LEARNING WHEN TO TRUST GEOMETRY: PROBE-RELATIVE SOURCE PREFERENCE FOR CAMOUFLAGED OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive fusion assumes a gate can learn where one representation is more trustworthy than another. We test that assumption. We supervise a spatial gate with the detached pixelwise loss difference between two auxiliary segmentation probes—the exact minimizer of an entropy-regularized two-source risk—and instantiate it in CLOAK, a camouflaged object detector fusing frozen SAM 2.1 semantics with frozen Depth Anything V2 geometry. CLOAK is strong: after SHA-256 auditing removes seven contaminated training images, three seeds reach S_α of 0.9402, 0.9174, 0.9264, and 0.9327 on CHAMELEON, CAMO, COD10K, and NC4K, and matched-protocol inference beats ZoomNeXt-B4 on all four datasets with non-overlapping paired bootstrap intervals in S_α, F_β^w, and E_ϕ. The gate nevertheless fails to identify trustworthy geometry: AUROC for the probe-preference event is 0.6057–0.6429, and as a selector it incurs 2.8–4.5$\times$ the regret of the trivial rule "always trust geometry." We trace this to the target, not the gate. Because both probes are accurate, $|e_s - e_g| \leq 0.05$ at 75.9–81.9% of pixels, which at $\tau = 0.25$ confines the target to $[0.450, 0.550]$ and leaves the gate a near-constant signal. Three instruments confirm the mechanism: the margin distribution; a five-condition corruption ladder in which mean gate response is perfectly rank-correlated with induced damage ($\rho = -1.00$ on every dataset) and therefore blind wherever damage is small; and oracles showing 25.4–33.7% of output-space complementarity unclaimed. Margin-based competence supervision degenerates precisely as its experts co-improve—a general obstacle for loss-difference routing, not an artifact of one architecture.

1 INTRODUCTION

Camouflaged object detection segments foreground that matches its surroundings in color, texture, and structure. Monocular depth estimated from the same RGB image supplies a complementary view: inferred surfaces and discontinuities can expose an object that appearance conceals. Depth also misleads, most often under the conditions camouflage creates—flat textured scenes, occlusion, thin structures. Useful geometric evidence therefore varies spatially, and monotonically increasing geometry's contribution is the wrong fusion rule.

The standard response is a learned gate. That response presumes something rarely tested: that a gate can be taught *where* one source deserves more weight. We make the presumption explicit, supervise it as directly as possible, and measure whether it holds.

We formulate the decision as *probe-relative source preference*. Two auxiliary predictors estimate foreground from semantic and geometric features, and their detached pixelwise loss difference supervises a gate that interpolates between the representations. The target asks which probe currently segments the training mask better; it needs no depth ground truth and never treats foreground membership as a reliability label. We show it is the unique minimizer of an entropy-regularized two-source risk, which fixes its sign and temperature rather than leaving them heuristic.

The formulation separates three questions the literature conflates: whether combining representations improves segmentation, whether this target improves fusion over alternatives, and whether the resulting gate distinguishes useful from harmful geometry. Our answers differ. Geometry improves

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

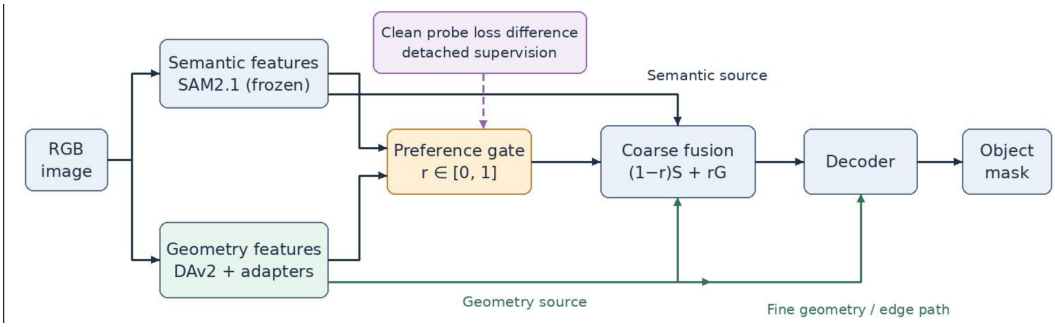

Figure 1: The preference-learning mechanism. Clean semantic and geometry probes supply a detached relative-loss target. Training corrupts a separate geometry copy for fusion. Fine geometry and edge conditioning also reach the decoder; Appendix **??** makes these paths explicit.

the semantic-only system by $+0.037$ to $+0.065$ S_α, and CLOAK is competitive with published geometry-assisted detectors under matched inference. The target confers no segmentation advantage over mask supervision or cross-attention. And on all 6,473 test images the gate is a poor selector: AUROC 0.6057–0.6429 for the probe-preference event, worse than the constant rule "always trust geometry."

