THE GEOMETRY OF AP DEGRADATION: A CLOSED-FORM, ZERO-PARAMETER THEORY OF LOCALIZATION SENSITIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Average Precision (AP) is discontinuous in box coordinates because it thresholds a continuous quantity, IoU — but exactly *where* that discontinuity falls, and why it produces the empirical fragility practitioners observe (a single-pixel shift can drop mAP by double digits, hitting small objects hardest), has been missing. We derive closed-form IoU expressions under translation and corner-anchored scaling giving the exact shift at which a detection flips from correct to incorrect, $t^*(\tau, W) = W \cdot \frac{1-\tau}{1+\tau}$, a zero-parameter formula in object size W and IoU threshold τ. At the dataset level, the full degradation curve is then an exact, rescaled survival function of the object-size distribution, so its shape can be read off a dataset's own size statistics before any detector is evaluated. This single formula unifies four otherwise-disconnected phenomena as corollaries of one mechanism: why small objects degrade fastest, why diagonal shifts hurt more than axis-aligned ones, why enlarging and shrinking have near-identical effect, and why averaged mAP degrades faster than AP_{50}. We validate against official `COCOeval` on real COCO annotations (MAE ≤ 0.032, correlation ≥ 0.997, transferring with no refitting to LVIS), and show the same closed form predicts real detectors' (Faster R-CNN, YOLOv11) *relative* degradation shape, once rescaled by their own baseline, under both fixed offsets and random Gaussian localization noise. The practical stakes are concrete: a one-pixel systematic shift of the kind a coordinate-convention or rounding mismatch between evaluation codebases can introduce, with no change to the detector, costs Faster R-CNN 2.5 mAP points — comparable to many reported state-of-the-art gains. The residual gap on real detectors is explained, not fitted: composing each true positive's own matched-IoU margin into the theory cuts the relative-curve error from 0.136 to 0.031. The theory is prescriptive as well as explanatory: inverting the formula yields a size-adaptive threshold that cuts the size bias of AP degradation roughly twentyfold on the idealized protocol ($5.4\times$ on five real detectors), and a smoothed variant removes the discontinuity while leaving detector rankings unchanged.

1 INTRODUCTION

Mean Average Precision (mAP) is the currency of progress in object detection: it ranks entries on the COCO leaderboard (Lin et al., 2014), decides which architectural change is kept, and serves as the headline number by which a decade of detectors has been compared. That entire edifice rests on a single primitive: a predicted box is a true positive iff its Intersection-over-Union (IoU) with a ground-truth box meets a fixed threshold τ. This hard threshold is convenient and interpretable, but it also makes the metric a step function of box geometry — and the consequences of that fact for the reliability of AP as a measure of detection quality are not well understood.

A growing body of evidence suggests the consequences are severe. As we show empirically in Section 3, perturbing predicted boxes by a single pixel — a change far below the threshold of visual perception — can drop mAP by double digits, small objects are hit disproportionately hard, diagonal shifts appear to hurt more than axis-aligned ones, and averaged mAP degrades faster than AP_{50}. Individually, each is an alarming but isolated observation; collectively, they raise a question empirical measurement alone cannot answer: are these fragilities artifacts of a particular dataset,

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.

annotation protocol, or detector, or intrinsic to how AP is defined? Without knowing *why* they occur, it is impossible to say whether they generalize, or to design a metric that avoids them.

In this paper we show that all of these behaviors are consequences of the geometry of IoU itself, derivable in closed form and free of any dataset, detector, or fitted parameter. That AP is discontinuous in box coordinates is, on its own, an immediate consequence of thresholding a continuous quantity, true of any hard threshold applied to any continuous score, and not the contribution we claim. The contribution is quantitative: we obtain exact expressions for IoU under the two perturbations studied empirically — translation and corner-anchored scaling — and use them to derive the precise shift magnitude at which a detection flips from correct to incorrect,

$$t^*(\tau, W) = W \cdot \frac{1-\tau}{1+\tau},$$

a closed form in object size W and IoU threshold τ alone. This one formula does more than locate the jump: it subsumes the four scattered empirical findings above as corollaries of a single mechanism, and, under an idealized ground-truth-as-prediction protocol, its zero-parameter predictions reproduce the *absolute* output of official `COCOeval` on real COCO annotations to within 0.032 mean absolute error (correlation ≥ 0.997) across every metric and transform we test. For real detectors, where confidence ranking and sub-unity initial IoU depart from that idealized protocol, the same closed form instead predicts the *relative* degradation shape once rescaled by the detector's own measured baseline (Section 5). The upshot is a reframing: AP's fragility under small localization errors is a structural, quantitatively predictable property of IoU thresholding, not a quirk of any dataset or detector, pointing to concrete remedies (smoothed or size-normalized alternatives) for more reliable detector evaluation.

Contributions. We derive closed-form expressions for IoU under bounding-box translation (Lemma 1) and corner-anchored scaling (Lemma 2), verified against brute-force geometry, giving the central threshold-crossing formula $t^*(\tau, W) = W\frac{1-\tau}{1+\tau}$ (Corollary 2); at the dataset level, the entire degradation curve is an exact, rescaled survival function of the object-size distribution (Corollary 4), so its shape is predictable from a dataset's own size statistics. We show that four empirical phenomena observed on real COCO boxes (Section 3) — inverse size dependence, directional asymmetry, enlarge/shrink equivalence, and the excess fragility of averaged mAP — are corollaries of these two lemmas, and we identify and correct a distance-versus-direction confound in how we measure directional sensitivity (Remark 2). We validate the zero-parameter theory against official `COCOeval` on real COCO boxes under the ground-truth-as-prediction protocol, obtaining *absolute*-level MAE ≤ 0.032 (correlation ≥ 0.997) across mAP, AP_{50}, AP_{75}, and $\mathrm{AP}_{\mathrm{small}}$, and, with no refitting, on LVIS (Section 5), a second dataset with a much smaller median object size; for an off-the-shelf Faster R-CNN with real confidence ranking, under both fixed and random localization error (Sections 5, 5), the same theory predicts the *relative* degradation shape once rescaled by the detector's own measured baseline. Finally, the analysis is prescriptive: inverting the threshold-crossing formula yields a size-adaptive threshold $\tau(s, c) = \frac{s-c}{s+c}$ (Section 6) that reduces the size bias of AP degradation roughly twentyfold, and a smoothed variant that makes the metric Lipschitz while leaving dataset-level ranking essentially unchanged.

Section 3 presents the motivating empirical observations, Section 4 develops the closed-form theory, Section 5 validates it against `COCOeval`, and we conclude with implications for metric design.

2 Related Work

Average Precision and its variants. AP became the standard detection metric through PASCAL VOC (Everingham et al., 2010) at a single threshold ($\tau = 0.5$); COCO (Lin et al., 2014) generalized it to mAP averaged over ten thresholds and introduced the size-stratified $\mathrm{AP}_{\mathrm{small/medium/large}}$ breakdown we use. Others propose alternatives: Localization-Recall-Precision (Oksuz et al., 2018; 2021) adds an explicit localization-error term, Probabilistic Detection Quality (Hall et al., 2020) models spatial and semantic uncertainty, and Optimal Correction Cost (Otani et al., 2022) replaces ranking-based AP with an optimal-transport correction cost. Closest to our motivation, Dave et al. (2021) show that even holding the metric's definition fixed, AP remains sensitive to implementation-level choices. We are orthogonal: rather than propose a new metric or audit implementation choices, we give a closed-form account of why the existing IoU-thresholded AP's fragility should be expected in the first place.

Table 1: Empirical AP degradation under each transform, measured with official `COCOeval` on real COCO boxes ($N = 929$, ground-truth-as-prediction). All three transforms erase double-digit mAP within a single pixel; $\mathrm{AP}_{\mathrm{small}}$ collapses fastest of all.

	Translation				Enlarge	Shrink
o (px)	mAP	AP_{50}	AP_{75}	$\mathrm{AP}_{\mathrm{small}}$	mAP	mAP
0	1.000	1.000	1.000	1.000	1.000	1.000
1	0.874	0.992	0.919	0.672	0.920	0.910
2	0.753	0.949	0.782	0.412	0.835	0.807
3	0.662	0.892	0.665	0.251	0.759	0.720
5	0.533	0.783	0.551	0.087	0.644	0.600
10	0.370	0.580	0.384	0.012	0.472	0.421

IoU as a geometric quantity. IoU also underlies training, where its non-differentiability motivated surrogate losses — IoU loss (Yu et al., 2016), Generalized IoU (Rezatofighi et al., 2019), and Distance/Complete IoU (Zheng et al., 2020) — that reshape it to supply gradients during box regression. These study IoU's geometry to improve *optimization*; we characterize the same geometry's *evaluation*-time discontinuities, with closed-form IoU-under-perturbation expressions (Section 4) absent from this literature.

Diagnosing and stress-testing detectors. A complementary line dissects *where* detectors fail: Hoiem et al. (2012) introduced an error taxonomy and TIDE (Bolya et al., 2020) decomposes mAP into error types, and box-perturbation stress tests establish *that* AP is fragile to small localization error — but all take the metric as given; we go further and explain *why*, deriving these patterns from IoU geometry with no free parameters. More recently, Yoo et al. (2025) propose evaluating detectors without ground truth at all, using cross-prediction consistency as a reliability proxy; we instead characterize exactly where the ground truth AP assumes, however accurate, is insufficient to make AP stable.

3 Empirical Observations

We begin from a simple, reproducible experiment that isolates the *localization* sensitivity of AP from every other moving part of a detection pipeline: rather than train a detector, we treat each ground-truth box as its own prediction (confidence 1.0), apply a controlled geometric perturbation, and score against the original annotations with the official COCO evaluator, so any change in AP is due solely to the perturbation (Figures 4–5, Appendix E).

3.1 Protocol and Findings

We use the publicly available `coco128` subset of COCO train2017 (128 images, $N = 929$ ground-truth boxes spanning 71 of the 80 COCO categories; min side 1.2–640 px, with 36.3% of boxes "small" by the COCO convention area $< 32^2$). Each box is copied to a prediction, perturbed, and scored against the unperturbed ground truth with `pycocotools` `COCOeval` in `bbox` mode, reporting the four headline COCO metrics — mAP (IoU $0.50{:}0.05{:}0.95$), AP_{50}, AP_{75}, and $\mathrm{AP}_{\mathrm{small}}$ — as a function of perturbation magnitude $o \in \{0, \ldots, 10\}$ px, under three transforms. **Translation** shifts each box by o px along one of the eight canonical directions $\{(\pm 1, 0), (0, \pm 1), (\pm 1, \pm 1)\}$, drawn uniformly at random per box and held fixed across o (the "same-offset" convention: a diagonal shift moves each coordinate by o, hence travels $o\sqrt{2}$). **Enlarging** and **shrinking** hold the top-left corner fixed and grow or shrink the box by o px in width and height. Table 1 reports the measured curves; four patterns stand out, and the rest of the paper explains them from first principles.

F1: a single pixel is expensive. A one-pixel shift drops mAP from 1.000 to 0.874 (a 12.6-point loss) and $\mathrm{AP}_{\mathrm{small}}$ to 0.672 (32.8 points); one-pixel enlarging and shrinking cost 8.0 and 9.0 mAP points — sub-visual perturbations moving the metric by amounts comparable to years of benchmark progress. **F2: small objects degrade fastest.** $\mathrm{AP}_{\mathrm{small}}$ reaches 0.251 by $o = 3$ px and effectively vanishes (0.012) by $o = 10$, while overall mAP is still 0.370. **F3: averaged mAP is more fragile than** AP_{50}, lying well below it at every offset (0.533 versus 0.783 at $o = 5$) as the stricter thresholds inside the average flip first. **F4: enlarging and shrinking behave alike** (0.920 vs 0.910 at $o = 1$; 0.472 vs 0.421 at $o = 10$), both milder than diagonal translation.

Diagonal shifts appear worse than axis-aligned ones ($\mathrm{mAP} = 0.455$ vs. 0.647 at $o = 5$ px), but a diagonal offset o travels $o\sqrt{2}$ px: holding Euclidean distance fixed splits the 0.192 penalty into $\sim$44% distance and $\sim$56% genuine direction (Remark 2). In Section 4 all four findings — and this directional effect — follow as corollaries of two closed-form IoU expressions with no free parameters, confirmed in Section 5 against `COCOeval` directly.

