When to Trust Context: A Causal Evidence Gate for Occlusion-Aware Object Detection

Anonymous authors
Paper under double-blind review

Abstract

Transformer-based object detectors aggregate scene context through cross-attention causing query representations to entangle an object's direct visual evidence with contextual co-occurrence priors. Traditionally treated as distinct challenges, we demonstrate that this contextual entanglement underpins two opposing failure modes: (i) *co-occurrence bias*, where models over-rely on spatial surroundings, impairing generalization under distribution shifts and long-tail categories; and (ii) *partial observability*, where context serves as the sole signal for recovering heavily occluded objects. We unify these phenomena through a single causal framework and introduce **Evidence-Gated Backdoor Adjustment (EG-BA)** for Open-Set Object Detection (OSOD). EG-BA employs a learned per-query evidence score to dynamically interpolate between two regimes: when visual evidence is reliable, it executes a backdoor adjustment to strip confounding context; when evidence is degraded, it leverages context for identity recovery. To ensure the internal mechanism behaves in the exact way, we regularize the latent space via a variational information bottleneck and propose the *Gate-Occlusion Calibration Error (GOCE)* a diagnostic evaluation tool to understand the gate aligns with true physical occlusion. We investigated that EG-BA improves calibration and unknown-class recall over a vanilla Deformable DETR baseline model on Pascal VOC to MS-COCO dataset and its evidence score tracks synthetic occlusion severity when properly regularized, without this regularization the same information-bottleneck term drives the gate to a collapsed input-independent constant state instead, considered as a failure mode, we report as a central finding of independent interest. We further find that this occlusion-robustness gain does not transfer to real, naturally occurring occlusion and that the mechanism remains steady for co-occurrence-driven performance degradation, the bias it is designed to remove.

1 Introduction

Object detectors built on transformer decoders represent each candidate object as a query embedding refined through repeated self- and cross-attention over encoded scene features (Carion et al., 2020; Zhu et al., 2021). This design introduces a structural property that is rarely made explicit: because attention aggregates information across every object attended to during decoding, a query's final representation is not purely a function of its target object's local appearance. Instead, it forms a mixture of that object's direct visual evidence and the co-occurrence context contributed by neighboring objects. While this contextual mixing is precisely the mechanism that enables transformers to exploit scene context, it inherently binds the representation to two opposing failure modes that prior literature has treated in isolation.

The first failure mode is *co-occurrence bias*, where a detector relies on typical spatial surroundings rather than intrinsic appearance. This shortcut impairs generalization under distribution shifts and disproportionately penalizes rare or long-tailed categories Rosenfeld et al. (2018). The second failure mode represents its converse: *partial observability*. When an object's direct visual evidence is degraded by occlusion, the exact contextual signal responsible for bias becomes the sole remaining lifeline for identity recovery. Consequently, a model that aggressively suppresses context to mitigate bias loses its capacity to recover occluded instances, whereas a model that exploits context to handle occlusion inevitably reintroduces spurious bias. Treating these phenomena as unrelated

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.

problems creates a false dichotomy, forcing a hand-tuned trade-off between debiasing and context compensation.

The existing works fail to resolve this issues, for example, causal debiasing methods proposed in (Zhang et al., 2020; Yang et al., 2021) apply backdoor or front-door adjustments at the image or region level to remove spurious correlations, but are unequipped to recover missing evidence and are evaluated primarily on classification or grounding rather than detection. Conversely, occlusion-robustness methods (Kortylewski et al., 2020) leverage contextual or part-based reasoning to recover missing evidence without examining whether that same reasoning induces bias under clean visual conditions. Furthermore, Open-Set Object Detection (OSOD) Singh & Lobo (2026) and Open-World Object Detection (OWOD) methodologies (Joseph et al., 2021; Zohar et al., 2023; Gupta et al., 2022) target the orthogonal problem of identifying unknown categories absent at training time, without modeling co-occurrence context as a causal confound. To our knowledge, no prior work unifies debiasing and occlusion recovery into a single, bidirectional mechanism modulated by evidence reliability.

To address this, we introduce **Evidence-Gated Backdoor Adjustment (EG-BA)**, a causal module for OSOD that computes a unified attention-based estimate of a query's co-occurrence context and dynamically regulates its usage via a learned per-query evidence score. When direct visual evidence is reliable, the score steers the module toward a backdoor adjustment, stripping the confounding context contribution from the query. When evidence is degraded by occlusion, the same context estimate is leveraged directly for feature recovery. Thus, debiasing and context compensation cease to be competing objectives; rather, they emerge as two distinct operating points of a single, continuous mechanism governed per query.

Verifying internal mechanistic behavior requires looking beyond aggregate downstream accuracy, as modules can sometimes improve performance metrics through uninterpretable internal mechanics. To ensure the internal mechanism functions reliably, we regularize the evidence score using a variational information bottleneck (IB) and introduce the **Gate-Occlusion Calibration Error (GOCE)** module - a novel diagnostic metric evaluating whether the score's implied reliance on context tracks true physical occlusion severity. Through this diagnostic process, we uncover an important representation dynamic: while IB regularization is theoretically motivated to prevent over-reliance on context, unconstrained tuning risks driving the evidence score toward an uninformative constant. We analyze this collapse regime and establish that direct occlusion supervision successfully stabilizes the gate, yielding a well-calibrated evidence mapping.

Our primary contributions are summarized as follows:

- We identify co-occurrence bias and occlusion-induced evidence loss as dual manifestations of query-level contextual entanglement, introducing EG-BA to dynamically switch between backdoor adjustment and context recovery on a per-query basis.
- We propose Gate-Occlusion Calibration Error (GOCE) as an independent diagnostic metric to quantify whether learned evidence scores reliably track physical occlusion severity without relying solely on downstream task accuracy.
- We evaluate EG-BA on VOC-to-COCO and targeted co-occurrence shifts and demonstrate that EG-BA substantially reduces Wilderness Impact and improves unknown-class recall, while providing critical empirical insights into the interaction between information-bottleneck regularization and gate collapse.

2 Related Work

2.1 Open-Set and Open-World Object Detection

Open-World Object Detection (OWOD) introduced in Joseph et al. (2021) combines energy-based unknown-aware proposals with incremental class learning across sequential tasks. OW-DETR (Gupta et al., 2022) extends this to a deformable-DETR backbone using attention-mined pseudo-labels, while PROB (Zohar et al., 2023) replaces energy scoring with a probabilistic objectness head. These methods evaluate on an incremental VOC-to-COCO split. In contrast, we adopt a single-pass, non-incremental OSOD formulation (Hosoya et al., 2025; Singh & Lobo, 2026), training once on

known classes to isolate our causal mechanism from catastrophic forgetting. None of these methods model co-occurrence context as a causal confound.

2.2 Causal Debiasing via Backdoor and Front-Door Adjustment

Backdoor adjustment has been used to remove context bias in weakly-supervised segmentation: Zhang et al. (2020) (CONTA) approximate the interventional distribution via a learned dictionary and the Normalized Weighted Geometric Mean (NWGM), while Yang et al. (2021) (CATT) apply front-door adjustment in vision-language models Singh et al. (2020). Other methods like Ma et al. (2023) use co-occurrence as a positive alignment signal. However, these image-level approaches are designed solely to remove bias, failing to recognize that context is necessary when evidence is degraded. Unlike them, our backdoor adjustment operates at the individual query level, enabling query-specific context adjustments based on evidence reliability. More recently applies causal intervention in (Tang et al., 2020) and for attention-based visual recognition (Wang et al., 2021).

