monotone-envelope CRC baseline explicitly uses a monotone loss envelope and complete image-block calibration.

Adaptive FDR baselines: Storey-BH (Storey, 2002) and BKY two-stage (Benjamini et al., 2006). Both apply within-family BH after estimating the null proportion $\hat{\pi}_0$ (Storey, $\lambda = 0.5$) or running stage-1 BH at $\alpha/(1+\alpha)$ (BKY). Across all six settings at $\alpha = 0.10$, both methods *simultaneously* violate pooled and per-family targets (e.g. Storey-BH on COCO/GDino: pooled FDP 39.7%, per-family 20.3%; BKY: 30.2%, 18.2%). Applying these adaptive procedures requires checking their p-value and dependence assumptions for the selected candidates.

Calibration-tuned top-k release. Within each family, this rule returns the top-k candidates by detector score. Calibration selects the largest $k \in \{1, 2, 3, 5, 10\}$ with observed mean-family FDP at most α. On COCO/GDino at $\alpha = 0.10$, most splits select $k = 1$. Pooled FDP reaches 67.4%: every nonempty absent-query family contributes a false release.

Family-size weighting (IHW-lite) (Ignatiadis et al., 2016). We group families by size K into bins $\{1, 2\text{–}5, 6\text{–}10, 11\text{–}20, >20\}$. In each bin b, we estimate the null proportion $\hat{\pi}_{0,b}$ with Storey's estimator and apply BH at $\alpha/\hat{\pi}_{0,b}$. We call this simplified comparison IHW-lite. Detector-dependent family size requires a separate check of the null-independence condition used by IHW. At $\alpha = 0.10$, the observed pooled FDP ranges from 6.5% to 15.6%, and recall ranges from 21% to 76% (Table 35). VOC/GDino exceeds the pooled target in all 20 splits.

Related adaptive methods. Adaptive p-value thresholding (AdaPT) (Lei & Fithian, 2018) and learned-score FDR procedures (Marandon et al., 2022) use side information under assumptions on null p-values and adaptation. Our evaluated side-information comparison uses family size. Its results accompany Storey-BH and BKY in the candidate-pooled study.

Config	Rej	Pooled FDP	Per-fam FDP	Recall	Pass (pool)	Pass (fam)
COCO/GDino	209	8.07%	7.90%	27.5%	0.80	0.80
COCO/OWL-ViT	146	7.77%	6.55%	25.1%	0.80	0.90
COCO/YOLO-World	212	10.35%	8.21%	47.6%	0.60	0.85
VOC/GDino	281	15.59%	19.13%	41.8%	0.00	0.00
VOC/OWL-ViT	100	6.50%	5.49%	21.3%	0.85	0.95
VOC/YOLO-World	322	7.95%	7.21%	76.1%	0.80	0.90

Table 35: IHW-lite with family-size covariates, $\alpha = 0.10$, 20 seeds. Pass is the fraction of seeds with FDP $\leq \alpha$.

G.4.4 Error–recall curves

Figure 22 visualises the FDP–recall trade-off of the baseline panel. The grey curve sweeps an oracle score threshold across all test candidates (threshold chosen with knowledge of the test labels); the coloured markers show each method's operating point at $\alpha = 0.10$, averaged over 20 splits.

Candidate-pooled e-BH (blue circle) is the highest-precision measured operating point in this diagnostic plot. The calibration-tuned score threshold (brown cross) is near the pooled-FDP line on average but exceeds it on 13 of 20 splits. Pooled empirical p-BY has higher recall and lower pooled precision.

G.4.5 Matching-overlap thresholds

We vary the matching threshold $\tau \in \{0.3, 0.5, 0.7\}$ on COCO/GDino (5 seeds). A lower τ increases the TP pool and makes the task easier; a higher τ shrinks it.

τ	FDP ($\alpha = 0.10$)	Recall	$\phi_{\max}$
0.3	$3.39 \pm 1.63\%$	10.9%	274
0.5	$4.66 \pm 2.67\%$	19.3%	316
0.7	$6.98 \pm 2.69\%$	33.2%	295

Table 36: Matching-IoU sensitivity on COCO/GDino, $\alpha = 0.10$, 5 seeds. Stricter matching reduces the TP denominator.

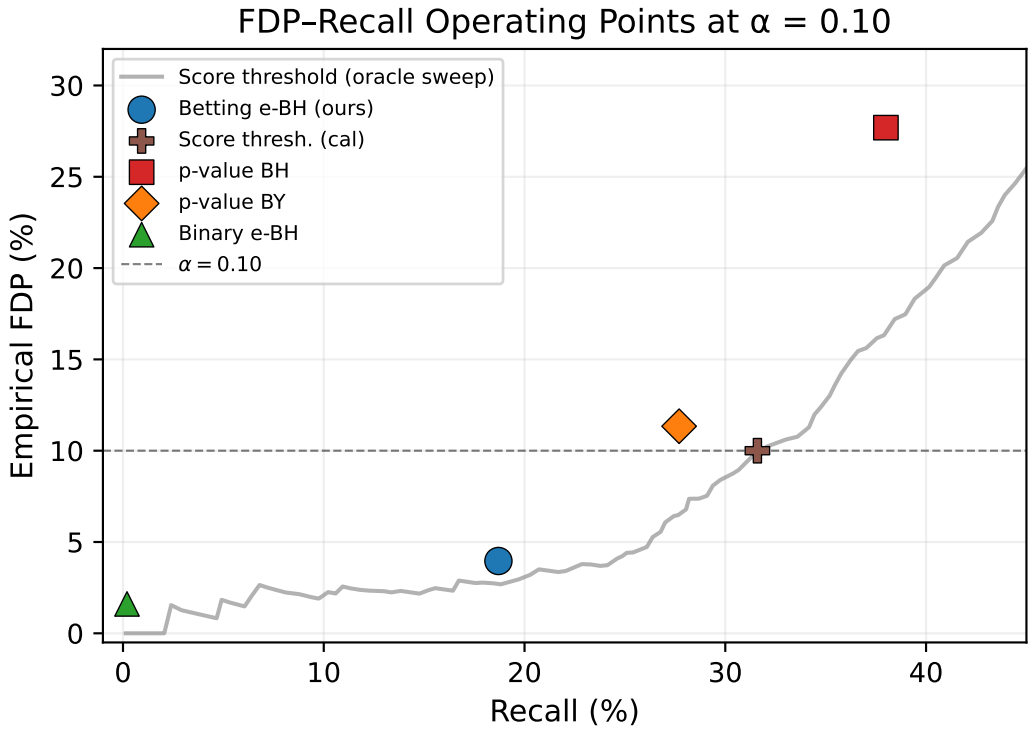

Figure 22: Candidate-pooled FDP–recall curves on COCO/GDino, $\alpha = 0.10$, 20 splits. Gray: threshold chosen with test labels. Dashed: nominal target.