We then explain the failure rather than reporting it, and the explanation is arithmetic on the target. When both probes are accurate their losses nearly coincide, so the loss margin—the entire supervision signal—collapses. At $\tau = 0.25$ the 75.9–81.9% of pixels with $|e_s - e_g| \leq 0.05$ receive targets in $[0.450, 0.550]$, indistinguishable in practice from a constant 0.5. The gate learns what it was taught: its clean mean is 0.485–0.493 although geometry helps in 89.2–91.7% of images. A corruption ladder confirms the mechanism from the other side—response is monotone in induced damage on every dataset ($\rho = -1.00$ over five conditions), rejecting shuffled and zeroed geometry decisively while staying wholly inert to noise that leaves segmentation error unchanged. The gate is not undertrained or undersized; it correctly fits a target carrying almost no information in the regime where the system succeeds. This generalizes: relative-competence signals—loss differences, per-expert error, validation regret—are natural supervision for routing in mixtures of experts and multimodal fusion, and all inherit a signal that vanishes as experts co-improve. Demonstrating this requires a strong system, since a weak one's routing failure would be uninformative.

- **A margin-collapse account of relative-competence routing supervision.** We derive the probe-relative target as the exact minimizer of an entropy-regularized two-source risk, then show its supervision content is governed entirely by the probe loss margin: at $\tau = 0.25$, the 75.9–81.9% of pixels with $|e_s - e_g| \leq 0.05$ receive targets inside $[0.450, 0.550]$. Margin-based competence targets degenerate as experts co-improve—a limitation of the objective family, not of one gate.
- **Instruments separating segmentation accuracy from trustworthy selection.** On all 6,473 test images, a gate inside a state-of-the-art detector attains AUROC 0.6057–0.6429, incurs 2.8–4.5$\times$ the selection regret of "always trust geometry," and leaves 25.4–33.7% of output-space complementarity unclaimed. A five-condition corruption ladder shows gate response perfectly rank-correlated with induced damage ($\rho = -1.00$, every dataset), pinning the blindness to small margins rather than capacity.
- **An audited, matched-protocol detector that makes the diagnosis credible.** SHA-256 byte auditing removes seven contaminated training images, including renamed duplicates, before any split. Three seeds reach S_α 0.9402/0.9174/0.9264/0.9327, and identical-protocol inference beats ZoomNeXt-B4 on all four datasets with non-overlapping paired intervals in S_α, F_β^w, and E_ϕ. The gate's failure is reported inside a system whose accuracy is not in question.

2 RELATED WORK

Geometry-assisted camouflage detection. Source-free Depth for Object Pop-out adapts pre-trained depth through mask supervision and a contact-surface representation (Wu et al., 2023). DaCOD learns complementary appearance and depth features through asymmetric cross-modal fusion (Wang et al., 2023). DSAM transfers depth into SAM by distillation and bias correction (Yu et al., 2024). DepthSAM, a distinct method, combines sparse adaptation with geometric–semantic fusion (Han et al., 2026). ZoomNeXt instead aggregates appearance across scales (Pang et al., 2024). These methods learn *how* to combine sources and evaluate the combination by segmentation accuracy; we isolate the prior question of whether the combining mechanism knows *where* each source is trustworthy, and show segmentation accuracy does not answer it.

Gating, competence signals, and calibration. Input-dependent expert allocation dates to adaptive mixtures (Jacobs et al., 1991); gating is not our contribution. What supervises the gate is. Loss- and competence-based signals are attractive because they need no auxiliary ground truth, and our analysis bounds what they can teach: their informativeness is the probe loss margin, which shrinks as experts improve. Diagnostic probes measure information accessible in a representation (Alain & Bengio, 2017); ours are jointly trained and additionally *supply* the signal, so probe construction partly determines what preference means—a confound we quantify rather than assume away. Calibration concerns agreement between probabilities and events (Guo et al., 2017) and is distinct from both ranking and segmentation quality; we keep the three separate throughout.

3 LEARNING PROBE-RELATIVE PREFERENCE

3.1 A TASK-LOSS TARGET FOR SOURCE SELECTION

Frozen SAM 2.1 and Depth Anything V2 supply semantic and geometry features from RGB (Ravi et al., 2024; Yang et al., 2024). Trainable geometry adapters and projections produce aligned features $S_\ell, G_\ell \in \mathbb{R}^{192\times H_\ell \times W_\ell}$. Clean-source probes output logits z_s, z_g, giving unreduced pixelwise losses $e_b = \mathrm{BCEWithLogits}(z_b, Y)$ for source $b \in \{s, g\}$ and training mask Y.

For finite losses and $\tau > 0$, the unique minimizer of the entropy-regularized two-source risk is

$$\begin{aligned} q^* &= \arg\min_{q\in[0,1]} \{(1-q)e_s + qe_g + \tau[q\log q + (1-q)\log(1-q)]\} \\ &= \sigma((e_s - e_g)/\tau), \qquad r^* = \mathrm{stopgrad}(vq^*), \quad \tau = 0.25. \end{aligned} \tag{1}$$

Here v flags whether the fusion geometry is valid; $v = 0$ imposes rejection, $r^* = 0$. The identity fixes the target's direction and temperature; Appendix **??** gives the derivation. Detachment prevents the gate objective from inflating probe losses to manufacture an easier target.

