SAM-RISE: SIGNED BLACK-BOX SALIENCY FOR VISION-LANGUAGE MODELS VIA SEMANTIC REGION ABLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The most capable vision-language models (VLMs) are closed and reachable only through APIs, which rules out interpretability tools that need gradients, attention maps, or activations. Black-box occlusion saliency, as in RISE and D-RISE, needs only the model's output, but it does not transfer to VLMs unchanged: a VLM reads a tight occlusion mask as an object in the scene (masking a cat's eyes yields "the cat has black eyes"), and probability-weighted masks saturate on a confident model. We introduce **SAM-RISE**, a black-box saliency method built around both failures. It removes a few dozen semantic regions from Segment Anything, each as a grey bounding box, scores the change in the unsaturated logit margin, and keeps both signs: *positive evidence*, regions whose removal lowers confidence in the answer, and *negative evidence*, regions whose removal raises it. When an API returns no token probabilities, sampled answers take their place. Semantic regions pay off in three ways. At the small query budgets a paid API imposes, they localise objects as well as random masks with 2 to $4\times$ fewer queries. Negative evidence is actionable: subtracting it corrects over-confident "yes" answers on NaturalBench for two model families. And because the same part can be matched across images, per-image maps aggregate into class-level maps that show whether one image's explanation is typical of its class, and that separate classes the model recognises from pixels from those it answers from language priors. We evaluate on MS-COCO, ImageNet, and NaturalBench, compare against white-box Grad-CAM and attention rollout and against human judgments, and run SAM-RISE on the closed Gemini model.

1 INTRODUCTION

Vision-language models (VLMs) are increasingly deployed as closed APIs. The models most users interact with, such as GPT-4o, Gemini, and Claude, expose no weights, gradients, attention maps, or activations, so the standard interpretability toolkit does not apply: Grad-CAM (Selvaraju et al., 2017) needs gradients, attention rollout (Abnar & Zuidema, 2020) needs attention weights, causal tracing and activation patching (Meng et al., 2022) need weight access, and TCAV (Kim et al., 2018) needs internal representations. Some APIs (Gemini 3, Claude) do not even return the probability of an answer token.

Black-box occlusion saliency looks like the natural fallback: RISE (Petsiuk et al., 2018) and its detection extension D-RISE (Petsiuk et al., 2021) perturb the input, measure the change in the output, and need nothing else. We find that it does not transfer to VLMs unchanged, for two reasons. First, the ablation operator misleads: a VLM reads a tight black silhouette not as missing content but as a black *object* in the scene, so the score barely moves and the region looks unimportant (Sec. 3.2). Second, the aggregation saturates: a confident VLM answers "yes" with probability 1.0 under almost every random mask, so D-RISE's probability-weighted sum collapses to a count of how often each pixel was kept (Sec. 3.3).

We introduce **SAM-RISE**, a black-box saliency method built around both failures (Sec. 4). It removes a few dozen semantic regions proposed by Segment Anything (Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2026), each as a grey bounding box that reads as occlusion, and scores the change in

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.

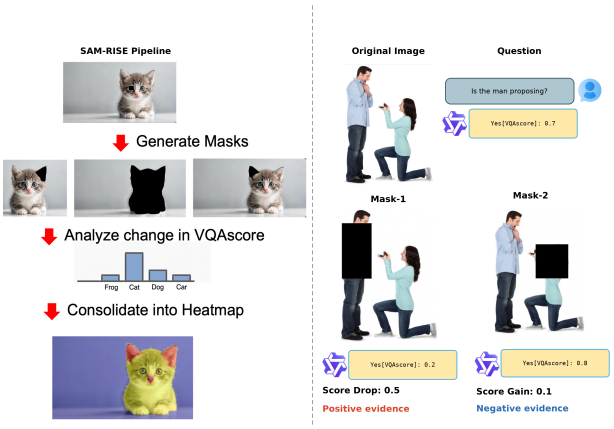

Figure 1: **Signed region ablation on a black-box VLM.** Left: SAM proposes semantic regions; each is ablated and the VLM re-queried; the per-region score changes form a signed heatmap. Right: asked "Is the man proposing?", masking the man lowers the score (*positive evidence*), while masking the woman *raises* it (*negative evidence*: content that suppresses the answer). Classical occlusion saliency keeps only the first signal; SAM-RISE keeps both, and the second corrects over-confident answers (Sec. 6.4).

the unsaturated logit margin. It keeps both signs of that change: *positive evidence*, regions whose removal lowers confidence in the answer, and *negative evidence*, regions whose removal raises it (Fig. 1), the signal behind the hallucinations (Li et al., 2023) and spurious correlations that adversarial benchmarks expose. Where an API hides token probabilities, the fraction of Yes among sampled answers takes their place, which lets us explain Gemini 3.8 Flash (Fig. 6).

Semantic regions pay off in three ways. At the small query budgets a paid API imposes, they localise objects as well as random masks with 2 to $4\times$ fewer queries, although random masks win given hundreds (Sec. 6.1). Negative evidence is actionable: subtracting it from the score corrects over-confident "yes" answers on NaturalBench for two model families (Sec. 6.4). And because the same part can be matched across images, per-image maps aggregate into class-level maps that test whether one image's explanation is typical of its class (Fig. 7) and separate classes the model recognises from pixels from those it answers from language priors (Sec. 6.5). Signed attributions themselves are not new (SHAP values and LIME weights are signed by construction (Lundberg & Lee, 2017; Ribeiro et al., 2016)); to our knowledge no VLM saliency method uses the score-gain half to correct the model's prediction or aggregates it across a class, and we scope our claims to that.

This paper makes five contributions. First, we diagnose two measurement failures of occlusion saliency on VLMs, mask re-interpretation and probability saturation, and derive a fix for each (Sec. 3). Second, we introduce SAM-RISE, which needs only the model's answers, and show that it runs on closed APIs that hide token probabilities (Secs. 4 and 6.3). Third, we benchmark it against D-RISE on MS-COCO and ImageNet under an identical model, prompt, and read-out, against white-box Grad-CAM and attention rollout, and against human judgments, and we state where it wins and where it does not (Secs. 6.1 and 6.2). Fourth, we show that negative evidence is actionable on two model families (Sec. 6.4). Finally, we lift per-image attribution to class-level population maps, which only semantic regions allow (Sec. 6.5).

2 Related Work

White-box VLM interpretability. Grad-CAM (Selvaraju et al., 2017) and gradient saliency (Simonyan et al., 2014; Sundararajan et al., 2017) need gradients; attention rollout (Abnar & Zuidema, 2020) needs every layer's attention weights. Causal tracing and activation patching (Meng et al., 2022), including VLM-specific interventions for hallucination (Wan et al., 2025), need to read and write activations, and TCAV (Kim et al., 2018) fits probes on them. None apply to closed APIs; we use Grad-CAM and rollout on an open model as reference points (Sec. 6.2).

Black-box saliency. RISE (Petsiuk et al., 2018) and D-RISE (Petsiuk et al., 2021) aggregate thousands of random binary masks; the maps do not follow object boundaries and the query cost is large against paid APIs. LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) fit signed local surrogates over superpixels, but do not feed the signed signal back into the prediction. PixelSHAP (Goldshmidt, 2025) estimates Shapley values over objects segmented by a text-prompted detector and SAM, and GLIMPSE (Shen, 2025) builds heatmaps from gradient-weighted attention,

which needs model internals. We differ in three ways: our regions come from prompt-free segmentation, our signed attribution re-enters the score, and our unit of analysis extends from images to class populations.

VLM evaluation. NaturalBench (Li et al., 2024) pairs two images with two questions whose yes/no answers alternate, defeating language-only shortcuts; we use it as our failure-case benchmark.

3 Occlusion Saliency Misleads VLMs

3.1 Black-Box Setup and the Per-Region Score Change

We assume API-level access to a VLM that, given an image I and a prompt q, returns the probability $p(y \mid I, q)$ of an answer token, summed over capitalisation and whitespace aliases. Following VQAScore (Lin et al., 2024), we score an answer by the probability of Yes, normalised over Yes and No: $s_0(I,q) = p(Yes \mid I,q)/[\,p(Yes \mid I,q) + p(No \mid I,q)\,] \in [0,1]$, with logit margin $m_0 = \sigma^{-1}(s_0)$. Modern VLMs saturate ($s_0 = 1.0$ in float32 on typical images), so every difference we compute is taken in margin space (App. A). For a region $R \subseteq I$, let $I \setminus R$ denote I with R ablated. The *per-region score change* is

$$\Delta_R \;:=\; m_0(I \setminus R, q) \;-\; m_0(I, q) \;\in\; \mathbb{R}. \tag{1}$$

$\Delta_R < 0$ means R supported the prediction (*positive evidence*, the only signal classical occlusion saliency keeps). $\Delta_R > 0$ means removing R *raised* the margin: content in R suppressed the answer (*negative evidence*).

3.2 VLMs Read Tight Masks as Objects, Not as Missing Content

The ablation operator inherited from classifier saliency overlays a tight black silhouette along the region boundary. On VLMs this fails silently. The model re-interprets the silhouette as a black object inside the scene: masking a cat's eyes with a tight outline prompts the description "the cat has black eyes," and the score drops by only 0.005 and 0.016 (Fig. 2, left). A saliency method built on this operator concludes that the eyes do not matter.

Two mundane explanations fail. The ragged geometry is not the cause: area-matched solid shapes leave the drops at outline level, on the probe image (-2.4 vs. -2.5 margin units) and at benchmark scale (-0.17 to -0.29 vs. -0.20, $\sim$88k ablations per shape). The fill value is not the cause either: nine constant fills from black to white move the mean drop by less than 0.04 with no trend (App. G). The fix follows from the diagnosis: present a shape that cannot be an object part. The axis-aligned bounding box of the region, filled with constant grey, reads as occlusion. On identical masks the box yields drops of 0.096 and 0.082 where the outline gave 0.005 and 0.016 (Fig. 2); at benchmark scale it doubles the deep-drop tail (5th percentile -4.3 vs. -2.2, 66k paired masks). The box also covers context beyond the region (median box-to-mask area ratio 1.55), but at matched box area the share of context does not predict the effect (Spearman $\rho_s = 0.03$), so the signal tracks the part inside the box. Photorealistic inpainting is no better: it restores the predictable content whose importance we want to measure, and "removing" a region has no ground-truth counterfactual (App. F).

3.3 Probability-Weighted Aggregation Saturates

D-RISE weights each random mask M_i by a detection similarity whose localisation and objectness terms have no analogue for a yes/no VLM; what survives is the class term, so the D-RISE map becomes $H = \sum_i P(\text{yes} \mid I \odot M_i)\, M_i$, i.e. RISE with the VLM as classifier (App. C). A confident VLM breaks this estimator. On several COCO images the standard deviation of $P(\text{yes})$ across $4{,}000$ masks is 0.0000 to four decimals, and the resulting map correlates at 0.9994 with the plain count of how often each pixel was kept. Its peak is then an artifact of RISE's random-crop step, not an explanation. Both remedies follow from Sec. 3.1: weight by the logit margin, which does not saturate and carries the same information ($P(\text{yes}) = \sigma(m_0)$), and subtract the mean weight, which removes the mask-count pedestal. We call this variant ***D-RISE (our adaptation)***: the same masks and weighted-sum aggregation, with each mask weighted by $m_0(I \odot M_i) - \frac{1}{N}\sum_j m_0(I \odot M_j)$ instead of $P(\text{yes} \mid I \odot M_i)$. It needs no additional queries, since both weights come from the same forward passes, and we use it as the strongest fair baseline; Sec. 6.1 measures both versions.

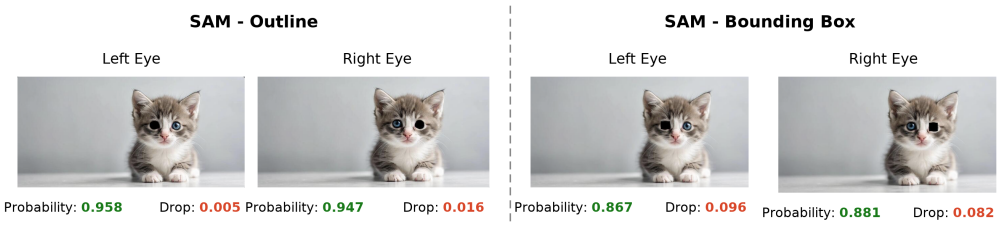

Figure 2: **Outline ablation under-reports; bounding-box ablation of the same masks restores the signal.** "Is this a cat?", two SAM masks on the eyes. The outline (left) reads as "the cat has black eyes" and moves the score by 0.005 and 0.016, so an outline-based method would conclude the eyes do not matter; the bounding box of the same masks (right) reads as occlusion and moves it by 0.096 and 0.082.

Finding. Occlusion saliency does not port to VLMs unchanged. Tight masks are read as objects (bounding boxes of the same masks move the score up to 19× more), and probability weights saturate (the published D-RISE map correlates 0.9994 with the mask-count field), so both the operator and the read-out must change.