G.5 Query composition and evidence estimation

G.5.1 Present and absent query proportions

We compare present-to-absent query ratios of 1:1 and 1:2 on the 1,000-image COCO/GroundingDINO subset at $\alpha = 0.10$. Both use the same detector outputs, 20 random family-level splits, and an unsmoothed 80-bin histogram. The family list is resampled to change the query ratio.

Present:Absent	Families	Cal null	Rej.	Per-fam FDP	Pooled FDP	Recall	Recall@0.5	LOO	Pass
1:1	1844	4527.4	127.5	0.884%	3.79%	18.0%	54.3%	1.000	20/20
1:2	2686	5288.7	134.0	0.648%	3.96%	18.7%	56.4%	1.000	20/20

Table 37: Present/absent query ratios on COCO-1k/GDino, $\alpha = 0.10$, 20 seeds. Cal null: calibration null candidates. Recall@0.5 uses TPs with score ≥ 0.5 as its denominator.

G.5.2 Fixed-template prompts and score calibration

We evaluate fixed-template prompts on 2,000 COCO/GDino images. Each image receives one present and two absent queries drawn from seven templates. Examples include "a visible {category} in the scene" and "a photo of a {category}." This produces 5,984 image–query families and 167,515 candidates, with mean family size $\bar{K} = 28.0$. We repeat the fit, calibration, and test split for 20 seeds.

Fitting each template separately improves release quality in this experiment. A global density-ratio model produces 3.2 releases at 41% pooled FDP. Per-template fitting produces 98.6 releases at 3.9% pooled FDP. The following analysis examines the evidence and family sizes behind this difference.

Prompt-conditional activation-frontier diagnostic. The release thresholds explain the difference between global and template-specific fitting. The global fitted evidence peaks at $\phi_{\max} \approx 14.5$. At $\alpha = 0.10$, the first-rank threshold is $T_1(C_m) \approx 283$ for mean family size $\bar{K} \approx 28$. Their ratio

Method	Rej.	FDP	Recall	Pass	LOO mean
Global betting e-BH	3.2	41.0%	0.11%	5/20	1.000
Per-template betting e-BH	**98.6**	**3.9%**	**5.6%**	**20/20**	**1.000**
Score oracle†	56.5	9.4%	3.0%	—	—
p-value BH	520.8	60.7%	12.1%	0/20	—
Binary e-BH	0.0	—	0.0%	20/20	—

†Score threshold selected using test labels.

Table 38: Fixed-template pilot on COCO/GDino: 2,000 images, 5,984 families, $\alpha = 0.10$, 20 seeds. Global/per-template fitting uses one shared/separate density-ratio model. LOO is a self-normalization diagnostic.

is approximately 0.051. Table 39 shows how template-specific fitting raises evidence toward the release thresholds.

Template	$\bar{K}$	$\phi_{\max}^{\text{global}}$	$\phi_{\max}^{\text{tmpl}}$	Rank-1 T_1	Ratio	Below?
a small {cat} near center	52.6	14.5	14.5*	535	0.027	✓
a {cat} in the scene	49.5	14.5	14.5*	503	0.029	✓
a {cat} object	25.0	14.5	77.0	252	0.306	✓
a photo of a {cat}	21.1	14.5	18.9	212	0.089	✓
a visible {cat} in scene	14.9	14.5	**119.3**	150	**0.796**	✓
the {cat}	14.5	14.5	**95.0**	145	**0.655**	✓
other	14.6	14.5	14.5*	147	0.099	✓

Table 39: Prompt-conditional rank-1 diagnostic, $\alpha = 0.10$, 20 seeds and 2,000 images. Peaks $\phi_{\max}^{\text{global}}/\phi_{\max}^{\text{tmpl}}$ use shared/per-template fitting; ratio is $\phi_{\max}^{\text{tmpl}}/T_1$. *Absent-only templates fall back to global fitting.

The template experiment changes both family size and evidence strength. Mean family size rises from about 10 for category names to about 28 for fixed templates. At fixed calibration evidence, this raises the rank thresholds. The global fitted evidence peak also falls from 357 to 14.5. These observations motivate separating the templates during calibration and fitting.

Separating calibration pools by template. We first keep the global score map and calibrate each template separately. Releases remain near zero at $\alpha = 0.10$. The within-template LOO means are 1.000 for every template and seed, as expected from self-normalization. Separating the calibration pools leaves the global score map's weak evidence unchanged.

Fitting the score map by template. The next comparison fits a separate density-ratio model using each template's fit data and calibration-null pool. Templates with no correct fit detections use the global model. The single-score result then assumes exchangeable null scores within each template. The block construction supplies the separate argument for detector-dependent families.

Per-template fitting boosts $\phi_{\max}$ by 5–8× for templates that contain true positives (Table 39): e.g. "a visible {cat} in the scene" reaches $\phi_{\max} = 119.3$ (global: 14.5). This brings small-K families close to or past the activation frontier, recovering substantial power. At $\alpha = 0.10$, per-template betting e-BH achieves 98.6 rejections at 3.9% pooled FDP and 5.6% recall, with *all 20 seeds* passing the FDP$\leq \alpha$ check (Table 38). This surpasses even the (non-deployable) score oracle (56.5 rejections, 9.4% FDP), because per-template ϕ captures template-specific score structure that a single score threshold cannot.

The template bank is shared by fitting, calibration, and testing throughout this comparison.

Calibration support per template. The seven-template pilot improves with per-template fitting. A separate 500-image sweep increases the template count while holding the image budget fixed. At $T = 50$, each template receives about six calibration families. The evidence peak falls to about 5.7, compared with 10.4 at $T = 7$ on the same image budget. Mean releases fall to 0.05. At $T = 100$ and $T = 200$, mean releases remain at or below 0.1. This comparison shows the cost of sparse template-specific calibration data.

Runtime. These diagnostic experiments ran on a single NVIDIA RTX 3090. Candidate generation (GroundingDINO forward pass on 1,000 images) takes $\sim$280 s; the post-hoc statistical pipeline (density-ratio fitting, e-value computation, and e-BH testing across all families) takes $\sim$13 s for 5 seeds ($\sim$2.6 s per seed). The method requires no gradient-based training.