The target is conditional on learned probes and their current representations. It is neither a depth-correctness probability nor a causal effect of geometry on the decoder. Even under BCE-only population regression the optimum is $\mathbb{E}[r^* \mid X]$ for gate inputs X, which generally differs from $\Pr(e_g < e_s \mid X)$; we therefore evaluate the soft target and the hard event separately.

3.2 MARGIN COLLAPSE: HOW MUCH CAN THIS TARGET TEACH?

Equation 1 makes the target a deterministic function of the loss margin $\Delta = e_s - e_g$, so the supervision is exactly as informative as Δ is large. Inverting the sigmoid bounds the band a margin threshold δ admits:

$$|\Delta| \le \delta \implies r^* \in [\sigma(-\delta/\tau),\, \sigma(\delta/\tau)]\,, \qquad \delta = 0.05,\ \tau = 0.25:\ r^* \in [0.450, 0.550]. \tag{2}$$

This is adversarial to the method's own premise. Two well-trained probes rarely disagree by much, so most pixels fall inside the band: Section 4.5 measures $|\Delta| \le 0.05$ at 75.9–81.9% of pixels. Four fifths of the supervision sits within ± 0.05 of the uninformative midpoint, and the residual fifth carries whatever discriminative content exists.

Two falsifiable predictions follow. The gate should output values near 0.5 on clean geometry regardless of whether geometry helps—confirmed in Section 4.4 (clean mean 0.485–0.493 against an

89.2–91.7% helpful rate). And gate response should scale with the margin a perturbation induces, making it sensitive to gross corruption and blind to subtle harm—confirmed in Section 4.5. Neither invokes gate capacity, and neither is specific to depth: any relative-competence target inherits Equation 2 with its own τ.

3.3 Corruption and fusion computation graph

We retain clean G for the geometry probe and construct a separate fusion copy $(\widetilde{G}, \widetilde{E})$, where E is a normalized Sobel map of predicted depth. Independent image-level shuffle and dropout events have probabilities 0.15 and 0.10. Shuffling replaces all geometry levels and edges jointly with another batch image; dropout zeros them and takes precedence. Clean, shuffle-only, and zeroed copies therefore occur with probabilities 0.765, 0.135, and 0.100. Any corruption sets $v = 0$. RGB, labels, and the clean probes stay unchanged, so the target itself is never corrupted. Inference uses uncorrupted geometry.

To measure feature disagreement we normalize across channels *at each spatial location*:

$$N(X)_{c,h,w} = \frac{X_{c,h,w}}{\max\{(\sum_j X_{j,h,w}^2)^{1/2}, 10^{-12}\}}, \qquad A_\ell = |N(S_\ell) - N(\widetilde{G}_\ell)|. \tag{3}$$

This applies only to A_ℓ; fusion uses the unnormalized projected features. At each of two coarse levels,

$$\begin{aligned} r_\ell &= \sigma\big(h_\ell[S_\ell, \widetilde{G}_\ell, A_\ell, \widetilde{E}_\ell, U_\ell]\big), \\ Z_\ell &= (1 - r_\ell) \odot S_\ell + r_\ell \odot \widetilde{G}_\ell, \qquad F_\ell = Z_\ell + \phi_\ell(Z_\ell). \end{aligned} \tag{4}$$

U_ℓ is a learned semantic uncertainty feature and carries no separate calibration claim. The gate applies a 3×3 convolution to 96 channels, group normalization, GELU, and a 1×1 sigmoid projection. Finer geometry features and edge conditioning also reach the decoder, so zero coarse gates do not eliminate every geometry path—a fact we invoke when interpreting incomplete rejection.

3.4 Training

The gate loss compares bilinearly upsampled coarse gates with the full-resolution target:

$$\mathcal{L}_{\text{rel}} = \tfrac{1}{2} \sum_{\ell=1}^{2} \text{BCE}(\text{up}(r_\ell), r^*), \qquad \mathcal{L} = \mathcal{L}_{\text{seg}} + 0.20\mathcal{L}_{\text{rel}} + 0.20\mathcal{L}_{\text{probe}} + \mathcal{L}_{\text{aux}}. \tag{5}$$

The probe term averages the two probes' structure losses. Main segmentation combines weighted BCE/IoU and Tversky terms; boundary, distance, intermediate-mask, entropy, and routing terms support optimization. The gate also receives segmentation gradients. Appendix **??** records weights and decoder details. Sparse adapters, skip alignment, and high-frequency refinement are supporting choices whose necessity our controls do not establish, and we claim none of them as contributions.