2.3 Occlusion-Robust Detection and Recognition

Compositional part-based models (Kortylewski et al., 2020) and detection-specific occlusion modules (Zhang et al., 2018) recover occluded objects by representing them as part mixtures and localizing occluders, demonstrating robustness on PASCAL3D+ and COCO. Amodal benchmarks such as KINS (Qi et al., 2019), alongside Occluded and Separated COCO (Zhan et al., 2022), evaluate real-world partial occlusion and mask separation. These methods treat context exclusively as a recovery resource, overlooking that the same relationships induce bias when evidence is reliable.

2.4 Information Bottleneck for Representation Learning

The variational Information Bottleneck (Alemi et al., 2017) frames representation learning as a trade-off between input compression and target predictability via a tractable variational bound. We adapt this objective to penalize excess context retention in high-evidence queries, validating the resulting gate behavior empirically via GOCE (Section 3) rather than assuming theoretical convergence.

3 Method

3.1 Problem Formulation and Causal Model

Let a Deformable-DETR-style decoder produce N object queries $\{Q_1, \ldots, Q_N\}$, $Q_i \in \mathbb{R}^d$, from an input image X through repeated self- and cross-attention over multi-scale encoder features. Because attention aggregates information across every object attended to during decoding, Q_i is not purely a function of object i's own appearance: it is a mixture of an unobserved true evidence signal Z_i and the co-occurrence context C contributed by the remaining objects in the scene. We formalize this with the causal graph

$$Z_i \to Q_i \leftarrow C, \qquad C \to Y_i, \qquad Q_i \to Y_i, \tag{1}$$

where Y_i denotes the classification decision for query i. The edge $C \to Y_i$ constitutes a backdoor path $Q_i \leftarrow C \to Y_i$: naive estimation of $P(Y_i \mid Q_i)$ is confounded by C, since C is a common cause of both the information entering Q_i and the label decision itself. A detector trained under this confound under performs whenever an object appears outside its typical context, and relies most heavily on context precisely where that context is least reliable, notably for rare classes and atypical scenes.

We are interested in the interventional distribution $P(Y_i \mid \mathrm{do}(Q_i))$, which severs this confounding path. By the backdoor adjustment formula, provided C satisfies the backdoor criterion relative to (Q_i, Y_i),

$$P(Y_i \mid \mathrm{do}(Q_i)) = \sum_c P(Y_i \mid Q_i, c)\, P(c). \tag{2}$$

Exact enumeration over c is intractable; Section 3.2 introduces a tractable, differentiable approximation to Equation 2.

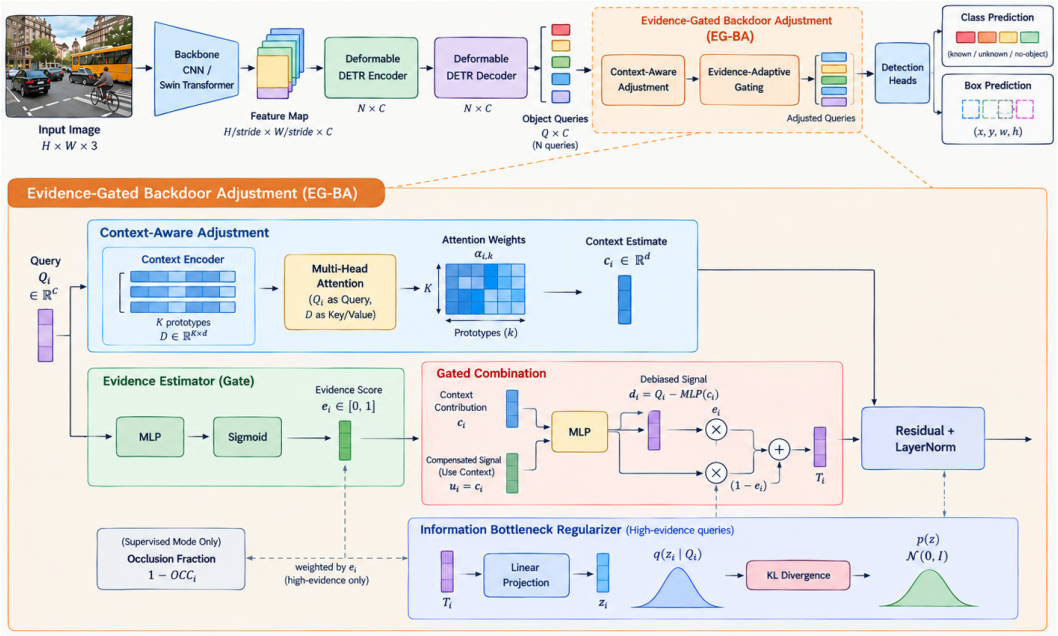

Figure 1: Overview of Evidence-Gated Backdoor Adjustment (EG-BA). A shared backbone produces per-query features, which feed two parallel branches: a context-dependent branch that predicts directly from the observed features, and a debiased branch that applies backdoor adjustment over a learned confounder dictionary.

3.2 Evidence-Gated Backdoor Adjustment

Context estimation. We approximate the marginal $P(c)$ with a learned dictionary $D \in \mathbb{R}^{K\times d}$ of K representative context prototypes, approximating the backdoor summation via a single attention operation under the NWGM approximation (Zhang et al., 2020; Yang et al., 2021):

$$c_i = \text{Attention}(Q_i, D, D) = \sum_{k=1}^{K} \alpha_{i,k}\, D_k, \qquad \alpha_{i,k} = \frac{\exp\big(Q_i^\top W_q (D_k^\top W_k)^\top / \sqrt{d}\big)}{\sum_{k'} \exp\big(Q_i^\top W_q (D_{k'}^\top W_k)^\top / \sqrt{d}\big)}. \tag{3}$$

Evidence score. A per-query score $e_i \in [0, 1]$ estimates the proportion of Q_i constituted by reliable direct visual signal versus occlusion-induced noise. Computed via a clamped logit to prevent gradient saturation near 0 or 1, the score is defined as:

$$e_i = \sigma\big(4 \tanh(\text{MLP}_e(Q_i)/4)\big), \qquad \tau = 4. \tag{4}$$

Gated combination. The context estimate c_i is deployed dynamically based on e_i: for high evidence reliability, c_i undergoes subtraction for backdoor adjustment; for low reliability, it serves as direct replacement for missing evidence:

$$d_i = Q_i - \text{MLP}_{\text{debias}}(c_i) \qquad \text{(dominant as } e_i \to 1\text{)}, \tag{5}$$

$$u_i = c_i \qquad \text{(dominant as } e_i \to 0\text{)}, \tag{6}$$

$$T_i = \text{LayerNorm}\big(Q_i + e_i\, d_i + (1 - e_i)\, u_i\big). \tag{7}$$

Here, T_i supersedes Q_i downstream. The residual connection in Equation 7 ensures the module refines, rather than substitutes, the decoder representation.

