Confidence Scores in Open-Vocabulary Detection Are a Biased Mixture of Scale and Semantics

Yi Tang Soon[1][0009−0009−8889−8989] and Jun-Wei Hsieh[2][0000−0002−5477−4891]

[1] Institute of Intelligence Systems, National Yang Ming Chiao Tung University, Tainan, Taiwan
esthersoon2002.ai14@nycu.edu.tw

[2] College of Artificial Intelligence and Green Energy, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
jwhsieh@nctu.edu.tw

Abstract. Foundation models such as CLIP [21] have enabled open-vocabulary object detectors that generalise to novel categories via vision-language similarity. However, the confidence scores these detectors produce are not reliable localization probability estimates: they conflate visual scale and semantic query specificity with the true detection signal. Through controlled experiments on COCO across three foundation-model-based detectors (GroundingDINO, OWL-ViT, YOLO-World), with the scale-bias finding further replicated on LVIS (1,203 categories) using GroundingDINO, we show that $s = \cos(v, t)$ is a biased mixture of two effects. Scale bias ($\hat{\alpha} = +0.064$, $r = 0.579$, $p = 1.29 \times 10^{-58}$) systematically inflates scores for large objects. Semantic bias ($\hat{\beta} = -0.705$, $p = 5.23 \times 10^{-41}$) suppresses scores for generic queries. Both biases are structurally inevitable from CLIP's image-level pretraining. Threshold adjustment cannot remove them: oracle per-scale thresholding yields ΔF1 = +0.001 for small objects versus +0.102 for large. A parameter-free temperature scaling correction improves small-object Recall@10 by 19.6% ($p < 0.01$) without retraining. This comes at a modest, measurable cost to pooled-ranking precision, so the bias is partially, not freely, reversible at inference time. These findings reveal a fundamental limitation of adapting image-level foundation models to region-level detection tasks.

Keywords: Foundation models · Open-vocabulary detection · CLIP · Confidence calibration · Scale bias · Prompt sensitivity

1 Introduction

Foundation models such as CLIP [21], ALIGN [8], and SigLIP [29] have transformed computer vision by providing powerful image-text representations that generalise across tasks. Open-vocabulary object detectors build on these models to localise objects described by arbitrary text queries. GroundingDINO [14], OWL-ViT [17], and YOLO-World [4] all achieve strong zero-shot performance this way. The standard output of these systems is a scalar confidence score $s \in [0, 1]$ per detection, derived from cosine similarity in CLIP space. Practitioners use this

score to filter detections by a fixed threshold. The implicit assumption is that s measures localization certainty: a detection with $s = 0.7$ should be correct 70% of the time, regardless of object size or query specificity. We show this assumption is fundamentally false.

Consider applying such a detector to a scene. At a threshold of 0.3, small objects are systematically discarded. This is not because the detector failed to localise them: it produced correct bounding boxes, but with confidence scores of 0.08, 0.11, and 0.09. Lowering the threshold to rescue small objects floods the output with large-object false positives. The root cause is not a model failure but a structural property of $s = \cos(v, t)$ that conflates two independent signals.

We identify and quantify two systematic biases:

Scale bias ($\hat{\alpha} = +0.064$, $r = 0.579$, $p = 1.29 \times 10^{-58}$). Large objects receive systematically higher confidence scores than small objects for identical queries. Averaging CLIP features over fewer pixels yields a noisier, less concentrated region direction, which pulls the normalised cosine similarity away from its noise-free value.

Semantic bias ($\hat{\beta} = -0.705$, $p = 5.23 \times 10^{-41}$). For fixed visual content, generic queries ("an object") produce lower scores than specific queries ("a dog"), because specific queries occupy tighter regions in CLIP embedding space.

Scale bias is replicated across three detectors and on LVIS. Semantic bias is demonstrated on GroundingDINO across 8 categories. Both derive theoretically from CLIP's image-level contrastive pretraining objective. Our contributions are:

1. A formal additive decomposition $s \approx \alpha \cdot \phi(a) + \beta \cdot \psi(d_{\text{sem}}) + \varepsilon$ with identified coefficients, showing that foundation model confidence scores are structurally biased for region-level tasks.
2. Controlled experiments confirming scale bias across three detectors ($r \in [0.36, 0.64]$) and semantic bias on GroundingDINO across 8 categories (paired $t > 16$).
3. A 2×2 factorial experiment revealing a floor effect that drowns out the localization signal for small objects.
4. A parameter-free test-time correction that improves small object Recall@10 by 20% without retraining, with its precision and calibration costs measured rather than assumed.