G.5.3 Growing the prompt bank

We test how validity and power change as the prompt bank grows. We use saved COCO/GDino outputs for 2,000 images and eight templates, such as "a photo of a {category}." The bank size is $T \in \{1, 2, 3, 4, 8\}$. Each setting uses the same 60/20/20 fit/calibration/test proportions, 20 seeds, and an unsmoothed 80-bin histogram. Pooled-template calibration fits one map across all templates. Per-template calibration fits a separate map ϕ_t for each template. Table 40 reports the $\alpha = 0.10$ results.

T	Mode	Cal/T	Rej	Pooled FDP	Per-fam FDP	Recall	LOO mean
1	pooled $\equiv$ per-tmpl	10165	0.2	0.0%	0.0%	0.04%	1.000
2	pooled	9190	1.3	0.0%	0.0%	0.13%	0.998
2	per-tmpl	9190	0.25	0.0%	0.0%	0.03%	0.996
3	pooled	7921	3.4	0.0%	0.0%	0.25%	0.997
3	per-tmpl	7921	0.15	0.0%	0.0%	0.01%	0.998
4	pooled	7254	14.0	1.81%	1.81%	0.80%	0.982
4	per-tmpl	7254	1.25	0.0%	0.0%	0.07%	0.987
8	pooled	19865	65.6	**19.65%**	19.65%	3.12%	1.001
8	per-tmpl	19865	1.1	0.0%	0.0%	0.07%	0.186

Table 40: Candidate-pooled prompt scaling on COCO/GDino, $\alpha = 0.10$, 20 seeds. Modes share image splits and 80-bin histograms without smoothing or clipping; LOO measures self-normalization.

Reading the table. The pooled-template experiment shows increasing observed FDP as the bank grows, while per-template calibration makes few releases. Sparse support and absent-only fit groups affect the evidence map and power. Proposition 3 describes the population ratio under common mixture weights. The measured error also depends on calibration sampling. Appendix E evaluates the image-block prompt-bank method and reports both mean-family and pooled diagnostics.

G.5.4 Score-map estimator comparison

Table 41 compares score-map estimators on COCO/GroundingDINO over five splits. The alternatives are a histogram, logistic regression, kernel density estimation (KDE), and isotonic regression. The full target sweep across detectors appears in Table 22.

Estimator	Rejections	FDP	Recall
Unsmoothed histogram	132.8 ± 15.7	$4.66 \pm 2.67\%$	$19.27 \pm 3.01\%$
Logistic regression	122.4 ± 6.4	$4.68 \pm 3.11\%$	$17.75 \pm 2.03\%$
KDE	122.2 ± 20.3	$3.99 \pm 2.94\%$	$17.69 \pm 2.38\%$
Isotonic	113.0 ± 47.3	$3.72 \pm 3.30\%$	$16.18 \pm 6.54\%$

Table 41: Estimator ablation on COCO-1000/GDino at $\alpha = 0.10$, 5 seeds. The unsmoothed histogram has the highest observed recall.

G.5.5 Score-map estimators and smoothing

We fit five alternative estimators of ϕ on the same fitting split and apply each map to the candidate-pooled rule. All maps are nonnegative. This ablation measures how evidence estimation changes releases and error rates. The alternatives include smoothed histograms, logistic regression, kernel density estimation (KDE), and isotonic regression. Table 42 reports COCO/GDino at $\alpha = 0.10$ across 5 seeds.

Estimator	Rejections	Pooled FDP	Recall
hist + smooth 0.5	130.8	4.57%	19.0%
hist (80-bin, default)	129.4	4.60%	18.8%
hist + smooth 1.0	129.4	4.60%	18.8%
hist + smooth 2.0	125.8	4.71%	18.2%
logistic	122.4	4.68%	17.8%
KDE	122.2	3.99%	17.7%
hist + smooth 5.0	116.0	4.60%	16.8%
isotonic	113.0	3.72%	16.2%

Table 42: Density-ratio estimators on COCO/GDino at $\alpha = 0.10$, 5 seeds. KDE denotes kernel density estimation.

G.5.6 CATEGORY-SPECIFIC CALIBRATION

A key assumption is that null scores are approximately exchangeable across families sharing the same calibration pool. If different prompt categories induce systematically different null-score distributions, the global calibration could over- or under-weight certain families.

To test this, we ran Mondrian (group-conditional) calibration on COCO/GDino, splitting the calibration null pool by prompt identity (each prompt gets its own C_m and self-normalized denominator; prompts with fewer than 200 nulls fall back to the global pool).

Calibration	Seeds	Pooled FDP	Family FDP	Recall
Global	20	$3.96 \pm 1.98\%$	0.65%	18.7%
Mondrian (prompt)	5	$4.73 \pm 2.86\%$	0.68%	17.5%

Table 43: Global versus per-prompt (Mondrian) calibration on COCO/GDino at $\alpha = 0.10$.

Per-prompt calibration has similar observed error and slightly lower recall in this comparison. Each prompt has a smaller calibration pool. The main block experiment uses complete images to preserve dependence across queries.

G.6 CALIBRATION SUPPORT AND ANNOTATION QUALITY

G.6.1 CALIBRATION SAMPLE SIZE

We vary the calibration pool size m on COCO/GDino at $\alpha = 0.10$ (5 seeds) to assess sensitivity to calibration data quantity. The mean observed FDP is below nominal for all tested pool sizes, including $m = 500$, while recall varies little across these settings.

m	FDP ($\alpha = 0.10$)	Recall	Rejections
500	$4.92 \pm 2.81\%$	18.0%	124.8
1000	$4.66 \pm 2.69\%$	18.5%	127.4
2000	$4.93 \pm 2.52\%$	18.6%	128.8
5000	$3.43 \pm 0.56\%$	19.5%	132.3
all ($\sim$5200)	$5.39 \pm 3.08\%$	18.3%	127.5

Table 44: Calibration-pool size on COCO/GDino at $\alpha = 0.10$, 5 seeds.

G.6.2 SMALL CALIBRATION POOLS

We subsample the calibration null pool to $m \in \{50, 100, 200, 500, 1000, 2000, \text{full}\}$ candidates and re-run within-family e-BH. Table 45 reports COCO/GDino at $\alpha = 0.10$ across 20 seeds.