4 The SAM-RISE Framework

The diagnosis sets the design. Ablations must read as occlusion (bounding boxes); scores must live in margin space; regions must be few, because each is a paid call, and nameable, because an attribution to an arbitrary blob explains nothing; and both signs of Δ_R must be kept.

Semantic region proposals. We prompt SAM 3 (Carion et al., 2026) in everything mode with a 32×32 point grid, keep masks that pass predicted-IoU, stability, and area filters, merge duplicates with IoU-based NMS, and cap $|\mathcal{M}|$ at 64. Because SAM segments whatever a point lands on, the masks mix whole objects with their parts (head, ears, paws; rotor, tail, cockpit). On NaturalBench this yields 31 proposals per image on average, and the per-image cost is $|\mathcal{M}|+1$ VLM calls (App. B). Fig. 3 contrasts these regions with DINOv2 clusters (Oquab et al., 2024) and uniform grids; Sec. 6.1 compares them with random masks and grids quantitatively, and App. H after area normalisation.

Signed, density-weighted saliency. SAM masks overlap (an ear lies inside the head mask and the whole-cat mask), so we visualise the per-pixel density-weighted aggregate $H(p) = \frac{1}{n(p)} \sum_{R \ni p} \Delta_R$, where $n(p)$ counts the covering masks, and split it by sign into positive and negative channels (Fig. 21 in App. N). The positive channel localises class-defining parts (the husky's ears, the eagle's head); the negative channel exposes cues that pull away from the class (the rock under a tabby cat); and the same region can switch sign with the question (Fig. 5).

Adjusted VQAscore from negative evidence. If ablating some region raised the margin by $\delta = \max(0, \max_{R\in\mathcal{M}} \Delta_R)$, the baseline contained δ units of misleading evidence, and a corrected score subtracts it:

$$\tilde{m}_0 \;=\; m_0 \;-\; w(\delta)\,\delta, \qquad \tilde{s}_0 \;=\; \sigma(\tilde{m}_0). \tag{2}$$

With $w \equiv 1$ the maximum over up to 64 masks subtracts noise; the proportional weight $w(\delta) = (\delta/(\max(m_0, 0) + \delta))^k$ leaves confident baselines almost untouched and recovers the full subtraction only when suppressive evidence rivals supporting evidence. All adjusted numbers use $k = 5$; both variants are compared in App. I.

APIs without token probabilities. Only m_0 touches the model, so any estimate of it suffices. We pose the same question $n = 20$ times at temperature 1, where the Yes fraction estimates the model's own $P(\text{yes})$, and set $\hat{m}_0 = \log[(n_{\text{yes}} + \frac{1}{2})/(n_{\text{no}} + \frac{1}{2})]$, a smoothed log-odds that stays finite when all answers agree; proposals, ablations, and heatmaps are unchanged (App. O).

Population statistics. A single map answers *where did the model look on this image?* To ask which parts the model grounds across a class, we take a query image I_q of class c, use its strongest positive-evidence regions as part anchors, and match each anchor π to the region $R_\pi(I)$ of every

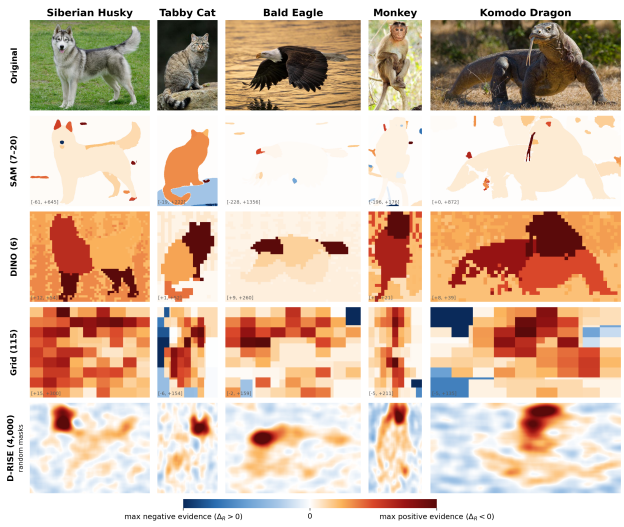

Figure 3: **Only semantic proposals attribute evidence to nameable parts.** Per-region score change for five ImageNet classes (red: positive, blue: negative evidence); row labels give each method's queries per image. SAM localises class-defining parts (the husky's ears, the eagle's head, the Komodo dragon's forked tongue) and marks the rock under the tabby as negative evidence. DINOv2 clusters cover whole foregrounds and grid cells cut across objects. D-RISE (our adaptation), which aggregates random masks, finds each head as a smooth blob, but even at 4,000 queries it cannot say which part carried the evidence; at SAM's budget its maps are noise (Fig. 8).

gallery image $I \in \mathcal{I}_c$ with the closest DINOv2 embedding (Oquab et al., 2024). The gallery-averaged drop $\bar{\Delta}_\pi(c) = \frac{1}{|\mathcal{I}_c|} \sum_{I \in \mathcal{I}_c} \Delta_{R_\pi(I)}$ measures how much part π contributes to recognising c across the dataset (formal procedure in App. K).

5 Experimental Setup

Models. Unless stated otherwise, results use Qwen3-VL-8B-Instruct, served with vLLM, greedy decoding, reading the top-20 log-probabilities of a single answer token. Grad-CAM and rollout additionally use the same weights in-process. We replicate the NaturalBench study on Gemma 4 (12B, instruction-tuned) and run the sampled estimator on Gemini 3.8 Flash, which returns no log-probabilities.

Saliency quality: MS-COCO and ImageNet. We build 500-image evaluation sets from MS-COCO val2017 (Lin et al., 2014) (masks and boxes) and ImageNet val (Deng et al., 2009) (boxes), keeping one target category per image whose ground truth covers 1 to 50% of the frame and that the model confirms ("Is there a {category} in the image?", $P(\text{yes}) \geq 0.5$). Every method explains the same predictions with the same prompt, 512 px JPEG input, and read-out; only the mask distribution differs. We report pointing-game accuracy (Zhang et al., 2018) (the saliency peak falls inside the ground-truth mask or box), insertion and deletion AUC on the per-image normalised margin, bootstrap 95% CIs, and exact McNemar tests on paired hits. Two controls need no queries: a random pixel, and the image centre. D-RISE uses its published mask generator with 4,000 seeded masks per image, whose first N form a valid N-mask run, so every budget is evaluated on the same queries (App. C).

White-box reference: NaturalBench. We compare SAM-RISE with Grad-CAM (last vision-encoder block, target: yes-minus-no logit) and attention rollout (all 36 decoder layers, answer-position row restricted to image tokens) on 10 NaturalBench images, each with both of its questions.

Negative evidence: NaturalBench. All 1,900 groups, every (I_i, q_j) baseline and every $(I_i \setminus R, q_j)$ ablation, for SAM, DINOv2 clusters, and 3×3, 5×5, 9×9 grids plus their unions. The Qwen release run has SAM and DINO masks for 534 groups (28%), corroborated by a full-coverage run on the first 720 (App. I); the Gemma 4 run covers all 1,900.

Population statistics: ImageNet. We rank all 1,000 classes by mean baseline VQAscore over 20 images each, and run the population pipeline on the top-5 and bottom-5 classes with galleries of up to 49 images, 5 anchors, and 10 regions per gallery image.

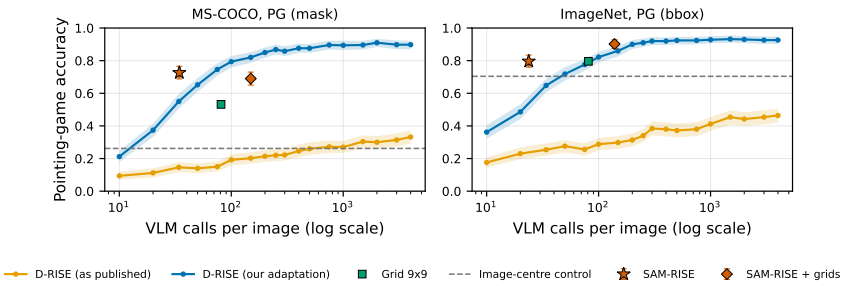

Figure 4: **At small budgets semantic proposals extract more per query; given hundreds of queries, random masks point more accurately but still cannot name a part.** Pointing-game accuracy vs. VLM calls per image (500 images each, shaded bands and bars are bootstrap 95% CIs). D-RISE as published (orange) stays near or below the image-centre control at every budget. SAM-RISE (star) sits above the D-RISE (our adaptation) curve (blue) at its own cost; that curve passes it at about 75 (COCO) and 100 (ImageNet) calls. On ImageNet the centre control scores 0.70, so pointing discriminates weakly there. Pointing rewards only the location of the peak: even at 4,000 masks the D-RISE map is a blob amid noise, while SAM-RISE's follows object parts (Fig. 3).

6 Results

6.1 Against D-RISE: Accuracy per Query

Fig. 4 and Table 3 (App. C) give the comparison. D-RISE as published is not a usable baseline on a confident VLM: at 4,000 queries per image it scores 0.444 PG (box) on COCO and 0.464 on ImageNet, barely above the image-centre control on COCO and far below it on ImageNet. This is the saturation of Sec. 3.3, so we compare against D-RISE (our adaptation).

At small budgets, semantic proposals extract more per query. At 34 calls on COCO, SAM-RISE reaches 0.726 PG (mask) against 0.550 for D-RISE (our adaptation) with the same number of calls (McNemar $p = 10^{-8}$); on ImageNet, 0.796 at 24 calls against 0.648 at 34 ($p = 10^{-7}$). D-RISE (our adaptation) needs about 2× the queries to match on COCO and 3.6 to 4.2× on ImageNet. On insertion, the pooled SAM-RISE + grids variant ties 4,000-mask D-RISE on both datasets (0.877 vs. 0.878 on COCO, 0.904 vs. 0.907 on ImageNet, overlapping CIs) with 27 to 29× fewer queries.

The advantage does not survive scaling. By 100 queries D-RISE (our adaptation) passes SAM-RISE's 34-call result, at 4,000 it reaches 0.898 PG (mask) on COCO against our 0.726, and deletion also favours the dense map. A map built from ∼30 regions ranks pixels only at region granularity, whereas thousands of random masks order individual pixels. Two choices here are conservative for SAM-RISE: it uses the mask-shaped occluder rather than the bounding box, and its pointing peak comes from the raw, not area-normalised, map (App. C).

Finding. D-RISE's published estimator fails on a confident VLM at every budget. Against D-RISE (our adaptation), SAM-RISE reaches the same pointing accuracy with 2 to 4× fewer queries and ties insertion faithfulness with 27× fewer, the regime a paid API imposes. With hundreds of queries random masks point more accurately, but their map stays a blob amid noise that names no part (Fig. 3).

6.2 Against White-Box Maps: Grad-CAM and Attention Rollout

On an open model, white-box maps are the natural reference. In the typical case of Fig. 5, Grad-CAM scatters over isolated patches, many in the background, and rollout concentrates on a few tokens at the border.

We test whether each map's peak is what the answer depends on, by occluding each method's top region and re-scoring. Over the 20 image-question pairs, occluding the SAM-RISE region changes the margin by a median of 22.3 logit units and flips the answer in 17 cases; occluding the top 10% of Grad-CAM or rollout pixels changes it by a median of 2.3 and 2.1 and flips one answer each. The comparison favours SAM-RISE by construction (examples selected by its effect size, regions chosen by the measurement that scores them, areas unmatched), so we read it qualitatively: white-box peaks

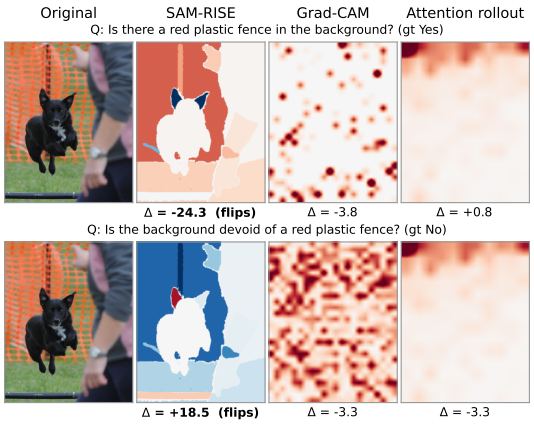

Figure 5: **Only SAM-RISE finds the region the answer depends on.** Rows: "Is there a red plastic fence in the background?" and its negation. Red marks regions that support the answer and blue regions that argue against it (white-box maps are red only). Below each map: the confidence change when its top region is occluded, bold if the answer flips. The fence is evidence for one question and against the other, so occluding it flips both; the white-box peaks flip neither. Over 20 pairs, SAM-RISE flips 17 answers, each white-box method one.

