OBJECT ABSENCE IN VISION-LANGUAGE MODELS IS A CRITERION SHIFT MORE THAN A BLIND SPOT

Anonymous authors
Paper under double-blind review

ABSTRACT

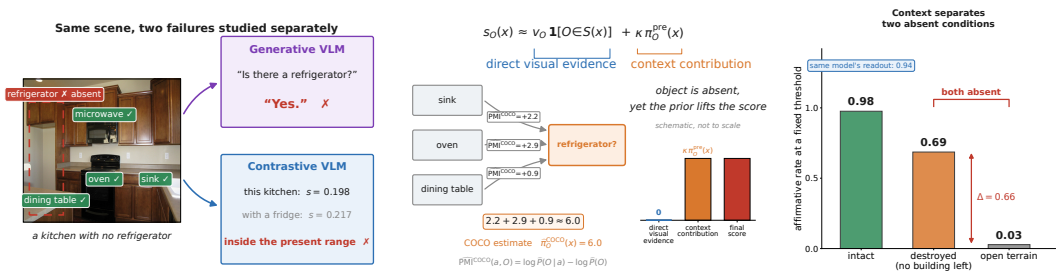
Object hallucination in generative vision-language models and failures to recognize object absence in contrastive encoders are usually studied separately and attributed to different causes. We show that both are better understood by separating two quantities that standard evaluations conflate, sensitivity to whether an object is present and a context-dependent shift of scores and decisions toward affirmation. Across five contrastive encoders and seven generative checkpoints, signal detection analysis shows that the two failures share a per-concept susceptibility axis whose stable component is the affirmative shift. Context raises absent-object scores in all fifteen encoder-by-specification fits, also when present and absent images are matched to receive the same context change, and in all seven generative checkpoints, whose free-form descriptions invent the objects that context predicts. We claim sensitivity invariance for neither family, since the sensitivity change depends on how the strata are cut and on the confusability proxy. Where sensitivity falls, the evidence remains in the representation. At matched context, a linear probe fitted on an encoder's embeddings in low-expectedness scenes separates present from absent images in high-expectedness scenes at 0.83 to 0.85 AUC, against 0.75 to 0.78 for the encoder's own zero-shot score. Negated queries, the usual evidence for absence blindness, stay highly correlated with affirmative ones and rank absent images below chance in all five syntaxes we test. On post-disaster satellite imagery, where city blocks at least ninety-five percent destroyed and open terrain both lack a building, a model affirms one on 69 percent of the former and 3 percent of the latter, although a linear probe on the hidden state its answer is read from separates intact from destroyed blocks of the same disaster and capture at 0.94 AUC. Object absence failures in these models are thus a criterion shift more than a blind spot, one that the context-conditioned threshold corrections we test cannot remove without sacrificing presence signal, and evaluations should report score movement, threshold behavior, sensitivity, and proxy dependence separately rather than accuracy alone.

1 INTRODUCTION

Show a vision-language model a photograph of a kitchen from which the refrigerator has been removed, and two things go wrong at once. Asked whether the image contains a refrigerator, a generative model tends to say yes, and the error grows with how strongly the scene suggests one (Li et al., 2023; Leng et al., 2024). Asked to retrieve that image among kitchens that do contain refrigerators, a contrastive encoder struggles to tell them apart, whether the query names the missing object or negates it (Yuksekgonul et al., 2022; Alhamoud et al., 2025). The first behavior is called object hallucination and is attributed to language priors and spurious co-occurrence. The second is called negation or absence insensitivity and is attributed to missing negative supervision in web image-text data. Each has its own diagnosis and its own remedy, and the two rarely cite each other.

We argue that both literatures measure a common context-dependent susceptibility axis, and we ask how that axis enters object-absence judgments. Two components may coexist, and they differ in kind. Under a representational-degradation component, a blind spot for absence, the representation itself separates present from absent images less well once context suggests the object, so the distinction is not there to be read out. Under an affirmative-bias component the representation still separates the classes, but the score or decision read from it sits closer to affirmation, so a threshold

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.



(a) Two failures, one scene. (b) A first-order decomposition. (c) Context separates two absences.