G.6.3 MISSING-POSITIVE ANNOTATION SWEEP

Table 46 reports the full contamination sweep from Figure 23. For each nominal contamination level η_{fam} (fraction of present-object families hidden), we compute η_{actual} (fraction of the null pool that is truly TP), the observed $\phi_{\max}$, true FDP (evaluated against clean labels), apparent FDP (evaluated against corrupted labels), and recall.

m	Rejections	Pooled FDP	Recall	LOO mean	No-rej rate	Pass rate
50	28.6	5.96%	3.9%	1.05 ± 0.45	96.1%	0.75
100	40.6	7.00%	5.5%	0.94 ± 0.35	94.4%	0.90
200	48.5	6.16%	6.6%	0.92 ± 0.27	93.2%	0.95
500	57.6	6.42%	7.8%	1.00 ± 0.27	92.0%	0.90
1000	58.2	6.27%	7.9%	0.94 ± 0.19	92.0%	0.90
2000	59.3	6.23%	8.1%	0.96 ± 0.23	91.8%	0.95
full	59.8	6.03%	8.1%	0.95 ± 0.19	91.8%	0.95

Table 45: Calibration-pool sweep on COCO/GDino, $\alpha = 0.10$, 20 seeds. LOO is the self-normalization diagnostic; m counts calibration null candidates.

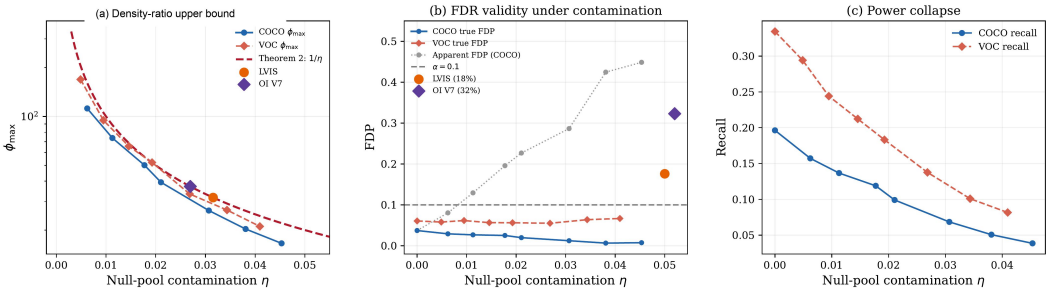

Figure 23: Controlled contamination sweep on COCO/VOC with GDino ($\alpha = 0.10$, 10 seeds): $\phi_{\max}$ tracks the $1/\eta$ upper bound, true FDP stays $\leq \alpha$ as apparent FDP rises, and recall collapses.

The key observations:

- $\phi_{\max}$ is consistently ∼75–80% of the theoretical upper bound $1/\eta_{\text{actual}}$, in this sweep; this is an observed relationship.
- True FDP remains well below $\alpha = 0.10$ even at $\eta_{\text{fam}} = 0.50$. Recall decreases across the sweep.
- The gap between apparent and true FDP grows with η: at $\eta_{\text{fam}} = 0.50$, COCO apparent FDP is 44.9% while true FDP is 0.7%. The corrupted labels count many correct detections as false.

Table 46: Contamination sweep on COCO/GDino and VOC/GDino ($\alpha = 0.10$, 10 seeds). η_{fam}: fraction of present families hidden. η_{act}: actual null-pool contamination fraction. $\phi_{\max}$ is the *mean* $\pm$ s.d. over seeds (Table 21 reports the *median* over 20 different seeds, accounting for the slight baseline difference).

η_{fam}	η_{act}	$\phi_{\max}$	$1/\eta_{\text{act}}$	True FDP	App. FDP	Recall
COCO / GroundingDINO						
0.00	0.000	327 ± 68	—	$3.7 \pm 2.2\%$	3.7%	19.6%
0.05	0.006	112 ± 28	161	$2.9 \pm 2.5\%$	8.0%	15.7%
0.10	0.011	74 ± 15	89	$2.7 \pm 1.6\%$	12.9%	13.7%
0.20	0.021	39 ± 4	48	$2.0 \pm 1.9\%$	22.7%	9.9%
0.30	0.031	26 ± 4	33	$1.2 \pm 1.5\%$	28.7%	6.8%
0.50	0.045	17 ± 1	22	$0.7 \pm 1.6\%$	44.9%	3.9%
VOC / GroundingDINO						
0.00	0.000	1577 ± 614	—	$6.1 \pm 1.7\%$	6.1%	33.4%
0.05	0.005	169 ± 47	205	$5.8 \pm 1.6\%$	10.7%	29.4%
0.10	0.009	95 ± 17	106	$6.2 \pm 1.9\%$	15.6%	24.4%
0.20	0.019	52 ± 7	52	$5.6 \pm 1.6\%$	22.3%	18.3%
0.30	0.027	33 ± 3	37	$5.5 \pm 2.0\%$	31.9%	13.8%
0.50	0.041	21 ± 2	24	$6.7 \pm 3.1\%$	53.9%	8.2%

Note that $\eta_{\text{actual}} \ll \eta_{\text{fam}}$ because the null pool is dominated by candidates from genuinely absent families. For example, hiding 50% of COCO's present families contaminates only ∼4.5% of the null pool, because absent families contribute ∼20× more candidates.

G.6.4 Sensitivity to a shifted null reference

We remove the highest-scoring $\rho\%$ of calibration null candidates before computing evidence. The test set stays fixed, creating a shift between calibration and test null scores. Table 47 measures the resulting change in FDP on COCO/GroundingDINO at $\alpha = 0.10$ across 20 seeds.

ρ	m_{cal}	Rej	Pooled FDP	Per-fam FDP	Recall	LOO mean	Pass (P/F)
0%	5128	59.8	6.03%	6.82%	8.1%	0.95 ± 0.19	0.95/0.95
5%	4872	88.3	7.98%	8.81%	11.8%	1.74 ± 0.30	0.75/0.45
10%	4615	97.5	8.77%	10.04%	12.9%	2.05 ± 0.31	0.70/0.45
20%	4102	105.3	9.63%	11.06%	13.8%	2.40 ± 0.34	0.50/0.40
30%	3590	113.1	10.53%	12.13%	14.7%	2.75 ± 0.36	0.50/0.20

Table 47: Calibration-null truncation on COCO/GDino, $\alpha = 0.10$, 20 seeds. Pass rates summarize pooled (P) and mean-family (F) FDP; LOO measures reference-pool normalization.

The LOO mean checks the normalization within the calibration pool. This experiment separately measures the effect of a changed null-score distribution on test predictions.

G.7 Computation and controlled family experiments

G.7.1 Postprocessing time

The release filter consists of one histogram fit on the fit split, one calibration aggregate ($\sum_j \phi(S_j^{\text{cal}})$), one ϕ evaluation per test candidate, and one within-family sort plus threshold check per test family. Detector inference is the dominant cost (~ 280 s for 1k images on a single RTX 3090) and is unchanged by our release filter. Table 48 reports release-filter wall time across the 6 settings (mean over 5 seeds, single CPU thread, NumPy 1.26 / pandas 2.0).