4 Theoretical Analysis

Translation. We formalize the empirical sensitivity of IoU and AP to bounding box perturbations observed in Section 3. Throughout, let A be an axis-aligned box of width W and height H, and let B denote a perturbed version of A under *translation* (this section) or *corner-anchored scaling*, i.e. enlarging/shrinking (Section 4), matching the two transforms studied empirically. Let B be A shifted by magnitude $t \geq 0$ along direction $\theta \in [0, \pi/2]$, so that the corner offsets are $\delta x = t\cos\theta$, $\delta y = t\sin\theta$. We restrict to the regime $\delta x < W$, $\delta y < H$, i.e. A and B still overlap, which covers all perturbation magnitudes used in Section 3 (≤ 10 px) for all but the very smallest COCO boxes.

Lemma 1 (Exact IoU under diagonal translation)**.** *For $t\cos\theta < W$ and $t\sin\theta < H$,*

$$\mathrm{IoU}(t,\theta) \;=\; \frac{(W - t\cos\theta)(H - t\sin\theta)}{2WH - (W - t\cos\theta)(H - t\sin\theta)}. \tag{1}$$

Proof. Since A and B have identical dimensions and are shifted only along $(\delta x, \delta y)$, their intersection is itself an axis-aligned rectangle of width $W - \delta x$ and height $H - \delta y$, giving $\mathrm{Area}(A \cap B) = (W - \delta x)(H - \delta y)$. As $\mathrm{Area}(A) = \mathrm{Area}(B) = WH$, the union follows from inclusion–exclusion: $\mathrm{Area}(A \cup B) = 2WH - \mathrm{Area}(A \cap B)$. Substituting $\delta x = t\cos\theta$, $\delta y = t\sin\theta$ gives (1). □

We verified (1) numerically against brute-force rectangle intersection over randomly sampled $(W, H, \delta x, \delta y)$; it agrees to numerical precision (see supplementary code).

Remark 1 (Independent axis shifts)**.** *The proof above never ties δx and δy to a common magnitude: it holds verbatim for any independent pair $(\delta x, \delta y) \in [0, W) \times [0, H)$,*

$$\mathrm{IoU}(\delta x, \delta y) \;=\; \frac{(W - \delta x)(H - \delta y)}{2WH - (W - \delta x)(H - \delta y)}, \tag{2}$$

of which (1) *is the special case $\delta x = t\cos\theta$, $\delta y = t\sin\theta$. This already covers a purely horizontal or vertical shift and any non-uniform combination a real regressor's independent x- and y-coordinate errors would produce, with no new derivation required.*

Corollary 1 (Directional sensitivity at zero shift)**.**

$$\left.\frac{\partial\,\mathrm{IoU}}{\partial t}\right|_{t=0} = -2\left(\frac{\cos\theta}{W} + \frac{\sin\theta}{H}\right). \tag{3}$$

Proof. Differentiate (1) with respect to t and evaluate at $t = 0$, where $\mathrm{Area}(A \cap B) = WH$ and $\mathrm{Area}(A \cup B) = WH$; a direct computation (quotient rule) gives (3). □

Corollary 1 has three immediate consequences that together account for the qualitative behavior reported in Section 3. **Inverse size dependence:** the axis-aligned slopes are $-2/W$ (horizontal) and $-2/H$ (vertical), so local sensitivity scales as $O(1/\text{size})$ — a distribution-free, detector-free explanation for why small objects suffer disproportionately from fixed-pixel noise, as a property of the IoU functional itself. **Directional asymmetry:** for a square box ($W = H$), the diagonal ($\theta = \pi/4$) slope has magnitude $2\sqrt{2}/W$, versus $2/W$ for axis-aligned shifts — 41% steeper, explaining why diagonal perturbations degrade AP more than axis-aligned ones of the same *offset* value (see Remark 2 for a caveat on "same offset"). **Non-smoothness:** the $|\cos\theta|, |\sin\theta|$ structure means IoU is only piecewise differentiable in shift-space, with kinks along the coordinate axes.

Because AP thresholds IoU at a fixed level τ (i.e. a prediction counts as a true positive iff $\mathrm{IoU} \geq \tau$), AP is not merely non-Lipschitz in box coordinates — it is *discontinuous*, with jumps located exactly where $\mathrm{IoU}(t, \theta) = \tau$. We characterize these jump locations exactly.

Corollary 2 (Threshold-crossing distance, axis-aligned case)**.** *For axis-aligned translation ($\theta = 0$), the shift magnitude at which* IoU *first crosses threshold τ is*

$$t^*(\tau, W) = W \cdot \frac{1-\tau}{1+\tau}. \tag{4}$$

Proof. Set $\theta = 0$ in (1): $\mathrm{IoU}(t) = (W-t)/(W+t)$. Solving $(W-t)/(W+t) = \tau$ for t gives (4). □

Equation (4) is the central quantitative result of this section. Since t^* scales *linearly* with W, at $\tau = 0.5$ a COCO "small" object ($W, H \lesssim 32$ px) has $t^* \approx 10$ px — squarely inside the 0–10 px perturbation range studied empirically, which is why such boxes routinely flip from true positive to false positive under sub-visual-pixel noise. Since $t^*(\tau, W)$ is also strictly decreasing in τ, the higher-IoU components of mAP flip first, giving a mechanistic explanation for why averaged mAP degrades faster under perturbation than AP_{50} alone. The perturbation range we study is also not an artificially gentle stress test: Papadopoulos et al. (2017) report that a 4 px PASCAL VOC coordinate displacement reproduces ordinary human-annotator disagreement (88% mIoU either way), yet Table 1 shows AP treats the same magnitude as a double-digit failure. Applying Lemma 1 to our `coco128` boxes at the same 4 px predicts a *lower* mean IoU (0.72) than their 88% on PASCAL VOC, exactly as Corollary 1's $O(1/W)$ scaling predicts, since COCO's boxes are smaller (median width 51.8 px) (Lin et al., 2014) — an independent, out-of-sample cross-check against a different dataset and measurement protocol.

Corollary 2 also has a population-level consequence that Section 3.1's aggregate curves make visible but do not yet explain: under a common axis-aligned shift, the dataset-level pass rate at offset t is exactly $P(t) = 1 - F_N\left(\frac{1+\tau}{1-\tau}t\right)$, a rescaled, reflected view of the dataset's own width CDF F_N (Corollary 4, Appendix A), so the steepness of the aggregate degradation curve is governed by the local density of the width distribution, not by any individual box. COCO's right-skewed width distribution explains why the empirical curves drop fastest immediately off zero rather than degrading linearly.

Remark 2 (Distance versus direction)**.** *Our protocol applies the* same offset *o per coordinate — $(o, 0)$ horizontal versus (o, o) diagonal, which travels $o\sqrt{2}$ — so part of the "diagonal hurts more" effect is just the box moving farther. Holding distance fixed via Lemma 1 separates the two: for a 10×10 box at $o = 1$,* IoU *drops $0.818 \to 0.752$ (horizontal, distance-matched) $\to 0.681$ (diagonal), roughly half distance and half direction (Figure 6, Appendix E); both comparisons should be reported.*

Corner-anchored scaling. We now analyze the enlarging/shrinking transforms of Section 3: the top-left corner of A is held fixed and the bottom-right corner shifted outward (enlarging) or inward (shrinking) by $o \geq 0$ in both coordinates.

Lemma 2 (Exact IoU under corner-anchored scaling)**.** *For $0 \leq o < \min(W, H)$,*

$$\mathrm{IoU}_{\mathrm{enlarge}}(o) = \frac{WH}{(W+o)(H+o)}, \qquad \mathrm{IoU}_{\mathrm{shrink}}(o) = \frac{(W-o)(H-o)}{WH}. \tag{5}$$

Proof. Because both boxes share the top-left corner, one is always nested inside the other. For enlarging, $B \supseteq A$, so $\mathrm{Area}(A \cap B) = \mathrm{Area}(A) = WH$ and $\mathrm{Area}(A \cup B) = \mathrm{Area}(B) = (W+o)(H+o)$. For shrinking, $B \subseteq A$, so $\mathrm{Area}(A \cap B) = \mathrm{Area}(B) = (W-o)(H-o)$ and $\mathrm{Area}(A \cup B) = \mathrm{Area}(A) = WH$. Both expressions in (5) follow directly. □

As with Lemma 1, we verified (5) against brute-force geometric intersection. The same nesting argument holds verbatim for independent offsets $o_w \neq o_h$ (of which (5) is the symmetric case used empirically), covering aspect-ratio-changing errors as the special cases $o_h = 0$ or $o_w = 0$ with no new derivation required.

Corollary 3 (Sensitivity of scaling transforms)**.**

$$\left.\frac{\partial\, \mathrm{IoU}_{\mathrm{enlarge}}}{\partial o}\right|_{o=0} = \left.\frac{\partial\, \mathrm{IoU}_{\mathrm{shrink}}}{\partial o}\right|_{o=0} = -\left(\frac{1}{W} + \frac{1}{H}\right). \tag{6}$$

Table 2: Agreement between closed-form theory and official `COCOeval` across perturbation offsets $o = 0, \ldots, 10$ px.

	Translation		Enlarge		Shrink	
Metric	MAE	Corr.	MAE	Corr.	MAE	Corr.
mAP	0.026	0.9999	0.028	1.0000	0.032	0.9999
AP_{50}	0.019	0.9979	0.009	0.9987	0.020	0.9988
AP_{75}	0.027	0.9992	0.027	0.9987	0.031	0.9988
AP_{small}	0.012	0.9998	0.006	0.9999	0.005	0.9999

Proof. Direct differentiation of each expression in (5) at $o = 0$, using the quotient rule. □

Equation (6) shows the two transforms have *identical* first-order sensitivity, formally confirming the empirical observation of Section 3 that enlarging and shrinking "have similar effects." Comparing to Corollary 1, a square box's scaling sensitivity ($-2/W$) exactly equals its axis-aligned translation sensitivity, but is smaller in magnitude than diagonal translation sensitivity ($-2\sqrt{2}/W$) — yielding a single unified ranking,

$$\text{diagonal translation} \;>\; \text{axis-aligned translation} \;\approx\; \text{scaling},$$

in place of three separately-motivated empirical claims.

Solving $\mathrm{IoU}_{\mathrm{enlarge}}(o) = \tau$ and $\mathrm{IoU}_{\mathrm{shrink}}(o) = \tau$ for o gives closed-form threshold-crossing offsets for both scaling transforms (Corollary 5, Appendix A): for a 10×10 box at $\tau = 0.5$, $o^*_{\mathrm{shrink}} \approx 2.93$ and $o^*_{\mathrm{enlarge}} \approx 4.14$ against $t^*_{\mathrm{translate}} \approx 3.33$ — comparable but not identical fragility, refining "similar effects" into an exact ranking.

Together, Lemmas 1 and 2 and their corollaries give a closed-form, detector-independent account of every pattern reported empirically in Section 3; Section 5 next tests these predictions against real annotated boxes and detector outputs.

5 Empirical Validation

Setup and results. The closed-form predictions of Section 4 are exact statements about a single isolated box; we now test whether they hold on real, cluttered COCO images evaluated with the official `pycocotools` implementation of AP. We reuse the same $N = 929$ `coco128` ground-truth boxes as Section 3.1, following the ground-truth-as-prediction protocol of Section 3: each GT box is treated as its own prediction with confidence 1.0, then perturbed. This isolates the localization-sensitivity question from confidence ranking and detector-specific behavior, and is the regime in which Section 4's formulas apply directly.

For each magnitude $o \in \{0, \ldots, 10\}$ px we compute two quantities independently: the *empirical* value scores the perturbed boxes against the unperturbed ground truth with `COCOeval` (`bbox` mode), exactly as a COCO submission would be scored, reporting mAP, AP_{50}, AP_{75}, and $\mathrm{AP}_{\mathrm{small}}$; the *theory* value evaluates the closed-form IoU (Lemma 1 or 2) from each box's width, height, direction, and o alone — no detector, no fitted parameters, since under self-matching AP at threshold τ is just the fraction of boxes with $\mathrm{IoU} \geq \tau$. For translation, each box is assigned one of the eight canonical directions uniformly at random, held fixed across o.