2 Related Work

Foundation Models for Detection. Open-vocabulary detection builds on vision-language foundation models. Early work [28] used captions for novel categories. Subsequent methods include ViLD [5], GLIP [10], RegionCLIP [32], OWL-ViT/v2 [17,16], GroundingDINO [14], YOLO-World [4], and Bangalath et al. [1]. Concurrently, [31] proposes a training-free confidence aggregation scheme that boosts open-vocabulary detection accuracy by combining scores across proposals. All of these share the same confidence interface: cosine similarity in CLIP space. Our work questions whether this scalar is a meaningful quantity

for downstream applications. It also provides a structural account of why such aggregation-style fixes are needed in the first place.

Limitations of Image-Level Pretraining. CLIP's image-level contrastive objective has known spatial limitations. Prior work shows failures on compositional reasoning [27], spatial understanding [30,20], and fine-grained visual tasks [26]. DeCLIP [11] and FLAVA [24] explore alternatives but retain image-level alignment. We show that the confidence score aggregation mechanism introduces additional systematic biases independent of feature quality.

Confidence Calibration. Calibration is studied for classification [6,18], dataset shift [19], and detection [9]. Temperature scaling [6] and label smoothing [25] are post-hoc logit fixes that cannot address structural entanglement. No prior work identifies the structural sources of miscalibration in foundation-model-based open-vocabulary detectors.

Small Object Detection. Small object detection is addressed through feature pyramids [12], single-stage detectors [15,22], and specialised architectures [3,23,2]. We provide a new structural explanation for why foundation-model-based open-vocabulary detectors systematically underperform on small objects.

3 Formal Formulation

Let $v \in \mathbb{R}^d$ denote the visual region feature and $t \in \mathbb{R}^d$ the text embedding of query q. Foundation-model-based open-vocabulary detectors compute:

$$s = \cos(v, t) = \frac{v \cdot t}{\|v\| \, \|t\|}. \tag{1}$$

In practice, each detector implements Eq. (1) through a learned, sigmoid-activated projection head rather than an unprocessed cosine value. We use $\cos(v, t)$ as the shared formal abstraction of this interface throughout, since all three detectors reduce to a monotone function of region-text similarity. Throughout, a denotes the candidate region's ground-truth pixel area, the geometric proxy for object scale. d_{sem} denotes the semantic specificity of query q. We operationalise it as the CLIP text-embedding cosine distance $d_{\text{sem}} = 1 - \cos(t_q, t_{q_0})$ between q's text embedding and the embedding of the category's most specific query q_0 (e.g. `"a dog"`). By this definition, $d_{\text{sem}} = 0$ for the most specific wording and increases for progressively more generic rewordings of the same category (Sec. 4, Experiment 2).

Definition 1 (Ideal Confidence Score). *The ideal confidence score is:* $s^* = P(\text{IoU}(\hat{b}, b^*) \geq 0.5 \mid v, t)$*, where* $\hat{b}$ *is the predicted box and* b^* *the ground truth. A score is well-calibrated if* $s \approx s^*$*.*

We present two observations showing s is not well-calibrated.

Observation 1 (Scale Bias). *With category and query fixed,* $\mathbb{E}[s \mid a_{\text{small}}] \ll \mathbb{E}[s \mid a_{\text{large}}]$. *Across three detectors, Pearson* $r \in [0.36, 0.64]$ *between* $\log a$ *and* s *(*$p < 10^{-20}$*).*

Observation 2 (Semantic Bias). *With visual content fixed,* $\mathbb{E}[s \mid t_{\text{specific}}] \gg \mathbb{E}[s \mid t_{\text{generic}}]$. *Paired* $t = 16.5$, $p = 9.83 \times 10^{-49}$ *across 8 categories.*

Empirical Finding 1 (Additive Decomposition). *Let* $\phi(a) = \log(a+1)$ *be the log-area scale term, and let* ψ *be the identity map,* $\psi(d_{\text{sem}}) = d_{\text{sem}}$*, so the semantic specificity level enters linearly. The observed score decomposes as:*

$$\boxed{s \approx \alpha \cdot \phi(a) + \beta \cdot \psi(d_{\text{sem}}) + \varepsilon}, \tag{2}$$

with $\hat{\alpha} = +0.064$ *(from Exp. 1) and* $\hat{\beta} = -0.705$ *(from Exp. 2), both* $p < 10^{-40}$*. The residual* ε *contains the localization signal* s^* *and noise. Equation* (2) *is a first-order (leading-term) approximation that linearises the multiplicative mechanism derived in Sec. 3.1 (Eq.* (3)*). It is intended to identify and quantify the two bias coefficients, not to claim that* s *is exactly additive in* $\phi(a)$ *and* $\psi(d_{\text{sem}})$ *everywhere. The non-parallel interaction observed in Experiment 3 is consistent with, rather than contradicting, this multiplicative origin. As* a *shrinks, the scale term drives* s *toward zero and compresses the room available for the semantic term to act. This produces the floor effect reported in Sec. 4.*