Config	Fit (ms)	Cal. (ms)	Evidence (ms)	e-BH (ms)	Total (ms)	Family (μs)
COCO/GDino	1.89	0.72	0.22	10.25	13.08	19.5
COCO/OWL-ViT	1.09	0.43	0.14	4.34	6.01	18.0
COCO/YOLO-World	0.74	0.32	0.10	3.13	4.29	15.4
VOC/GDino	1.69	0.71	0.22	10.21	12.83	18.7
VOC/OWL-ViT	0.85	0.34	0.10	3.58	4.88	14.6
VOC/YOLO-World	0.69	0.30	0.09	3.27	4.35	14.2

Table 48: Postprocessing wall time for each 1,000-image configuration, 5 seeds and one CPU thread. Total is per test split; Family is per query. Detector inference is excluded.

G.7.2 Controlled dependence and family size

We also study dependent families in a synthetic gene-set-style experiment. Family sizes are sampled from $K \in \{5, 10, 20, 40\}$ with probabilities $(0.35, 0.30, 0.25, 0.10)$; null scores are marginally $N(0, 1)$; and 15% of hypotheses are non-null with scores shifted by 2.5. We use 6000 fit nulls, 2000 fit alternatives, $m = 3000$ calibration nulls, the same 80-bin ϕ, and $\alpha = 0.10$. We compare independent families, equicorrelated Gaussian families ($\rho = 0.60$), and a bursty-null dependence condition in which null p-values are marginally uniform but share a family-level small-p burst state.

To check the activation frontier quantitatively, we sweep K under block-correlated families and compute the number of discoveries predicted by Theorem 2. The predicted and observed recall match exactly up to rounding, while the rank-1 threshold T_1 grows quickly with family size.

H Additional related work

OVD reliability and known prompts. OVD methods expand category coverage through vision–language pretraining, prompt adaptation, and open-world modeling (Gu et al., 2021; Zhong et al., 2022; Li et al., 2022; Du et al., 2022; Cho & Krähenbühl, 2024). Recent evaluations show that this flexibility can produce hallucinated or fine-grained false detections (Ruis et al., 2026; Bianchi et al., 2024). We study statistical control of predictions released by a frozen detector.

Setting	Method	Rej./fam.	Per-family FDP	Pooled FDP	Recall
Independent	Betting e-BH	0.252 ± 0.081	0.96 ± 0.42	5.90 ± 2.26	11.51 ± 3.69
Independent	p-BH	1.526 ± 0.050	8.43 ± 0.65	14.18 ± 0.95	63.59 ± 1.12
Independent	p-BY	0.927 ± 0.034	2.94 ± 0.47	5.94 ± 0.82	42.35 ± 1.19
Independent	Storey-BH	1.923 ± 0.063	14.01 ± 0.82	25.71 ± 1.40	69.33 ± 1.11
Independent	IHW-lite	1.639 ± 0.054	9.89 ± 0.75	16.23 ± 1.05	66.63 ± 1.14
Block corr.	Betting e-BH	0.299 ± 0.093	1.03 ± 0.40	10.33 ± 4.87	13.08 ± 4.18
Block corr.	p-BH	1.685 ± 0.099	6.81 ± 0.68	25.53 ± 2.72	60.98 ± 1.92
Block corr.	p-BY	0.988 ± 0.057	2.53 ± 0.42	11.40 ± 2.26	42.56 ± 1.84
Block corr.	Storey-BH	3.825 ± 0.207	21.87 ± 1.07	65.37 ± 1.37	64.36 ± 1.96
Block corr.	IHW-lite	1.832 ± 0.123	7.94 ± 0.82	28.51 ± 2.87	63.59 ± 2.28
Bursty null	Betting e-BH	0.251 ± 0.082	0.98 ± 0.40	6.81 ± 2.41	11.40 ± 3.76
Bursty null	p-BH	1.707 ± 0.071	8.49 ± 0.72	23.17 ± 1.74	63.84 ± 1.24
Bursty null	p-BY	0.938 ± 0.038	2.94 ± 0.43	7.13 ± 0.97	42.43 ± 1.33
Bursty null	Storey-BH	3.833 ± 0.217	17.58 ± 1.09	61.75 ± 1.70	71.23 ± 1.26
Bursty null	IHW-lite	1.937 ± 0.103	10.03 ± 0.95	28.80 ± 2.28	67.05 ± 1.33

Table 49: Non-vision family-FDR benchmark, 200 repetitions. FDP/recall are percentages (mean $\pm$ s.d.); pooled FDP weights individual releases.

K	T_1	mean $\phi_{\max}$	$\Pr(\phi_{\max} \geq T_1)$	Emp. recall	Frontier recall
5	32.3	23.7	20.3	31.8	31.8
10	65.7	38.1	27.9	25.3	25.3
20	136.2	58.6	18.1	17.9	17.9
40	293.2	81.0	0.0	9.4	9.4
80	693.1	103.7	0.0	4.1	4.1

Table 50: Frontier sweep on the synthetic benchmark. Recall values are percentages. T_1 is the rank-1 frontier diagnostic; the "Frontier recall" column uses the full step-up sequence T_r and matches e-BH exactly.

Conformal/e-value FDR tools. Conformalized multiple testing reconstructs calibration data that emulate the selected distribution before applying BH (Wang et al., 2024). Our construction keeps score-dependent candidate generation inside each complete image block and calibrates these blocks. Conformal e-values support FDR-controlled novelty detection (Bashari et al., 2023). False-positive-controlled prediction sets constrain the number of incorrect outputs (Fisch et al., 2022). Conformal risk control calibrates expected loss for black-box predictors (Angelopoulos et al., 2024). Visual applications quantify box uncertainty (Timans et al., 2024) or calibrate zero-shot prediction sets (Silva-Rodríguez et al., 2025). Our block construction aligns detector-selected prediction families with a mean-family FDR target. The exact step-up condition characterizes when those predictions are released.

Missing labels and contamination. Non-exhaustive annotation changes the null pool that calibrates an e-value. Missing-label detection work (Xu et al., 2019; Yang et al., 2020) concerns training or dataset quality, while conformal contamination results (Bates et al., 2023; Barber et al., 2023) characterize validity under impure nulls. Our release condition connects annotation errors to the evidence required for selection.

Operating points and adaptive baselines. Adaptive FDR and false-positive-control methods can improve power when their calibration target matches the release decision (Gibbs et al., 2025; Marandon et al., 2022; Lei & Fithian, 2018; Fisch et al., 2022; Montreuil et al., 2024). We compare block e-BH with calibrated release rules under a shared complete-image protocol.