Figure 1: Overview. Panel (a) shows one COCO kitchen without a refrigerator eliciting both symptoms, and panel (b) the decomposition into target evidence and an additive context prior. Panel (c) shows the behavioral consequence on post-disaster satellite tiles. A generative model says a building is present on 0.69 of blocks at least ninety-five percent destroyed, where no undamaged building is left but the street grid survives, and on 0.03 of open terrain. A linear probe on the hidden state its answer is read from still separates intact from destroyed tiles of the same disaster and capture at 0.94 area under the curve.

that once rejected an absent image now accepts it. In the vocabulary of signal detection the first lowers sensitivity and the second moves the criterion. These are not two names for one thing, because a decision variable can separate two classes less well than the representation it is read from, and Sections 5 and 6 measure exactly that case. Nor is the difference merely terminological, since hard negatives and negated captions are meant to supply a distinction the model does not yet encode, whereas a criterion needs no new distinction, only a different reading of the one already present, so the two components imply different targets for repair. The empirical question, sketched in Figure 1, is whether affirmative movement occurs and whether any accompanying sensitivity change is negative, null, or positive. If context adds the same increment to present and absent scores and leaves their spread unchanged, only the criterion moves. If the increment differs by class, or the spread changes, sensitivity moves as well, in either direction.

Calling the second component a bias requires a standard, because context is genuine evidence about presence. It separates the classes by four units of summed log co-occurrence inside a single expectedness stratum, and a model that weighs it is not thereby irrational. We call a response biased when two conditions hold. First, the evidence that would settle the question is recoverable from the same representation the decision is read from, so the model is not missing the information. Second, the movement does not separate the two classes, since an update that carried target information would raise present scores more than absent ones. Section 5 establishes both conditions for the encoders on natural images, and Section 6 tests the first for a decoder far from them. Magnitude cannot serve as the standard. Under every summary we compute, the observed movement is several times smaller than the one an ideal observer holding the corpus base rates would make, while the balanced design we measure it in justifies none at all.

That context drives these failures is not in dispute. Recent work already suppresses the prior by editing weights or attention (Yang et al., 2025; Chen et al., 2025), but the accuracy it is scored by cannot show which component moved. We make three contributions.

- The two failures share a common axis (Section 4). Across 77 COCO categories a hallucination index and a high-prior probe-error index correlate at a mean rank correlation of 0.61 over four of our five encoders, the association transfers to the seven generative checkpoints and an eighth measured only there, and correcting both for measurement noise leaves it at 0.60 to 0.71, so the axis is shared rather than single.
- Affirmative movement is the stable component of that axis (Section 5). It appears under specifications built from three detectors, under strata that give both classes the same context change, in free-form descriptions, and under a controlled intervention, whereas the sign of the sensitivity change depends on the design and the proxy.
- The evidence the decision misses is still in the representation (Sections 4 to 6). A probe on an encoder's embeddings separates the classes where its zero-shot score loses ground,

negated queries are reversed rather than insensitive, and on post-disaster imagery a probe on the hidden state a decoder answers from beats the answer itself.

A first-order model of the score ties the three together and explains why the context-conditioned threshold corrections we test cannot remove the shift without also removing presence signal. What holds under every specification we state as a finding, what depends on a design or a proxy we state with it named, and what one checkpoint or domain supplies we call a sign test.