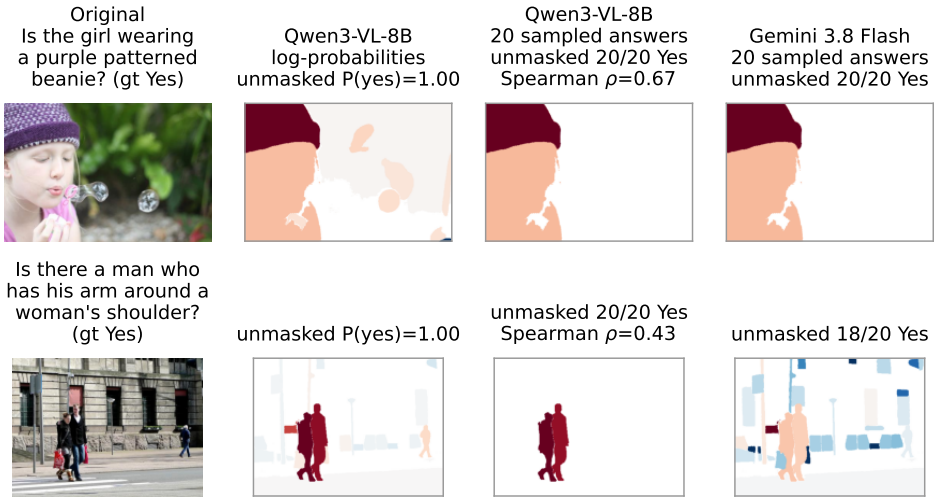

Figure 6: **Sampling answers matches log-probabilities where it matters, so SAM-RISE works on closed models such as Gemini.** The same SAM regions scored on Qwen3-VL-8B from log-probabilities and from 20 sampled answers, and on Gemini 3.8 Flash, which returns no log-probabilities, from 20 sampled answers (red: positive, blue: negative evidence). On Qwen, sampling keeps the region that flips the answer in all 16 test pairs, dropping only background that shifts the margin but not the answer (top). Gemini's maps pick the same regions (more in Fig. 23).

are rarely decisive for the answer. Both maps are also unsigned, so neither can express negative evidence.

Human preference. In a blinded plausibility user study, eleven raters picked, for 14 showcase pairs, the unlabelled heatmap that "best shows the area that actually explains the answer" (App. E): SAM-RISE in 76% of judgments against D-RISE (our adaptation), a 3×3 grid, and DINOv2 clusters (chance 25%), and in 98% against Grad-CAM and rollout (chance 33%).

6.3 Closed Models Without Log-Probabilities

Gemini 3 and Anthropic models return no token probabilities (OpenAI models return the top 20), so there SAM-RISE falls back to sampled answers, which we first test where log-probabilities give the truth. On 608 unmasked NaturalBench items, Qwen3-VL's sampled $P(\text{yes})$ agrees with its log-probability value at Pearson 0.996 and gives the same answer on 98.9% of them. On 16 image-question pairs, the region that moves the log-probability margin furthest toward the opposite answer is among the sampled map's top regions in all 16 (a set of 2 at the median). The full ranking agrees less (median Spearman 0.58): a region that flips all 20 answers scores the maximum however

Table 1: **Subtracting negative evidence improves NaturalBench accuracy with every proposal scheme, so cost decides.** Accuracy (%, all four answers of a group correct), raw and after subtracting negative evidence (Eq. 2). Gains are significant on Gemma 4 ($p = 0.001$ pooled) and within noise on Qwen3-VL, which gains in true negatives; SAM matches the pooled grids at 32 rather than 116 calls per image (Apps. I, J).

Method	Qwen3-VL	Gemma 4
Baseline (raw s_0)	23.5	25.2
Grid 3^2 / 5^2 / 9^2	23.5 / 23.7 / 23.8	26.1 / 26.4 / 26.4
SAM, outline / bbox	23.6 / 23.6	26.1 / 26.3
DINO, outline / bbox	23.6 / 23.6	26.2 / 25.7
All methods (union)	23.7	26.7

far it moves the margin, the saturation of Sec. 3.3 in another form (App. O). On Gemini 3.8 Flash the maps pick the same regions as Qwen's (Fig. 6), and where Gemini is less certain its map is graded; five pairs cost 1,760 requests and US$1.13 through the batch API.

Finding. Without log-probabilities, 20 sampled answers per query recover the region that changes the answer (16 of 16 pairs against log-probability ground truth), which is enough to run SAM-RISE on a closed model such as Gemini 3.8 Flash; ranking regions that each flip the answer still needs log-probabilities.

6.4 Negative Evidence Is Actionable on NaturalBench

The adjusted score $\tilde{s}_0$ lowers the "yes" confidence exactly when some region was found to suppress it, which bites where NaturalBench is designed to bite: pairs whose correct answer is "no" and on which the baseline over-confidently says "yes." With the pooled mask set, true-negative accuracy rises from 72.2 to 74.7% and from 74.3 to 76.6% (McNemar $p < 10^{-12}$), and the adjustment never flips a correct "no"; true-positive accuracy falls by 1.0 to 1.3 points ($p \leq 0.02$), so the joint metric moves by only $+0.2$ points, within noise (Table 1; Fig. 13 in App. I). On Gemma 4, with full SAM coverage, the pooled set raises true negatives from 89.0 to 90.6% ($p < 10^{-13}$), again never flipping a correct "no", at a 2.7-point true-positive cost, and joint accuracy from 25.2 to 26.7% ($p = 0.001$; App. J). SAM is also the cheaper route: on the 534 Qwen groups with SAM masks, SAM with bounding boxes matches the 115-mask grid pool ($+2.8$/$+2.1$ vs. $+2.8$/$+2.4$ true-negative points) with 32 calls per image instead of 116. Damping is necessary on Qwen: undamped, the grid pool gains up to 6.6 true-negative points but loses up to 5.6 true-positive points, and joint accuracy falls from 23.5 to 20.6%, worsening with pool size as subtracting the maximum of noise predicts.

Finding. Part of VLM over-confidence is attributable to specific, nameable distractor regions: subtracting the damped negative-evidence term raises NaturalBench true-negative accuracy on Qwen3-VL ($+2.3$ to 2.6 points) and Gemma 4 ($+1.6$), both $p < 10^{-12}$, at a 1 to 3-point true-positive cost, and raises Gemma 4's joint accuracy by 1.5 points; undamped, it lowers Qwen's joint accuracy from 23.5 to 20.6%.

6.5 Population Statistics: Which Classes Are Grounded in Pixels

A single map cannot say whether its explanation is typical of the class. Semantic regions can: Fig. 7 matches each part of a query to its DINOv2-nearest counterpart in 49 other images (Fig. 15) and averages the score drop: the neck, secondary on this black swan, is as important as the whole swan across the class.

Table 2 makes the separation quantitative. For the five most confidently recognised classes, removing the query's most important part lowers the margin by 13.4 logits on average across the class; for the five least confident classes it moves the margin by at most 1.2 logits in any class, the signature of a prediction made from priors rather than pixels. The exception is the prairie chicken, whose top query part, the field behind the bird (13.3), matters little across the class (2.9). The original probability-scored run, which also covers DINOv2 clusters and grids, shows why semantic proposals are needed: a grid cell is not a part, so its best match in another image carries no consistent evidence (0.019 for 3×3 grids and zero for finer ones, against 0.080 for SAM; Table 9). Population maps also expose shortcuts, such as a fur-texture cue standing in for an Egyptian cat (App. L).

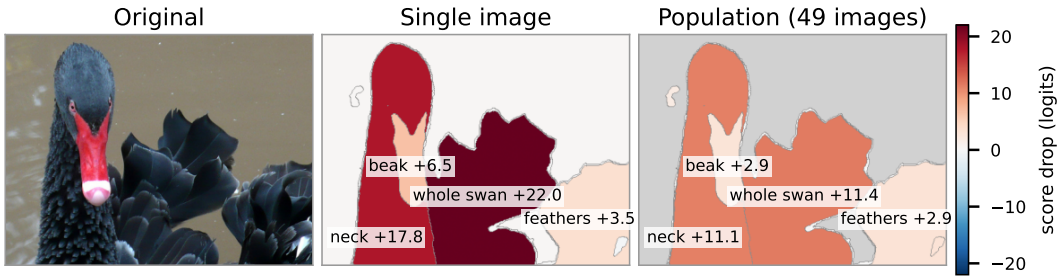

Figure 7: **Semantic regions let one image's explanation be checked against its class.** Left to right: a black-swan query; the score drop from removing each SAM region; the mean drop from removing the matched region in 49 other black-swan images (grey: not an anchor). On the single image the neck looks less important than the whole swan (17.8 vs. 22.0), but across the gallery it is equally important for recognising a black swan (11.1 vs. 11.4). One image may not be typical, so single-image maps invite cherry-picking; averaging over the class removes that doubt, and only semantic regions allow it, since a random or grid mask has no counterpart in another image.

Table 2: **Recognised classes rely on their parts across the class; language-prior classes rely on none.** For each class, the most important part of one query image (the SAM region whose removal lowers the logit margin most) is matched to its counterpart in 49 other images of the class. Columns: how much removing that part lowers the margin on the query image, and on average across the 49 matches; each row averages five ImageNet classes.

Classes	On the query image	Averaged over the class
Five most confidently recognised	17.3	13.4
Five least confidently recognised	1.3	0.2

Finding. Population-level attribution separates classes the model recognises from pixels (removing the key part costs 13.4 logits on average across the class) from language-prior classes (at most 1.2), shows when a single image's explanation is not typical of its class, and needs semantic proposals: grid cells carry no part evidence across a gallery.

7 Discussion and Limitations

Limitations. Saliency-quality and population results use one open model (NaturalBench adds Gemma 4); Gemini maps cannot be checked for fidelity, and sampling costs 20× the queries. Against random masks our advantage is confined to small budgets (Sec. 6.1); the white-box comparison and the preference study use selected showcase pairs (Sec. 6.2). The bounding box is itself out of distribution; a dilated-outline control and an inpainting measurement are future work (Apps. F and G). The population study covers 10 classes and depends on the query image that seeds its anchors.

Implications. Occlusion probes of VLMs must control for mask re-interpretation and score saturation, or they under-measure grounding. Closed VLMs can then be audited for shortcut reliance from their answers alone, and negative evidence points to inference-time correction without retraining, which should be validated independently before deployment.

Reproducibility, AI Use, and Ethics Statements

Reproducibility. Protocols and settings: Sec. 5, App. P; Code is provided in the supplementary material and will be released publicly. **AI use.** AI tools helped with code and editing; the authors reviewed all of it and take full responsibility. **Ethics.** We audit existing models on licensed public benchmarks; the preference study collected only multiple-choice judgments from 11 volunteers. Negative-evidence correction changes model outputs; validate it before deployment.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4190–4197, 2020. doi: 10.18653/v1/2020.acl-main.385.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2511.16719.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Roni Goldshmidt. Attention, please! PixelSHAP reveals what vision-language models actually focus on. *arXiv preprint arXiv:2503.06670*, 2025.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pp. 2668–2677, 2018.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023. doi: 10.1109/ICCV51070.2023.00371.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. NaturalBench: Evaluating vision-language models on natural adversarial samples. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, pp. 17044–17068, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 292–305, 2023. doi: 10.18653/v1/2023.emnlp-main.20.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision (ECCV)*, pp. 366–384, 2024.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018.

Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I. Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11438–11447, 2021. doi: 10.1109/CVPR46437.2021.01128.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Guanxi Shen. GLIMPSE: Holistic cross-modal explainability for large vision-language models. *arXiv preprint arXiv:2506.18985*, 2025.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations (ICLR) Workshop*, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, volume 70 of *PMLR*, pp. 3319–3328, 2017.

Zifu Wan, Ce Zhang, Silong Yong, Martin Q. Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. ONLY: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3225–3234, 2025. doi: 10.1109/ICCV51701.2025.00309.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126 (10):1084–1102, 2018. doi: 10.1007/s11263-017-1059-x.

Appendix

A Working in Margin Space

The probability-space score s_0 saturates: we observe $\sum p(Yes^*) \,/\, \sum p(No^*) > 10^9$ on typical NaturalBench images, so $s_0 = 1.0$ in float32 and $\partial s_0/\partial m_0 = s_0(1 - s_0) \approx 0$. Differences in probability space collapse to machine epsilon, while the same differences in margin space remain exact. We define the logit margin

$$m_0(I, q) \;=\; \log \sum_{t \in Yes^*} p(t \mid I, q) \;-\; \log \sum_{t \in No^*} p(t \mid I, q) \;=\; \sigma^{-1}(s_0), \tag{3}$$

where $Yes^* = \{\mathsf{yes}, \mathsf{Yes}, \mathsf{YES}, \dots\}$ and No^* collect tokeniser variants. Every Δ is computed in margin space; $\sigma(\tilde{m}_0)$ is only for display. The same saturation is what breaks probability-weighted D-RISE (Sec. 3.3).