I VISUAL ATLAS OF RELEASE BEHAVIOR

The atlas preserves the evaluation protocols and shows all tested configurations: lines use the main-figure settings; background points and envelopes show the remaining configurations and their range. Query-slot methods use known present/absent roles. Table 9 gives cluster counts; Table 10 and Figure 7 give the worked example.

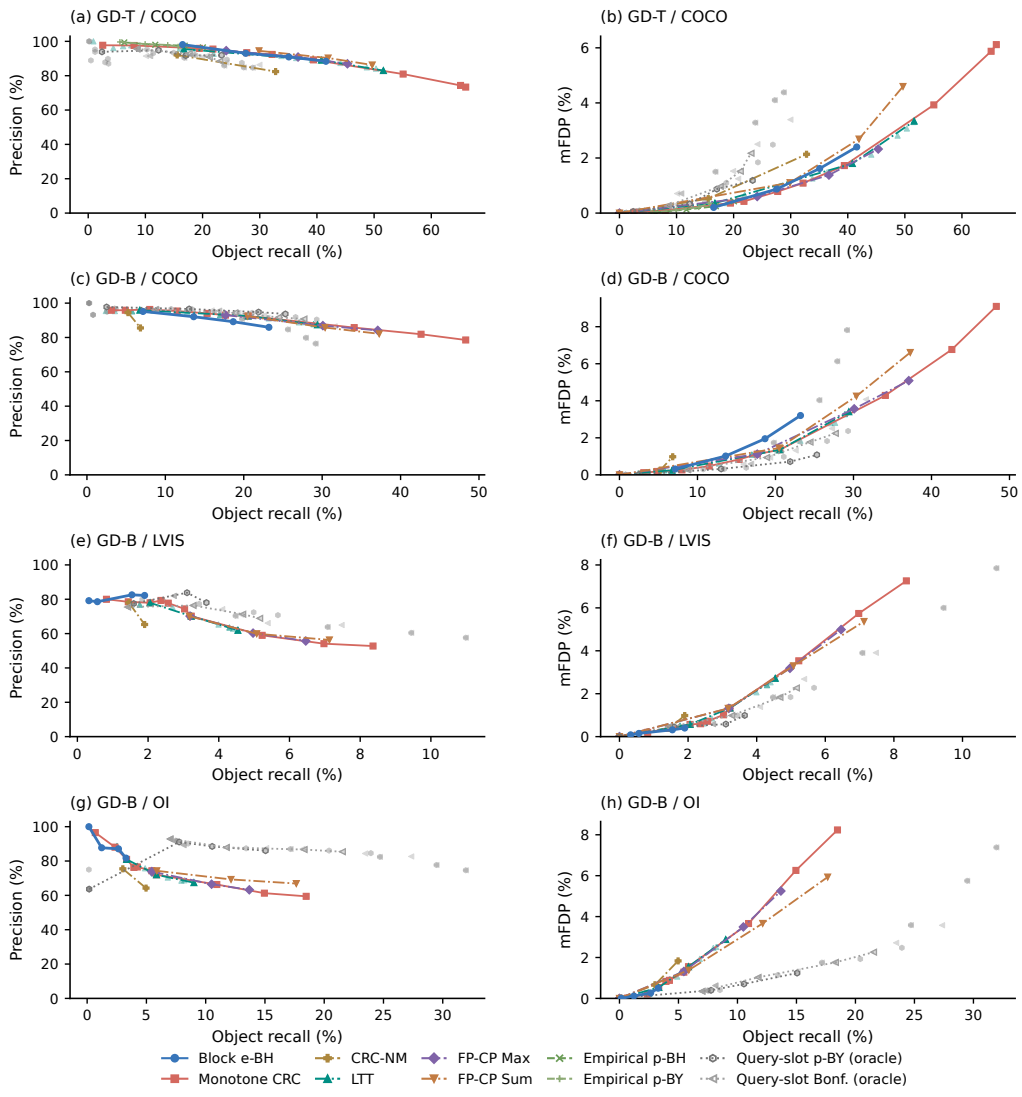

Figure 24: GroundingDINO operating curves: pooled precision (left) and mFDP (right) versus recall. Lines use LTT $\delta = 0.01$, FP-CP cap 20, and query-slot cap 3; points include all tested configurations. Empirical p-BH/p-BY appear only for COCO/tiny.

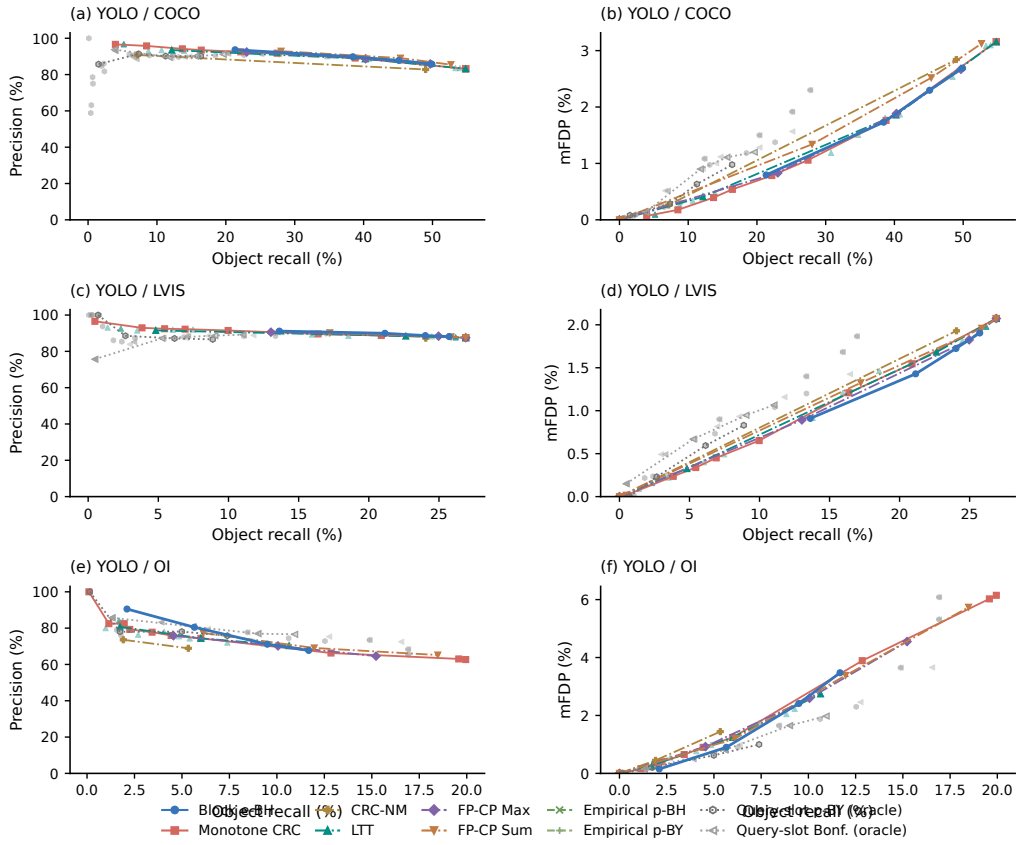

Figure 25: YOLO-World operating curves; encodings match Figure 24. Precision is undefined for zero releases.