2 Related work

The POPE protocol queries models about absent objects sampled at random, by frequency, or adversarially by co-occurrence (Li et al., 2023), extending earlier caption-level measurements (Rohrbach et al., 2018), and visual contrastive decoding is a representative mitigation (Leng et al., 2024; Zhou et al., 2024), which we measure inside our design. Recent work treats the language prior as an established cause and intervenes on it directly (Yang et al., 2025; Chen et al., 2025), so that diagnosis is common ground rather than our claim. On the contrastive side, benchmarks report near-chance handling of negated queries and insensitivity to what is missing (Yuksekgonul et al., 2022; Alhamoud et al., 2025), which motivates a representational account and remedies built from negated captions or hard negatives. Neither literature routinely decomposes its failure into sensitivity and criterion components, and neither joins the two families on a common per-concept axis. Their mitigations remove the prior implicitly rather than estimating and subtracting it, which places the double dissociation of Section 5.2 out of their reach. Our adjustments follow the long-tail recipe of subtracting a log prior from logits (Menon et al., 2020) and its zero-shot extensions (Zhu et al., 2023; Peng et al., 2026), used here as diagnostic controls, while Section 5.2 estimates that term from the encoder's own scores rather than from a class prior. Signal detection theory has long separated discriminability from response bias (Green et al., 1966), so our contribution is not the vocabulary but a design that keeps sensitivity and proxy dependence visible where accuracy would collapse them. Appendix A extends the survey to human scene-context research and to model coverage.

3 Two components, one instrument, and a first-order model

3.1 Setup and measured quantities

Let O range over object categories and x over images with unit-norm embedding $f(x)$. Let t_O be the unit-norm text embedding of O from a fixed ensemble of affirmative prompt templates, $s_O(x) = \langle f(x), t_O \rangle$ the similarity score, and $S(x)$ the set of categories annotated as present in x. From corpus statistics we define the conditional co-occurrence $P(O \mid a)$ as the fraction of images containing category a that also contain O, and the expectedness

$$E_O(x) = \frac{1}{|S(x) \setminus \{O\}|} \sum_{a \in S(x) \setminus \{O\}} P(O \mid a), \tag{1}$$

a context-only quantity because it never consults whether O is present. For each category we partition images by the presence of O and by terciles of E_O into cells, writing $p \in \{\mathrm{lo}, \mathrm{hi}\}$ for the prior level of a cell. Class labels and estimator identities appear as superscripts and the concept and prior level as subscripts, so $\mu^{\mathrm{P}}_{O,p}$ and $\mu^{\mathrm{A}}_{O,p}$ are the mean present and absent scores of O at level p, and omitting the subscript denotes pooling, as Appendix B details.

Two indices carry each literature's phenomenon into the other's model class, so that hallucination is measured on a contrastive score and absence blindness as a probe error, namely

$$H(O) = \frac{\mu^{\mathrm{A}}_{O,\mathrm{hi}} - \mu^{\mathrm{A}}_{O,\mathrm{lo}}}{\mu^{\mathrm{P}}_{O} - \mu^{\mathrm{A}}_{O,\mathrm{lo}}}, \qquad B(O) = 1 - \mathrm{AUC}_{O,\mathrm{hi}}. \tag{2}$$

The hallucination index measures how far contextual expectation alone pushes absent-image scores toward genuinely present ones, so that $H(O) = 1$ would mean expectation fully impersonates presence. Its denominator is a per-concept scale, so a concept whose two classes nearly coincide yields an unstable, occasionally negative ratio. Every claim we build on H therefore uses rank statistics,

and dropping the fifth of concepts with the smallest denominators still leaves Section 4's association at 0.471 to 0.574. The high-prior probe-error index uses the cross-validated area under the receiver operating characteristic curve of a logistic probe on image embeddings separating present from absent images inside the high-expectedness stratum. It operationalizes what prior work calls absence blindness, since a large value means that even a probe on the full embedding cannot separate the classes once the scene expects the object, without identifying its source. Because $H(O)$ reads only the projection of $f(x)$ onto t_O while $B(O)$ uses the full embedding, we also write $B_\perp(O)$ and $B_\parallel(O)$ for probes on the component of $f(x)$ orthogonal to t_O and on that projection alone, since a correlation between H and $B_\perp$ cannot arise from reading one projection twice.