Consequence 1 (Loss of Discriminative Power). *For small objects,* $\text{Var}(\varepsilon \mid a_{\text{small}}) \approx \text{Var}(s \mid a_{\text{small}}) - \hat{\alpha}^2 \text{Var}(\phi(a_{\text{small}})) = 0.0138 - 0.0054 = 0.0084$*, insufficient to separate true from false positives (confirmed empirically: oracle thresholding yields* $\Delta\text{F1} = +0.001$ *for small vs.* $+0.102$ *for large).*

3.1 Why Entanglement Is Structurally Inevitable

Scale bias: directional concentration. Since $s = \cos(v, t)$ is invariant to the norm of v by construction, scale bias cannot arise from $\|v\|$ itself. It must arise from how scale affects the direction of v. Write the unnormalised region feature as $v_{\text{raw}} = \mu + \bar{\eta}(R)$, where μ is the noise-free, category-aligned direction and $\bar{\eta}(R) = \frac{1}{|R|} \sum_{p \in R} \eta(p)$ is the average of zero-mean, variance-σ_f^2 pixel-level feature noise over region R. By the law of large numbers, $\text{Var}(\bar{\eta}(R)) \propto \sigma_f^2 / |R|$, so smaller regions yield a noisier estimate of μ. After normalisation, the resulting direction $v = v_{\text{raw}} / \|v_{\text{raw}}\|$ deviates from $\mu / \|\mu\|$ by a small, zero-mean angle δ, with $\text{Var}(\delta) \propto \sigma_f^2 / (|R| \, \|\mu\|^2)$. Writing $\theta = \angle(\mu, t)$ for the noise-free category-query alignment (the localisation signal s^* is monotone in $\cos\theta$) and expanding $\cos(\theta + \delta)$ to second order in δ:

$$\mathbb{E}[s \mid a] \;\approx\; \cos\theta \left(1 - \tfrac{1}{2}\text{Var}(\delta)\right) \;\approx\; \cos\theta \left(1 - \frac{\kappa}{2a}\right), \tag{3}$$

where $a = |R|$ is the region's pixel area and $\kappa \propto \sigma_f^2/\|\mu\|^2$ is a constant governing the local feature-noise level, treated as approximately fixed for a given category-query pair. For $\cos\theta > 0$ (the relevant regime for true matches), Eq. (3) is monotonically increasing in a and asymptotes to the noise-free alignment $\cos\theta$ as $a \to \infty$. This is exactly the angular-concentration behaviour of a von Mises–Fisher-distributed direction estimate, whose concentration parameter grows with the number of pixels averaged. Scale bias is therefore a structural consequence of spatial pooling acting on the direction of the region embedding.

Semantic bias: CLIP alignment geometry. CLIP's contrastive loss trains specific query embeddings (e.g., "a dog") to align tightly with concept features. Generic queries ("an object") instead align diffusely with all objects. For any fixed v: $\cos(v, t_{q_0}) > \cos(v, t_{q_k})$ for $k > 0$, predicting the confidence drop confirmed in Observation 2.

Both biases arise from the same root: $s = \cos(v, t)$ optimises image-level semantic similarity, not region-level localization certainty.

4 Experiments

Setup. We evaluate GroundingDINO-SwinT [14], OWL-ViT-B/32 [17], and YOLO-World-S [4] off-the-shelf on COCO 2014 val [13], 8 categories, 80 images per category (640 pairs total). Scale groups: small ($a < 32^2$), medium (32^2–96^2), large ($a \geq 96^2$). Representative instances at each scale include, e.g., a distant bird (small), a chair viewed across a room (medium), and a close-range car or truck filling much of the frame (large). Unless otherwise noted, inference uses box/text threshold 0.01, which retains the weak detections needed for the precision–recall and oracle-thresholding analyses of Sec. 5. Experiment 1's GroundingDINO correlation (Table 1, Fig. 1) instead uses threshold 0.05/0.10, which is sufficient for the sanity-check correlation it reports.

Experiment 1: Scale Bias. With text query fixed to `"a [category]"`, mean confidence rises monotonically: 0.180 (small), 0.317 (medium), 0.520 (large) (ANOVA $F = 146.92$, $p = 3.40 \times 10^{-53}$; Fig. 1). Per-category $r \in [0.296, 0.666]$, all $p < 0.01$. Table 1 shows the bias is universal across all three detectors.[3]

Table 1. Scale-confidence Pearson r across three detectors. All values positive and significant ($p < 10^{-20}$).