Table 2 reports the agreement: the closed-form theory tracks the official `COCOeval` output closely across the full perturbation range for both translation and scaling, despite using no information beyond box geometry — mean absolute error is ≤ 0.032 for every metric and every transform (Pearson correlation ≥ 0.997). We lead with MAE, denominated directly in AP points, since every curve here is monotonic in o and so close to guaranteed a high correlation regardless of whether the theory gets the absolute numbers right; correlation is a secondary check on the curve's *shape*. The corresponding curves are plotted in Figures 7–8 (Appendix E). **No parameter is fit to `COCOeval`'s output anywhere in this comparison**: every theoretical value is the closed-form expression of Lemma 1 or 2, evaluated directly on each box's measured width, height, and applied offset.

The agreement shows the size- and direction-dependent sensitivity derived in Section 4 for a single idealized box carries over quantitatively to real annotated objects with COCO's size distribution, and tests an implicit independence assumption — `COCOeval` matches per image and could rematch a

perturbed box onto a different nearby instance of the same category — that leaves agreement intact on `coco128`; we probe it directly at scale next. One scope condition: every box here is randomly assigned one of eight directions, held fixed across the offset sweep, and this is not incidental to the agreement's tightness — Appendix A repeats the comparison with every box sharing the *same* direction instead and finds correlation drops to 0.87 for mAP with a systematic residual we do not yet fully explain, so the ≥ 0.997 correlations reported throughout this paper are a property of the randomized-direction protocol, not a direction-independent guarantee.

Scaling to full COCO `val2017` and a second dataset. To rule out that the near-perfect agreement of Table 2 is an artifact of the 128-image sample, we repeat the protocol on all of `val2017` (36,334 non-crowd boxes across 5,000 images, 42% small). Absolute error stays modest but somewhat larger than on `coco128` (up to 0.081 MAE for AP_{75}), and correlation stays ≥ 0.999 for every metric and transform (Table 4, Appendix E), confirming the theory reproduces the degradation *shape* at $40\times$ the box count; Appendix F traces the gap, which is not dense-scene rematching as the theory's per-box independence assumption might suggest, to a tied-confidence tie-break artifact of scoring ground truth as its own prediction. To test whether the theory is a property of IoU geometry or an artifact of one dataset's box statistics, we also repeat it, unchanged and with no refitting, on LVIS v1 val (Gupta et al., 2019) — a federated relabeling of a subset of COCO's own images with a substantially smaller median object size (33 vs. 51.8 px). MAE stays ≤ 0.067 for every metric (Table 9, Appendix H; correlation ≥ 0.994), and, exactly as Corollary 1's $O(1/W)$ scaling predicts, LVIS degrades faster than coco128 at every matched offset (0.834 vs. 0.874 at $o = 1$ px): the theory transfers to a dataset, protocol, and evaluator it has never seen.

Validation with a real detector. The protocol so far uses ground-truth boxes as predictions, isolating localization but exercising neither confidence ranking nor imperfect ($\mathrm{IoU} < 1$) boxes. We therefore perturb the predictions of a real detector — a pretrained Faster R-CNN (ResNet-50 FPN v2), 3,976 detections (score ≥ 0.05) on the same 128 images, unperturbed $\mathrm{mAP} = 0.574$. A single-pixel systematic shift of every predicted box — the kind of discrepancy a half-pixel coordinate convention or a resize-and-rescale rounding difference between two evaluation codebases can introduce, with no change to the detector at all — already drops this to 0.549 (2.5 mAP points), a swing on the order of what many papers report as their entire improvement over a baseline. The phenomena survive intact (Figure 9, Appendix E): small objects still collapse fastest (at 10 px $\mathrm{AP}_{\mathrm{small}}$ retains 7% of baseline versus 52% for mAP) and averaged mAP falls faster than AP_{50} (relative 0.52 versus 0.68), so findings F2–F3 are not artifacts of the ground-truth-as-prediction design. Rescaled by the single scalar $\mathrm{AP}(0)$ — the detector's own measured, unperturbed score, not a quantity fit to the degradation curve — the parameter-free theory reproduces the degradation shape across all four metrics, slightly over-predicting fragility because a detector's true positives are matched with margin rather than at a perfect $\mathrm{IoU} = 1$; composing each detection's own margin into the theory instead (Appendix D) cuts the relative-curve MAE for mAP from 0.136 to 0.031, confirming margin as the cause. This is a claim about *relative* degradation *shape* only: the stronger *absolute* agreement of Table 2 is established under the idealized protocol, not for real detector outputs, where sub-unity initial IoU and confidence-based ranking interact with the metric in ways the isolated-box theory does not model. This complements Dave et al. (2021), who show AP rankings move under implementation choices in the metric itself: a coordinate-convention or rounding difference of a single pixel moves them by itself. Two further robustness checks, detailed in Appendix D: perturbing pre-NMS candidates instead of final detections leaves the headline curve essentially unchanged, and repeating the protocol on the modern anchor-free YOLOv11s (Jocher & Qiu, 2024) reproduces the same story, so the threshold law is not an artifact of one detector family or generation.

Robustness to random localization noise. Every real-detector result so far uses a single, adversarially-fixed per-box offset, not what localization error looks like in practice, where it is better modeled as independent, zero-mean noise on each box's position. Reusing the same 3,976 Faster R-CNN detections, we add i.i.d. Gaussian jitter $(\delta x, \delta y) \sim \mathcal{N}(0, \sigma^2)$ independently to every detection — the general translation $\mathrm{IoU}(\delta x, \delta y)$ of Remark 1 evaluated at a random rather than fixed offset, so no new theory is required, only its expectation over the noise distribution (Monte Carlo). Figure 11 (Appendix E) shows the same phenomena survive: at $\sigma = 10$ px, $\mathrm{AP}_{\mathrm{small}}$ retains only 8% of its $\sigma = 0$ value versus 53% for mAP, again reproducing findings F2–F3, and the parameter-free theory

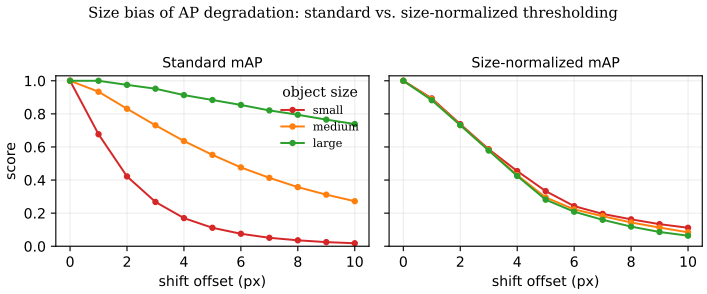

Figure 1: Standard vs. size-normalized mAP under translation, by object size. **Left:** a fixed IoU threshold makes the small/medium/large curves diverge (the size bias, F2). **Right:** the size-adaptive threshold $\tau(s, c) = \frac{s-c}{s+c}$ of (7) collapses them onto one.

again tracks the relative degradation curve closely (correlation ≥ 0.982) as a mild upper bound on fragility — the fragility we study is not an artifact of using a fixed, hand-chosen offset either.

6 Toward Size-Robust Evaluation

The analysis of Section 4 does more than explain AP's fragility; it prescribes remedies, one for each pathology the theory exposes — the size bias (F2) and the discontinuity — which are orthogonal and composable. Taking continuity, scale invariance, robustness, and calibration as a minimal checklist for a localization metric, Appendix B states which standard IoU-thresholded AP violates and this paper's response to each (Table 3). In brief: the size-normalized and smoothed remedies below fix scale invariance and continuity directly; robustness and calibration are only partially addressed, with a gaming vulnerability and a tied-confidence calibration bias left open (Section 7, Appendix F).

Size-normalized thresholding. The size bias (F2) is rooted in Corollary 2: the tolerated shift $t^*(\tau, W) = W\frac{1-\tau}{1+\tau}$ scales linearly with size, so a fixed threshold τ grants a small object far less absolute slack than a large one. To instead judge every object against the same absolute tolerance of c pixels, we invert the formula — setting $t^*(\tau, s) = c$ and solving — to obtain a size-adaptive threshold

$$\tau(s, c) = \frac{s - c}{s + c}, \qquad s = \sqrt{WH}. \tag{7}$$

We take $s = \sqrt{WH}$, the geometric mean of width and height, since it is the size scale implicit in COCO's own "small/medium/large" convention (other choices, e.g. the diagonal, are equally natural and untested; Appendix B). A prediction is correct at tolerance c iff $\mathrm{IoU} \geq \tau(s, c)$; for a square box under translation this is exactly "displaced by fewer than c pixels," independent of size. Averaging over tolerances $c \in \{1, \ldots, C\}$, in analogy to COCO's averaging over IoU thresholds, yields a *size-normalized* mAP. This is obtained by inverting the *axis-aligned* threshold-crossing formula ($\theta = 0$), so the guarantee is direction-dependent (only $\sim 0.7c$–$0.75c$ pixels in the worst diagonal case, not c), quantified exactly in Appendix B, which also derives a direction-uniform alternative and a resolution-invariant rescaling of c.

Figure 1 contrasts the two metrics on the real COCO boxes of Section 5. Under standard mAP the small-, medium-, and large-object curves fan apart (large–small gap averaging 0.62 at 10 px); under the size-normalized metric they almost coincide (gap 0.03, a ∼twentyfold reduction), with degradation shared equally across scales. On five off-the-shelf COCO detectors (Appendix G), the metric preserves the standard-mAP ranking exactly — re-weighting rather than reshuffling — while dropping the weakest detector from 46% to 14% of the leader's score.

Smoothed thresholding. Equation (7) removes the size bias but keeps a hard threshold, hence the discontinuity of Section 4, which is intrinsic to the indicator $\mathbf{1}[\mathrm{IoU} \geq \tau]$: as Figure 2 (left) shows, a single box's mAP credit — the mean of ten indicators — is a descending *staircase*, with ten jumps of 0.1 at the threshold-crossing distances $t^*(\tau)$ of Corollary 2. Replacing each indicator with a ramp of half-width δ,

$$s_\tau(u) = \mathrm{clip}\left(\frac{u - (\tau - \delta)}{2\delta}, 0, 1\right), \tag{8}$$

and averaging over the ten thresholds yields a *soft* mAP that is Lipschitz in box coordinates — the jumps vanish (Figure 2, left) — while, crucially, tracking standard mAP almost exactly at the dataset level (mean absolute difference 0.002 across offsets; Figure 2, right). Smoothing thus removes the

discontinuity without perturbing the aggregate ranking the metric is used for, and as $\delta \to 0$ it recovers COCO mAP exactly.

The choice of ramp is not load-bearing: a logistic sigmoid with steepness matched to the ramp's slope at τ gives the same picture (Figure 3, Appendix E; mean absolute difference 0.005 versus 0.002), so any monotone smoothing of the indicator removes the discontinuity while leaving ranking essentially unchanged.

The two remedies are orthogonal and compose — $\tau(s, c)$ inside the soft reward (8) is both size-fair and Lipschitz — and we offer them as a proof of concept, not a finished replacement for mAP. For standardized benchmarking we anchor both remedies' free parameters to COCO's own convention: $\delta = 0.05$, the largest ramp half-width for which neighboring thresholds' smoothing bands do not overlap, and $c \in \{1, \ldots, 10\}$ pixels, mirroring COCO's own ten-threshold averaging. These are reasonable defaults, not uniquely optimal ones; smaller δ tracks standard mAP more closely at the cost of a steeper transition, recovering hard-threshold mAP exactly as $\delta \to 0$.

7 Limitations

Perturbation scope. We cover axis-aligned translation and corner-anchored scaling only; rotations, aspect-ratio changes, and compound perturbations are not treated in closed form, though the same IoU geometry extends to them in principle. This matches the axis-aligned parameterization most detectors regress; oriented-box detectors fall outside this argument. We also lack a direct measurement decomposing real detectors' localization error into translation, scale, and aspect-ratio components.

Residual absolute-level gaps. The theory treats each box in isolation: on `val2017` it tracks degradation *shape* almost exactly but departs from `COCOeval` by up to 0.081 MAE in absolute level (Appendix F traces this to tie-breaking, not rematching); LVIS shows the same shape and size-dependence but a larger, opposite-signed residual we cannot yet explain. Our transferability claim is about shape, not exact absolute-level agreement.