4 Experiments

4.1 Audited data, protocols, and uncertainty

Before splitting 4,040 training pairs, we audit image stems and SHA-256 hashes of image-file bytes against all test sets; hashing raw bytes detects renamed duplicates that stem matching misses. Seven contaminated images are removed, leaving 4,033 pairs yielding 3,832 training and 201 validation images (Appendix **??**). Test images never inform checkpoint selection or calibration fitting.

We evaluate every CHAMELEON (76), CAMO (250), COD10K (2,026), and NC4K (4,121) image (Skurowski et al., 2017; Le et al., 2019; Fan et al., 2020; Lv et al., 2021). Training runs 300 epochs with AdamW, effective batch size eight, 512-pixel RGB and 518-pixel geometry input. The geometry-only control uses FP32 after a numerical restart; other source controls use mixed precision. Seed variation spans the validation partition as well as optimization randomness.

Primary inference uses one nominal scale, no flips, and no explicit post-inference min–max normalization; the shared COD metric backend separately normalizes nonconstant masks per image. We

Table 1: Standard inference, mean ± sample SD across three training seeds. CA denotes the matched cross-attention control.

Model	Dataset	MAE ↓	S ↑	Fw ↑	Emean ↑
CLOAK	CHAMELEON	0.0158 ± 0.0002	0.9402 ± 0.0017	0.9031 ± 0.0069	0.9694 ± 0.0057
CLOAK	CAMO	0.0302 ± 0.0003	0.9174 ± 0.0009	0.8882 ± 0.0010	0.9516 ± 0.0008
CLOAK	COD10K	0.0134 ± 0.0005	0.9264 ± 0.0011	0.8774 ± 0.0024	0.9631 ± 0.0013
CLOAK	NC4K	0.0203 ± 0.0003	0.9327 ± 0.0001	0.9038 ± 0.0015	0.9611 ± 0.0009
CA	CHAMELEON	0.0163 ± 0.0005	0.9379 ± 0.0024	0.9003 ± 0.0089	0.9689 ± 0.0044
CA	CAMO	0.0294 ± 0.0016	0.9181 ± 0.0019	0.8905 ± 0.0055	0.9534 ± 0.0029
CA	COD10K	0.0133 ± 0.0002	0.9268 ± 0.0015	0.8798 ± 0.0023	0.9652 ± 0.0016
CA	NC4K	0.0204 ± 0.0004	0.9305 ± 0.0037	0.9024 ± 0.0013	0.9607 ± 0.0007

Table 2: Fresh common-protocol inference on all 6,473 images per model: nominal 512-pixel RGB, FP16, no extra test augmentation, and identical native-resolution output/scoring. Both models retain their intrinsic computation. Training and pretrained sources remain unmatched.

Dataset	Model	S ↑	Fw ↑	MAE ↓	Emean ↑
CHAMELEON	CLOAK	0.9398	0.9049	0.0159	0.9724
CHAMELEON	ZoomNeXt-B4	0.9258	0.9022	0.0174	0.9624
CAMO	CLOAK	0.9174	0.8870	0.0305	0.9514
CAMO	ZoomNeXt-B4	0.8810	0.8523	0.0438	0.9333
COD10K	CLOAK	0.9269	0.8774	0.0134	0.9634
COD10K	ZoomNeXt-B4	0.9036	0.8490	0.0167	0.9483
NC4K	CLOAK	0.9327	0.9030	0.0205	0.9605
NC4K	ZoomNeXt-B4	0.8981	0.8628	0.0280	0.9405

report MAE, S_α, F_β^w, and mean E_ϕ (Fan et al., 2017; Margolin et al., 2014; Fan et al., 2018), at native ground-truth resolution for main metrics and 512×512 for gate diagnostics. Three seeds (3407, 1337, 2027) support full-model and cross-attention mean ± sample SD; twelve further controls have one seed. Paired image-bootstrap intervals use 10,000 resamples and quantify test-image uncertainty, not training variance.

4.2 The detector is competitive, so its gate failure is diagnostic

Full fusion exceeds semantic-only S_α by 0.0370, 0.0512, 0.0651, and 0.0457 at seed 3407 and improves geometry-only F_β^w on every dataset (Table 1). Cross-attention edges ahead in mean S_α on CAMO and COD10K while CLOAK leads on CHAMELEON and NC4K; we draw no superiority claim from these gaps.

We rerun CLOAK and the official ZoomNeXt-B4 checkpoint on all four test sets under one protocol: bicubic 512-pixel RGB resize, FP16 inference, sigmoid, native-size probability resize, uint8 export, and a single evaluator (Table 2). There are no flips and no extra base scales. Each architecture retains its intrinsic computation—ZoomNeXt its 256/512/768 pyramid, CLOAK its internal 518-pixel depth input. Published/native-protocol results and shared scoring of the original released masks stay separate in Appendix **??**.