Detector	r	p	Small mean	Large mean
GroundingDINO	0.579	1.29×10^{-58}	0.180	0.520
OWL-ViT	0.363	3.18×10^{-21}	0.085	0.279
YOLO-World	0.637	4.56×10^{-72}	0.145	0.615

[3] Even the smallest p-value (4.56×10^{-72} for YOLO-World) is far from machine epsilon ($\approx 1.1 \times 10^{-16}$). It comes from the analytic Student-t CDF, not floating-point subtraction, and reflects a strong correlation ($|r| > 0.6$) over hundreds of samples.

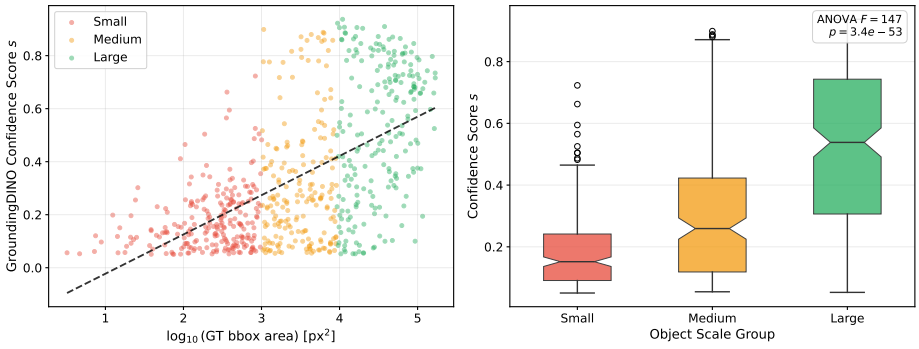

Fig. 1. Scale bias in GroundingDINO (COCO, 8 categories). *Left*: Confidence vs. log(area); regression line $r = 0.579$, $p = 1.29 \times 10^{-58}$. *Right*: Confidence distributions by scale group; ANOVA $F = 146.92$, $p = 3.40 \times 10^{-53}$. Text query held constant throughout.

Replication on LVIS, Fig. 2 To test whether scale bias generalises beyond 8 hand-picked categories, we replicate Experiment 1 on LVIS v1 val [7], which covers 1,203 categories across 19,809 images. Using GroundingDINO on 1,816 sampled annotations, scale bias replicates strongly: $r = 0.369$, $p = 1.08 \times 10^{-59}$, ANOVA $F = 135.90$, $p = 1.01 \times 10^{-55}$. Mean confidence follows the same direction as COCO: 0.110 (small), 0.167 (medium), 0.301 (large). Notably, scale bias is stronger for rare categories ($r = 0.491$) than for frequent ($r = 0.386$) or common ($r = 0.351$) categories. This suggests that objects with fewer training examples in CLIP's pretraining corpus are more susceptible to scale-induced confidence distortion. This finding is particularly relevant for foundation model deployment in long-tail recognition scenarios.

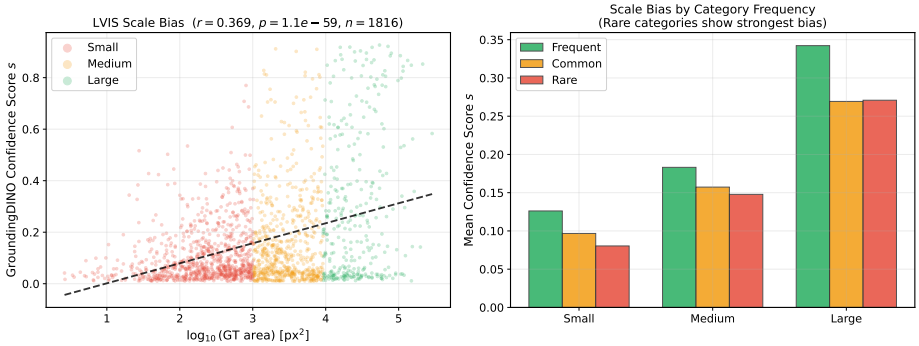

Fig. 2. Scale bias replication on LVIS v1 val (1,203 categories, $n = 1{,}816$). *Left*: Confidence vs. log(area); $r = 0.369$, $p = 1.08 \times 10^{-59}$. *Right*: Mean confidence by scale group and category frequency. Rare categories show the strongest bias ($r = 0.491$), suggesting that objects underrepresented in CLIP pretraining data are more susceptible to scale-induced score distortion.