3.3 Information-Theoretic Regularization

We regularize the evidence score's behavior using a variational Information Bottleneck objective (Alemi et al., 2017), applied asymmetrically as a function of e_i: high-evidence queries are penalized

for retaining context-derived information, while low-evidence queries are not. We project T_i to the parameters of a diagonal Gaussian posterior and sample via the reparameterization trick such that the sampled latent z_i directly consumed by the detection heads instead of T_i. This ensures z_i is load-bearing for the detection loss rather than a quantity the network can disregard without penalty:

$$(\mu_i, \log \sigma_i^2) = \mathrm{Linear}(T_i), \tag{8}$$

$$z_i = \mu_i + \sigma_i \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \qquad (z_i = \mu_i \text{ at inference}), \tag{9}$$

$$\mathrm{KL}_i = \tfrac{1}{2} \sum_{j=1}^{d} \Big(\mu_{i,j}^2 + \sigma_{i,j}^2 - \log \sigma_{i,j}^2 - 1 \Big). \tag{10}$$

3.4 Gate-Occlusion Calibration Error

Downstream detection accuracy alone is insufficient to establish that e_i is calibrated to occlusion severity, since improved accuracy is consistent with an internally arbitrary score. Given a ground-truth occlusion fraction $o_i \in [0, 1]$ for a matched query, available under the synthetic occlusion construction described in Section 4, we define

$$\mathrm{GOCE} = \mathbb{E}_i \Big[\big| (1 - e_i) - o_i \big| \Big], \tag{11}$$

computed overall and per occlusion-severity bin. A lower value indicates that the score's implied reliance on context, $1 - e_i$, tracks true occlusion severity.

The total loss combines a standard Hungarian-matched detection loss $\mathcal{L}_{\mathrm{det}}$ with three terms on the evidence score: $\mathcal{L}_{\mathrm{IB}} = \mathbb{E}_i[e_i \cdot \mathrm{KL}_i]$, weighting Equation 10 by e_i; $\mathcal{L}_{\mathrm{evidence}} = \mathbb{E}_i[\mathrm{BCE}(e_i, 1 - o_i)]$ against ground-truth occlusion where available ($\lambda_{\mathrm{ev}} = 0$ otherwise); and an anti-collapse term $\mathcal{L}_{\mathrm{gate}} = (\bar{e} - 0.5)^2$. As shown in Section 4.2, this term and the clamp in Equation 4 do not prevent collapse under the default information-bottleneck weighting. Following this, the total loss is represented in Equation 12.

$$Total : \mathcal{L} = \mathcal{L}_{\mathrm{det}} + \lambda_{\mathrm{IB}} \mathcal{L}_{\mathrm{IB}} + \lambda_{\mathrm{ev}} \mathcal{L}_{\mathrm{evidence}} + \lambda_{\mathrm{gate}} \mathcal{L}_{\mathrm{gate}} \tag{12}$$

4 Experiments and Results

We follow the standard non-incremental VOC $\rightarrow$ COCO OSOD protocol and trained the model on Pascal VOC 2007+2012 trainval set (20 known classes) and evaluated on COCO val2017 (60 unknown categories grouped into three super-category sets G_1, G_2, G_3). For co-occurrence shift analysis, test subsets are partitioned based on whether an object's frequent companion class is present. Synthetic occlusion ground truth is constructed via instance mask-masking to provide supervision (λ_{ev}) and GOCE diagnostics, trading natural realism for controlled precision.

Our implementation builds upon a ResNet-50 Deformable DETR backbone ($N = 100$ queries, $d = 256$, $K = 20$ confounder prototypes), optimized via AdamW. We benchmark against an identically trained vanilla Deformable DETR baseline and our evaluation metrics include known mAP (IoU $= 0.5$), Wilderness Impact (WI), Absolute Open-Set Error (A-OSE), unknown-class recall (U-Recall), and Gate-Occlusion Calibration Error (GOCE). All the experiments are performed on NVIDIA H100 GPU.

4.1 Main Result

Our Table 1 summarizes primary evaluation on the VOC $\rightarrow$ COCO benchmark. EG-BA significantly improves open-set calibration and discovery metrics over the vanilla Deformable DETR baseline, reducing Wilderness Impact from 0.0181 to 0.0097 while raising average U-Recall from 0.0452 to 0.0573. The most pronounced unknown-recall gain is observed within the *outdoor/accessory/appliance/truck/animal* group (0.1126 vs. 0.0927). As dictated by the fundamental trade-off between closed-world shortcut exploitation and open-world generalization, these gains incur a known-mAP cost of 2.87 points (51.20 vs. 54.07) alongside a minor increase in A-OSE (1224 vs. 1079).

Table 1: Open-set detection performance under the VOC→COCO protocol. Unknown classes are split into three 20-class super-categories (G1: outdoor/accessory/appliance/truck/animal; G2: sports/food; G3: electronic/indoor/kitchen/furniture). Best value is reported per column in **bold**.

Model	Known mAP	U-Recall (↑)				WI	A-OSE
	(↑)	Avg	G1	G2	G3	(↓)	(↓)
Vanilla Deformable DETR	**54.07**	0.0452	0.0927	0.0205	0.0224	0.0181	**1079**
EG-BA (Ours)	51.20	**0.0573**	**0.1126**	**0.0226**	**0.0367**	**0.0097**	1224

Table 2: Effect of confounder dictionary size K on open-set detection performance (VOC→COCO, adaptive λ_{IB}). WI improves monotonically with K; known mAP and U-Recall are non-monotonic.

K	**Known mAP** (↑)	**WI** (↓)	**A-OSE** (↓)	**U-Recall avg** (↑)
10	49.25	0.0329	1115	0.0497
20	**53.07**	0.0146	1272	**0.0381**
50	51.40	**0.0128**	**1207**	0.0412

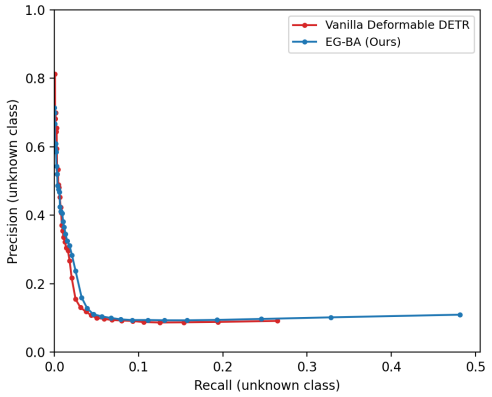

Figure 2: Precision-recall curves for unknown-class detection, obtained by sweeping the objectness threshold at fixed known_thresh$= 0.40$.

The two curves closely track each other through most of the shared recall range; EG-BA extends to a higher maximum recall (0.48 vs. 0.27), consistent with its higher average U-Recall. To verify that these improvements are robust to threshold selection, Figure 2 illustrates the precision-recall curves for unknown-class detection across a swept objectness threshold (obj_thresh $= 0.30$, known_thresh $= 0.40$). Both curves track closely through the shared operating domain, but EG-BA successfully extends the upper recall frontier to a maximum of 0.48 (compared to 0.27 for the baseline). This confirms that EG-BA's superiority in U-Recall is a structural property of the learned representation rather than an artifact of a single isolated operating point.

4.2 Ablation Studies

To dissect the contribution of individual design components within EG-BA, we perform a series of targeted ablations examining the confounder dictionary capacity, the information bottleneck regularization strength, branch routing configurations, and explicit occlusion supervision.