Table 3 measures both models on the same physical H200. CLOAK uses 8.81 times the total parameters and 1.12 times the forward latency under this protocol; freezing reduces trainable parameters, not the inference footprint. We report this cost plainly: the diagnosis below concerns what the gate learns, not efficiency. DepthSAM-L checkpoint inference and timing remain unavailable, and no external baseline has been retrained on the audited CLOAK split. Under identical 512-pixel inference, CLOAK improves on ZoomNeXt-B4 across all four datasets, with paired intervals excluding zero for S_α ($+0.014$ to $+0.036$), F_β^w on three datasets, and mean E_ϕ on all four (Table **??**); MAE intervals exclude zero on CAMO, COD10K, and NC4K and narrowly cross it on CHAMELEON's 76 images. Published-protocol context places CLOAK at or above DepthSAM-L on the three datasets it reports (Table **??**). This accuracy is not the paper's claim but its precondition: a routing diagnosis inside an uncompetitive detector would be attributable to the detector.

Table 3: Same-H200 forward timing: batch one, FP16 AMP, ten warmups, 50 timed forwards of one fixed image; mean ± iteration SD. Includes internal geometry/pyramid computation; excludes input/output processing and transfers. *DepthSAM-L counts the official source architecture without a verified trained checkpoint; its latency is unmeasured.

Model	Total M	Trainable M	ms/image	Peak MiB
CLOAK	576.249	16.487	60.13 ± 2.95	3388.6
ZoomNeXt-B4	65.374	65.364	53.66 ± 3.09	624.1
DepthSAM-L*	362.837	27.521	Pending	Pending

Table 4: Consolidated controls, seed 3407. Dataset columns report S-measure; macro values weight datasets equally. These rows do not have three-seed uncertainty.

Control	CHAM	CAMO	COD10K	NC4K	Macro Fw	Macro MAE
Semantic only	0.9028	0.8662	0.8617	0.8870	0.8091	0.0374
Geometry only	0.9395	0.9157	0.9242	0.9305	0.8795	0.0234
Ungated fusion	0.9357	0.9097	0.9217	0.9230	0.8926	0.0201
Mask target	0.9398	0.9179	0.9260	0.9315	0.8941	0.0197
No corruption	0.9370	0.9143	0.9250	0.9270	0.8941	0.0195
Shuffle only	0.9382	0.9158	0.9259	0.9283	0.8948	0.0197
Dropout only	0.9382	0.9185	0.9264	0.9337	0.8911	0.0200
Full model	0.9398	0.9174	0.9268	0.9327	0.8931	0.0201

4.3 Controlled source and fusion ablations

Table 4 consolidates source, fusion, target, and corruption controls at seed 3407. Three findings matter, including one that cuts against the method.

Geometry dominates, and fusion buys boundary quality rather than S_α. Geometry-only reaches macro S_α 0.92747 against the full model's 0.92917, a gap within seed noise. Fusion's real gain lies elsewhere: macro F_β^w rises 0.8795 to 0.8931 and macro MAE falls 0.0234 to 0.0201, a 14.1% relative reduction. Complementarity is established in boundary-sensitive metrics, not in S_α, and we state it that way.

Gating helps S_α but not MAE. Gated fusion exceeds equal-weight fusion in S_α on all four datasets, yet macro MAEs are indistinguishable (0.02008 versus 0.02014).

The relative-loss target confers no segmentation advantage. Mask supervision attains macro MAE 0.01968 and F_β^w 0.89410 against 0.02008 and 0.89308 for the relative-loss target, and removing corruption lowers macro MAE to 0.01949 while lowering macro S_α to 0.92581. Corruption and target choice have metric-dependent effects with no consistent winner. These controls are single-seed and we do not over-read them; the load-bearing result below does not require the target to win.

4.4 The gate does not learn when to trust geometry

On all 6,473 images, the gate's AUROC for $e_g < e_s$ is 0.6057–0.6429 (Table 5). Helpful-minus-nonhelpful mean-gate gaps are 0.0260–0.0374 on a unit scale—about 3% of the gate's range separating the two classes it exists to distinguish. The soft target correlates with the decoder's zero-geometry BCE degradation ($r = 0.5524$–0.6238), yet the gate discriminates that intervention's benefit barely above chance (AUROC 0.5448–0.5923, Table **??**).

The clean mean gate is 0.485–0.493 (Table 6) while geometry benefits 89.2–91.7% of images. A gate that had learned trust would sit well above 0.5; this one sits at the midpoint, nearly independent of an event with a $\sim$90% base rate. Soft-target BCE also exceeds the constant-0.5 value $\log 2 = 0.6931$, so miscalibration to the hard event does not explain the weak fit—the gate tracks the soft target poorly too.

Table 5: Archived full-test gate diagnosis, seed 3407. Gap is helpful-minus-nonhelpful mean gate; r correlates probe target with decoder zero-geometry BCE degradation.