B SAM Proposal Generator and Cost

Single-pass generation (all reported runs). We prompt SAM 3 in everything mode with a 32×32 grid $\mathcal{P}_{32}(I)$ and filter:

$$\begin{aligned} \mathcal{M} = \mathrm{NMS}_{\tau_{\mathrm{NMS}}}\Big(\{R \in \mathrm{SAM}(I, \mathcal{P}_{32}(I)) : q_{\mathrm{IoU}}(R) \geq \tau_q,\; q_{\mathrm{stab}}(R) \geq \tau_s, \\ \alpha_{\min} \leq |R|/(HW) \leq \alpha_{\max}\}\Big), \end{aligned} \tag{4}$$

with $\tau_{\mathrm{NMS}} = \tau_q = 0.7$, $\tau_s = 0.85$, $\alpha_{\min} = 0.001$, $\alpha_{\max} = 0.95$, and $|\mathcal{M}| \leq 64$.

Optional refinement passes (implemented, disabled in all reported runs). A zoom-out pass tiles I with overlapping crops at coarser levels (overlap 0.34, grid downscale 2 per level); a zoom-in pass re-prompts densely (24×24) inside the 5%-padded boxes of the K largest masks. Both feed the same filter and NMS (Algorithm 1).

Algorithm 1 SAM proposal generation (optional passes in brackets)

Require: image I; SAM; $\tau_q, \tau_s, \tau_{\mathrm{NMS}}, \alpha_{\min}, \alpha_{\max}, M_{\max}$ [overlap, downscale, L, K].
1: $\mathcal{M} \leftarrow \mathrm{SAM}(I, \mathcal{P}_{32}(I))$ {all reported runs skip lines 2–3}
2: [**zoom-out:** for $\ell \in \{1, \dots, L-1\}$ and each level-ℓ crop C: $\mathcal{M} \leftarrow \mathcal{M} \cup \mathrm{SAM}(C, \mathcal{P}_\ell(C))$]
3: [**zoom-in:** for the K largest $R \in \mathcal{M}$: $\mathcal{M} \leftarrow \mathcal{M} \cup \mathrm{SAM}(\mathrm{bbox}(R) \oplus 5\%, \mathcal{P}_{24})$]
4: $\mathcal{M} \leftarrow \mathrm{NMS}_{\tau_{\mathrm{NMS}}}(\mathrm{filter}(\mathcal{M}; \tau_q, \tau_s, \alpha_{\min}, \alpha_{\max}))$; truncate to $M_{\max}$; **return** $\mathcal{M}$

Cost (measured). The single pass issues 1,024 SAM point prompts (about a GPU-second). Images covered by the generation pass carry 31 proposals on average (median 28, maximum 64). The dominant cost is the $|\mathcal{M}| + 1$ VLM calls per image. The NaturalBench release run issued 1.04M VLM calls across all proposal schemes and baselines, in about 13 hours against a single-node vLLM server. In the saliency benchmark, SAM-RISE averaged 34 (COCO) and 24 (ImageNet) calls per image.

C D-RISE Benchmark: Adaptation, Protocol, and Full Results

From D-RISE to a yes/no VLM. D-RISE (Petsiuk et al., 2021) samples RISE masks (an $s \times s$ grid, each cell kept with probability p, bilinearly upsampled and randomly cropped; $s = 16$, $p = 0.5$), runs the detector on $I \odot M_i$, and weights M_i by $w_i = \max_j \mathrm{IoU}(L_t, L_j) \cdot \cos(P_t, P_j) \cdot O_j$ over proposals j. A VQA model emits no boxes or objectness, so only the categorisation term survives, which for a yes/no query is $P(\text{yes} \mid I \odot M_i)$; the map $H = \sum_i w_i M_i$ is then exactly RISE with the VLM as classifier. Mask generation, masking, and aggregation follow the reference code (we omit the $1/(Np)$ normalisation, which rescales the map without changing its ranking). Masked images are rounded to uint8 and JPEG-encoded (quality 90) before querying, as are SAM-RISE's. If a returned top-20 list lacks "yes" or "no", the missing log-probability is set to -20.

Table 3: **D-RISE as published fails at every budget; with margin weights it wins given hundreds of queries, and SAM-RISE wins per query below that.** Qwen3-VL-8B, 500 images per dataset. PG: pointing game; Ins./Del.: insertion (higher is better) and deletion (lower is better) AUC on the normalised margin. "As published" weights masks by $P(\text{yes})$; "our adaptation" by the mean-centred logit margin (Sec. 3.3). Calls are VLM queries per image (COCO / ImageNet). Full budget grids with CIs below.

		MS-COCO				ImageNet		
Method	Calls	PG mask	PG box	Ins. ↑	Del. ↓	PG box	Ins. ↑	Del. ↓
Image-centre control	0	0.262	0.402	–	–	0.704	–	–
D-RISE, as published	34	0.146	0.232	–	–	0.254	–	–
D-RISE, as published	4000	0.332	0.444	0.614	0.237	0.464	0.632	0.251
D-RISE, our adaptation	34	0.550	0.694	–	–	0.648	–	–
D-RISE, our adaptation	149	0.820	0.934	–	–	0.860	–	–
D-RISE, our adaptation	4000	**0.898**	**0.966**	**0.878**	**−0.138**	**0.926**	**0.907**	**−0.049**
Grid 9×9	81	0.532	0.748	0.741	0.157	0.796	0.806	0.193
SAM-RISE	34 / 24	0.726	0.802	0.782	0.245	0.796	0.797	0.232
SAM-RISE + grids	149 / 139	0.690	0.872	0.877	−0.035	0.902	0.904	0.004

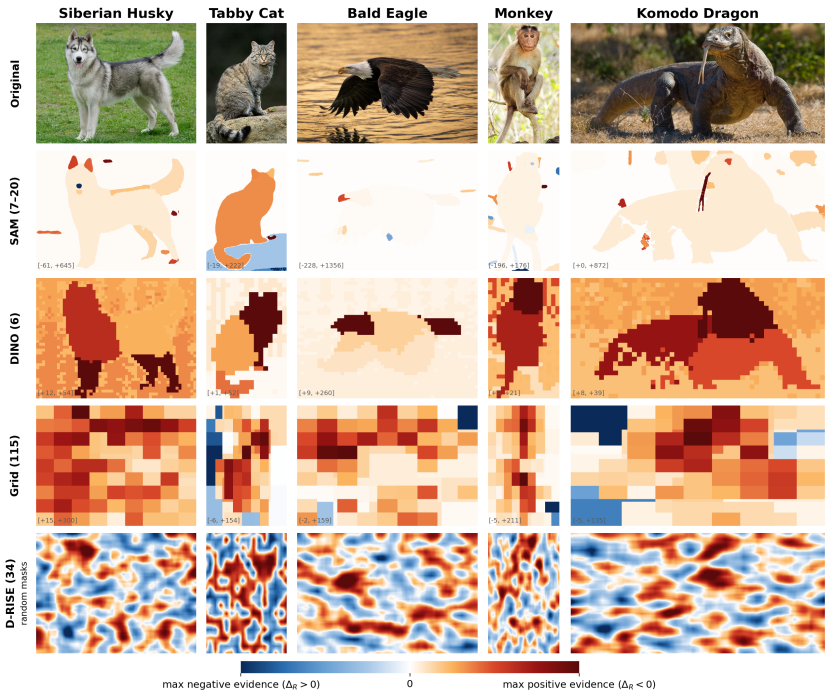

Figure 8: **At SAM-RISE's budget, random masks carry no object structure.** Layout as in Fig. 3, with D-RISE (our adaptation) built from the first 34 of the same 4,000 random masks, the per-image budget of SAM-RISE on MS-COCO (Table 3).

Weight variants. Because masks are generated from a stored per-image seed and every per-mask score is stored, any other weighting of the same masks is rebuilt offline without new queries. We report four: $P(\text{yes})$ (as published); $P(\text{yes})$ minus its mean over masks (centred); the logit margin (unsaturated); and the centred margin (*D-RISE (our adaptation)*). Centring does most of the work (COCO PG mask at $N = 4{,}000$: $0.332 \to 0.860$ with probability weights), because it removes the constant mask-count pedestal; the margin adds a further gain at $N \geq 1{,}000$. The uncentred margin alone is also weak, since margins cluster near $+12$ and the pedestal dominates again.

Protocol. Images are resized to at most 512 px. PG (mask) counts a hit when the saliency argmax lies in the ground-truth segmentation, PG (box) when it lies in the ground-truth box; a 15-pixel tolerance variant is also computed. Region-based maps are piecewise constant, so ties are broken at the argmax pixel nearest the centroid of all tied pixels; the default raster-order argmax returns the top-left corner of the winning region and penalises only region-based methods (0.463 vs. 0.726 PG mask for SAM-RISE). SAM-RISE's pointing map uses raw Δ_R; the area-normalised density map scores 0.336 PG mask, because density over-weights tiny regions. SAM-RISE regions are occluded with the mask-shaped operator (flat fill), not the bounding box of Sec. 3.2. Insertion starts from a heavily blurred image and reveals the most salient pixels first; deletion blanks them first from the original; both use 50 steps and the margin normalised per image between the blank-image and original scores, so deletion AUC can be negative. The COCO pool starts from 14,631 candidate (image, category) pairs, 9,526 after the 1 to 50% area filter; the first 500 confirmed by the model are kept (576 screened).

Why our D-RISE numbers differ from the published ones. The D-RISE paper reports PG (box) 0.966 and PG (mask) 0.846 explaining YOLOv3 on COCO; D-RISE (our adaptation) reaches 0.966 and 0.898. Restricting to the 275 single-instance images does not remove the gap (0.909 PG mask), nor does restricting to the smallest objects (0.874 for 1 to 5% of the frame). The estimators differ (the published estimator scores 0.332 here), and so do the model and the target, so we cite the published numbers as context, not as a baseline.

Table 4: **Full COCO results: margin weighting, not the mask count, repairs D-RISE.** 500 images; "our adaptation" weights masks by the mean-centred logit margin. As published, D-RISE reaches only 0.332 PG (mask) at 4,000 masks; the same masks with margin weights reach 0.898.

Method	Calls	PG mask	PG box	Del. ↓	Ins. ↑
Random point / image centre (controls)	0	0.108 / 0.262	0.212 / 0.402	–	–
Grid 3×3 / 5×5 / 9×9	9 / 25 / 81	0.422 / 0.528 / 0.532	0.606 / 0.738 / 0.748	– / – / 0.157	– / – / 0.741
D-RISE, N=34 (published / our adaptation)	34	0.146 / 0.550	0.232 / 0.694	–	–
SAM-RISE	34	0.726	0.802	0.245	0.782
D-RISE, N=50 (published / our adaptation)	50	0.140 / 0.652	0.244 / 0.776	–	–
D-RISE, N=100 (published / our adaptation)	100	0.192 / 0.794	0.298 / 0.904	–	–
D-RISE, N=149 (published / our adaptation)	149	0.202 / 0.820	0.300 / 0.934	–	–
SAM-RISE + grids	149	0.690	0.872	−0.035	0.877
D-RISE, N=500 (published / our adaptation)	500	0.260 / 0.876	0.342 / 0.956	– / −0.082	– / 0.832
D-RISE, N=4000 (published / our adaptation)	4000	0.332 / 0.898	0.444 / 0.966	0.237 / −0.138	0.614 / 0.878

Table 5: **Full ImageNet results: the same ordering as COCO, on a benchmark where centre bias is strong.** 500 images, boxes only. The image-centre control scores 0.704, so pointing is a weak discriminator here, and D-RISE as published (0.464 at 4,000 masks) falls well below it.

Method	Calls	PG box	Del. ↓	Ins. ↑
Random point / image centre (controls)	0	0.250 / 0.704	–	–
Grid 3×3 / 5×5 / 9×9	9 / 25 / 81	0.720 / 0.798 / 0.796	– / – / 0.193	– / – / 0.806
SAM-RISE	24	0.796	0.232	0.797
D-RISE, N=34 (published / our adaptation)	34	0.254 / 0.648	–	–
D-RISE, N=50 (published / our adaptation)	50	0.276 / 0.718	–	–
D-RISE, N=100 (published / our adaptation)	100	0.288 / 0.822	–	–
SAM-RISE + grids	139	0.902	0.004	0.904
D-RISE, N=149 (published / our adaptation)	149	0.298 / 0.860	–	–
D-RISE, N=500 (published / our adaptation)	500	0.372 / 0.924	– / 0.008	– / 0.852
D-RISE, N=4000 (published / our adaptation)	4000	0.464 / 0.926	0.251 / −0.049	0.632 / 0.907

D White-Box Baselines: Implementation and Further Examples

Grad-CAM. We hook the output of the last vision-encoder block, back-propagate the yes-minus-no logit margin for the question at hand, take the ReLU of the channel mean of activation times gradient per patch, reshape to the patch grid, and upsample bilinearly.

Attention rollout. With eager attention, at each of the 36 language-model layers we average attention over heads, mix it $50/50$ with the identity to account for residual connections, renormalise

rows, and multiply across layers (Abnar & Zuidema, 2020). The row of the answer position, restricted to image-token columns, gives the per-patch importance.