Confounder Dictionary Size (K). Our Table 2 evaluates the sensitivity of EG-BA to the dictionary size $K \in \{10, 20, 50\}$. We observe that Wilderness Impact (WI) scales monotonically with dictionary capacity, improving from 0.0329 at $K = 10$ down to 0.0128 at $K = 50$. Conversely, known mAP, Absolute Open-Set Error (A-OSE), and average unknown recall (U-Recall) exhibit non-monotonic trends, peaking or optimizing at $K = 20$ (achieving 53.07 known mAP and 1272 A-OSE). Consequently, $K = 20$ is selected as the balanced default configuration for all subsequent experiments.

Information Bottleneck Weight (λ_{IB}) and Gate Collapse. Our, Table 3 and Figure 3 reveal a critical failure mode: setting $\lambda_{IB} = 0.1$ causes the evidence gate to collapse to a near-constant value (mean 0.0179, std 0.0000) within the initial epoch, persisting through training convergence. Conversely, eliminating the bottleneck ($\lambda_{IB} = 0$) preserves a healthy, distributed gate across the full spectrum (mean 0.4890, std 0.1804). While the collapsed variant ($\lambda_{IB} = 0.1$) matches the unconstrained model on known mAP (53.07 vs. 51.20), it suffers catastrophic degradation on open-

Table 3: Effect of the information bottleneck weight λ_{IB} on evidence gate behavior and open-set detection performance (VOC→COCO, $K = 20$). A non-zero λ_{IB} collapses the gate (mean ≈ 0.018, std ≈ 0); $\lambda_{IB} = 0$ keeps the gate distributed throughout training.

λ_{IB}	**Known mAP** (↑)	**WI** (↓)	**A-OSE** (↓)	**U-Recall avg** (↑)	**Gate (mean/std)**
0.1 (default)	**53.07**	0.0146	1272	0.0381	0.0179 / 0.0000
0.0	51.20	**0.0097**	**1224**	**0.0573**	0.4890 / 0.1804

Table 4: Effect of forcing the evidence gate to a fixed value, versus learning it adaptively (VOC→COCO, $K = 20$, $\lambda_{IB} = 0$ for the adaptive row).

Configuration	**Known mAP** (↑)	**WI**	**A-OSE** (↓)	**U-Recall avg** (↑)
Fixed $e_i = 1$ (pure debiasing)	7.79	−0.1129	29	0.0204
Fixed $e_i = 0$ (pure context)	52.82	0.0141	1366	0.0689
Adaptive gate ($\lambda_{IB} = 0$)	51.20	0.0097	1224	0.0573

world metrics (WI of 0.0146 vs. 0.0097; U-Recall of 0.0381 vs. 0.0573). This demonstrates that the information bottleneck penalty, under standard configurations, inadvertently incentivizes gate collapse rather than meaningful compression; hence, $\lambda_{IB} = 0$ is adopted as our headline setting.

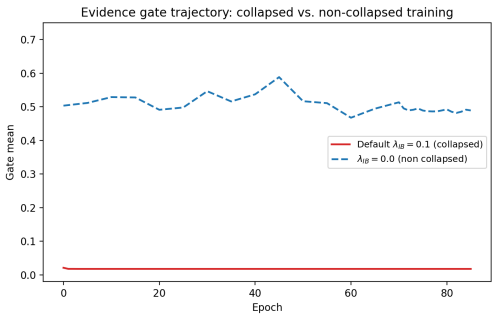

Figure 3: Evidence gate trajectory for $\lambda_{IB} = 0.1$ (collapsed) vs. $\lambda_{IB} = 0.0$ (non-collapsed).

Branch Interpolation and Routing Strategy. The Table 4 contrasts our learned adaptive gate against static routing extremes ($e_i = 1$ for pure debiasing and $e_i = 0$ for pure context). Forcing pure debiasing ($e_i = 1$) triggers near-total prediction collapse, yielding only 29 valid detections and rendering open-set metrics uninterpretable. Conversely, fixing $e_i = 0$ (retaining pure context without adjustment) trains stably and achieves the highest U-Recall in this study (0.0689) at a modest known-mAP expense (52.82). Nevertheless, the adaptive gate ($\lambda_{IB} = 0$) remains the sole architecture capable of balancing these dynamics via dynamic, input conditioned evaluation rather than relying on a rigid hyperparameter.

Impact of Evidence Supervision (λ_{ev}). Our, Table 5 and Figure 4 isolate the role of explicit supervision on occlusion tracking. Omitting the supervision term ($\lambda_{ev} = 0$) degrades occlusion-aware calibration—measured via Gate-Occlusion Calibration Error (GOCE)—by a factor of 5.7 (0.0991 vs. 0.5632). Crucially, general confidence calibration (ECE, (Guo et al., 2017)) remains virtually unaffected (0.0072 vs. 0.0099), confirming that direct supervision specifically aligns the evidence score with physical occlusion geometry rather than indiscriminately skewing overall network calibration.

4.3 Occlusion Robustness

To comprehensively evaluate how effectively contextual compensation handles missing visual evidence, we test both models under controlled synthetic occlusion as well as real-world occlusion benchmarks. Table 6 summarizes performance under synthetic occlusion across five distinct severity tiers ranging from 0% to 80%. While known mAP degrades at comparable relative trajectories for both architectures

Table 5: Effect of evidence supervision on occlusion-aware calibration (GOCE) and general confidence calibration (ECE). Supervision improves occlusion-aware calibration specifically.

Config	**GOCE** (↓)	**ECE** (↓)
Supervised ($\lambda_{ev} = 1$)	**0.0991**	0.0072
Unsupervised ($\lambda_{ev} = 0$)	0.5632	**0.0099**

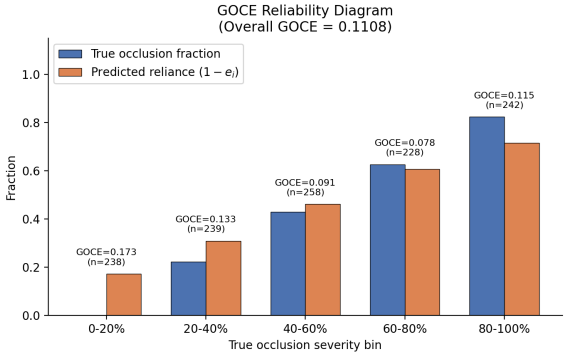

Figure 4: GOCE reliability diagram: predicted evidence vs. true occlusion, binned into five occlusion ranges.

Table 6: Detection performance across synthetic occlusion severity bins. Baseline U-Recall degrades sharply with increasing occlusion while EG-BA's unknown-class recall remains essentially flat at a moderate cost in known-class mAP.

Occlusion	**Vanilla Deformable DETR**		**EG-BA (Ours)**	
	Known mAP	**U-Recall**	**Known mAP**	**U-Recall**
0%	54.20	0.0456	51.51	0.0579
20%	51.80	0.0430	49.51	0.0545
40%	48.32	0.0381	46.24	0.0587
60%	45.61	0.0350	41.75	0.0557
80%	41.68	0.0347	38.14	0.0577

(experiencing a 23.1% relative drop for the baseline versus 26.0% for EG-BA), unknown-class recall (U-Recall) exhibits sharply contrasting behaviors. Specifically, the baseline U-Recall drops by 23.9% relative (shifting from 0.0456 down to 0.0347), whereas EG-BA maintains an exceptionally stable profile (0.0579 to 0.0577, representing a minimal 0.35% relative variation). Consequently, EG-BA outperforms the baseline across all severity levels, with the absolute performance margin widening progressively as occlusion severity intensifies—growing from a 0.0123 gap at 0% occlusion to 0.0230 at the 80% threshold.