Dataset	n	AUROC ↑	ECE ↓	Gate gap	Pearson r
CHAMELEON	76	0.6395	0.2463	0.0302	0.5524
CAMO	250	0.6057	0.1902	0.0260	0.6238
COD10K	2026	0.6273	0.2465	0.0307	0.5581
NC4K	4121	0.6429	0.1181	0.0374	0.5838

Table 6: Full-test stress evaluation. Each cell reports mean gate / segmentation MAE at 512×512.

Condition	CHAMELEON	CAMO	COD10K	NC4K
clean	0.485 / 0.0157	0.493 / 0.0304	0.488 / 0.0132	0.488 / 0.0204
shuffle	0.201 / 0.0460	0.207 / 0.0705	0.193 / 0.0412	0.190 / 0.0480
zero	0.022 / 0.0478	0.026 / 0.0723	0.021 / 0.0414	0.028 / 0.0487
noise	0.485 / 0.0157	0.492 / 0.0305	0.487 / 0.0132	0.487 / 0.0204
blur	0.448 / 0.0175	0.455 / 0.0313	0.446 / 0.0145	0.440 / 0.0221
misalign	0.325 / 0.0412	0.320 / 0.0572	0.312 / 0.0332	0.308 / 0.0399

Section 3.2 predicts this. Three architectural confounds remain worth separating from it: the semantic probe averages two resized logit maps and the geometry probe four, entangling probe construction with representation quality; fine geometry bypasses the coarse gates; and gate capacity may contribute, though no ablation isolates it. None explains why the gate is inert to perturbations that leave segmentation error unchanged, which we show next.

4.5 Margin collapse explains the failure

The margin distribution. Across the full test sets, $|e_s - e_g| \leq 0.05$ at 75.9–81.9% of pixels. By Equation 2 those pixels carry targets in $[0.450, 0.550]$. Supervision is near-constant over four fifths of the signal, which accounts quantitatively for a gate whose clean mean is 0.488 and whose class separation is 0.03.

The corruption ladder. Table 6 stress-tests all 6,473 images under feature noise, blur, and spatial misalignment alongside clean, shuffled, and zeroed geometry (settings and intervals in Appendix **??**). Gate response is *perfectly rank-correlated with the damage each condition inflicts* on every dataset (Spearman $\rho = -1.00$ over five perturbations). On NC4K, noise leaves MAE unchanged at 0.0204 and the gate unchanged at 0.487; blur costs $+0.0017$ MAE and gives 0.440; misalignment costs $+0.0195$ and gives 0.308; shuffling costs $+0.0276$ and gives 0.190; zeroing costs $+0.0283$ and gives 0.028. The other datasets follow the same ordering.

This is the mechanism observed directly. The gate responds in proportion to the loss margin a perturbation creates, exactly as Equation 1 instructs: it rejects gross corruption because gross corruption produces a large margin, and is blind to subtle harm because subtle harm produces none. Natural geometry failures live at the blind end of this ladder. Reported alone, "strong rejection of shuffled geometry" would flatter the method; read against the ladder it is evidence for the limitation. This evaluates one severity per unseen corruption and does not establish severity-independent robustness.

Calibration separates confidence from ranking. We fit $c(r) = \sigma(a \operatorname{logit}(r) + b)$ with $a > 0$ on 1,024 pixels sampled uniformly without replacement from each of the original 201 validation images (seed 3407), then freeze it before any test evaluation; the checkpoint and decoder stay unchanged. Table 7 reports lower ECE and Brier error on CHAMELEON, CAMO, and COD10K. On NC4K, ECE rises from 0.1181 to 0.1512 and Brier from 0.2406 to 0.2426. The fitted map is almost purely an intercept shift ($a = 0.995$, $b = 1.025$), and its transfer failure is consistent with NC4K's lower helpful-event prevalence. Because a strictly increasing transformation preserves AUROC exactly, calibration changes how diagnostic scores are read and nothing about the ranking or the trained routing mechanism. Post-hoc calibration cannot repair a signal that was never discriminative.

Table 7: Validation-fitted post-hoc diagnostic calibration, unchanged seed-3407 checkpoint. ECE uses 256 bins; Brier targets the hard probe-helpful event. Ranking is preserved.

Dataset	ECE raw	ECE cal.	Brier raw	Brier cal.
CHAMELEON	0.2463	0.0659	0.2496	0.1905
CAMO	0.1902	0.1005	0.2467	0.2137
COD10K	0.2466	0.0779	0.2490	0.1897
NC4K	0.1181	0.1512	0.2406	0.2426

Table 8: Internal probes from the same unchanged checkpoint: raw sigmoid-probability MAE at 512×512, without min–max normalization. Hard selection thresholds the original averaged gate at 0.5. Regret is selected loss minus pixel-oracle loss. These are probe-output selectors, not oracle feature-fusion forwards.