Detector-side coupling and remedy gaps. Our real-detector experiments perturb detections after NMS; perturbing before NMS (Appendix D) changes candidate survival but not the degradation curve. The twentyfold size-bias reduction (Section 6) is smaller ($5.4\times$) on five real detectors (Appendix G), so the idealized figure is an upper bound. Neither remedy is evaluated against human judgment or downstream performance, and the choice of c and δ, and robustness to metric gaming (Section 6), remain open. The discontinuity itself is unsurprising; the contribution is characterizing exactly where it falls.

8 Conclusion

AP's fragility under small localization errors is a structural consequence of IoU thresholding: the closed-form account of where the discontinuity falls, $t^*(\tau, W) = W\frac{1-\tau}{1+\tau}$, explains the size, direction, scaling, and threshold effects as corollaries, and matches `COCOeval` absolutely under the idealized protocol or relatively (rescaled by baseline) for real detectors. Inverting t^* yields a size-adaptive threshold that cuts the size bias ∼twentyfold on the idealized protocol ($5.4\times$ on five real detectors), and a smooth reward removes the discontinuity while leaving ranking unchanged. The size-dependence is not an artifact of COCO's own box statistics: on LVIS — an independently collected dataset with a different size distribution and annotation protocol — the same degradation shape and the same size-adaptive remedy both reproduce closely, cutting the large–small gap $8.9\times$ (Appendix H). The remedy's one free parameter, the tolerance range C, is not fully inert, either: rankings among close detectors are stable under a finer grid of C but can invert under a substantially wider one (Appendix G), so we recommend reporting rankings at more than one C rather than treating the choice as settled. Appendix C further shows IoU-family regression losses (Yu et al., 2016; Rezatofighi et al., 2019; Zheng et al., 2020) — GIoU, DIoU, CIoU — inherit the same $O(1/W)$ local sensitivity as $t \to 0$, so this size bias has a training-time analogue these losses do not resolve, though we do not test whether it affects measured training dynamics. Confirming that empirically, extending to rotated and compound perturbations, and benchmarking on live detectors are natural next steps.

Ethics Statement

This work analyzes a standard evaluation metric (COCO/LVIS Average Precision) theoretically and empirically; it involves no human subjects, no personally identifiable information, and no new data collection. All experiments use publicly released annotations (COCO, LVIS) and openly available pretrained detectors (`torchvision`, Ultralytics YOLO11); no model is trained and no private or sensitive data is accessed. We are not aware of a direct negative societal impact. If anything, the size-adaptive remedy of Section 6 works against a systematic bias against small objects in standard evaluation, which is disproportionately relevant to safety-critical applications — e.g. detecting distant pedestrians or small road hazards — where small objects are already underrepresented in benchmark performance.

Reproducibility Statement

All results are exactly reproducible from box geometry alone: the closed-form expressions (Lemmas 1 and 2) are self-contained, and every empirical number, table, and figure is generated by short scripts operating on public COCO and LVIS annotations and pretrained `torchvision`/`ultralytics` detectors, with fixed random seeds, no model trained, and no parameter fitted anywhere in the paper.

AI Use Statement

A large language model (Claude, Anthropic) was used as a writing assistant: drafting, editing, and condensing prose in the manuscript, including tightening the abstract, limitations, and conclusion. It was not used to derive the theory, design the experiments, or write the reproduction code; all technical content, derivations, and analysis are the authors' own.

References

Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. TIDE: A general toolbox for identifying object detection errors. In *European Conference on Computer Vision (ECCV)*, pp. 558–573. Springer, 2020.

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

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5356–5364, 2019.

David Hall, Feras Dayoub, John Skinner, Haoyang Zhang, Dimity Miller, Peter Corke, Gustavo Carneiro, Anelia Angelova, and Niko Sünderhauf. Probabilistic object detection: Definition and evaluation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1031–1040, 2020.

Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *European Conference on Computer Vision (ECCV)*, pp. 340–353. Springer, 2012.

Glenn Jocher and Jing Qiu. Ultralytics YOLO11, 2024. URL `https://github.com/ultralytics/ultralytics`.

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 (ECCV)*, pp. 740–755. Springer, 2014.

Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. Localization recall precision (LRP): A new performance metric for object detection. In *European Conference on Computer Vision (ECCV)*, pp. 504–519, 2018.

Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. One metric to measure them all: Localisation recall precision (LRP) for evaluating visual detection tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9446–9463, 2021.

Mayu Otani, Riku Togashi, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Optimal correction cost for object detection evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 658–666, 2019.

Seungju Yoo, Hyuk Kwon, Joong-Won Hwang, and Kibok Lee. Automated model evaluation for object detection via prediction consistency and reliability. *arXiv preprint arXiv:2508.12082*, 2025.

Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. UnitBox: An advanced object detection network. In *ACM International Conference on Multimedia (ACM MM)*, pp. 516–520, 2016.

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-IoU loss: Faster and better learning for bounding box regression. In *AAAI Conference on Artificial Intelligence*, volume 34, pp. 12993–13000, 2020.

A General-Direction Threshold-Crossing Distance

Dataset-level degradation as a rescaled size-survival function. Section 4 states that Corollary 2 has a population-level consequence. We give it here in full.

Corollary 4 (Dataset-level degradation is a rescaled size-survival function)**.** *Let $W_1, \ldots, W_N$ be the widths of N boxes under a common axis-aligned translation of magnitude t, and let F_N denote their empirical CDF. The fraction still passing threshold τ is exactly*

$$P(t) \;=\; 1 - F_N\!\left(\frac{1+\tau}{1-\tau}\,t\right), \tag{9}$$

and where F_N has density ρ_N,

$$\frac{dP}{dt} \;=\; -\frac{1+\tau}{1-\tau}\,\rho_N\!\left(\frac{1+\tau}{1-\tau}\,t\right). \tag{10}$$

Proof. By Corollary 2, box i still passes at shift t iff $t \le t^*(\tau, W_i) = W_i \frac{1-\tau}{1+\tau}$, i.e. iff $W_i \ge \frac{1+\tau}{1-\tau}t$. The fraction of boxes satisfying this is, by definition of the empirical CDF, $1 - F_N\big(\frac{1+\tau}{1-\tau}t\big)$; differentiating with the chain rule gives (10). □

Equation (9) is exact for a population perturbed by a common axis-aligned shift: the pass-rate curve is a rescaled, reflected view of the dataset's own width distribution. Table 1 instead mixes eight per-box directions (Section 3.1); the general-direction formula below extends the same argument to $P(t) = 1 - F_N^{t^*}(t)$, the survival function of the per-box crossing distances $\{t_i^*\}$ rather than of width alone, of which (9) is the axis-aligned special case. Either way, the steepness of the aggregate curve at offset t is governed by the *local density* of threshold-crossing values near t, not by any individual box — concretely, in the axis-aligned case, the object-width density at scale $\frac{1+\tau}{1-\tau}t$. COCO's width distribution is heavily right-skewed toward small sizes, so $|dP/dt|$ is largest exactly where our 0–10 px range begins, consistent with the empirical curves dropping fastest immediately off zero.

This also separates what is structural from what is dataset-specific: the *existence* of a per-box t^* is intrinsic to IoU thresholding, but *how steep* the aggregate curve looks is inherited from the dataset's own size distribution — a differently-composed dataset would show the same discontinuities per box yet a visibly different aggregate curve.

Threshold-crossing offset for scaling. Section 4 solves for the offset at which each scaling transform crosses threshold τ.

Corollary 5 (Threshold-crossing offset for scaling)**.**

$$o^*_{\text{enlarge}}(\tau) = \frac{-(W+H) + \sqrt{(W^2+H^2)\tau + 4WH\left(1-\frac{\tau}{2}\right)}}{2\sqrt{\tau}}, \tag{11}$$

$$o^*_{\text{shrink}}(\tau) = \frac{(W+H) - \sqrt{W^2+H^2+2WH(2\tau-1)}}{2}. \tag{12}$$

Proof. Set each expression in (5) equal to τ and solve the resulting quadratic in o, retaining the physically valid root (positive, and $< \min(W, H)$ for the shrinking case). □

General-direction threshold-crossing distance. For completeness we also give the threshold-crossing distance for an arbitrary translation direction θ, referenced in Section 4.

Corollary 6 (Threshold-crossing distance, general direction)**.**

$$t^*(\tau,\theta) = \frac{\sqrt{1+\tau}\,(H\cos\theta + W\sin\theta) - \sqrt{2HW(\tau-1)\sin 2\theta + (1+\tau)(H\cos\theta + W\sin\theta)^2}}{2\sqrt{1+\tau}\,\sin\theta\cos\theta}. \tag{13}$$

Proof. Setting $\mathrm{IoU}(t,\theta) = \tau$ in (1) and clearing denominators yields $(W - t\cos\theta)(H - t\sin\theta) = \frac{2WH\tau}{1+\tau}$, a quadratic in t. Expanding and applying the quadratic formula, then retaining the smaller (first-crossing) root, gives (13). Taking $\theta \to 0$ recovers (4), confirming consistency with Corollary 2. □

Empirical validation across continuous direction. The experiments of Section 3 and the validation of Section 5 exercise eight canonical directions, each box independently and randomly assigned one and held fixed across the offset sweep — sufficient to establish Corollary 6 covers arbitrary θ in closed form, but not, on its own, a test of the closed form against `COCOeval` across the continuum θ actually ranges over. We run that test directly: at a fixed offset ($t = 6$ px) on the same `coco128` boxes, we shift *every* box in the dataset by the *same* angle θ, swept over 19 values from $0°$ to $90°$ in $5°$ steps (symmetry covers the rest), and compare `COCOeval` against Lemma 1's closed form at each angle. Agreement is weaker than the randomized, per-box-direction protocol's ≥ 0.997 correlation — mAP correlation is 0.87 (MAE 0.044), AP_{50} correlation is 0.79 (MAE 0.029) — and the residual is systematic rather than noise: for mAP, `COCOeval` sits above the theory at every angle, by as much as 0.070 near $\theta = 0°$ and shrinking to 0.009 near $\theta = 90°$; for AP_{50} the residual changes *sign* near $\theta \approx 70°$. We do not have a mechanistic account of this pattern (a natural hypothesis is that per-box direction randomization, used everywhere else in this paper, itself does some of the averaging work that keeps the randomized-direction agreement so tight, by mixing over whatever secondary effect makes a *shared*-direction perturbation of the whole dataset behave differently from a per-box-random one); we report it rather than suppress it, since it qualifies how far the near-perfect agreement reported elsewhere in this paper should be expected to generalize to every possible perturbation protocol.

Testing one candidate mechanism. A natural hypothesis is that this residual is the same tied-confidence tie-break effect identified in Appendix F: every self-matched detection here also scores 1.0, so `COCOeval`'s stable sort also leaves ties in construction order. We test this directly by building a "hits-first" accumulator at each angle — for every IoU threshold τ, sorting each category's tied detections with $\mathrm{IoU} \geq \tau$ before those with $\mathrm{IoU} < \tau$, which Appendix F shows recovers the raw pass rate almost exactly in the offset-sweep, binary hit/miss setting. It does *not* do so here: the hits-first reconstruction diverges from theory by *more* (mean $|\text{gap}| = 0.100$) than official `COCOeval`

does (mean $|\text{gap}| = 0.044$), at every one of the 19 angles, though the two gap series are highly correlated ($r = 0.996$). The tied-confidence mechanism that fully explains the val2017 and density-decomposition residuals (Appendix F) is therefore not the (or not the whole) explanation here: unlike that setting, every box's hit/miss status varies continuously across the ten IoU thresholds averaged into mAP rather than being fixed regardless of τ, and hits-first sorting does not isolate a clean tie-break-free baseline under that condition. The high correlation between the two residual series suggests a shared underlying driver tied to how pass rate varies with θ, but we do not have a closed-form account of it, and flag the discrepancy as a genuinely open question rather than force a fit to an existing mechanism.

B Extended Analysis of the Size-Normalized Threshold

Desiderata for a localization metric. Section 6 asks what a localization-sensitive detection metric should look like, taking four properties as a minimal checklist — continuity, scale invariance, robustness, and calibration. Table 3 states which of them standard IoU-thresholded AP violates, using results established in Sections 4–5, and this paper's response to each. This connects to the broader metric-design literature of Section 2: LRP (Oksuz et al., 2018; 2021) and PDQ (Hall et al., 2020) can be read as responses to the calibration and robustness gaps below, and Optimal Correction Cost (Otani et al., 2022) as a response to the continuity gap, though none derive the gaps themselves in closed form from IoU geometry.