Occlusion read-out. Images are resized to at most 448 px so that all layers' attention tensors fit in memory alongside the gradients. For Grad-CAM and rollout we occlude the top 10% of pixels of the upsampled map in black; for SAM-RISE we occlude its single highest-effect mask. The 10 NaturalBench images were chosen by the size of their largest SAM-RISE effect, which biases this comparison toward SAM-RISE (Sec. 6.2). Per-pair values: median $|\Delta|$ of 22.3 (SAM-RISE), 2.3 (Grad-CAM), 2.1 (rollout) logit units; answer flips 17, 1, 1 of 20; SAM-RISE has the largest $|\Delta|$ in 18 of 20 pairs (one tie with rollout). These statistics come from the original run; the figures are re-rendered from a rerun of the same protocol, in which bf16 nondeterminism moves individual values by up to about one logit (beanie question: -31.75, -1.62, -8.75 originally) without changing which answers flip.

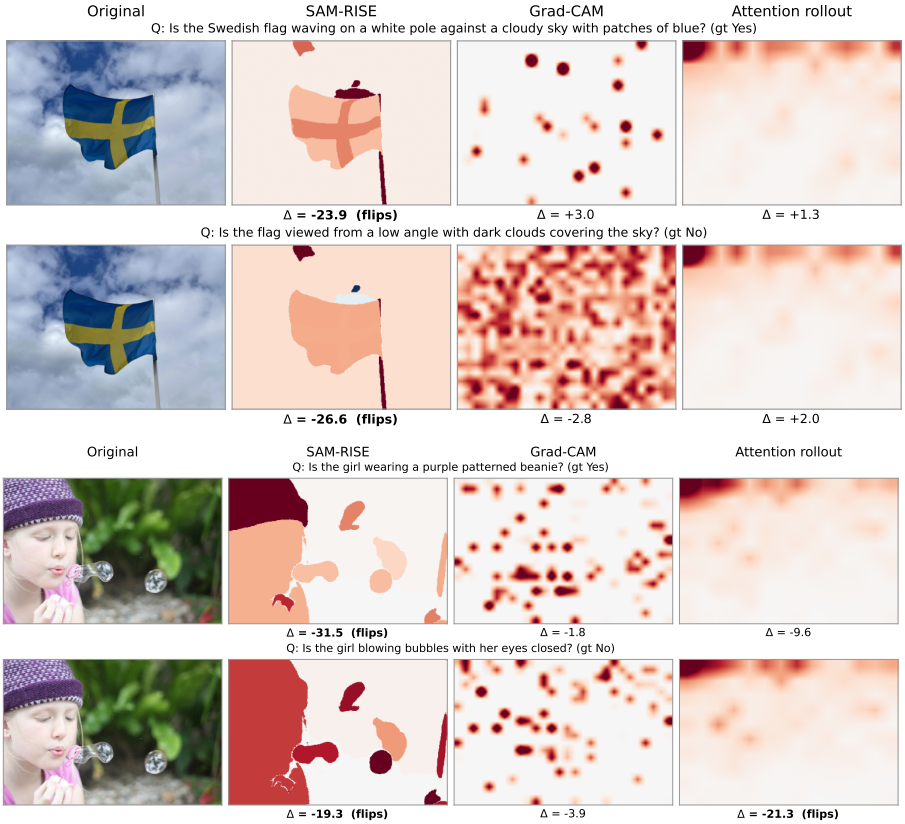

Figure 9: **White-box peaks can even point away from the evidence.** Further NaturalBench comparisons; layout as in Fig. 5. Top: occluding the Grad-CAM peak *raises* the margin for the flag question ($+3.0$), while the SAM-RISE flag region flips the answer (-23.9). Bottom: for the beanie question, the SAM-RISE region flips the answer (-31.5) while the Grad-CAM and rollout regions barely move it (-1.8, -9.6).

E Human Preference Study

Stimuli. The 14 image-question pairs come from our single-image study: airplane, a bird in the same airplane image, car, eagle, helicopter, Komodo dragon, man, minibus, monkey, ostrich, Siberian husky, tabby cat, and two images of a woman, each asked as "Is there a {class} in the image?" of Qwen3-VL-8B. Each trial showed the original image and the candidate heatmaps on a plain background, without the photo blended in, identified only by letter. The SAM-RISE map averages each region's Δ_R over three occluder shapes (square, circle, hexagon) and three fills (black, grey, white), painted on the SAM masks.

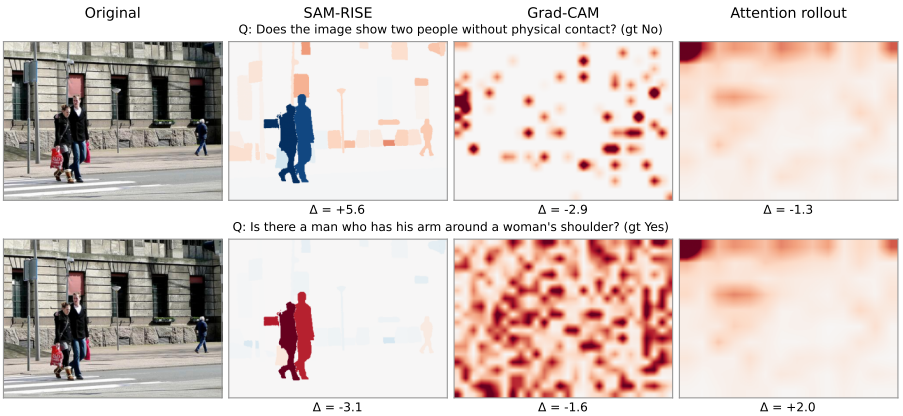

Figure 10: **A failure case: relations spread over two objects.** A pair on which no method's top region flips the answer (the relational question "Is there a man who has his arm around a woman's shoulder?"): SAM-RISE's single region drops the margin by only 3.1. Relations spread over two objects are hard for single-region ablation.

Protocol. The black-box track compared SAM-RISE with D-RISE (our adaptation), a 3×3 grid, and DINOv2 clusters; the white-box track compared it with Grad-CAM and attention rollout. Raters read: "A vision-language model was asked yes/no questions about images. Several methods each produced a heatmap claiming to show why the model answered as it did. For each image, pick whichever heatmap best shows the area that actually explains the answer." Eleven raters gave 362 judgments, of which 9 were skipped (Table 6).

Table 6: **Raters prefer SAM-RISE in both tracks, far above chance.** Picks per method (share of answered judgments) from 11 raters on 14 showcase pairs. The preference is strongest against the white-box maps, whose peaks often sit on background or border tokens (Fig. 5); DINOv2 clusters, the other semantic scheme, are the most-picked alternative.

Track	Method	Picks
Black-box (184 judgments, chance 25%)	SAM-RISE	140 (76%)
	DINOv2 clusters	36 (20%)
	D-RISE (our adaptation)	8 (4%)
	Grid 3×3	0 (0%)
White-box (169 judgments, chance 33%)	SAM-RISE	165 (98%)
	Grad-CAM	3 (2%)
	Attention rollout	1 (1%)

Caveats. The pairs are showcase images chosen for the paper's figures, not a random sample. Rollout maps are dominated by attention-sink artifacts at the image border (Sec. 6.2). Judgments by the same rater are not independent, so we report counts rather than a test. The study measures plausibility to people; Secs. 6.1 and 6.2 measure faithfulness to the model.

F Why Inpainting Does Not Fix the Operator

A grey box is out-of-distribution content, and a photorealistic removal can look preferable. It fails twice. First, a generative inpainter fills a region with its most probable completion given the context; asked to inpaint a dog's ear, it would paint an ear back, the evidence would never leave, and the region would be declared unimportant. We argue this from the mechanism and leave its measurement to future work. Second, "remove region R" has no ground-truth counterfactual: cat ears inject evidence for another class, background grass rewrites the silhouette, a coffee mug adds evidence of its own, and each answers a different question. The fixed grey box is the intervention whose semantics the

analyst controls, and holding it fixed across regions and images lets its bias cancel in the comparisons the paper makes (between regions, schemes, and signs of Δ_R).

G Operator Ablations: Shape, Fill, and Context

Shape at matched area. On a Siberian-husky probe image (baseline margin 12.9), we ablate the same 15 SAM masks with the outline, the bounding box, and nine solid shapes area-matched to each mask. Every shape behaves like the outline (per-shape means -1.8 to -3.0; average -2.4; outline -2.5); only the box restores the deep drops (-4.5 mean, -25.2 strongest vs. -14.2 for the outline). At benchmark scale (corrected run, first 720 groups, $\sim$88k ablations per operator) the shapes' means are -0.17 to -0.29 against the outline's -0.20; box vs. outline on 66k paired masks: means -0.285 vs. -0.161 and 5th percentiles -4.28 vs. -2.16.

Fill value. Nine constant fills (0 to 255) move the mean drop only between -0.17 and -0.21 at benchmark scale, with no trend. Per mask, the strongest fill differs: it is not black for 71.8% of $87{,}574$ signal-bearing mask-questions, and the best-to-worst swing is $4.0\times$ in median, which motivates choosing the fill per region (with held-out validation, since a maximum over fills inflates $|\Delta_R|$ under the null).

Context coverage. Across $66{,}018$ scored boxes, the Spearman correlation between $|\Delta_R|$ and the box's background fraction is 0.17 raw and 0.03 within box-area deciles. This bounds the concern correlationally; a dilated-outline control (the outline grown to the box's area) and a random-box null of matched area would be the causal versions.

H Proposal Schemes: Grids, DINOv2 Clusters, and SAM

Fig. 11 compares heatmaps from SAM, DINOv2+PCA+KMeans clusters, uniform grids, and their union. Grid cells split faces and merge parts, so their drops attribute to no nameable concept; DINO clusters lump the foreground into one region; SAM isolates parts such as the husky's ears and nose. Larger regions trivially cause larger drops, so we normalise by area, $\widetilde{\Delta}_R = \Delta_R \cdot HW/|R|$. On the panel of Fig. 11, SAM's deepest raw drop is about $4\times$ the densest grid's, and more than $10\times$ after normalisation (Fig. 12), consistent with semantic targeting rather than occluded area. The quantitative comparison with grids at scale is Table 3 (pointing) and Table 9 (cross-gallery transfer).

I Adjustment Variants, Significance, and Calibration

The general form of Eq. (2) is $\tilde{m}_0 = m_0 - w(\delta;\tau,k)\,\delta$ with $w = 1$ (default), $\mathbf{1}[\delta \geq \tau]$ (threshold), $(\delta/(m_0^+ + \delta))^k$ (proportional), or their product. Table 7 compares the default and proportional variants on all $1{,}900$ groups; the threshold variants were not run.

Table 7: **The undamped adjustment over-corrects; damping keeps the gain.** NaturalBench accuracy (%) per proposal scheme. TP: $\mathrm{abs}_{i_0q_0}$, $\mathrm{abs}_{i_1q_1}$; TN: $\mathrm{abs}_{i_1q_0}$, $\mathrm{abs}_{i_0q_1}$; All: joint metric.

Variant	Scheme	TP i_0q_0	TP i_1q_1	TN i_1q_0	TN i_0q_1	All
–	Baseline (raw s_0)	58.6	59.4	72.2	74.3	23.5
Default	Grid 3×3 / 5×5 / 9×9	54.7 / 54.3 / 53.6	56.4 / 55.8 / 55.3	77.2 / 77.6 / 77.9	78.3 / 78.9 / 79.3	21.4 / 21.3 / 21.0
Default	SAM / SAM + bbox	57.7 / 57.7	58.6 / 58.4	73.7 / 73.7	75.7 / 75.7	23.1 / 23.0
Default	DINO / DINO + bbox	57.9 / 57.6	58.7 / 58.9	73.6 / 73.0	75.6 / 75.5	23.4 / 23.1
Default	All grids	53.1	54.7	78.8	80.0	20.6
Proportional	Grid 3×3 / 5×5 / 9×9	57.7 / 57.9 / 57.6	58.9 / 58.7 / 58.5	74.0 / 74.1 / 74.1	75.8 / 76.0 / 76.1	23.5 / 23.7 / 23.8
Proportional	SAM / SAM + bbox	58.5 / 58.5	59.3 / 59.2	72.7 / 72.8	74.8 / 74.7	23.6 / 23.6
Proportional	DINO / DINO + bbox	58.5 / 58.5	59.2 / 59.2	72.7 / 72.7	74.8 / 74.7	23.6 / 23.6
Proportional	All grids	57.5	58.6	74.6	76.5	23.7
Proportional	All methods	57.4	58.5	74.7	76.6	23.7

Significance. Bootstrap over the $1{,}900$ groups ($2{,}000$ resamples) and exact two-sided McNemar tests. Pooled proportional adjustment: true-negative deltas $+2.58$ $[+1.89, +3.32]$ and $+2.32$