Experiment 2: Semantic Bias. With visual content fixed (same image, same box, large objects only), queries vary from specific (q_0) to generic (q_3) along a fixed four-level hierarchy designed to be increasingly generic. For category `dog`, this hierarchy is L0 = `"a dog"` (category name), L1 = `"a pet"` (functional category), L2 = `"an animal"` (superordinate class), and L3 = `"an object"` (maximally generic). The same category-to-functional-to-superordinate-to-generic pattern is applied to all 8 categories. The semantic-distance term d_{sem} entering the decomposition (Eq. (2)) is the CLIP cosine distance from each level's query to the category's L0 query (Sec. 3), not the ordinal level itself. Confidence drops across all 8 categories: L0: 0.403 → L3: 0.194 (paired $t = 16.5$, $p = 9.83 \times 10^{-49}$; Fig. 3). CLIP distances are non-monotone for 3 of 8 categories, reflecting corpus statistics rather than taxonomic hierarchy.

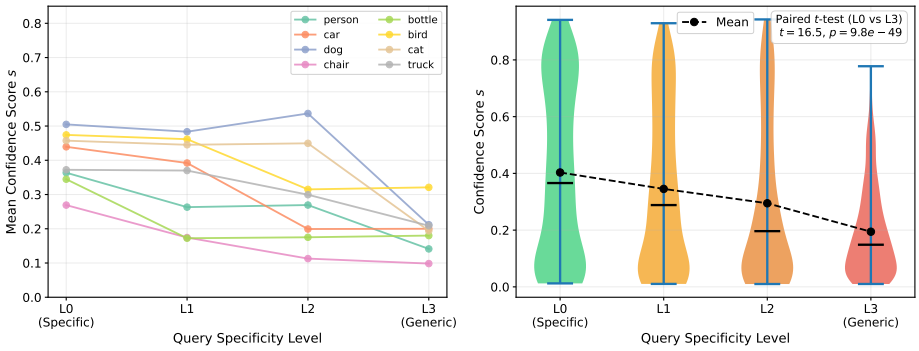

Fig. 3. Semantic bias. *Left*: Per-category mean confidence across query levels. *Right*: Pooled violin plots. Paired $t = 16.5$, $p = 9.83 \times 10^{-49}$; visual content fixed.

Experiment 3: Two-Signal Interaction. A 2×2 factorial design (scale $\times$ specificity) confirms both main effects ($p < 10^{-58}$; Fig. 4, Table 2). The semantic effect is 2.2$\times$ larger for large objects ($\Delta = +0.223$) than small ($\Delta = +0.100$). Scale bias saturates small-object scores near zero, suppressing the semantic signal.

Table 2. 2×2 cell means: mean confidence score s (GroundingDINO), averaged over the same 8 COCO categories used in Experiments 1–2, for each combination of object-scale group (large, $a \geq 96^2$; small, $a < 32^2$) and query specificity (specific query L0 vs. generic query L3, defined in Experiment 2). Both main effects significant ($p < 10^{-58}$).

	Specific query	Generic query
Large objects	0.409	0.187
Small objects	0.171	0.071

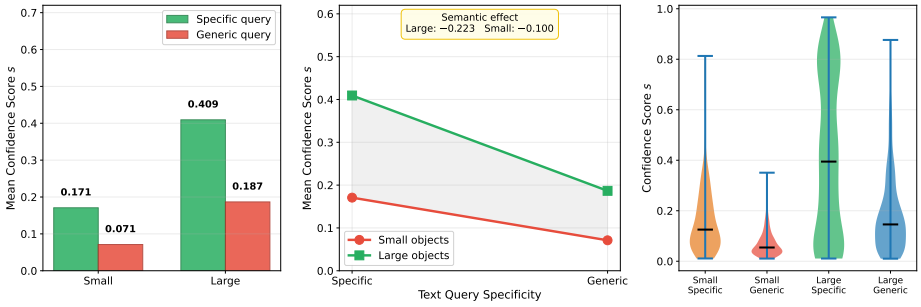

Fig. 4. 2×2 interaction. *Left*: Cell means. *Center*: Interaction plot; non-parallel lines reveal a floor effect for small objects. *Right*: Full distributions per cell.

5 Practical Consequences

Precision-Recall and Threshold Gap. We summarise detection quality per scale group with Average Precision (AP): the area under the precision–recall curve, equivalently the mean precision averaged over recall levels from 0 to 1. AP is the standard scalar summary of a precision–recall curve.[4] AP gaps and the confidence threshold τ required to reach a target recall R, per scale group, are shown in Table 3 and Fig. 5. The AP gap (large − small) is 0.27. Achieving 50% recall for small objects requires a threshold 3.7× lower than for large objects (0.129 vs. 0.476). Oracle per-scale thresholding provides no benefit for small objects (ΔF1 = +0.001) but substantial benefit for large objects (ΔF1 = +0.102). This confirms the localization signal is unrecoverable by threshold selection alone.