Dataset	Semantic	Geometry	Pixel oracle	Gate select	Regret
CHAMELEON	0.07208	0.01744	0.01063	0.04154	0.03091
CAMO	0.09399	0.03210	0.02107	0.05506	0.03399
COD10K	0.06618	0.01493	0.00971	0.03142	0.02171
NC4K	0.06748	0.02171	0.01318	0.03728	0.02410

4.6 Selection headroom and the cost of a bad gate

For source predictions p_s, p_g, a pixel oracle takes the lower ground-truth loss at each pixel and an image oracle commits one source per image. For MAE or BCE (clipped for quantized saved masks),

$$
\begin{aligned}
L_{\text{pix}} &= \frac{1}{N} \sum_i \frac{1}{|\Omega_i|} \sum_{x \in \Omega_i} \min_{b \in \{s,g\}} \ell(p_b(i,x), Y(i,x)), \\
L_{\text{img}} &= \frac{1}{N} \sum_i \min_{b \in \{s,g\}} \frac{1}{|\Omega_i|} \sum_{x \in \Omega_i} \ell(p_b(i,x), Y(i,x)).
\end{aligned}
\tag{6}
$$

Table 8 prices the gate's decisions. Using probes from the same unchanged checkpoint, hard gate selection incurs MAE regret 0.03091/0.03399/0.02171/0.02410 against the pixel oracle, while the constant rule "always take the geometry probe" incurs 0.00681/0.01103/0.00522/0.00853—the learned gate is 2.8–4.5$\times$ worse than ignoring the gate entirely, and always-geometry also beats both original and calibrated hard selectors on every dataset. A learned selector that loses to a constant is the sharpest available statement of the diagnosis, and the comparison rule uses no ground truth.

Headroom persists in output space (Table **??**). For independently trained source-only models, pixel selection lowers raw-mask MAE relative to CLOAK by 25.4%, 25.9%, 27.7%, and 33.7% with paired intervals positive on all four datasets (Appendix **??**); whole-image selection is worse than CLOAK on all four, so the unclaimed complementarity is genuinely sub-image. These selectors consult ground truth and bound selection between the two specified outputs, not every decoder that could synthesize a different value. One implication deserves stating: CLOAK's fusion gains come from the decoder's feature combination, not from the gate's decisions, since the gate is a worse selector than a constant.

4.7 Qualitative failures

At the fixed descriptive threshold raw MAE ≥ 0.1, severe-case counts are 1/2, 12/22, 27/39, and 109/187 for CLOAK/ZoomNeXt; NC4K's 109 CLOAK failures split 54 false-positive-dominant and 55 false-negative-dominant. Fewer severe cases coexist with a worse maximum error (0.922 versus 0.705).

NC4K 1004 shows the diagnosis in a single image. CLOAK predicts 96.26% foreground against 2.71% ground truth, with 93.65% of pixels false positive. Its mean target is 0.0386—the probes agree geometry is failing—yet the mean gate is 0.5808, and deleting geometry improves final BCE by 4.3218. The supervision was correct and the gate ignored it, which is what a signal compressed into $[0.450, 0.550]$ over most of training produces. Cases 1654 and 3402 instead lose foreground ex-

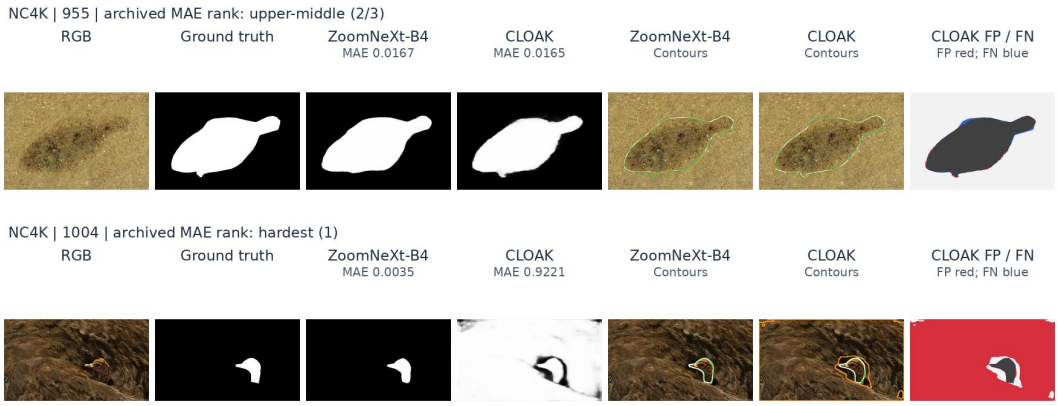

Figure 2: Fresh CLOAK and ZoomNeXt predictions under the common 512-pixel protocol. Image IDs are fixed from the archived upper-middle CLOAK-MAE selection for the first three datasets and the severe NC4K failure 1004. Green/orange contours denote ground truth/prediction, with white overlap; red/blue errors denote false positives/negatives. Appendix **??** retains additional archived-protocol examples and source decompositions.

tent; Appendix **??** reports source masks, boundaries, pixel-error decompositions, and within-model intervention statistics, kept separate from independently trained source-only models.