Table 3: Four desiderata for a localization metric, whether standard IoU-thresholded AP satisfies each, and this paper's response.

Property	Standard AP	This paper
Continuity	Discontinuous: flips at $t^*(\tau, W)$ (Cor. 2)	*Fixed* — smoothed reward, Lipschitz in coordinates (Section 6)
Scale invariance	$t^* \propto W$: same τ, size-dependent pixel tolerance (F2)	*Fixed* — size-adaptive threshold, $\sim 20\times$ gap reduction (Section 6)
Robustness	A single-pixel coordinate-convention shift moves mAP by 2.5 points with no detector change (Section 5)	*Partial* — ranking preserved across five detectors (Appendix G); robustness to an adversary gaming the metric directly is not certified (Section 7)
Calibration	Tied-confidence tie-breaking biases AP against the raw pass rate by up to 0.081 MAE even with no localization error at all (Appendix F)	*Open* — not addressed by either remedy here

The size-normalized and smoothed remedies of Section 6 address the first two rows directly. The remaining two are only partially addressed: on robustness, $s = \sqrt{WH}$ in $\tau(s, c)$ is computed from the matched *ground-truth* box, not the predicted one, so a detector cannot loosen its own threshold by uniformly resizing its predictions, but a detector or loss that specifically targets localization error just under c pixels could still score well without being genuinely better — a gaming vulnerability we have not tested for (Section 7); on calibration, Appendix F's tied-confidence tie-break bias is untouched by either remedy here.

This appendix also collects three extensions of Section 6's size-normalized threshold $\tau(s, c) = \frac{s-c}{s+c}$, $s = \sqrt{WH}$: the sensitivity of that choice of size measure, the exact cost of anchoring the threshold to axis-aligned rather than worst-case error, and a resolution-invariant rescaling of c.

Choice of size measure. $s = \sqrt{WH}$ is the size scale implicit in COCO's own "small/medium/large" convention: that convention stratifies by *area* ($< 32^2$, 32^2–96^2, $> 96^2$ px^2), and $\sqrt{\text{area}} = \sqrt{WH}$ is the corresponding linear scale, so $\tau(s, c)$ uses the same notion of size COCO already stratifies by, rather than introducing a new one. It is not the only defensible choice: the diagonal $\sqrt{W^2 + H^2}$ or the perimeter $2(W + H)$ are equally natural linear size scales and would give a numerically different $\tau(s, c)$ for non-square boxes, and we have not tested whether the qualitative results of Section 6 (twentyfold gap reduction, rank preservation) are sensitive to that choice.

Remark 3 ($\tau(s, c)$ guarantees c pixels only along the axes)**.** *Equation* (7) *is obtained by inverting the* axis-aligned *threshold-crossing formula* $t^*(\tau, W) = W\frac{1-\tau}{1+\tau}$ *(Corollary 2,* $\theta = 0$*), not the general-direction one (Corollary 6). We do this because a single, closed-form threshold that depends only on a box's size* s *— not on an error direction no evaluator observes at test time — is what makes* $\tau(s, c)$ *usable as a drop-in replacement for a fixed IoU threshold. The cost is that the resulting guarantee is direction-dependent: since diagonal shifts degrade* IoU *faster than axis-aligned ones of the same Euclidean magnitude (Remark 2), a box displaced diagonally by exactly* c *pixels has* $\mathrm{IoU} < \tau(s, c)$ *and so* fails *a criterion nominally built to tolerate* c *pixels of error. Substituting* $\tau(s, c)$ *into the general-direction formula shows the worst-case (diagonal,* θ=45°*) direction is actually guaranteed only* $t^*(\tau(s, c), 45°) = \sqrt{2}\, s\left(1 - \sqrt{1 - c/s}\right)$ *pixels of tolerance, which is* 7.47 *px — not* 10 *— for a representative* $s = 50$*,* $c = 10$ *px, and approaches* $c/\sqrt{2} \approx 0.707\, c$ *as* $s \to \infty$*. The size-normalized metric therefore inherits, rather than removes, AP's directional asymmetry (F3): it is exactly* c*-pixel-tolerant for axis-aligned error and only* $\sim 0.7c$*-to-*$0.75c$*-pixel-tolerant in the worst diagonal direction.*

Corollary 7 (A direction-uniform size-normalized threshold)**.** *Inverting the general-direction formula at* $\theta = 45°$ *instead of* $\theta = 0$ *gives a threshold that guarantees* c *pixels of tolerance in* every *direction, not just axis-aligned:*

$$\tau_{\mathrm{worst}}(s, c) \;=\; \frac{k}{2 - k}, \qquad k \;=\; \left(1 - \frac{c}{s\sqrt{2}}\right)^2. \tag{14}$$

Proof. Setting $t^*(\tau, 45°) = c$ in the square-box ($W = H = s$) specialization of (13) and solving for τ gives (14) directly. □

Remark 4. $\tau_{\mathrm{worst}}(s, c)$ *answers Remark 3's gap by construction, at a cost: it is* stricter *than* $\tau(s, c)$ *everywhere, including for axis-aligned error, since it is calibrated to the worst case rather than the axis-aligned one — for* $s = 50$*,* $c = 10$ *px,* $\tau_{\mathrm{worst}} = 0.584$ *against* $\tau(s, c) = 0.667$*. It grants every object a uniform* c*-pixel guarantee* regardless of error direction*, the property* $\tau(s, c)$ *lacks, but does so by discarding leniency for the (empirically more common) axis-aligned case to cover the (rarer, more damaging) diagonal one. We have not evaluated* $\tau_{\mathrm{worst}}(s, c)$ *empirically — whether it still preserves detector ranking, and by how much it changes the large–small gap reduction of Section 6 — and leave that comparison, and the choice between the two thresholds' different trade-offs, to future work.*

Resolution-invariant rescaling of c. Because c is a pixel count, its meaning depends on the coordinate system boxes are expressed in. We follow `COCOeval`'s own convention throughout: boxes live in the image's native annotation resolution, not whatever resolution a detector internally resizes to for inference — a detector that runs on a resized 640×640 copy of a 4096×2160 image still reports boxes rescaled back to the original frame, since that is what `COCOeval` expects. c pixels is therefore c native-resolution pixels, unaffected by a detector's internal input resolution. This holds regardless of how that internal resize is performed — uniform letterboxing that preserves aspect ratio, or the non-uniform (aspect-ratio-distorting) stretch to a fixed square input such as 640×640 common in multi-scale training and testing — because `COCOeval` always scores boxes after they are mapped back to native coordinates; the internal resize changes what the detector's regressor sees during a forward pass, not the coordinate system c is measured in. Framing c as a native *sensor*-pixel tolerance, rather than a fraction of some resized canvas, is what makes it portable across domains with very different native resolutions and aspect ratios — a $c = 3$ px tolerance means the same physical thing on a 12-megapixel aerial frame as on a microscopy tile, whatever input size the detector itself trains or infers at. Comparing tolerances *across* datasets or images of different native resolution is a separate question — a c-pixel tolerance on a 4096×2160 aerial image is a far finer relative tolerance than the same c on a 640×480 web image — and admits a direct fix: rescale c by a reference resolution before applying (7),

$$c'(I) \;=\; c \cdot \frac{\sqrt{\mathrm{area}(I)}}{\sqrt{\mathrm{area}(I_{\mathrm{ref}})}}, \tag{15}$$

where $\mathrm{area}(I)$ is image I's native width×height and I_{ref} is an arbitrary reference resolution (taking $I_{\mathrm{ref}} = I$ throughout, i.e. COCO's own convention, recovers the tolerances used in this paper);

equivalently, $c/\sqrt{\text{area}(I)}$ is itself a resolution-free tolerance comparable across datasets of any native resolution. We do not exercise (15) empirically here, since every dataset in this paper shares COCO's own resolution convention, but it removes the size-normalized metric's remaining resolution dependence for anyone applying it to imagery of substantially different scale (e.g. aerial or satellite imagery).

C IoU-Family Regression-Loss Gradients at Perfect Overlap

Section 8 asks whether the GIoU, DIoU, and CIoU regression losses (Rezatofighi et al., 2019; Zheng et al., 2020) inherit raw IoU's $O(1/W)$ local sensitivity (Corollary 1) as training converges toward a match, as distinct from the vanishing-gradient behavior at zero overlap those losses were designed to fix. We answer this in closed form for the diagonal-translation setup of Section 4: yes, exactly, because each loss's correction term to IoU is $O(t^2)$ or identically zero near $t = 0$.

Lemma 3 (GIoU and DIoU under diagonal translation)**.** *With $\delta x = t\cos\theta$, $\delta y = t\sin\theta$ as in Section 4, and C the smallest axis-aligned box enclosing A and B,*

$$\text{GIoU}(t,\theta) = \text{IoU}(t,\theta) - \frac{2t^2\sin\theta\cos\theta}{(W + t\cos\theta)(H + t\sin\theta)}, \tag{16}$$

$$\text{DIoU}(t,\theta) = \text{IoU}(t,\theta) - \frac{t^2}{(W + t\cos\theta)^2 + (H + t\sin\theta)^2}. \tag{17}$$

Under pure translation, $\text{CIoU}(t,\theta) = \text{DIoU}(t,\theta)$ *identically, since the aspect-ratio-consistency term* $v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2$ *vanishes whenever* $w = w^{gt} = W$ *and* $h = h^{gt} = H$*, as they are throughout a translation-only perturbation.*

Proof. Place $A = [0, W] \times [0, H]$ and $B = [\delta x, W + \delta x] \times [\delta y, H + \delta y]$ (WLOG $\delta x, \delta y \geq 0$), so $C = [0, W + \delta x] \times [0, H + \delta y]$ and $\text{Area}(C) = (W + \delta x)(H + \delta y)$. From Lemma 1's proof, $\text{Area}(A \cup B) = 2WH - (W - \delta x)(H - \delta y)$. Expanding both products,

$$\text{Area}(C) - \text{Area}(A \cup B) = \big[WH + W\delta y + H\delta x + \delta x\delta y\big] - \big[WH + W\delta y + H\delta x - \delta x\delta y\big] = 2\delta x\delta y,$$

so $\text{GIoU} = \text{IoU} - \frac{\text{Area}(C) - \text{Area}(A\cup B)}{\text{Area}(C)} = \text{IoU} - \frac{2\delta x\delta y}{\text{Area}(C)}$; substituting $\delta x = t\cos\theta$, $\delta y = t\sin\theta$ gives (16). For DIoU, the box centers are $(W/2, H/2)$ and $(\delta x + W/2, \delta y + H/2)$, giving center-distance term $\rho^2 = \delta x^2 + \delta y^2 = t^2$ and enclosing-box diagonal $c^2 = (W + \delta x)^2 + (H + \delta y)^2$; substituting gives (17). Since box dimensions are unchanged by translation ($w = w^{gt} = W$, $h = h^{gt} = H$ for every t), $v \equiv 0$ and $\text{CIoU} = \text{DIoU} - \alpha v = \text{DIoU}$. □

Corollary 8 (GIoU/DIoU/CIoU match IoU's sensitivity at $t = 0$)**.**

$$\left.\frac{\partial\,\text{GIoU}}{\partial t}\right|_{t=0} = \left.\frac{\partial\,\text{DIoU}}{\partial t}\right|_{t=0} = \left.\frac{\partial\,\text{CIoU}}{\partial t}\right|_{t=0} = \left.\frac{\partial\,\text{IoU}}{\partial t}\right|_{t=0} = -2\left(\frac{\cos\theta}{W} + \frac{\sin\theta}{H}\right), \tag{18}$$

matching Corollary 1 exactly.