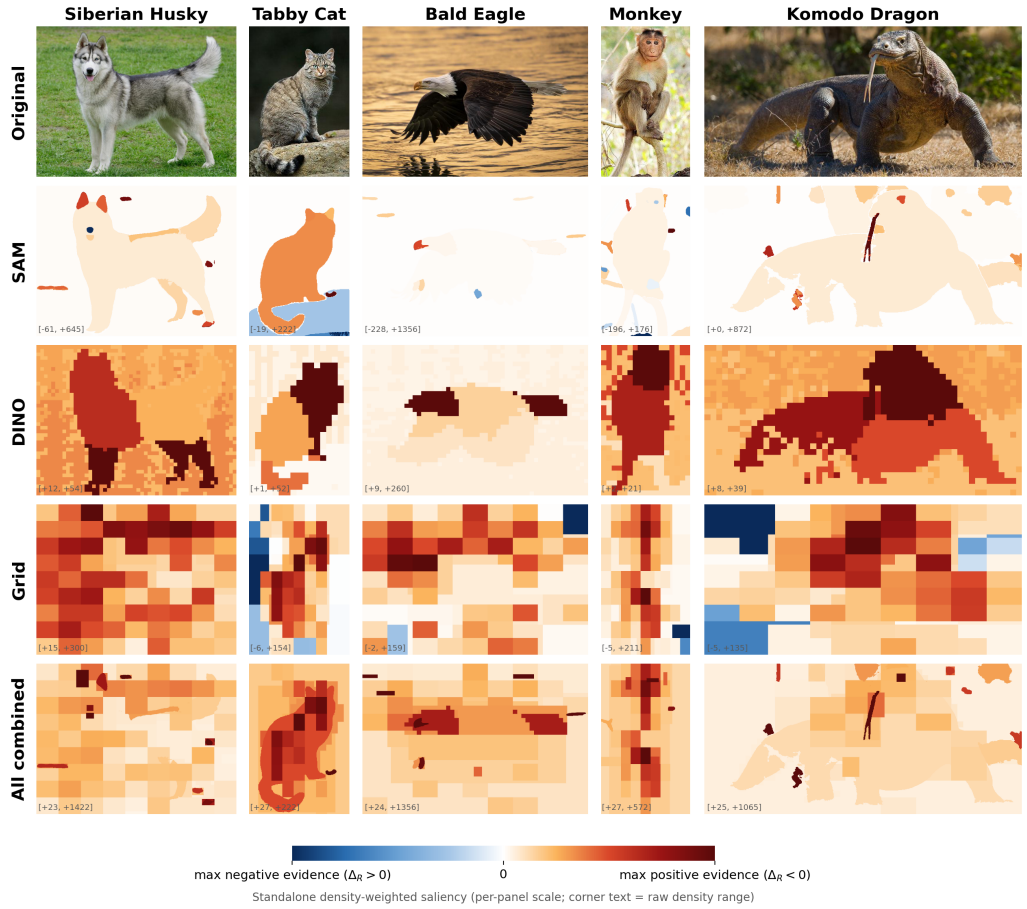

Figure 11: **Proposal schemes decide what ablation can attribute.** Area-normalised per-pixel score change for each scheme (rows: original; SAM; DINOv2 clusters; grids; union) on five ImageNet classes. Red: positive evidence; blue: negative evidence. Only SAM's regions are nameable parts: DINOv2 clusters cover whole foregrounds, and grid cells cut across objects, so neither can say which part carried the evidence.

$[+1.68, +3.00]$ points ($p = 3.6{\times}10^{-15}$, $1.1{\times}10^{-13}$; discordant pairs $0/49$ and $0/44$, so no correct "no" is flipped); true-positive deltas -1.26 $[-2.11, -0.42]$ ($p = 4.3 \times 10^{-3}$) and -0.95 $[-1.68, -0.21]$ ($p = 0.018$); joint $+0.21$ $[-0.69, +1.11]$ ($p = 0.73$). NaturalBench's relative metrics move the other way ($67.7 \rightarrow 65.1\%$ pooled), consistent with a one-sided confidence reduction.

Corrected run. On the first 720 groups with full SAM coverage, SAM alone yields true-negative gains of $+2.3/+1.5$ points at a true-positive cost of at most 0.4 (joint $25.1 \rightarrow 25.8\%$), and the undamped estimator again over-corrects ($25.1 \rightarrow 23.1\%$).

Calibration. ECE (15 bins over 7,600 questions): baseline 0.289; pooled proportional 0.304; SAM-only proportional 0.292. The adjustment corrects the decision threshold where a distractor exists, not the model's confidence at large. On the Qwen run, gains track mask coverage: on the 534 groups where SAM masks exist, SAM with bounding boxes matches the 115-mask grid pool ($+2.8/+2.1$ vs. $+2.8/+2.4$ true-negative points) with 32 calls per image instead of 116. Empirical accuracy stays between 42 and 70% in every confidence bin (Fig. 14).

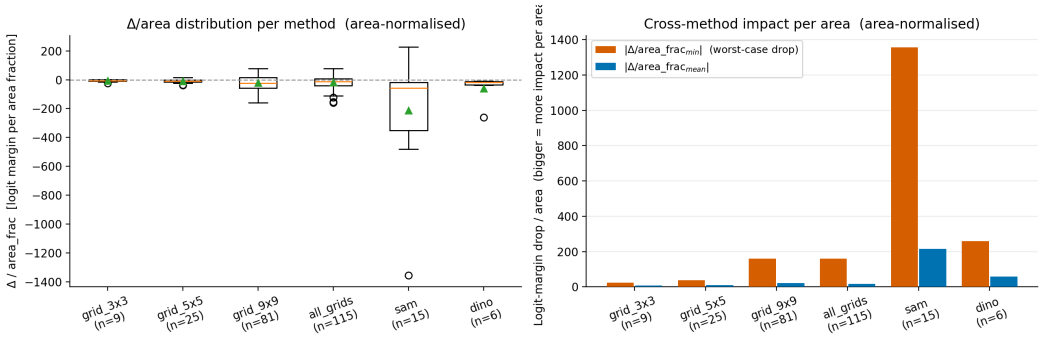

Figure 12: Area-normalised score drop per proposal scheme on the panel of Fig. 11.

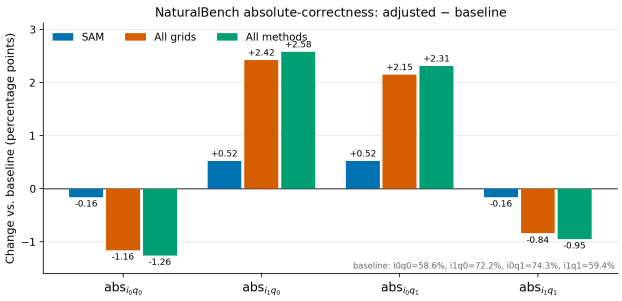

Figure 13: **Adjustment trades true positives for true negatives.** Change in the four NaturalBench sub-metrics on Qwen3-VL-8B (adjusted minus baseline, points). Subtracting negative evidence only lowers "yes" confidence: it fixes over-confident "yes" answers (TN $+2.3$ to 2.6) at a smaller cost in correct ones (TP -1.0 to -1.3).

J Gemma 4 Replication on NaturalBench

We rescored every NaturalBench group with Gemma 4 (`google/gemma-4-12B-it`, bfloat16, vLLM 0.26; float16 produced malformed outputs) with the prompt, 768-px input, and SAM, DINOv2, and grid masks of the Qwen run; SAM masks exist for all 1,900 groups. Table 8 gives joint accuracy under both adjustment variants. Under $k = 5$, SAM bounding boxes gain 33 groups and lose 13 (joint $25.2 \rightarrow 26.3\%$, McNemar $p = 0.005$); the pooled set gains 50 and loses 22 (26.7%, $p = 0.001$). Neither flips a correct "no"; true negatives rise from 89.0 to 90.2% (bbox) and 90.6% (pooled), and true positives fall from 61.0 to 59.4% and 58.3% (all $p < 10^{-10}$). Gemma's baseline already answers 89% of true negatives correctly, against 72 to 74% for Qwen, which leaves less room for the gain. Unlike on Qwen, the undamped adjustment also raises joint accuracy. The pairwise relative metric falls from 66.3 to 61.5% with the pooled set.

Table 8: **Gemma 4 gains under both adjustment variants.** NaturalBench joint accuracy (%, 1,900 groups). Unlike on Qwen (Table 7), even undamped subtraction raises joint accuracy; damping adds most for the pooled set ($26.1 \rightarrow 26.7$).

Method	Undamped ($w \equiv 1$)	Proportional ($k = 5$)
Baseline (raw s_0)	25.2	25.2
Grid 3×3 / 5×5 / 9×9	25.7 / 25.9 / 26.1	26.1 / 26.4 / 26.4
SAM, outline / bbox	26.6 / 26.4	26.1 / 26.3
DINO, outline / bbox	26.3 / 25.5	26.2 / 25.7
All methods (union)	26.1	26.7

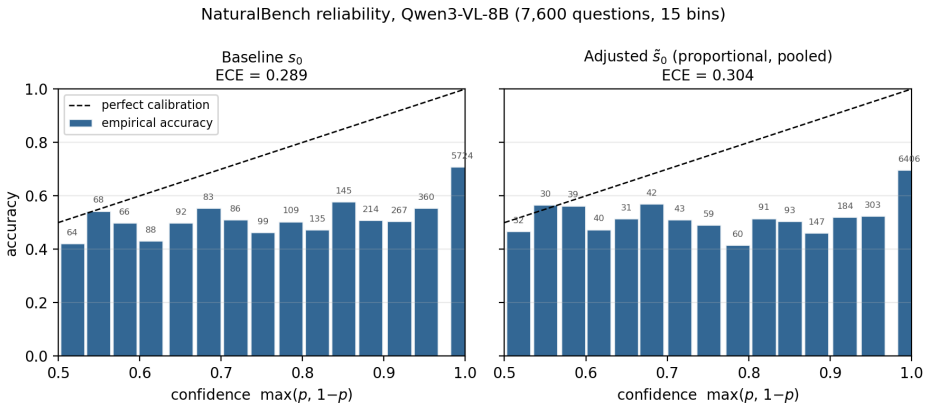

Figure 14: **Confidence is nearly uninformative on NaturalBench, before and after adjustment.** Reliability diagrams (15 bins; counts above bars).

K Population Statistics: Formal Procedure

For class c with gallery $\mathcal{I}_c$ and query I_q, the anchors are the K regions of I_q with the most negative Δ_R. Every gallery image is re-segmented with the same generator (no query conditioning), capped at its M largest regions, and scored. Each anchor is matched by a frozen DINOv2 embedding $\phi(R; I)$ of the region's foreground crop (background blacked, 224×224):

$$R_\pi(I) \;=\; \arg\max_{R \in \mathcal{M}(I)} \langle \phi(\pi; I_q),\ \phi(R; I) \rangle, \qquad \bar{\Delta}_\pi(c) = \frac{1}{|\mathcal{I}_c|} \sum_{I \in \mathcal{I}_c} \Delta_{R_\pi(I)}. \tag{5}$$

Cost: $|\mathcal{I}_c|(M+1) + (K+1)$ VLM calls per class; the run used 49 gallery images, $M = 10$, $K = 5$, i.e. 545 calls per class. The ten study classes come from a ranking of all 1,000 classes by mean baseline VQAscore (20 images per class, fixed seed). Scaling the population stage to all 1,000 classes and varying the query image are left to future work.

Margin rerun. The original run (Table 9, Figs. 15–17) scores $P(\text{yes})$, which saturates at 1.0 on confident queries, so most regions score exactly 0 and ties decide the anchors. Fig. 7 comes from a rerun with the same classes, query images, and galleries, scored in logit margin with the same outline ablation; SAM regions reproduce closely (on the first class, 25 of 28 query boxes are identical and gallery regions match at median IoU 1.00). A second rerun with bounding-box ablation is unsuitable here: a background region's box can span the whole frame, so removing it deletes the object too.

Table 9: **Cross-gallery transfer separates semantic from grid proposals, and grounded from language-prior classes.** Gallery-averaged matched-part drop $\bar{\Delta}_\pi(c)$ (VQAscore units, ∼245 anchor-gallery pairs per cell) for the top-5 and bottom-5 ImageNet classes by baseline confidence. Positive: removing the matched region lowers confidence in the class. †Amphibian, Egyptian cat, pirate, Madagascar cat, thunder snake.