Signal detection separates two questions that accuracy merges. Sensitivity asks how well the score separates the two classes, and the criterion asks where along that score the boundary between yes and no falls. For concept O at prior level p we define the raw class separation $D_{O,p} = \mu^{\mathrm{P}}_{O,p} - \mu^{\mathrm{A}}_{O,p}$ and the standardized sensitivity $d'_{O,p} = D_{O,p}/\sigma_{O,p}$ with $\sigma_{O,p}$ the pooled standard deviation. We take the criterion to be the equal-error threshold in a prior-independent standardized frame whose origin is the per-concept grand mean of present scores, and a class's position to be its mean score in that frame. The origin is one constant subtracted from both class positions, so it cancels from d' and constrains neither the size nor the sign of the criterion's movement. A degradation component lowers d' whereas an affirmative component raises absent-object scores, so sensitivity is an outcome we report, not our defining null. A contrastive encoder has no threshold of its own. Its criterion is therefore an observable, a context-conditioned translation of both class distributions along the score, not a rule inside the model, so what we claim for the encoders is its downstream cost. Any threshold held fixed across scenes inherits that translation as error, whatever rule placed it, which Section 5 prices at a threshold fitted before the context changes.

3.2 A FIRST-ORDER MODEL OF CONTEXT-ASSOCIATED SCORE SHIFT

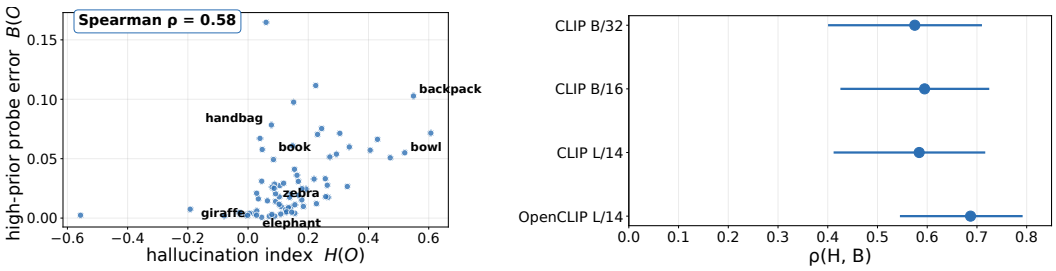
Rather than positing a bias, we ask how an additive context prior could arise from contrastive training. Three assumptions, stated formally in Appendix C, carry the argument. Image-text inner products recover pointwise mutual information up to an increasing affine transformation at the population optimum of the contrastive objective (Oord et al., 2018). Image embeddings compose approximately additively over scene constituents as $f(x) \approx \sum_{a \in S(x)} w_a u_a + \varepsilon$, with unit direction u_a, salience weight $w_a \geq 0$ per category, and residual ε (Trager et al., 2023). For recurring scene populations the cross term $\langle u_a, t_O \rangle$ increases approximately affinely with $\mathrm{PMI}^{\mathrm{pre}}(a, O)$, the association between observing a and mentioning O in the model's pretraining distribution. The first is asymptotic for the softmax objective, and the third cannot be verified without model-specific pretraining pairs, so we test the observable consequences instead, and one fails, since an encoder's own pretraining corpus does not predict its scores best (Appendix G). We accordingly read π^{pre}_O as a first-order description of an additive context component that earns its place by what it predicts, not as an identified pretraining quantity. The latent context prior and the score it implies are then

$$\begin{aligned} \pi^{\mathrm{pre}}_O(x) &= \sum_{a \in S(x) \setminus \{O\}} \mathrm{PMI}^{\mathrm{pre}}(a, O), \\ s_O(x) &\approx v_O \, \mathbf{1}[O \in S(x)] + \kappa \, \pi^{\mathrm{pre}}_O(x) + \varepsilon_O(x), \end{aligned} \tag{3}$$

where v_O collects target evidence, $\kappa > 0$ is a model-specific scale, and $\varepsilon_O(x)$ absorbs residuals together with an affine offset growing in the number of context objects, which the context-prior regressions carry as a covariate. We estimate it by $\widehat{\pi}^{\mathrm{COCO}}_O$, $\widehat{\pi}^{\mathrm{VG}}_O$, and $\widehat{\pi}^{\mathrm{CC}}_O$, replacing $\mathrm{PMI}^{\mathrm{pre}}$ with the image co-occurrence tables of COCO and Visual Genome and the caption co-mention table of Conceptual Captions. The stratification statistic E_O and $\widehat{\pi}^{\mathrm{COCO}}_O$ summarize the same estimated table, related term by term through $\widehat{\mathrm{PMI}}^{\mathrm{COCO}}(a, O) = \log \widehat{P}^{\mathrm{COCO}}(O \,|\, a) - \log \widehat{P}^{\mathrm{COCO}}(O)$.