Proof. In (16) and (17), each correction term has numerator $O(t^2)$ and a denominator that is smooth and nonzero at $t = 0$ ($\text{Area}(C)|_{t=0} = WH \neq 0$; $c^2|_{t=0} = W^2 + H^2 \neq 0$), so by the quotient rule its derivative at $t = 0$ is 0. Hence $\partial_t\text{GIoU}|_{t=0} = \partial_t\text{DIoU}|_{t=0} = \partial_t\text{IoU}|_{t=0}$, and CIoU inherits DIoU's value by Lemma 3. □

Remark 5. *The three losses differ from raw IoU only at second order near a match — through the $2\delta x\delta y$ enclosing-box penalty (GIoU) or the t^2 center-distance penalty (DIoU/CIoU) — so the training-time size bias Section 8 identifies (small objects receiving a proportionally sharper corrective gradient near convergence for the same pixel residual) applies identically to all four losses' local behavior at $t \to 0$; it is not resolved by switching among them. Resolving it would require a loss whose correction term itself depends on scale near perfect overlap, which none of GIoU, DIoU, or CIoU provide.*

D MARGIN-CORRECTED THEORY FOR REAL DETECTORS

Section 5 reports that the parameter-free theory slightly over-predicts a real detector's fragility, attributing this to real true positives being "matched with margin" rather than starting at a perfect $\mathrm{IoU} = 1$ the way the idealized ground-truth-as-prediction protocol's $t = 0$ baseline does. We quantify and model this margin directly, rather than leaving it as a qualitative gloss.

Quantification. Extracting `COCOeval`'s own IoU matrices for Faster R-CNN's 828 true positives at $\tau = 0.50$ ($o = 0$), the matched IoU has mean 0.829, median 0.863, and minimum 0.507 (so no detection sits at the threshold edge): real detections start well above, but not at, $\mathrm{IoU} = 1$.

Model. This margin can be composed into the closed-form theory directly. For each true positive, we convert its own matched IoU m into an equivalent pre-existing axis-aligned offset $t_0 = s(1 - m)/(1 + m)$ ($s = \sqrt{WH}$ of its matched GT box, inverting the same closed form behind $\tau(s, c)$), then combine t_0 — at a Monte Carlo-averaged unknown angle, since we do not observe the real error's actual direction — with the synthetic offset o applied in its assigned direction, and re-evaluate $\mathrm{IoU}(\delta x, \delta y)$ at the resulting combined displacement instead of assuming a fresh $\mathrm{IoU} = 1$ start. Concretely, with ϕ the unknown pre-existing angle (sampled uniformly, 200 Monte Carlo draws per box) and $(\hat{dx}, \hat{dy})$ the box's assigned synthetic direction, the corrected displacement at offset o is

$$(\delta x, \delta y) \;=\; \big(t_0 \cos\phi + o\,\hat{dx},\;\; t_0 \sin\phi + o\,\hat{dy}\big),$$

averaged over the ϕ draws before thresholding.

Result. This margin-corrected theory cuts the relative-degradation-curve MAE against the real detector from 0.136 (naive, fresh-start theory) to 0.031 for mAP — within the range of the idealized protocol's own agreement figures — confirming that a detector's existing localization margin, not some unmodeled architectural effect, accounts for the over-prediction. The correction does not help AP_{50} (MAE 0.035 either way): a single lenient threshold is largely insensitive to where in $[0.5, 1)$ a true positive's margin sits, so composing that margin only changes the finer-grained, higher-threshold components of averaged mAP.

Pre-NMS coupling. Section 5's headline experiment perturbs Faster R-CNN's final, already-NMS'd detections, so a perturbation cannot change which box non-maximum suppression kept. We test the alternative directly: intercepting the model's RoI head to expose its pre-NMS candidate boxes ($32{,}703$ of them, after score-thresholding but before suppression, versus $3{,}976$ final detections), perturbing *those*, re-running suppression on the perturbed candidates, and scoring the survivors. Coupling is real — more candidates survive suppression as offset grows (from $3{,}976$ at $o = 0$ to $6{,}631$ at $o = 10$ px, since perturbed near-duplicate boxes overlap each other less and so suppress one another less). But its effect on the headline degradation curve is small and, if anything, protective: perturbing before NMS gives *slightly milder* degradation than perturbing after ($\mathrm{mAP} = 0.307$ vs. 0.301 at $o = 10$; $\mathrm{AP}_{\mathrm{small}} = 0.018$ vs. 0.024), with the two curves within 0.02 of each other at every offset and metric we report. NMS interaction is therefore a real but second-order effect: it does not manufacture the fragility we study, nor does the post-NMS protocol used throughout this paper materially understate it.

A modern architecture. Faster R-CNN, RetinaNet, FCOS, and the two SSD variants of Appendix G are all several years old, so a natural concern is whether the effect is specific to that generation of architectures. We repeat the protocol on YOLOv11s (Jocher & Qiu, 2024), a current single-stage, anchor-free detector, under the identical perturbation, image set, and evaluation code (unperturbed $\mathrm{mAP} = 0.556$). Figure 10 shows the same qualitative story: $\mathrm{AP}_{\mathrm{small}}$ falls to 9% of baseline at 10 px versus 56% for mAP, and the parameter-free theory again reproduces the degradation shape once rescaled by $\mathrm{AP}(0)$ (mean absolute error 0.20 for mAP, comparable to Faster R-CNN's 0.16). The threshold law is thus not an artifact of any one detector family, old or new, within the architectures we test — two-stage (Faster R-CNN), one-stage anchor-based (RetinaNet), anchor-free (FCOS), and a modern anchor-free single-stage detector (YOLOv11s), all of which regress an axis-aligned box directly. We have not tested transformer-based set-prediction detectors (e.g. DETR and its variants), diffusion-based detectors, or promptable, segmentation-first detectors

(e.g. SAM-based pipelines), whose box-regression or box-extraction mechanics differ enough from the architectures above that the same closed-form agreement, while expected from the theory's dependence on box geometry alone rather than architecture, remains to be checked directly.

E Additional Figures

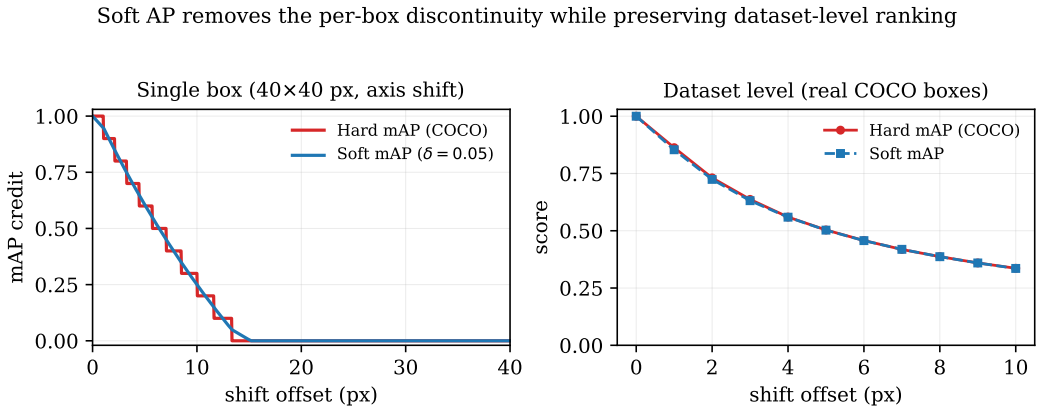

Figure 2: Smoothed AP (Section 6). **Left:** for a single 40×40 box under axis-aligned shift, standard mAP (red) is a ten-step staircase — the discontinuity of Section 4 — whereas soft mAP (8) (blue) is continuous. **Right:** on real COCO boxes the two are indistinguishable in aggregate (mean difference 0.002), so smoothing preserves dataset-level ranking while eliminating the per-box jumps.

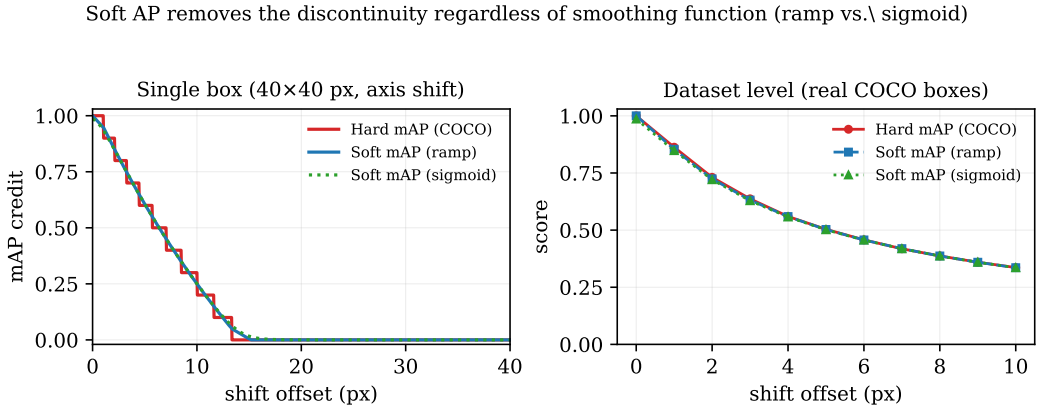

Figure 3: Soft AP with a logistic sigmoid $\sigma(k(u - \tau))$ in place of the ramp of Figure 2 (Section 6), steepness $k = 2/\delta$ matched to the ramp's slope at τ. **Left:** the sigmoid also turns the single-box staircase into a continuous curve. **Right:** at the dataset level it tracks standard mAP as closely as the ramp does (mean difference 0.005 vs. 0.002), showing the discontinuity fix does not depend on the specific smoothing function chosen.

F The Absolute-Level Gap in Section 5 Is Not a Density Effect

The density hypothesis, tested directly. A natural guess for why `COCOeval` sits below the isolated-box theory in Table 4 is the independence assumption: in a dense scene, a perturbed box could rematch onto a nearby ground-truth instance of the same category, altering its scored outcome in a way the single-box formula does not model. We test this by splitting all $36{,}334$ `val2017` boxes by (image, category) group size — *isolated* ($8{,}548$ boxes, the sole instance of their category in

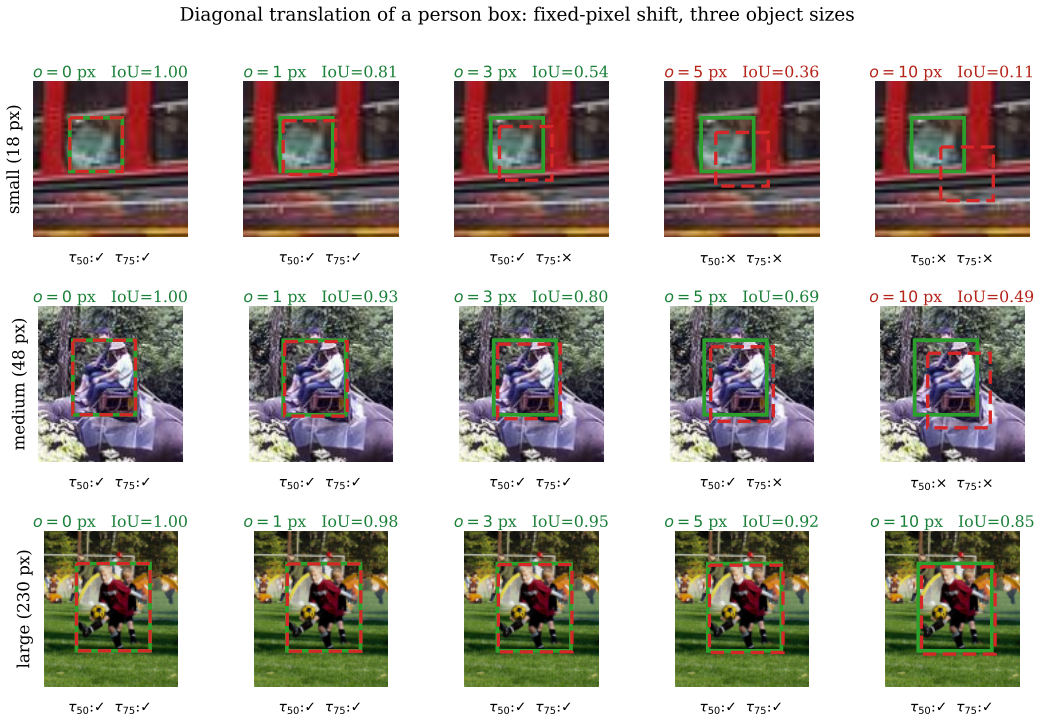

Figure 4: Diagonal translation of a real COCO person box (green: ground truth; red dashed: perturbed) at increasing pixel offsets o, for a small, medium, and large object (Section 3). Each panel is annotated with the resulting IoU and whether it still passes the COCO thresholds τ_{50} and τ_{75}. The same absolute shift that is harmless for the large box (IoU $= 0.85$ at $o = 10$) is fatal for the small one (IoU $= 0.11$), visualizing finding F2.