Table 3. AP and the confidence threshold τ required to reach target recall R (e.g. $\tau(R{=}0.3)$ is the score cutoff needed for 30% recall), by scale group.

Scale	AP	τ (R=0.3)	τ (R=0.5)	τ (R=0.7)
Small	0.430	0.186	0.129	0.088
Medium	0.566	0.302	0.234	0.184
Large	0.704	0.645	0.476	0.350

[4] Per-scale AP assigns each detection to a scale bucket using its own predicted box area, and matches it only against ground truth in that same bucket. A detection with no same-scale ground truth in its image therefore counts as a false positive. This isolates each scale's precision–recall behaviour, but differs from the COCO convention of matching against all ground truth in an image and filtering by ground-truth scale post hoc.

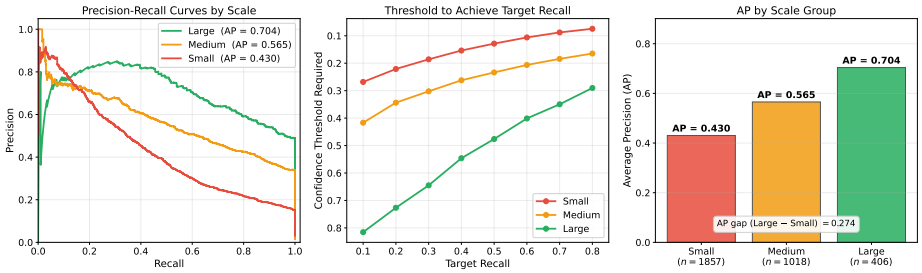

Fig. 5. Practical cost of entanglement. *Left*: PR curves (AP gap = 0.27). *Center*: Threshold for target recall per scale. *Right*: AP by scale group.

Calibration Analysis. Reliability diagrams (Fig. 6), computed across 56,630 detections ($s \geq 0.05$), show all groups are under-confident. The gap is scale-dependent: -0.384 for small objects versus -0.205 for large. This confirms that scale bias suppresses scores below their true localization probability.

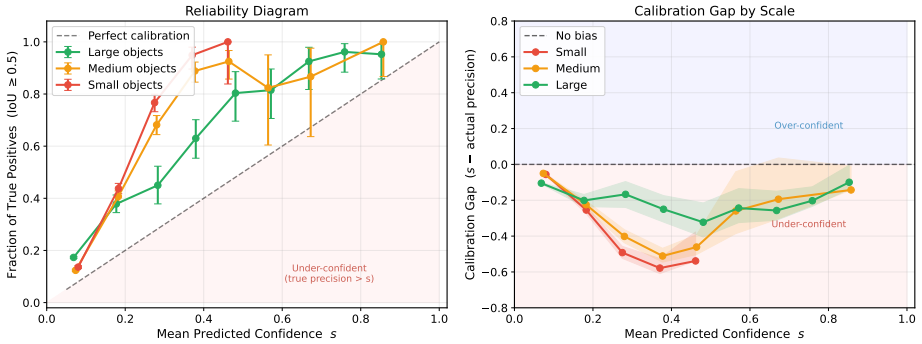

Fig. 6. Scale-dependent miscalibration. *Left*: Reliability diagram (90% CI); small objects are furthest from the diagonal. *Right*: Calibration gap; small objects show a larger negative gap (mean -0.384) than large (mean -0.205).

Parameter-Free Test-Time Correction. Scale bias creates a ranking unfairness in images with mixed-scale objects: large objects dominate the top-k detections regardless of localization quality. We propose variance-matched temperature scaling to correct this:

$$s_{\text{corr}} = s^{1/T_g}, \quad T_g = \sqrt{\frac{\text{Var}(\log s_g)}{\text{Var}(\log s_{\text{large}})}}, \tag{4}$$

estimated without ground-truth labels ($T_{\text{small}} = 1.28$, $T_{\text{medium}} = 1.06$, $T_{\text{large}} = 1.00$).

On $n = 79$ mixed-scale pairs, it improves small-object Recall@10 from 0.347 to 0.415 (+19.6%, $p < 0.01$) and overall Recall@10 from 0.544 to 0.566 (+4.1%, $p < 0.05$), with a non-significant large-object decrease (−6.0%, $p = 0.10$; Fig. 7).