Conversely, Table 7 extends the evaluation to Occluded-COCO and Separated-COCO (Zhan et al., 2022), which feature naturally occurring physical occlusion rather than synthetic mask corruptions. Across these real-world validation sets, all four known and unknown recall comparisons between EG-BA and the baseline remain constrained within a narrow margin of one percentage point. We interpret this outcome as statistical parity rather than a performance advantage: the pronounced robustness gains observed under controlled synthetic conditions (Table 6) fail to translate into a discernible margin in unconstrained, natural settings.

4.4 Co-occurrence Shift

To investigate whether the causal framework successfully mitigates contextual shortcut exploitation, Table 8 and Figure 5 quantify the average Average Precision (AP) gap between co-occurring (coupled) and isolated (decoupled) object instances across ten mined category pairs.

Contrary to expectations, both architectures yield a negative mean AP gap (recording -0.1186 for the baseline and -0.1344 for EG-BA), with nine out of the ten evaluated pairs exhibiting a negative delta in both instances. This indicates that EG-BA fails to contract the performance gap relative to the standard Deformable DETR baseline.

A qualitative audit of the data reveals a confounding structural artifact: eight out of the ten mined category pairs incorporate the *person* category as a primary companion class. In practice, co-occurrence with a person frequently introduces physical contact, dense crowds, or partial occlusion rather than

Table 7: Generalization to Occluded-COCO and Separated-COCO using box-IoU recall in place of the original mask-IoU metric.

Benchmark	Model	Known-class Recall	Unknown-class Recall
Occluded-COCO	Vanilla Deformable DETR	0.5350	0.0587
	EG-BA (Ours)	0.5356	0.0612
Separated-COCO	Vanilla Deformable DETR	0.3778	0.0879
	EG-BA (Ours)	0.3697	0.0871

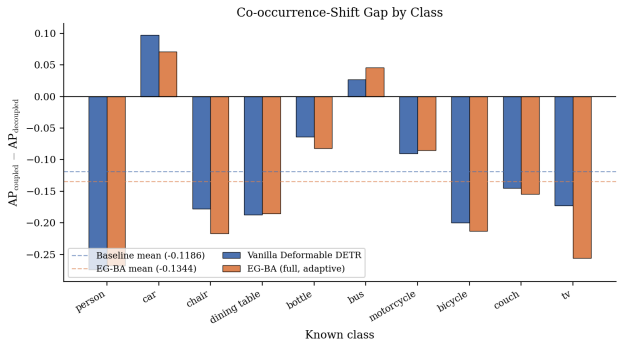

Figure 5: Per-pair AP gap (coupled − decoupled) for both models, across the 10 mined category pairs underlying Table 8.

acting as a purely spurious semantic prior. Consequently, this benchmark setup may conflate genuine physical occlusion dynamics with contextual co-occurrence bias, thereby restricting its capacity to isolate the specific causal vulnerability that EG-BA is engineered to address.

The performance gap between co-occurring (coupled) and isolated (decoupled) object instances, averaged over 10 category pairs, shows that EG-BA does not close this gap relative to the vanilla baseline. Both models exhibit negative mean gaps, with 9 of 10 pairs performing worse in the coupled setting.

Table 8: Detection performance gap between co-occurring (coupled) and isolated (decoupled) object instances.

Model	Mean gap	Neg. pairs
Vanilla DETR	-0.1186	9/10
EG-BA (Ours)	-0.1344	9/10

4.5 Comparison to Open-World Detection Methods

Our Table 9 compares EG-BA to open-world object detection methods where known/unknown split is based on our protocol (20 known VOC classes and 60 unknown COCO classes, prior to any incremental learning). Because these prior methodologies compute aggregate metrics across the complete unknown category pool, the fine-grained, group-stratified breakdown presented in Table 1 is omitted here.

As indicated by the results, EG-BA achieves substantially superior calibration, yielding dramatically lower Wilderness Impact (0.0097 vs. $0.0621/0.0571/0.0569$) and Absolute Open-Set Error (1224 vs. $10459/10240/5195$) compared to ORE, OW-DETR, and PROB. This strong calibration stems directly from the evidence gate's ability to isolate well-evidenced object instances from ambiguous predictions, thereby suppressing known-unknown confusion.

4.6 Qualitative Analysis

Figure 6 visually demonstrates the empirical impact of evidence supervision on a representative real-world COCO instance subjected to controlled, progressive occlusion. For the model trained with explicit evidence supervision, the predicted evidence score e_i decreases monotonically, shifting



Figure 6: Evidence score under progressive occlusion of the same real COCO instance (a person) on the same occlusion mechanism applied to the instance's real segmentation mask. Panel titles report the measured achieved occlusion fraction. Top row: evidence-supervised. Bottom row: default (collapsed-gate). The rightmost panel in each row plots evidence against measured occlusion directly.

Table 9: Comparison of various models where 20 VOC known classes and 60 COCO unknown classes are used. The finer 3-group unknown breakdown reported earlier is not available for prior work and is therefore omitted here.

Method	**Known mAP** (↑)	**U-Recall** (↑)	**WI** (↓)	**A-OSE** (↓)
ORE	56.0	4.9	0.0621	10459
OW-DETR	59.2	7.5	0.0571	10240
PROB	59.5	19.4	0.0569	5195
EG-BA (Ours)	51.20	5.73	0.0097	1224

smoothly from 0.98 under clean conditions (0% occlusion) down to 0.19 under heavy degradation (80% occlusion).

In sharp contrast, the evidence score for the default (collapsed-gate) checkpoint remains locked at a constant 0.018 across all five occlusion severity levels, directly corroborating the mathematical gate collapse documented in Table 3 and Figure 3. Because both model variations evaluate identical object instances, underlying images, and synthetic corruption masks, this divergence isolates the training objective configuration as the sole determinant of the gate's sensitivity to physical occlusion.

5 Conclusion

In this paper, we introduced Evidence-Gated Backdoor Adjustment (EG-BA), a causal framework designed to unify context debiasing and evidence recovery for open-set object detection via a learned, per-query evidence score. Our empirical evaluation demonstrates that EG-BA effectively improves open-set calibration and unknown-class recall over standard transformer baselines, maintaining robustness under synthetic occlusion. However, our investigation also uncovers notable limitations: the proposed information-bottleneck regularizer induces gate collapse, requiring supervision or a zero penalty to function; synthetic occlusion gains do not consistently transfer to real-world occlusion settings; and the model does not alleviate co-occurrence-driven performance degradation. We also highlight the crucial nuances regarding when evidence-gated causal mechanisms succeed, and where current formulations require further evolution.

AI Use Statement

In this work, we used generative AI tools for polishing and writing the research paper. We have not used generative AI tools for designing the core methodology, for generating any reported experimental values, or for interpreting results. We have reviewed all AI-assisted work and take full responsibility for the content of this submission.

Ethics Statement

Our work focuses on improving the robustness of object detection models through a causal evidence-gated mechanism. We utilize established, publicly available research datasets (Pascal VOC and MS-COCO). This research does not involve the collection of personal data, human subjects, or any applications that raise ethical concerns. While our analysis reveals that the 'person' category acts as a significant confounder in current detectors, our work is intended to help researchers understand and mitigate these biases, rather than to facilitate surveillance or identification of individuals. We believe our work contributes to the development of more reliable and interpretable AI systems.