Table 4: Agreement between closed-form theory and `COCOeval` on full COCO `val2017` (36,334 boxes, offsets $o = 0, \ldots, 10$ px; Section 5). Absolute error is slightly larger than on `coco128` (Table 2), traced in Appendix F to tied-confidence tie-breaking in `COCOeval`'s stable sort rather than cross-instance rematching; correlation stays ≥ 0.999 throughout.

	Translation		Enlarge		Shrink	
Metric	MAE	Corr.	MAE	Corr.	MAE	Corr.
mAP	0.070	0.9993	0.052	0.9997	0.058	0.9997
AP_{50}	0.049	1.0000	0.024	0.9999	0.038	1.0000
AP_{75}	0.081	0.9989	0.063	0.9998	0.067	0.9996
AP_{small}	0.029	0.9995	0.007	0.9999	0.014	0.9999

their image, so no rematch target exists at all) versus *crowded* (27,786 boxes, ≥ 2 same-category instances present). Re-running the Section 5 translation protocol on each group as a standalone dataset is a faithful decomposition, not an approximation: `COCOeval` matches independently per image and category, so keeping every instance of a kept group together reproduces exactly the matching behavior those boxes see in the full run. At a 10-px shift:

Metric	Isolated (no rematch possible)	Crowded (≥ 2 same-cat. neighbors)
mAP MAE	0.104	0.074
AP_{75} MAE	0.115	0.087

Table 5: Theory-vs-`COCOeval` MAE by local density. If rematching explained the residual, the isolated group should show the smaller gap. It shows the larger one.



Figure 5: The three transforms applied to the same medium (48 px) person box (companion to Figure 4). Diagonal translation degrades IoU fastest (IoU $= 0.49$ at $o = 10$), while enlarging (0.70) and shrinking (0.65) are milder and close to each other, consistent with finding F4 and the sensitivity ranking of Section 4.

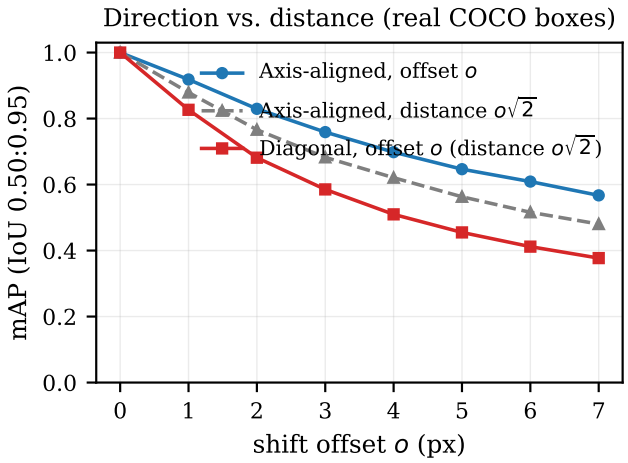

Figure 6: Disentangling direction from distance for translation on real COCO boxes (Section 3). A diagonal shift of offset o (red) is worse than an axis-aligned shift of the same offset (blue), but part of that gap is explained purely by the diagonal's longer Euclidean travel $o\sqrt{2}$: an axis-aligned shift matched to that distance (grey dashed) lands between the two.

The actual mechanism. Official AP integrates a 101-point precision–recall curve; it is not the same statistic as the theory's raw pass rate $\Pr[\text{IoU} \geq \tau]$. At a single threshold this difference has nothing to do with cross-box interaction: a detection's outcome there is a hard binary hit or miss

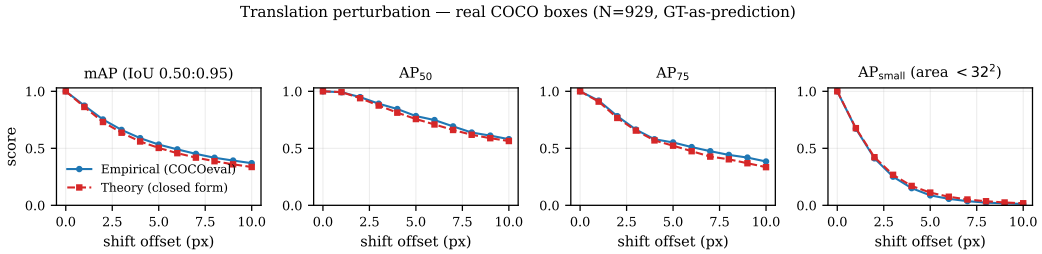

Figure 7: Translation perturbation on real COCO boxes ($N = 929$, ground-truth-as-prediction; Section 5). Solid: official `COCOeval` output. Dashed: closed-form prediction from Lemma 1, computed with no fitted parameters. Per-metric agreement is quantified in Table 2.

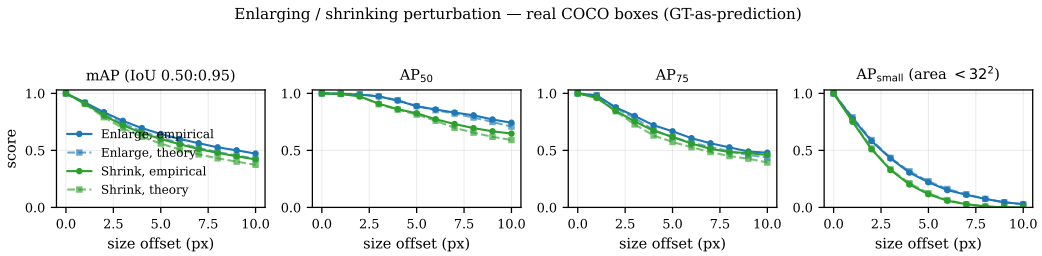

Figure 8: Enlarging/shrinking perturbation on the same real COCO boxes as Figure 7. Solid: `COCOeval`. Dashed: closed-form prediction from Lemma 2. The per-metric agreement is quantified in Table 2.

regardless of any competitor. We confirm this with a fully synthetic check — n detections, one per image (zero possibility of rematching), split cleanly into k exact hits (IoU $= 1$) and $n - k$ exact misses (IoU $= 0$):

k/n	0.10	0.30	0.50	0.70	0.90
Official AP	0.011	0.096	0.267	0.504	0.821
Gap (AP $-k/n$)	-0.089	-0.205	-0.233	-0.196	-0.079

Table 6: Official AP vs. raw hit rate k/n, $n = 1000$, no competition whatsoever. The gap is a property of the AP formula itself, largest near $k/n \approx 0.5$–0.7 and shrinking toward either extreme, because recall cannot exceed k/n and `COCOeval`'s accumulation assigns precision 0 to every unreached recall level beyond it.

An analytical bridge: tied confidence, not interpolation *per se*. Table 6's gap can be pinned down further than "a property of the AP formula." Every detection in our GT-as-prediction protocol shares identical confidence (score $= 1.0$), so `COCOeval`'s sort is a tie among all n detections; being a *stable* sort, it leaves them in whatever order they were constructed in, which (since hit/miss status is independent of construction order in our synthetic check) is effectively a uniformly random permutation of the k hits among the n slots. This — not the 101-point interpolation step itself — is what produces the gap. Two checks confirm it. First, an idealized best-case tie-break, all k hits ordered before all $n - k$ misses, gives AP $= k/n$ (up to 101-point grid rounding) exactly: with hits first, precision is 1.0 at every recall level up to k/n and undefined beyond it, so the interpolated AP is exactly the fraction of the 101-point grid at or below k/n — $0.098, 0.296, 0.494, 0.692, 0.890$ for $k/n = 0.1, 0.3, 0.5, 0.7, 0.9$, matching k/n to within grid resolution and leaving *no* unexplained gap. The entire deficit in Table 6, in other words, is attributable to tie-breaking order, not to the interpolation formula. Second, averaging Official AP over 20 independent random tie-break orders (same n, k, synthetic setup as Table 6) gives $0.012, 0.098, 0.260, 0.496, 0.810$ — matching Table 6's single-realization values of $0.011, 0.096, 0.267, 0.504, 0.821$ to within simulation noise (all within

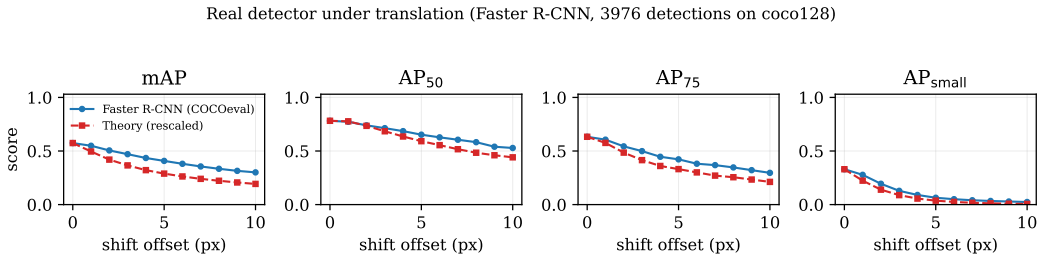

Figure 9: Translation perturbation of a real detector's predictions (Faster R-CNN ResNet-50 FPN v2, 3,976 detections; Section 5). Solid: official `COCOeval`. Dashed: the parameter-free theory, rescaled by the detector's unperturbed $\mathrm{AP}(0)$. The same size- and threshold-dependent fragility appears with real confidence ranking and imperfect localization; the idealized theory is a mild upper bound on the degradation.

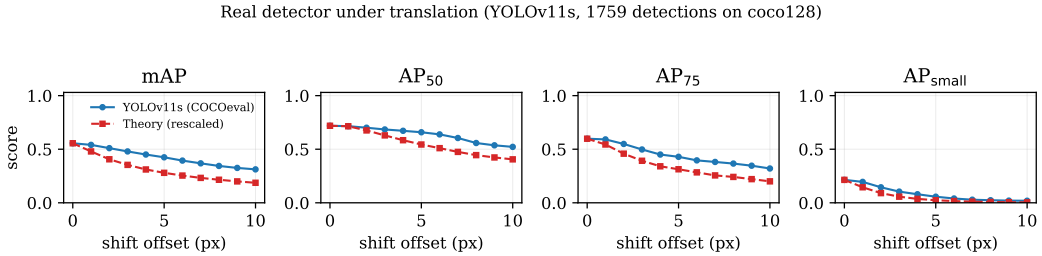

Figure 10: Translation perturbation of a modern single-stage detector's predictions (YOLOv11s (Jocher & Qiu, 2024), 1,759 detections; Section 5). Solid: official `COCOeval`. Dashed: the parameter-free theory, rescaled by the detector's unperturbed $\mathrm{AP}(0)$. Small objects again collapse fastest and the parameter-free theory again reproduces the shape, confirming that the fragility is not specific to the older two-stage/anchor-based detectors of Figure 9.

2 standard deviations of the 20-seed estimate). The mechanism, not merely the existence of a gap, is therefore identified: tied-confidence self-matching removes the ranking signal that AP otherwise relies on to order detections before interpolating, and a uniformly random tie-break is what the accumulation step effectively falls back to. A closed-form expression for $\mathbb{E}[\mathrm{AP} \mid k, n]$ under this random tie-break — an order-statistics quantity related to the expected running maximum of precision along a random hit/miss sequence — is a natural next step we leave to future work; the mechanism established here is what such a derivation would formalize.

Why isolated boxes show the larger gap. Larger boxes tolerate a fixed pixel shift better (Corollary 2), and isolated boxes are on average much larger than crowded ones (median 87 vs. 33 px in `val2017` — unsurprising, since small objects that co-occur many times per image, like place settings or crowd members, are exactly the ones classified as crowded, while large single-subject objects tend to appear alone). At a 10-px shift this gives isolated boxes a higher pass rate (0.556 vs. 0.258 at $\tau = 0.75$), which sits closer to the maximal-gap region of Table 6 than the crowded group's does — fully accounting for the asymmetry in Table 5 without any role for clutter or rematching. The residual absolute-level error of Section 5 is thus a consequence of comparing a pass-rate theory against a precision–recall-integrated metric, not evidence that the independence assumption fails in dense scenes; correlation stays ≥ 0.999 throughout precisely because this gap is a function of the aggregate pass rate, which the theory already tracks closely.