Pooling small-object and large-object detections into one ranking (necessary since within-group ranking is invariant to the correction), Average Precision decreases from 0.500 to 0.476 ($\Delta = -0.025$, 95% bootstrap CI $[-0.041, -0.007]$, $p = 0.002$): promoting small objects surfaces more false positives. This does not conflict with the recall gain: Recall@k only checks the top k, while AP integrates precision across the full ranking, including the low-confidence tail where reordering hurts most.

Calibration also improves (Table 4): ECE roughly halves for small (0.102→0.047) and medium (0.079→0.040) objects, and is unchanged for the large-object reference. Overall, the correction trades some ranking sharpness for better-calibrated, more equitable small-object detections, with no retraining or detector-specific assumptions.

Table 4. Calibration before and after the temperature-scaling correction, on the same detection set as Fig. 6. ECE: Expected Calibration Error. Large objects are the uncorrected reference group ($T_{\text{large}} = 1$) and are unaffected by construction.

Scale	Raw gap	Corrected gap	Raw ECE	Corrected ECE
Small	−0.384	−0.225	0.102	0.047
Medium	−0.281	−0.269	0.079	0.040
Large	−0.205	−0.205	0.110	0.110

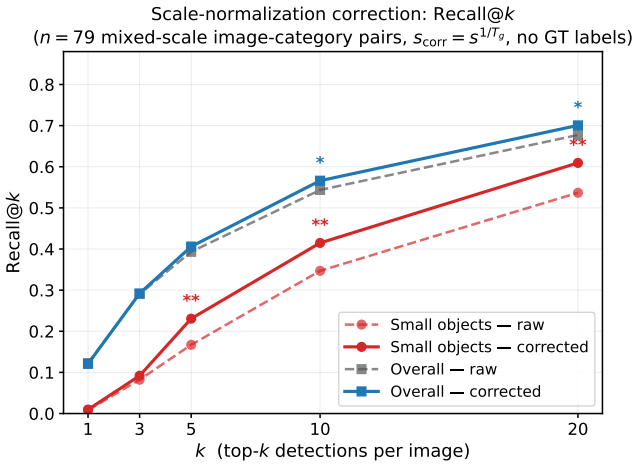

Fig. 7. Scale-normalization correction on $n = 79$ mixed-scale (image, category) pairs (images containing both small and large ground-truth objects). Solid: corrected; dashed: raw. Small-object Recall@k improves significantly at all k ($^{**}p < 0.01$); overall Recall@10 also improves ($^{*}p < 0.05$). Large-object recall decreases modestly and non-significantly (−6%, $p = 0.10$). No ground-truth labels required.

6 Discussion and Conclusion

Foundation models have enabled remarkable open-vocabulary detection capabilities, but their confidence scores carry a hidden cost. The same image-level contrastive objective that makes CLIP powerful for semantic matching also makes its cosine similarity scores unreliable as region-level localization probabilities. Taken in isolation, the direction of each bias is unsurprising. Large objects score higher, and generic queries score lower. What we contribute is the quantification beyond that direction. First, a consistent, statistically significant scale-bias effect across three architectures, corroborated by a large-scale replication on 1,203 LVIS categories, alongside a comparably significant semantic-bias effect on GroundingDINO. Second, a structural derivation (Sec. 3.1) showing both biases arise from the same angular-concentration mechanism, rather than being independent coincidences. Third, a demonstration that the resulting loss of discriminative power is irreversible by threshold selection alone (Consequence 1). Together, these results make the entanglement a structural consequence of adapting image-level representations to region-level tasks. Threshold tuning alone cannot fix it.

Scope and Limitations. Three limitations qualify these contributions. First, the ideal confidence score s^* (Sec. 3) is defined purely as localization correctness, $P(\text{IoU} \geq 0.5 \mid v, t)$. We choose this definition because it is the quantity a practitioner implicitly assumes when filtering detections with a single threshold. We do not claim semantic-grounding uncertainty is irrelevant. It is simply a distinct axis from the localization signal our diagnosis targets, and conflating the two inside one scalar is itself part of the problem we describe. Second, a potential confound is that large objects may be easier to detect due to lower occlusion. A dedicated experiment across 640 large objects rules this out: $r = -0.000$, $p = 0.99$ under our occlusion proxy (COCO segmentation area ratio). Other correlates of scale, namely local resolution, background clutter, candidate-proposal quality, and category frequency, are not individually disentangled from the mechanism we identify, and may contribute alongside it. Our LVIS frequency-stratified analysis (Fig. 2, right) is a first step toward separating category frequency from pure scale, but a full disentanglement is future work. Future work should also identify which architectural component, the spatial pooling stage or the contrastive alignment objective, is most responsible for each bias. Third, all controlled experiments use COCO 2014 val, with a GroundingDINO-based LVIS replication. Our three-detector cross-architecture replication (Table 1) and the structural argument in Sec. 3.1 both support the claim that the same direction holds for any detector built on image-level vision-language features. Broader validation across additional datasets, architectures, and deployment scenarios remains future work. These limitations narrow what specific numbers we can claim. They do not change the structural argument: it is not a sampling artefact, and its practical consequences follow directly.