Class	SAM	DINO	Grid 3×3	Grid 5×5	Grid 9×9
Goldfinch	0.160	0.148	0.032	0.001	0.000
Chickadee	0.066	0.116	0.027	0.000	0.000
Bald eagle	0.064	0.155	0.012	−0.001	−0.001
Prairie chicken	0.013	0.044	0.010	0.000	0.002
Black swan	0.096	0.068	0.014	0.000	0.000
Top-5 mean	*0.080*	*0.106*	*0.019*	*0.000*	*0.000*
Bottom-5 mean†	*0.000*	*−0.002*	*−0.001*	*0.001*	*0.001*

L Population Statistics: Further Figures

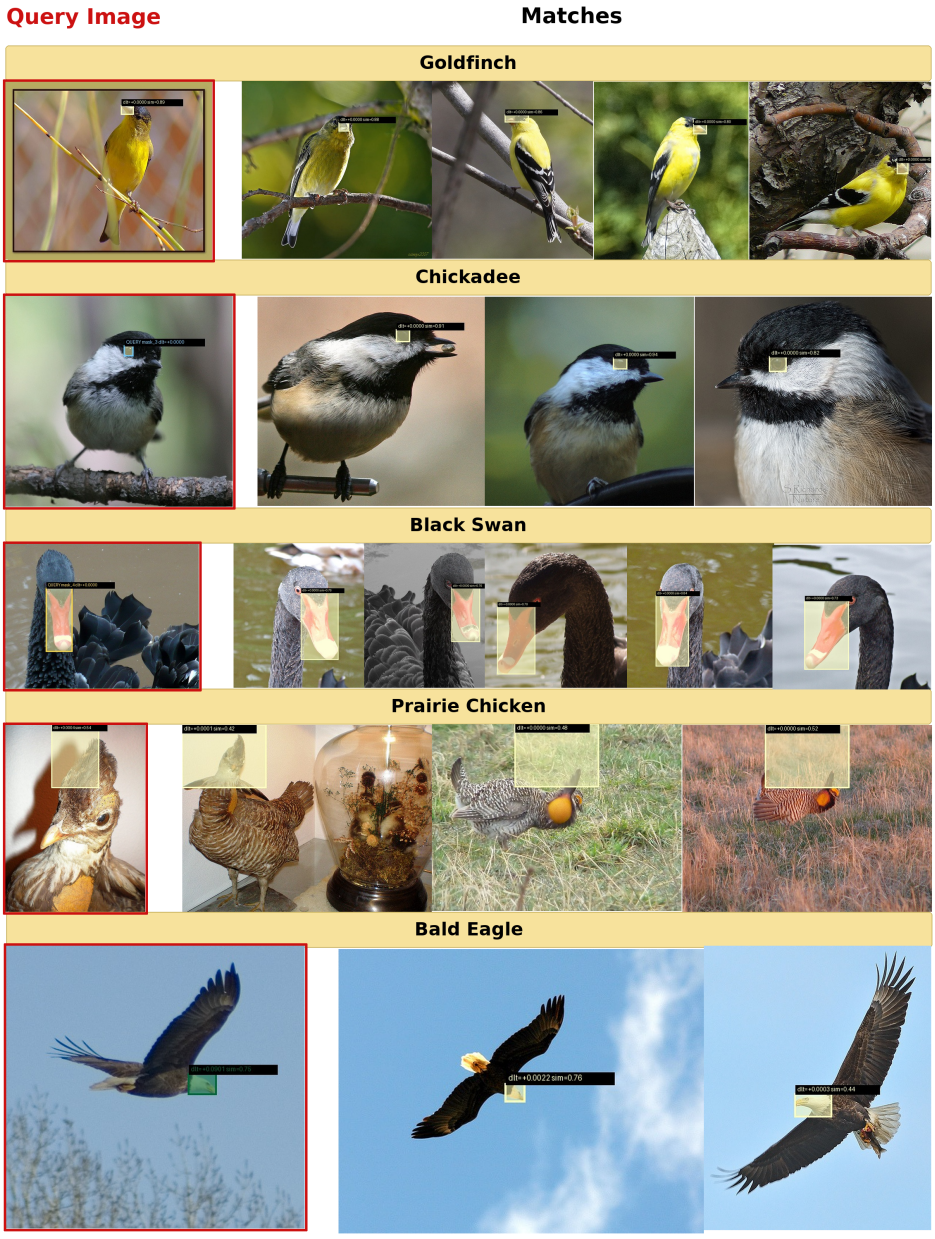

Figure 15: **The DINOv2 matcher localises the same part across a class gallery.** Five high-confidence classes; left column: query image with its top part anchor; other columns: gallery images with the matched region and cosine similarity. The match survives changes in pose, viewpoint, and scale, which is what makes a gallery average over "the same part" meaningful.



Figure 16: **On low-confidence classes, matches expose shortcuts.** Row 1: an Egyptian-cat anchor matches a woman's bare leg, a fur-texture cue that a single-image map could not reveal.

M NaturalBench Cake/Plate Example

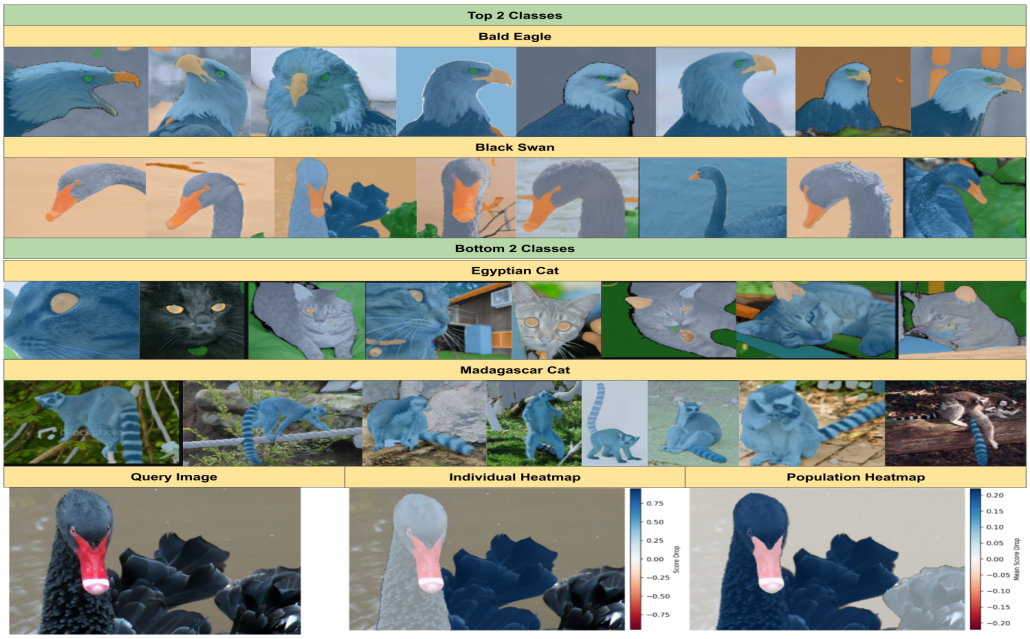

Figure 17: **Population maps average each part's evidence over a class gallery.** Top: gallery images of the top-2 and bottom-2 classes, coloured by the regions matched to the query's anchors. Bottom: the black-swan query with its single-image and population maps from the original run, scored in probability (blue: positive evidence; the red beak is the photo, not the map). Probability saturates on this confident query ($P(\text{yes}) = 1.0$): the single-image map credits the whole swan far above its neck (0.95 vs. 0.12) and gives the beak 0, while over the gallery the swan and neck average 0.22 each. Scored in logit margin (Fig. 7), the same comparison resolves the beak and shows the single image overstating every part.

N Additional Figures

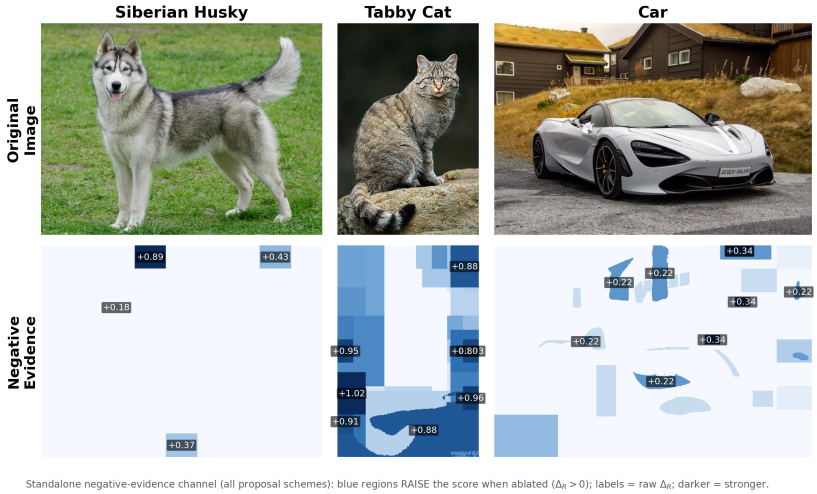

Figure 18: **Masking a class-defining part can raise confidence.** Negative-evidence channel: the dog's face (+0.89), the cat's lower body and the rock beneath it (+0.9 to +1.0), and car headlights and mirrors (+0.22 to +0.34), in logit-margin units. Removing a defining part should lower confidence; that it raises it instead suggests template matching (Sec. 6.5).

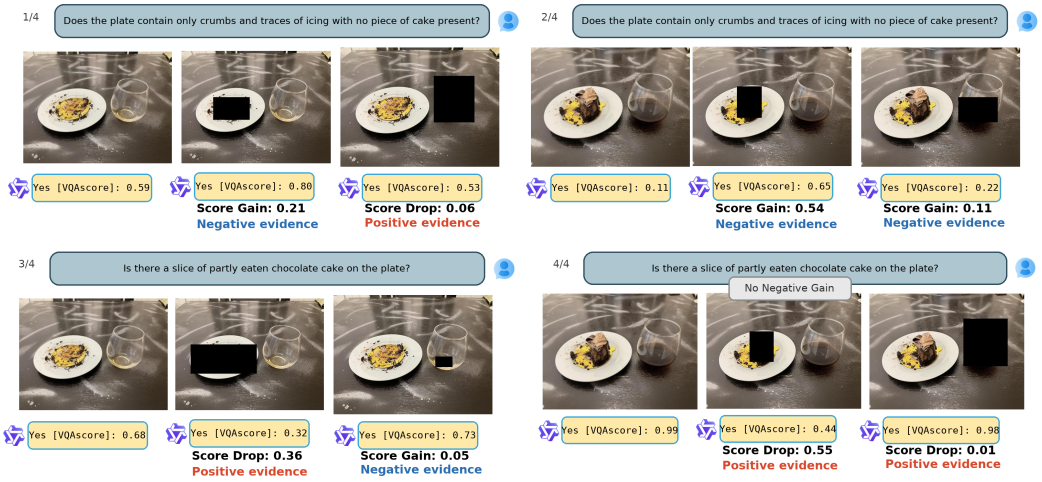

Figure 19: **The same pixels flip evidential sign across questions.** Two images crossed with two opposing questions; each panel shows the baseline VQAscore and two SAM-region ablations. Red: positive evidence; blue: negative evidence. The cake region is negative evidence for "empty plate?" (+0.54) and positive evidence for "cake on plate?" (−0.55).

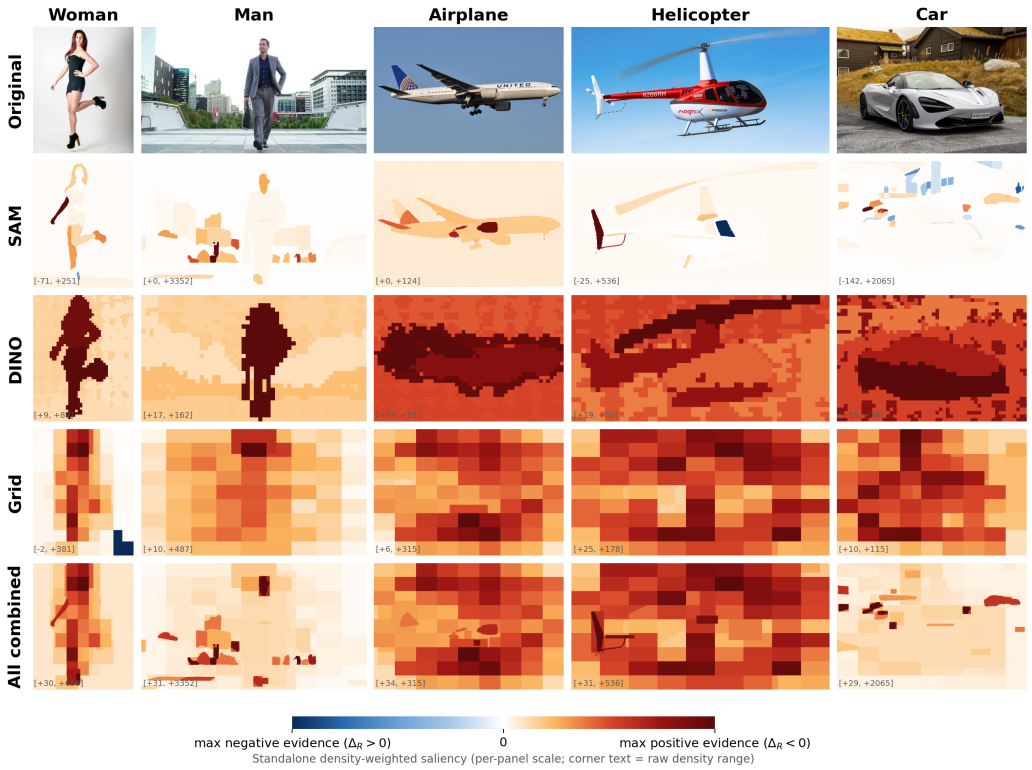

Figure 20: Proposal-scheme comparison, man-made and human classes; layout as in Fig. 11.