G Detector Leaderboard: Size-Normalized vs. Standard mAP

To test how the size-normalized metric of Section 6 behaves on real detectors, we evaluate five pretrained COCO detectors with deliberately different size profiles on a fixed 500-image subset

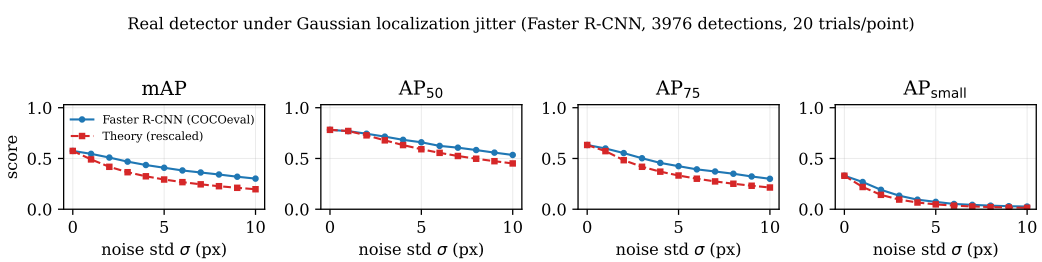

Figure 11: Gaussian localization jitter on a real detector's predictions (Faster R-CNN ResNet-50 FPN v2, same 3,976 detections as Figure 9; Section 5). Solid: official `COCOeval`, averaged over 20 independent noise draws per σ. Dashed: the parameter-free theory's expectation over the same noise distribution (Monte Carlo, 2,000 samples per box), rescaled by the detector's unperturbed $\mathrm{AP}(0)$. Replacing the fixed, worst-case offset of Figure 9 with independent, zero-mean noise on each box's position leaves the same size- and threshold-dependent fragility intact.

of `val2017` (3,504 boxes, 41.8% small). We report the official `COCOeval` mAP, its $\mathrm{AP}_{\mathrm{small}}$, and our size-normalized mAP (averaging the size-adaptive threshold $\tau(s,c) = \frac{s-c}{s+c}$ over $c \in \{1,\ldots,10\}$ px). Our size-normalized evaluator reduces to standard COCO AP when τ is held fixed, and in that mode reproduces official `COCOeval` mAP to within 0.003 for every detector, validating the implementation.

Table 7: Five COCO detectors ranked by standard mAP and by size-normalized mAP on the same 500-image `val2017` subset. The ranking is identical (Spearman $\rho = 1$): the size-normalized metric re-weights the score rather than reshuffling detectors. It is, however, far more sensitive to small-object localization — the weakest detector falls from 46% to 14% of the leader's score, and the RetinaNet/FCOS margin (a clear 0.018 under standard mAP) collapses to 0.002, indicating that part of their apparent difference is an artifact of how standard mAP weights object sizes.

Detector	Std. mAP	Rank	Size-norm. mAP	Rank
Faster R-CNN R50-FPNv2	0.513	1	0.264	1
RetinaNet R50-FPNv2	0.447	2	0.219	2
FCOS R50-FPN	0.428	3	0.217	3
SSD300 VGG16	0.289	4	0.063	4
SSDlite MobileNetV3	0.236	5	0.037	5

The rank agreement is itself informative: a principled re-weighting of the localization tolerance should not capriciously reorder detectors that are uniformly better or worse across sizes, and it does not. The value of the metric is instead its *sensitivity* — by refusing to let a fixed IoU threshold pre-collapse small objects, it exposes small-object localization quality that standard mAP compresses. We expect genuine reordering only among detectors that trade off performance across size regimes, which off-the-shelf general detectors do not; constructing such a comparison is left to future work.

The large–small gap reduction, measured on real detectors. Section 6's $\sim 20\times$ gap reduction (Figure 1) is demonstrated on the ground-truth-as-prediction protocol; we measure it directly on these five detectors' own real, unperturbed outputs instead of asserting it transfers. Restricting standard and size-normalized mAP each to `COCOeval`'s own small ($< 32^2$ px^2) and large ($> 96^2$ px^2) ground-truth populations, the mean |large$-$small| gap across all five detectors is 0.369 under standard mAP and 0.068 under size-normalized mAP, a $5.4\times$ reduction — smaller than the idealized protocol's $\sim 20\times$, but a real, measured collapse, not an assumption (Table 8, same five detectors as Table 7). As on LVIS (Appendix H), the correction overshoots for three of the five detectors (FCOS, Faster R-CNN, RetinaNet), whose size-normalized gap turns *negative* (small objects scoring higher than large): real detectors' localization error is not the isolated, single-box translation the theory models, so the exact reduction factor and its sign are empirical facts about a given detector population, not guaranteed by the theory itself. We report this measurement rather than the earlier assertion that the reduction "should hold in shape" on real detectors.

Table 8: Large–small AP gap under standard vs. size-normalized mAP, measured directly on each detector's own real, unperturbed outputs (same five detectors and cached detections as Table 7).

	Standard mAP			Size-normalized mAP		
Detector	Small	Large	Gap	Small	Large	Gap
Faster R-CNN R50-FPNv2	0.196	0.560	0.363	0.263	0.152	−0.111
RetinaNet R50-FPNv2	0.204	0.510	0.306	0.270	0.153	−0.118
FCOS R50-FPN	0.148	0.494	0.346	0.203	0.145	−0.058
SSD300 VGG16	0.030	0.443	0.413	0.050	0.060	+0.010
SSDlite MobileNetV3	0.004	0.423	0.419	0.015	0.058	+0.043

Sensitivity to the c-tolerance set. Table 7 averages $c \in \{1, \ldots, 10\}$ px; we check whether the ranking depends on that particular choice by recomputing it, on the identical cached detections, under a finer grid $c \in \{1, 1.5, \ldots, 10\}$ px and a wider one $c \in \{1, \ldots, 20\}$ px. The finer grid leaves the ranking and Spearman agreement with standard mAP unchanged (Spearman $\rho = 1$). The wider grid does not: RetinaNet and FCOS — already the closest pair in Table 7, with a size-normalized margin of 0.002 against a 0.018 standard-mAP margin — swap rank ($\rho = 0.9$, one pairwise inversion), since tolerating pixel errors up to 20 px rather than 10 reweights which of the two detectors' size-profile is favored. Reordering is thus not purely hypothetical: it already occurs among off-the-shelf detectors once the tolerance range is widened, precisely for the pair whose standard-mAP margin the size-normalized metric had already flagged as largely a size-weighting artifact. This makes C, the upper end of the tolerance range, a substantive design choice rather than an inert convention, and we have not identified a principled way to set it beyond anchoring to COCO's own ten-threshold precedent (Section 6). Absent such a principled rule, our practical recommendation is the same one COCO itself uses for the IoU threshold it replaces: do not commit to a single C and report the ranking (and the score gap between close detectors) at more than one, e.g. both $C = 10$ and $C = 20$ px, so a reordering that depends on the choice is visible rather than hidden behind whichever C happens to be reported. A dataset-driven rule — for instance, setting C from the annotation protocol's own measured localization noise (Section 4's comparison to Papadopoulos et al. (2017)) — is a natural candidate we have not tested.

H Second-Dataset Validation: LVIS

Every validation so far uses COCO. To test whether the theory transfers to a dataset it was never tuned against, we repeat the Section 5 protocol on LVIS v1 val (Gupta et al., 2019), a large-vocabulary (1,203 categories), long-tailed relabeling of COCO images. LVIS draws its val split from both COCO's train2017 and val2017 images; we use the 4,809 of its 19,809 val images that come from COCO val2017, since those are the only ones we have on disk, keeping every annotation on them (50,672 boxes). This is a substantially smaller and finer-grained object population than either coco128 or full COCO: median box width is 33 px, versus 51.8 px for coco128, and 61% of boxes are "small" by the COCO convention versus 36.3% for coco128 — itself a useful test of Corollary 1's $O(1/W)$ scaling on a size distribution shifted well below COCO's own.

LVIS is *federated*: unlike COCO, not every category is exhaustively annotated in every image, so its official evaluator (the `lvis` package's `LVISEval`, which we use here in place of `COCOeval`) excludes a category from scoring in images where it is marked `not_exhaustive_category_ids`, rather than assuming every instance is labeled. We otherwise follow Section 5 exactly: each GT box is its own prediction, perturbed by the same eight-direction, same-offset translation (identical seed and direction set as Table 1), and separately by corner-anchored enlarging and shrinking, then scored against the unperturbed LVIS ground truth.

Mean absolute error stays ≤ 0.067 for every metric and every transform (correlation ≥ 0.994) despite LVIS's federated annotation and much smaller median object size — the theory was fit to nothing here; every number in the "theory" column is the same closed-form expression (Lemma 1 or 2) evaluated on LVIS's own box geometry. As Corollary 1 predicts, LVIS also degrades *faster* than coco128 at every matched offset (e.g. translation mAP falls to 0.834 at $o = 1$ px and 0.246 at $o = 10$, versus coco128's 0.874 and 0.370): a dataset with smaller median object size shows more fragile AP under identical pixel perturbations, exactly as the $O(1/W)$ scaling requires, with

Table 9: Agreement between closed-form theory and the official `LVISEval` on LVIS v1 val (4,809 images, 50,672 boxes; Appendix H), offsets $o = 0, \ldots, 10$ px. No parameter is refit to LVIS; this is the same closed-form prediction as Table 2, applied to a dataset the theory has not seen.

	Translation		Enlarge		Shrink	
Metric	MAE	Corr.	MAE	Corr.	MAE	Corr.
mAP	0.038	0.9980	0.054	0.9982	0.055	0.9973
AP_{50}	0.057	0.9957	0.054	0.9973	0.067	0.9959
AP_{75}	0.042	0.9967	0.064	0.9946	0.062	0.9947
AP_{small}	0.018	0.9984	0.044	0.9982	0.033	0.9971

no per-dataset fitting involved. The residual MAE is somewhat larger than on coco128 (Table 2) and, unlike the full `val2017` case (Section 5), the empirical curve sits slightly *above* the theory rather than below it; we have not traced this sign flip to a specific mechanism the way Appendix F does for COCO, but a plausible contributor is LVIS's federated exclusion rule itself, which drops some box–category–image triples from scoring entirely rather than scoring every box as the isolated-box theory assumes. We report the gap rather than paper over it: MAE is the informative number here, and it is somewhat larger on LVIS than on COCO. Correlation merely checks that the degradation *shape* (not `COCOeval`'s or `LVISEval`'s exact aggregation arithmetic) is preserved, and is close to guaranteed for any monotonically decreasing curve; it remains high throughout, but we do not offer it as compensating evidence for the larger absolute-error gap.

Does the size-normalized remedy transfer to LVIS? Section 6's large–small gap collapse (Figure 1) is demonstrated on COCO; we check it directly on LVIS's own, much smaller object population using the identical raw-pass-rate statistic (mean indicator $\mathbf{1}[\mathrm{IoU} \geq \tau]$ per box, size-bucketed and averaged — the same construction behind Figure 1, which does not require running LVIS's federated accumulation at all). The qualitative result transfers: the mean $|\text{large} - \text{small}|$ gap under a fixed IoU threshold is 0.605 on LVIS, collapsing to 0.068 under the size-normalized threshold, an $8.9\times$ reduction — smaller than COCO's $\sim 20\times$, and, unlike COCO, the residual gap changes *sign* at larger offsets ($+0.001$ at $o = 1$ px, -0.121 at $o = 10$), a mild overcorrection under LVIS's more extreme, small-object-heavy size distribution rather than the near-zero residual COCO shows throughout. A further question is whether LVIS's federated not-exhaustive-category rule interacts with size-normalization specifically. Mechanistically, the `lvis` package's accumulation (which we inspected directly) applies that rule only to *unmatched* detections in a non-exhaustively-annotated category, suppressing their false-positive penalty; it never excludes a ground-truth box from being a valid match target. Since size-normalization's only effect is to make matching *more lenient* for small objects (a smaller $\tau(s, c)$ admits matches at lower IoU that a fixed threshold would have rejected), it should, if anything, reduce how often small-object detections in federated categories fall into the unmatched, rule-protected state in the first place — the two mechanisms interact on the precision side of the metric, not the recall side, and in a direction that would make the not-exhaustive rule matter *less*, not more, under size-normalization. We have reasoned through this from the official accumulation code rather than measured it with a custom size-normalized accumulator that replicates LVIS's exact not-exhaustive-category logic; confirming it directly is a natural next step we leave for future work.