Reproducibility Statement

We have taken several steps to ensure the reproducibility of our work:

- **Methodology:** Our causal framework, EG-BA, is fully defined in Section 3, including the mathematical formulation of the evidence gate (Eq. 4) and the information-theoretic regularization objective (Eq. 12).
- **Implementation Details:** We provide comprehensive details regarding our architecture (ResNet-50 Deformable DETR), hyperparameters (e.g., $N = 100$, $d = 256$, $K = 20$), and optimization strategy (AdamW) in Section 4.
- **Data:** All experiments were conducted using the standard, publicly available Pascal VOC 2007+2012 trainval sets and COCO val2017 dataset. We describe the synthetic occlusion construction and evaluation protocol in detail to enable verification.
- **Results and Ablations:** All primary metrics and ablation studies (including the impact of dictionary size and information bottleneck weights) are reported in Tables 1 through 12, with accompanying training trajectories provided in Figure 10 to demonstrate convergence behavior.

The source code and model implementation will be made publicly available upon publication to facilitate further research.

References

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, pp. 213–229. Springer, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1321–1330, 2017.

Akshita Gupta, Sanath Narayan, K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. OW-DETR: Open-world detection transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9235–9244, 2022.

Yusuke Hosoya, Masanori Suganuma, and Takayuki Okatani. Rethinking open-set object detection: Issues, a new formulation, and taxonomy. *International Journal of Computer Vision*, 133(9): 6145–6169, 2025.

K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5830–5840, 2021.

Adam Kortylewski, Ju He, Qing Liu, and Alan L. Yuille. Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8940–8949, 2020.

Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. CoDet: Co-occurrence guided region-word alignment for open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, 2023.

Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with KINS dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3014–3023, 2019.

Amir Rosenfeld, Richard Zemel, and John K. Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

Binay Kumar Singh and Niels Da Vitoria Lobo. Contextually enhanced unknown aware contrastive learning based co-occurring object detection. *IEEE Access*, 2026. doi: 10.1109/ACCESS.2026.3718368.

Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9847–9857, 2021.

Guanqi Zhan, Weidi Xie, and Andrew Zisserman. A tri-layer plugin to improve occluded detection. In *British Machine Vision Conference*, 2022.

Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 33, pp. 655–666, 2020.

Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Occlusion-aware r-cnn: Detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 637–653, 2018.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. PROB: Probabilistic objectness for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11444–11453, 2023.

A Appendix

A.1 Efficiency and Calibration

Our Table 10 details the inference throughput, parameter footprint, and calibration performance across three core training configurations. Here, notably, all configurations maintain a constant size of 28.70M parameters, confirming that the evidence-gating mechanism introduces zero parameter overhead relative to the base detector architecture. Across the variants, inference throughput ranges from 73.08 to 86.62 FPS. The higher frames-per-second metric observed for the standard VOC configuration stems strictly from variations in the evaluation pipeline rather than architectural scaling, given that the parameter count remains invariant across all rows. Furthermore, the Gate-Occlusion Calibration Error (GOCE) is omitted for the VOC configuration because the Pascal VOC dataset does not provide pixel-level occlusion ground truth by design.

Table 10: Inference efficiency and calibration across training configurations.

Config	GOCE	ECE	FPS	Params (M)	Train time/epoch (s)
VOC (default)	—	0.0202	86.62	28.70	1120.8
COCO-diagnostic (supervised)	0.0991	0.0072	73.08	28.70	1046.8
Unsupervised (evidence_sup_off)	0.5632	0.0099	73.34	28.70	1078.9

A.2 Discussion

The proposed EG-BA framework enhances open-set calibration—as measured by Wilderness Impact (WI) and Absolute Open-Set Error (A-OSE)—and improves unknown-class recall relative to the baseline, albeit at a marginal cost to known-class mAP (Table 1). Furthermore, the model demonstrates robust performance under synthetic occlusion (Table 6). These gains, however, are subject to two primary limitations: (i) the observed robustness does not generalize to real-world occlusion benchmarks (Table 7), suggesting that synthetic resilience is not a sufficient proxy for naturalistic occlusion; and (ii) EG-BA fails to mitigate co-occurrence-driven performance degradation (Table 8). Consequently, the evidence-gating mechanism improves calibration and recall within its specific evaluation protocol, yet it does not provide a fundamental, mechanism-level resolution to the underlying co-occurrence bias.

Our investigation identifies the gate-collapse phenomenon (Table 3) as a critical finding: configurations utilizing an information-bottleneck weight of $\lambda_{IB} = 0.1$ (for $K \in \{10, 20, 50\}$) result in a universal collapse of the evidence gate. Only by setting $\lambda_{IB} = 0$ is a distributed gate maintained throughout training. This suggests that the information-bottleneck objective, as currently formulated, incentivizes gate trivialization rather than meaningful representation regularization, rendering the model's efficacy contingent upon the removal of its intended regularization component.

Finally, comparative evaluations (Table 9) demonstrate that EG-BA achieves superior calibration (WI, A-OSE) compared to ORE, OW-DETR, and PROB, though it records lower U-Recall than specialized incremental-learning baselines. It is important to note that EG-BA is explicitly designed to address context-driven bias within known classes; the resulting open-set performance gains emerge as a secondary structural byproduct rather than a primary optimization objective.

A.3 Limitations

Despite the observed improvements in calibration and open-set performance, our method exhibits several limitations. First, EG-BA incurs a performance penalty of 2.87 points in known-class mAP (Table 1). As detailed in Table 12, this degradation is disproportionately concentrated in the *sheep* class, suggesting potential sensitivity in specific feature distributions that may hinder closed-world applicability.

Second, the evidence-gating mechanism is highly sensitive to the information-bottleneck regularization weight; as demonstrated in Table 3, the gate collapses under default settings, necessitating the removal of the regularization term ($\lambda_{IB} = 0$) to preserve an informative gate.

Third, our results highlight a gap between synthetic and natural robustness: the gains in occlusion-aware detection observed under synthetic mask corruption fail to generalize to real-world occlusions (Table 7). Additionally, the framework does not effectively mitigate co-occurrence-driven bias, as evidenced by the negative performance deltas in Table 8. Finally, while EG-BA achieves superior calibration, its unknown-class recall remains lower than that of purpose-built incremental learning baselines (Table 9), limiting its utility in scenarios where maximal discovery is the primary objective over calibrated prediction.

Note on diagnostic performance: As shown in Figure 7, while both models experience comparable relative declines in known-class mAP as occlusion severity increases, EG-BA maintains a markedly more stable U-Recall profile—exhibiting a 0.35% relative variation compared to 23.9% for the baseline—demonstrating the effectiveness of the gated mechanism in sustaining identity recovery under severe occlusion.