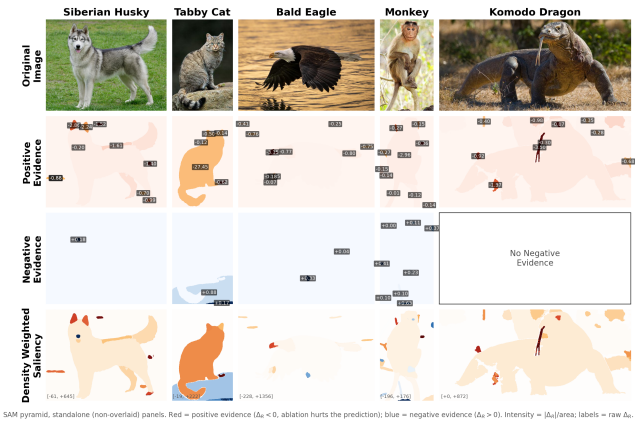

Figure 21: **Signed evidence separates what supports a class from what argues against it.** Rows: original; positive ($\Delta_R < 0$, red) and negative ($\Delta_R > 0$, blue) evidence; density-weighted composite; labels: per-mask Δ_R in logit margin. Positive evidence concentrates on a few class-defining parts: the husky's ears, the eagle's head, the Komodo dragon's forked tongue. Negative evidence exposes cues that pull away from the class, such as the rock under the tabby (+0.88); the Komodo dragon has none, so its answer rests on positive evidence alone. More in Fig. 22.

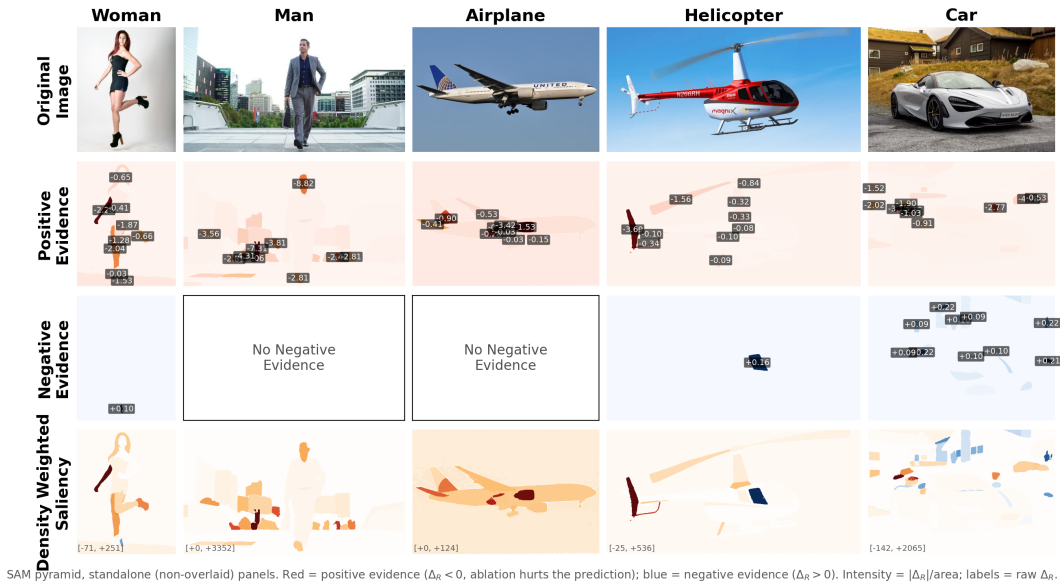

Figure 22: Signed evidence, man-made and human classes; layout as in Fig. 21. Two rows (Man, Airplane) have no negative evidence, so Eq. (2) leaves them unchanged.

O CLOSED MODELS WITHOUT LOG-PROBABILITIES

Some closed APIs return no token probabilities (Sec. 7); this appendix details the estimator of Sec. 4 and its validation (Sec. 6.3). For these we estimate the score by sampling: for the unmasked image and for each ablated region, we pose the same question n times at temperature 1, where the fraction of Yes answers estimates the model's $P(\text{yes})$, and set

$$\hat{s}_0 = \frac{n_{\text{yes}}}{n_{\text{yes}} + n_{\text{no}}}, \qquad \hat{m}_0 = \log \frac{n_{\text{yes}} + \frac{1}{2}}{n_{\text{no}} + \frac{1}{2}}, \tag{6}$$

a smoothed log-odds that stays finite when all answers agree. Proposals, ablations, Δ_R, and the heatmaps are unchanged. We validate the estimator on Qwen3-VL-8B, whose log-probabilities serve as ground truth, with $n = 20$, then apply it to Gemini 3.8 Flash, which returns no log-probabilities.

Scores. On unmasked images from 200 randomly drawn NaturalBench groups (152 after removing multiple-choice questions, 608 image-question pairs), $\hat{s}_0$ correlates with s_0 at Pearson 0.996 and gives the same answer on 98.9% of pairs; NaturalBench accuracy is 79.0% sampled versus 79.4% from log-probabilities (sampled numbers averaged over 20 random subsets of 50 draws).

Heatmaps. We score the eight white-box images of Sec. 6.2 that pose yes/no questions (the other two are multiple-choice) with both estimators, ablating bounding boxes, and paint each Δ_R on its SAM mask. Fig. 23 shows five of the 16 image-question pairs. Sampled maps keep the regions that change the answer and drop the rest. In all 16 pairs, the region whose ablation moves the log-probability margin furthest toward the opposite answer is among the sampled map's top-scoring regions, a set of 2 regions at the median and at most 4; it is the unique top region in 6 pairs. The full per-region ranking agrees less (median Spearman $\rho = 0.58$).

The cause is the resolution of 20 answers. A region whose ablation flips all 20 answers scores the extreme $|\Delta\hat{m}| = 7.4$ however far it moves the margin, so regions that each flip a confident answer tie; a region that leaves all 20 answers unchanged scores 0 whatever it does to the margin. This is the saturation of App. A in another form: sampling estimates s_0, not m_0. More samples resolve regions near the decision boundary but cannot rank regions that each flip the answer. For auditing a failure case, where the question is which region changes the answer, the sampled map suffices at 20 answers per query instead of one log-probability read-out.

Gemini 3.8 Flash. We run the same pipeline on Gemini 3.8 Flash through its Batch API at temperature 1 with low thinking effort, one answer per request because the API rejects multiple candidates. The five pairs of Fig. 23 need 88 scored images, i.e. 1,760 requests, which completed in about four minutes for US$1.13. With no log-probabilities to validate against, we read these maps qualitatively (Fig. 23, right column). On the jacket and beanie questions, Gemini's maps pick the same regions as Qwen's. On the flag question, its top region is the whole flag rather than the cross. On the relation question, Gemini is less confident (18 of 20 answers Yes), so its map is graded rather than binary: the couple is positive evidence and many background regions are weak negative evidence, the resolution near the decision boundary that sampling does provide.

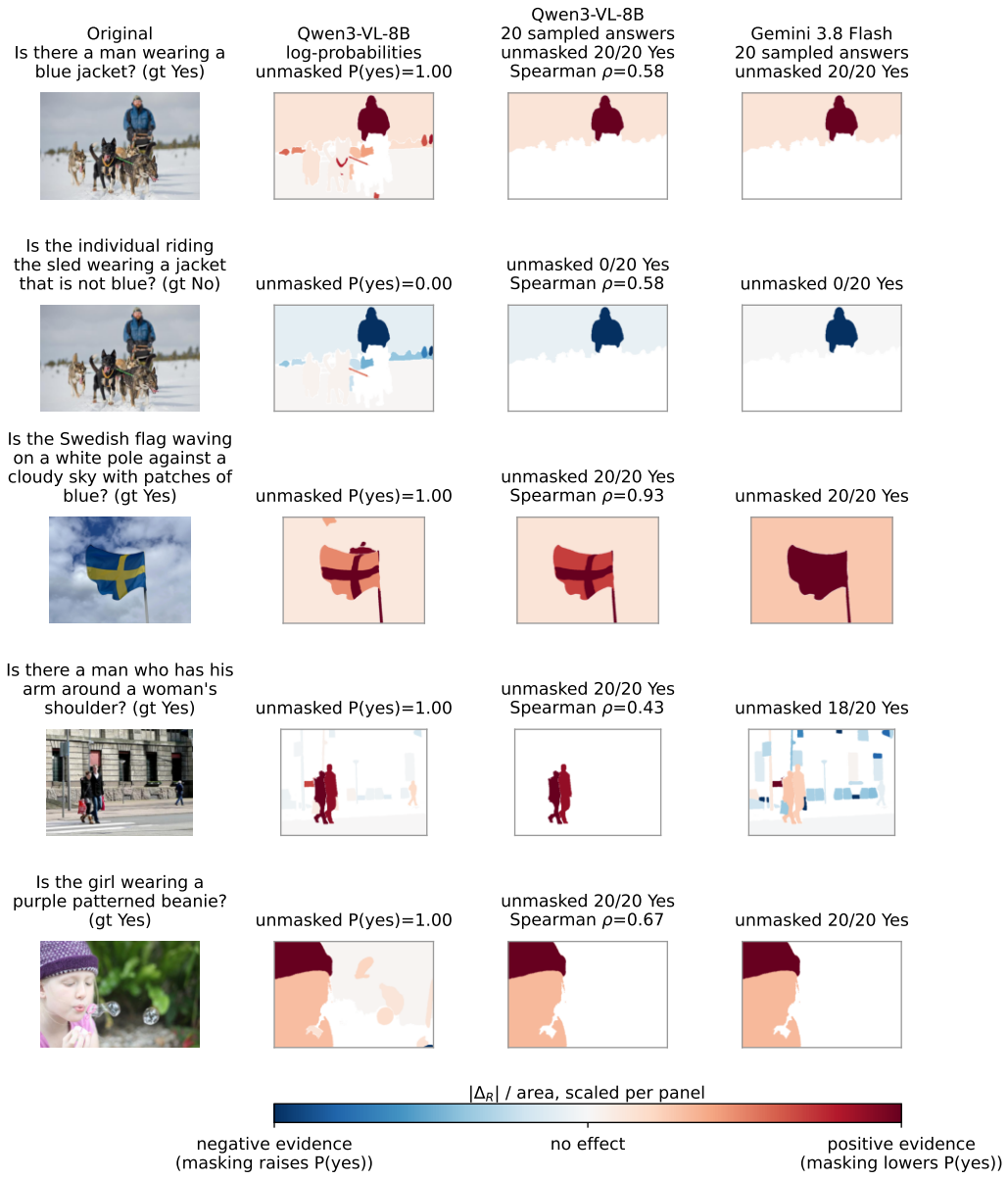

Figure 23: **Sampled answers recover the regions that change the answer, which lets SAM-RISE run on a closed model.** The same bounding-box ablations scored on Qwen3-VL-8B from log-probabilities (second column) and from 20 sampled answers (third), and on Gemini 3.8 Flash from 20 sampled answers (fourth); each Δ_R is painted on its SAM mask with intensity $|\Delta_R|$ per unit area, scaled per panel. Red (positive evidence): masking the region lowers $P(\text{yes})$; blue (negative evidence): masking raises it; white: no effect. Titles give $P(\text{yes})$ on the unmasked image (sampled: Yes answers out of 20) and, for Qwen's sampled map, the Spearman ρ against its log-probability map; Gemini has no log-probabilities to compare against. Sampling drops regions that shift the margin without changing the answer (background in the first two rows) and keeps the jacket, which is positive evidence for "blue jacket?" and negative evidence for "jacket that is not blue?"; Gemini's maps agree with Qwen's on these rows.

P Reproducibility Details and Hyperparameters

The code runs in three phases per pipeline: `mask` (region proposals per scheme), `score` (Yes/No log-probabilities for every $(I \setminus R, q)$ against an in-process model or any OpenAI-compatible endpoint), and `analyze` (Δ_R, $\tilde{s}_0$, H, and population aggregates). The saliency benchmark adds scripts to build the evaluation sets, run D-RISE with budget snapshots, rebuild weight variants and budgets offline, run insertion and deletion, and evaluate with bootstrap CIs and McNemar tests.

Hyperparameters. SAM 3 everything mode, single pass: 32 points per side, predicted-IoU 0.70, stability 0.85, NMS IoU 0.7, area fraction $[0.001, 0.95]$, $|\mathcal{M}| \leq 64$. DINO proposals: DINOv2 patch features, PCA to 16 dimensions, KMeans with 6 clusters. Maximum image side: 512 px for the NaturalBench release run and the D-RISE benchmark, 768 px for the corrected NaturalBench run and the ImageNet population runs, 448 px for the white-box comparison. Adjusted scores use the proportional variant with $k = 5$. Decoding in the D-RISE benchmark: temperature 0, one output token, top-20 log-probabilities. NaturalBench, population, and D-RISE runs query a vLLM-served Qwen3-VL-8B-Instruct; the single-image operator ablations and white-box baselines run in-process.