The entanglement has direct implications for deployment. Semantic bias means that confidence scores are exquisitely sensitive to prompt wording ("a dog" vs. "an animal"). The same detection can be accepted or rejected depending on how a user phrases the query. This prompt sensitivity, a known concern for foundation model applications, here has a concrete, quantifiable effect on detection reliability. Any downstream system in robotics, video surveillance, or medical imaging that relies on open-vocabulary confidence scores will systematically underserve small objects and over-represent large ones. Concrete examples include distant-pedestrian detection for autonomous driving, small-instrument tracking in surgical or industrial inspection footage, and rare, small-object categories in aerial or satellite imagery. In these long-tail settings, scale bias and the semantic effect we observe on LVIS (Fig. 2) compound. As the community moves toward lightweight detectors [4], this problem becomes more acute. Smaller models cannot afford post-hoc VLM verification, which makes well-calibrated scores a prerequisite rather than a luxury.

Our temperature scaling correction shows that scale bias is partially reversible at inference time without architectural change. This offers a practical mitigation while the deeper solution awaits: detectors that predict $P(\text{IoU} \geq 0.5 \mid v, t)$ directly via region-level supervision, rather than inheriting image-level similarity as a proxy. As Sec. 5 shows, this correction is not free. It trades a modest, statistically significant precision cost for substantially better calibration on the objects it targets. Extending it to a larger pool of images, across detectors, and on a downstream task remains future work.

Confidence scores from foundation-model-based open-vocabulary detectors should not be treated as calibrated localization probabilities without correction. A confidence score that depends on how large an object happens to be, or on which of several equivalent words a user chose, is not measuring confidence: it is measuring scale and wording, dressed up as certainty.

References

1. Bangalath, M., Maaz, M., Khattak, M.U., Khan, S., Khan, F.S.: Bridging the gap between object and image-level representations for open-vocabulary detection. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision (ECCV) (2020)
3. Chen, Q., Wang, Y., Yang, T., Zhang, X., Cheng, J., Sun, J.: You only look one-level feature. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
4. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: YOLO-world: Real-time open-vocabulary object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
5. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: International Conference on Learning Representations (ICLR) (2022)

6. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning (ICML) (2017)
7. Gupta, A., Dollar, P., Girshick, R.: LVIS: A dataset for large vocabulary instance segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
8. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning (ICML) (2021)
9. Küppers, F., Kronenberger, J., Shantia, A., Haselhoff, A.: Multivariate confidence calibration for object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2020)
10. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: GLIP: Grounded language-image pre-training. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
11. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J.: Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In: International Conference on Learning Representations (ICLR) (2022)
12. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision (ECCV) (2014)
14. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: European Conference on Computer Vision (ECCV) (2024)
15. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single shot multibox detector. In: European Conference on Computer Vision (ECCV) (2016)
16. Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
17. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., Houlsby, N.: Simple open-vocabulary object detection. In: European Conference on Computer Vision (ECCV) (2022)
18. Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring calibration in deep learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019)
19. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Lakshminarayanan, B., Snoek, J., Ghahramani, Z., Sohl-Dickstein, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
20. Parcalabescu, L., Cafagna, M., Muradjan, L., Frank, A., Calixto, I., Gatt, A.: Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting, colour, relative size and spatial relations. In: Workshop on Multimodal Pre-training for Vision and Language (MM-PT) (2021)

21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML) (2021)
22. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
23. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2015)
24. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: FLAVA: A foundational language and vision alignment model. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
25. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
26. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
27. Yüksekgönül, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags of words, and what to do about it. In: International Conference on Learning Representations (ICLR) (2023)
28. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
29. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
30. Zhang, Y., et al.: Beyond semantics: Exploring the visual spatial understanding failures of vision language models. arXiv preprint arXiv:2503.17349 (2025)
31. Zheng, Y., Liu, K.: Training-free boost for open-vocabulary object detection with confidence aggregation. arXiv preprint arXiv:2404.08603 (2024)
32. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: RegionCLIP: Region-based language-image pretraining. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)