5 Discussion and Conclusion

We set out to supervise a model's local preference for geometry and found the limit of the approach. The probe-relative target is principled—an exact entropy-regularized selection optimum—and yields a competitive detector: audited three-seed results and matched-protocol comparisons establish that semantic and geometric representations combine usefully, chiefly in boundary-sensitive metrics. But the gate it trains does not learn when to trust geometry. It discriminates the preference event at AUROC 0.61–0.64, sits at 0.488 on clean geometry while geometry helps 90% of the time, and loses to a constant selector by 2.8–4.5$\times$ in regret.

The cause is structural. A loss-difference target carries information only in proportion to the loss margin, and accurate experts produce small margins: at $\tau = 0.25$, four fifths of pixels are supervised within ± 0.05 of 0.5. The corruption ladder confirms it by ordering gate response perfectly by induced damage, so the gate is blind exactly where margins vanish. We therefore expect relative-competence supervision to degrade as its experts co-improve, in mixtures of experts and multimodal fusion as much as here. Accurate segmentation is not evidence of trustworthy routing, and the two must be measured separately.

The remedies this implies are margin-side, not capacity-side: temperature schedules or margin normalization preserving gradient where $|\Delta|$ is small, loss reweighting toward informative pixels, deliberately heterogeneous experts that keep margins wide, and targets defined by decoder intervention effect rather than probe loss difference. Our limitations are unequal probe construction, geometry paths outside the gates, single-seed coverage for most controls, and unmatched external training; DepthSAM-L checkpoint inference remains open. We contribute a formulation that makes geometry preference supervisable, instruments that reveal when such supervision fails, and a quantitative reason to expect that failure whenever the experts being routed are already good.

AI USE STATEMENT

An AI agent is used in manuscript restructuring and editing. Every quantitative claim traces to a saved output and a reproducible evaluation script. The authors verified the scientific claims, references, and final submission and retain full responsibility for them.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (Workshop Track)*, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4548–4557, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Fan_Structure-Measure_A_New_ICCV_2017_paper.html.

Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 698–704, 2018. doi: 10.24963/ijcai.2018/97. URL https://www.ijcai.org/proceedings/2018/97.

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2777–2787, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Fan_Camouflaged_Object_Detection_CVPR_2020_paper.html.

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 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Jinyu Han, Changguang Wu, Fuming Sun, and Jinhui Tang. Beyond appearance: Camouflaged object detection via geometric structure. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25830–25840, 2026. URL https://depthsam.github.io/.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79. URL https://doi.org/10.1162/neco.1991.3.1.79.

Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding*, 184:45–56, 2019. doi: 10.1016/j.cviu.2019.04.006. URL https://doi.org/10.1016/j.cviu.2019.04.006.

Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/papers/Lv_Simultaneously_Localize_Segment_and_Rank_the_Camouflaged_Objects_CVPR_2021_paper.pdf.

Ran Margolin, Lihi Zelnik-Manor, and Ayelet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2014. URL https://openaccess.thecvf.com/content_cvpr_2014/html/Margolin_How_to_Evaluate_2014_CVPR_paper.html.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. ZoomNeXt: A unified collaborative pyramid network for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9205–9220, 2024. doi: 10.1109/TPAMI.2024.3417329. URL https://doi.org/10.1109/TPAMI.2024.3417329.

Nikhila Ravi et al. Sam 2: Segment anything in images and videos, 2024. URL https://arxiv.org/abs/2408.00714.

Przemysław Skurowski, Hassan Abdulameer, Jakub Błaszczyk, Tomasz Depta, Adam Kornacki, and Przemysław Kozieł. Animal camouflage analysis: CHAMELEON database. Silesian University of Technology, dataset webpage, 2017. URL https://www.polsl.pl/rau6/chameleon-database-animal-camouflage-analysis/. Accessed 25 September 2026.

Qingwei Wang, Jinyu Yang, Xiaosheng Yu, Fangyi Wang, Peng Chen, and Feng Zheng. Depth-aided camouflaged object detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3297–3306. Association for Computing Machinery, 2023. doi: 10.1145/3581783.3611874. URL https://doi.org/10.1145/3581783.3611874.

Zongwei Wu, Danda Pani Paudel, Deng-Ping Fan, Jingjing Wang, Shuo Wang, Cédric Demonceaux, Radu Timofte, and Luc Van Gool. Source-free depth for object pop-out. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1032–1042, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/WU_Source-free_Depth_for_Object_Pop-out_ICCV_2023_paper.html.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/26cfdcd8fe6fd75cc53e92963a656c58-Abstract-Conference.html.

Zhenni Yu, Xiaoqin Zhang, Li Zhao, Yi Bin, and Guobao Xiao. Exploring deeper! segment anything model with depth perception for camouflaged object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*. Association for Computing Machinery, 2024. doi: 10.1145/3664647.3681119. URL https://doi.org/10.1145/3664647.3681119.