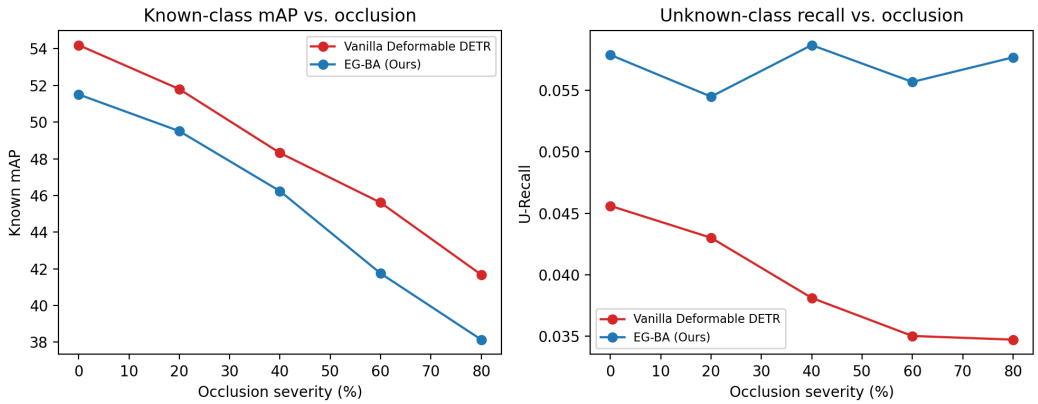

Figure 7: Known mAP (left) and U-Recall (right) as a function of synthetic occlusion severity, Vanilla Deformable DETR vs. EG-BA.

Qualitative Analysis of Evidence Patterns

Figure 8 provides a qualitative analysis of evidence patterns across 100 randomly sampled held-out images. Notably, all correct detections within this sample exhibit evidence scores above 0.5, suggesting a strong coupling between high evidence signals and detection accuracy. Conversely, the low U-Recall reported in Table 1 is corroborated by the disproportionate ratio of unknown-class misses (269) to unknown-class hits (5), underscoring the challenges of discovery in the current experimental setup.

Empirical Distribution of Gate Values

Figure 9 complements the epoch-level gate trajectory shown in Figure 3 by visualizing the distribution of per-query evidence scores at convergence. Under the collapsed regime, every query in the evaluation batch is assigned an identical gate value; this confirms that the observed collapse reduces the gate to a degenerate constant rather than merely shifting the mean. In contrast, with $\lambda_{IB} = 0$, the gate values span the full $[0, 1]$ range with a mode centered between 0.4 and 0.5, indicating that the mechanism successfully preserves input-dependent gate sensitivity.

Training Dynamics and Convergence

Figure 10 illustrates the detection loss and evidence gate trajectories for the five configurations with complete training logs. The collapsed configuration exhibits a rapid decline in the gate value, reaching a terminal state (0.018) within the first epoch and remaining invariant throughout the remainder of training. In contrast, the $e_i = 0$, $e_i = 1$, and $K = 10$ configurations converge smoothly regarding detection loss, without demonstrating such gate collapse. The `evidence_sup_off` configuration is plotted for detection loss only, as it employs a distinct evidence-logging convention that is not directly comparable to the `gate_mean` metric observed in the other four experimental runs.

Calibration Threshold Sensitivity Analysis

Table 11 details the comprehensive threshold sweep employed to determine the operating point (obj_thresh $= 0.30$, known_thresh $= 0.40$) utilized throughout the main text, covering the EG-BA headline configuration and three ablation studies ($e_i = 0$, $e_i = 1$, $K = 10$). Note that the reported known-class mAP is threshold-independent and calculated on a 500-image calibration subset; this differs from the primary mAP reported in the main text, which is evaluated on a disjoint 4,500-image held-out set. We observe that the $e_i = 1$ configuration exhibits consistently higher Wilderness Impact (WI) across all threshold combinations. This elevated WI is indicative of the near-total prediction collapse observed for this checkpoint (Table 4) rather than a measure of gen-

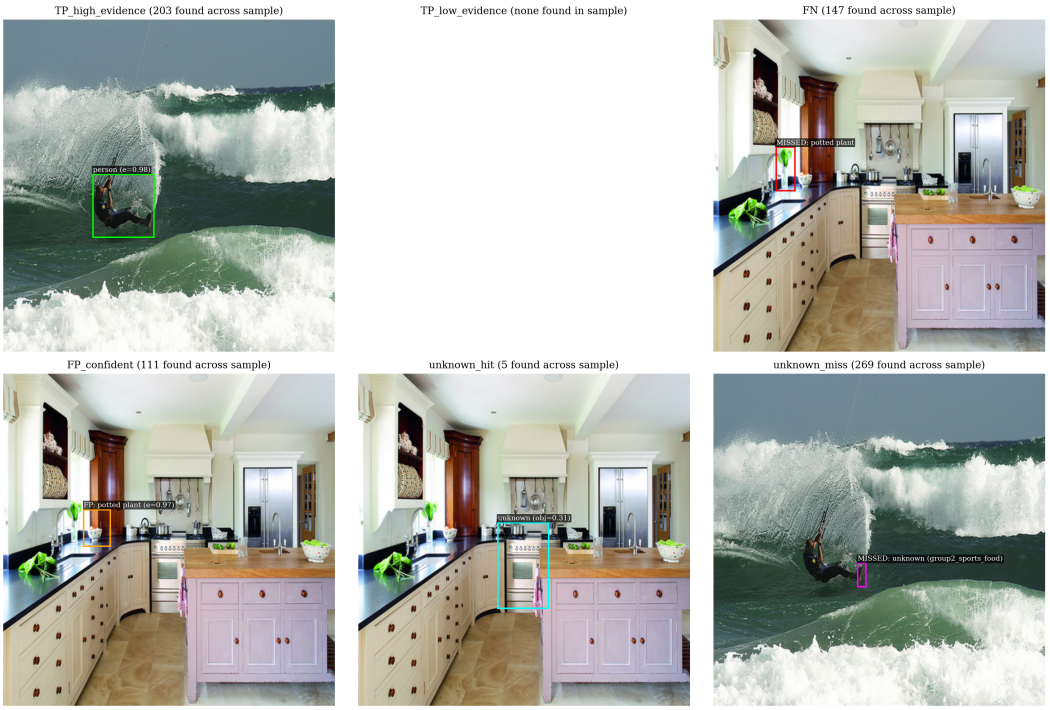

Figure 8: Evidence-pattern gallery across 100 randomly sampled COCO val2017 images (seed 0), evidence-supervised checkpoint, obj_thresh= 0.30, known_thresh= 0.40. Panels show a representative example of each pattern found; category counts across the full sample are given below. High-evidence true positive: 203. Low-evidence true positive: 0 (no correct detection in the sample carried evidence below 0.5). False negative: 147. Confident false positive: 111. Unknown-class hit: 5. Unknown-class miss: 269.

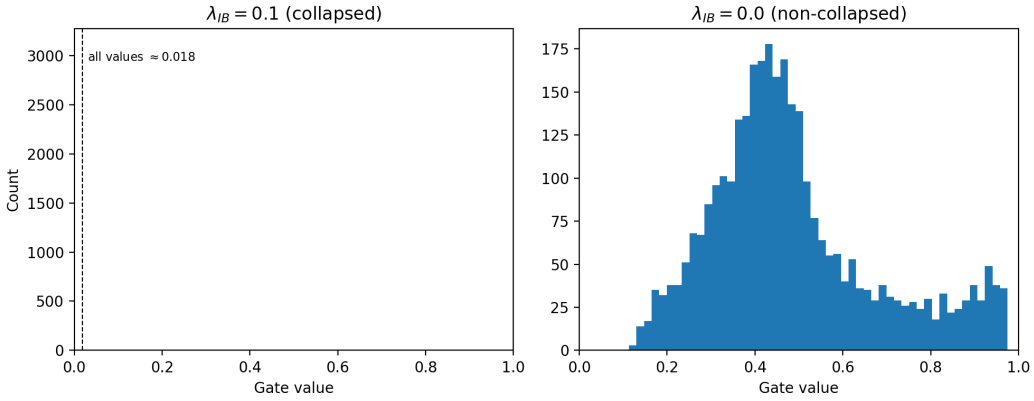

Figure 9: Distribution of per-query gate values at convergence, default $\lambda_{IB} = 0.1$ (collapsed, left) vs. $\lambda_{IB} = 0$ (non-collapsed, right), evaluated over a fixed 3,200-query batch. Left: all values equal 0.018 (std = 0.0000), shown as a delta spike. Right: values are broadly distributed (mean = 0.486, std = 0.191).

uine calibration quality, as the metric is computed over a near-empty prediction set at this operating point.