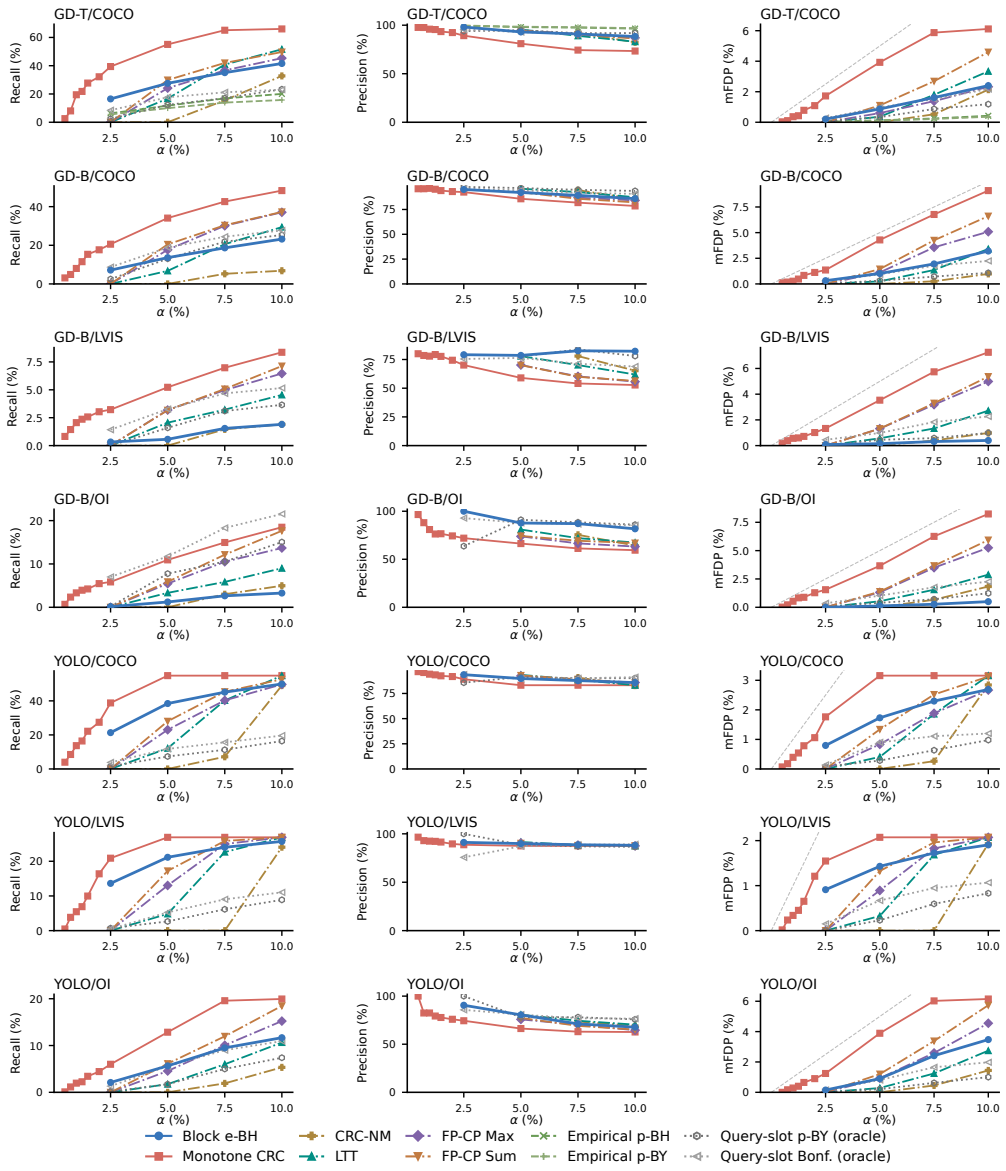

Figure 26: Nominal-level sweeps across detector–dataset settings. Diagonals mark mFDP $= \alpha$; configurations match the main figures. Query-slot rules use known present/absent roles.

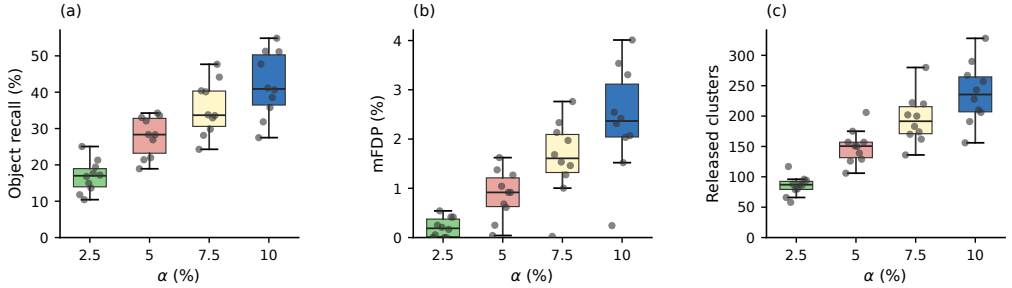

Figure 27: Block e-BH on COCO/GroundingDINO-tiny, ten overlapping splits per target. Boxes: median/quartiles. Whiskers: extreme points within 1.5 interquartile ranges. Points measure sensitivity to data partitioning.

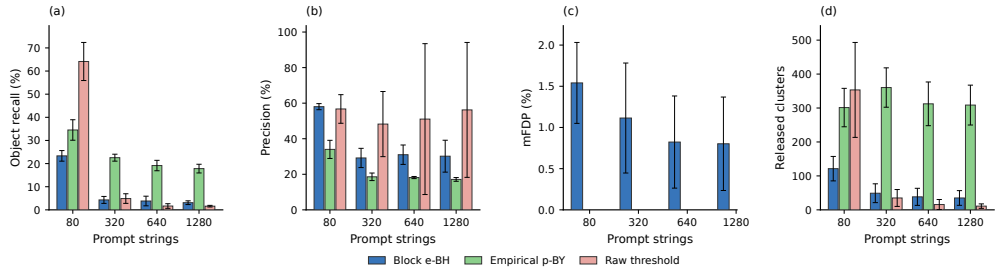

Figure 28: Cross-template cluster scaling, three splits. Bars/error bars: mean/sample s.d. Precision averages split-wise ratios. Empirical p-BY and raw thresholds are operating diagnostics; mFDP appears where measured.

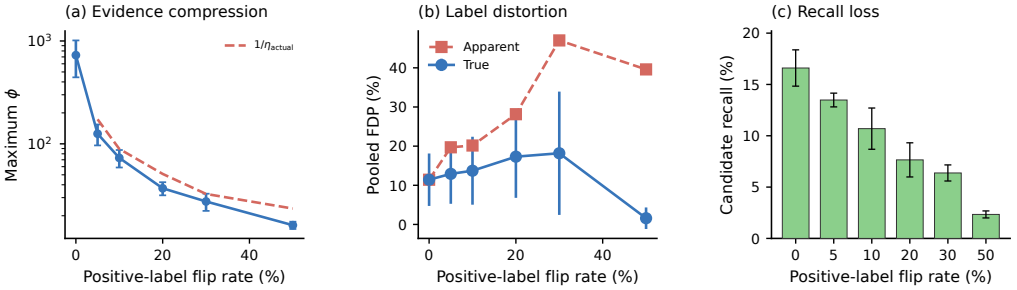

Figure 29: Candidate-pooled contamination: the flipped-positive fraction differs from null-pool contamination η_{actual}. Dashed: $1/\eta_{\text{actual}}$. FDP pools candidates; recall counts candidates. Error bars: split standard deviations.

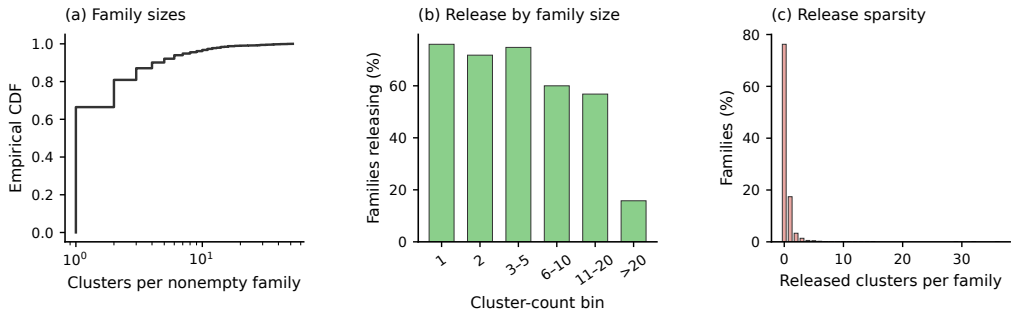

Figure 30: Three-query cluster study, floor 0.40 and $\alpha = 0.10$. (a) Sizes of nonempty families. (b) Release frequency by size. (c) Release counts over all 6,000 test-family occurrences, including empty families.

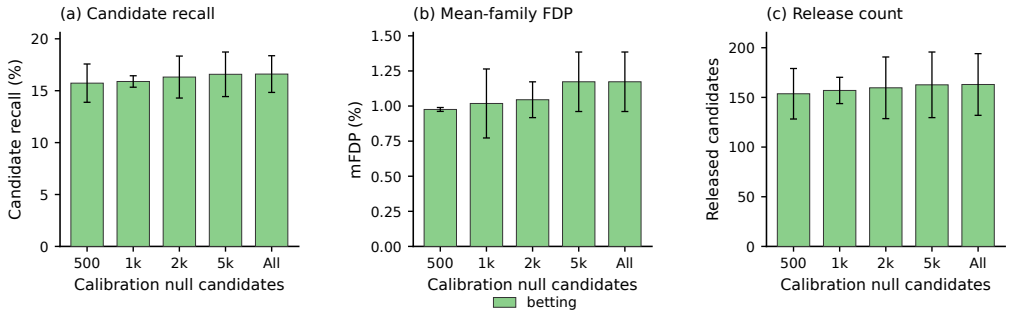

Figure 31: Candidate-pooled calibration support, measured in null candidates. Error bars: split standard deviations. Full pools contain 6,482–6,951 candidates per split; this protocol differs from image-block calibration.

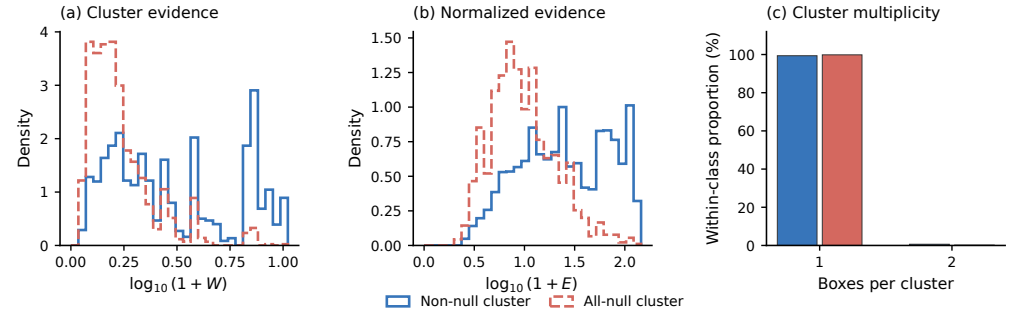

Figure 32: Evidence from 4,551 post-floor test clusters. (a,b) Within-class densities after $\log_{10}(1+x)$ transformation. (c) Within-class size proportions. Classes distinguish all-null clusters from those with a correct member.

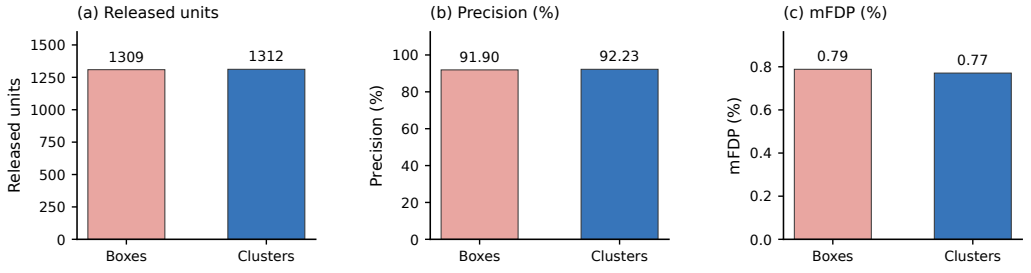

Figure 33: Box/cluster releases at $\alpha = 0.10$, ten block splits, no score floor. Counts, precision, and mFDP use the indicated unit. Candidate/object recall denominators differ (Table 8).