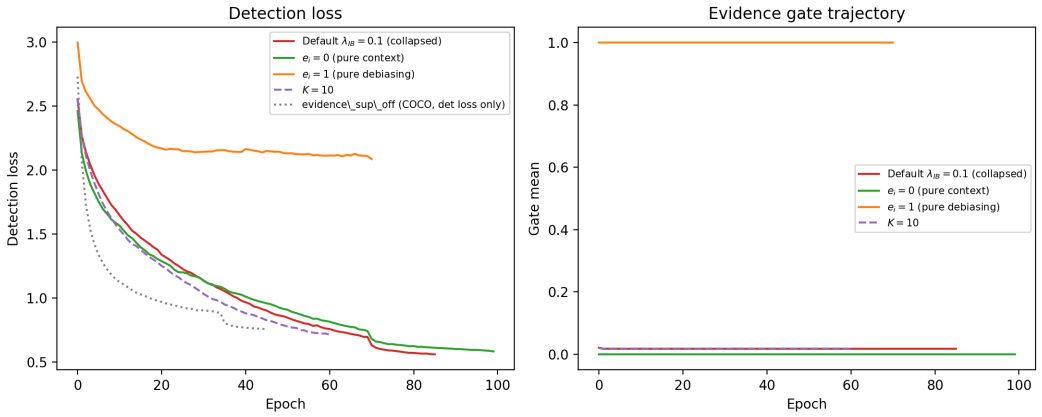

Figure 10: Full training curves for detection loss (left) and evidence gate trajectory (right), across five training configurations with complete epoch-by-epoch logs: the default $\lambda_{IB} = 0.1$ (collapsed), $e_i = 0$ (pure context), $e_i = 1$ (pure debiasing), $K = 10$, and `evidence_sup_off` (COCO-diagnostic detection loss only).

Per-Class Performance Analysis

Table 12 provides a per-class Average Precision (AP) comparison between the baseline and EG-BA. The observed degradation in known-class mAP is not distributed uniformly across categories. Specifically, the *sheep* class exhibits a performance drop of 19.83 points—an outlier nearly an order of magnitude larger than any other category—whereas the remaining 19 classes remain within a ± 5 point margin. Furthermore, five categories (*bicycle*, *train*, *boat*, *potted plant*, and *tv*) demonstrate performance improvements under the EG-BA framework. We do not observe a consistent correlation between this performance outlier and factors such as object category (e.g., animal, vehicle, or furniture) or instance scale, suggesting that the discrepancy for the *sheep* class may be attributable to specific feature distribution characteristics rather than systemic architectural bias.

Table 11: Full calibration threshold sweep for four checkpoints variations (500-image calibration subset, seed 0), sorted by average U-Recall within each checkpoint. Threshold-independent calibration-set known mAP shown per checkpoint; final reported mAP uses the disjoint 4500-image held-out set (see main text).

Checkpoint	obj_thresh	known_thresh	WI	avg_U_Recall
EG-BA headline ($\lambda_{IB} = 0$), calib. mAP $= 0.4249$				
	0.30	0.40	-0.0124	0.0760
	0.30	0.30	-0.0045	0.0385
	0.40	0.40	-0.0035	0.0312
	0.30	0.20	-0.0037	0.0146
	0.40	0.30	-0.0028	0.0146
	0.50	0.40	-0.0007	0.0140
	0.60	0.40	-0.0030	0.0101
	0.50	0.30	0.0009	0.0055
	0.60	0.30	-0.0001	0.0034
	0.40	0.20	-0.0016	0.0028
	0.50	0.20	-0.0005	0.0007
	0.60	0.20	0.0000	0.0000
Ablation $e_i = 0$ (pure context), calib. mAP $= 0.4545$				
	0.30	0.40	0.0023	0.0964
	0.30	0.30	-0.0013	0.0517
	0.40	0.40	-0.0014	0.0378
	0.50	0.40	-0.0018	0.0186
	0.40	0.30	-0.0019	0.0180
	0.30	0.20	-0.0010	0.0148
	0.60	0.40	-0.0008	0.0084
	0.50	0.30	-0.0010	0.0069
	0.40	0.20	-0.0002	0.0021
	0.60	0.30	-0.0001	0.0021
	0.50	0.20	-0.0004	0.0007
	0.60	0.20	0.0000	0.0000
Ablation $e_i = 1$ (pure debiasing), calib. mAP $= 0.0452$				
	0.30	0.40	0.1841	0.0399
	0.30	0.30	0.1085	0.0343
	0.40	0.40	0.2356	0.0266
	0.40	0.30	0.1405	0.0224
	0.50	0.40	0.2460	0.0168
	0.30	0.20	0.1567	0.0147
	0.50	0.30	0.1323	0.0126
	0.40	0.20	0.0861	0.0119
	0.60	0.40	0.0860	0.0091
	0.60	0.30	0.0691	0.0077
	0.50	0.20	0.0830	0.0035
	0.60	0.20	0.0600	0.0000
Ablation $K = 10$, calib. mAP $= 0.4387$				
	0.30	0.40	0.0407	0.0542
	0.30	0.30	0.0166	0.0334
	0.40	0.40	0.0115	0.0306
	0.40	0.30	0.0002	0.0164
	0.30	0.20	0.0021	0.0145
	0.50	0.40	0.0089	0.0135
	0.50	0.30	0.0029	0.0062
	0.60	0.40	0.0027	0.0061
	0.40	0.20	-0.0011	0.0048
	0.60	0.30	0.0002	0.0028
	0.50	0.20	0.0004	0.0007
	0.60	0.20	0.0004	0.0000

Table 12: Per-class known AP@0.5 (%), Vanilla Deformable DETR vs. EG-BA (VOC→COCO, obj_thresh= 0.30, known_thresh= 0.40, 4500-image held-out set). Class-level deltas average to the 2.87-point known-mAP gap. The sheep class shows a substantially larger drop (-19.83) than any other class; all remaining classes fall within a ± 8.5 range.

Class	**Vanilla**	**EG-BA**	Δ
person	68.19	67.63	-0.57
bicycle	27.89	29.48	$+1.59$
car	37.18	34.62	-2.56
motorcycle	56.00	50.01	-6.00
airplane	70.07	65.40	-4.67
bus	62.13	58.32	-3.80
train	85.43	85.72	$+0.28$
boat	19.56	23.86	$+4.31$
bird	26.14	21.76	-4.38
cat	92.67	90.67	-2.01
dog	78.35	76.64	-1.71
horse	62.32	61.33	-0.98
sheep	53.27	33.45	-19.83
cow	51.67	48.54	-3.13
bottle	22.32	19.65	-2.68
chair	43.78	41.89	-1.89
couch	73.42	69.32	-4.10
potted plant	30.71	33.69	$+2.98$
dining table	59.94	51.42	-8.52
tv	60.40	60.62	$+0.22$