The model implies that a class-common increment from the low to the high stratum leaves raw separation at $D_{O,\mathrm{hi}} = D_{O,\mathrm{lo}} + o(1)$ and shifts the equal-error criterion by that increment, whereas $d'_{O,\mathrm{hi}} = d'_{O,\mathrm{lo}} + o(1)$ further requires $\sigma_{O,\mathrm{hi}} = \sigma_{O,\mathrm{lo}} + o(1)$, with $o(1)$ the first-order remainder.

Section 5.2 requires one further identity. Let $\widehat{\pi}^{e}_O(x)$ be a context-prior estimate computed from content excluding O, with superscript e identifying its estimator, and inside a fixed evaluation cell write δ_O for the class gap between its means on present and on absent images. For the adjusted score



(a) Association on CLIP ViT-L/14.

(b) Rank correlation per encoder.

Figure 2: The two failure indices share a per-concept axis, with animals in the lower left of panel (a), context-bound objects in its upper right, and per-encoder intervals in panel (b).

$\tilde{s}_O(x) = s_O(x) - \gamma\widehat{\pi}^e_O(x)$ with strength $\gamma \geq 0$, the raw separation obeys

$$\tilde{D}_O(\gamma) = D_O(0) - \gamma\delta_O, \tag{4}$$

so subtraction preserves raw separation when $\delta_O = 0$ and reduces it linearly in γ otherwise, with the same first-order statement for standardized sensitivity when pooled dispersion is locally stationary. The two conditions speak to different quantities. The common-increment condition compares how far each class moves between strata, whereas Equation 4 compares the level of the estimated prior across classes inside one stratum. Natural scenes can satisfy the first while violating the second, because a scene holding an object also holds its companions, so the estimated prior runs higher on present images than on absent ones. The context prior therefore acts as both bias and evidence.

4 The two failures share a common axis

Protocol. We use the 2017 COCO detection set, keeping 77 of the 80 categories after a cell-size filter, with cells of up to 500 images, a fixed seed, and four encoders, CLIP ViT-B/32, ViT-B/16, and ViT-L/14 (Radford et al., 2021) and OpenCLIP ViT-L/14 on LAION-2B (Schuhmann et al., 2022). CoCa ViT-L/14 (Yu et al., 2022) joins from Section 5 on, so four-encoder statements exclude it, and CLIP ViT-L/14 is the anchor for single-encoder results. The two strata are formed inside each class rather than from one shared cut, with absent images split at the terciles of E_O over that concept's absent pool and present images at those over its present pool. These within-class strata balance the classes in every cell, at the cost of the two high strata not being one absolute expectedness bin, a cost Section 5 removes by matching. Every contrast below is read within a concept.

The shared axis. Across the four encoders the rank correlation between H and B runs from 0.575 to 0.687, every level below 10^{-7}, as Figure 2 shows. It is not a detectability artifact, since partialling out low-stratum baseline detectability, object area, and category frequency leaves it unchanged, and so does adding the index's own denominator, itself a detectability scale, to those controls. Nor is it one projection read twice, since the index recomputed orthogonally to t_O still predicts hallucination after partialling out $B_{\|}$. It also crosses model families, because a decoder's false-affirmation rate on high-expectedness absent images predicts $B_{\perp}$ at 0.500 to 0.637 across LLaVA-1.5 (Liu et al., 2023) and the seven checkpoints of Section 5.1, every level below 10^{-5}. Put plainly, the concepts a contrastive encoder struggles with in expected scenes tend to be the ones a generative decoder most often hallucinates there. The axis is shared rather than single. Were H and B noisy readings of one quantity, correcting for their split-half reliabilities, which exceed 0.92, would carry the correlation toward one, and it stays well short.

The negated instrument is reversed rather than near chance. The absence literature scores negated text, so we score a negated ensemble on the same images and keep it outside every other measurement. It ranks absent images above present ones, at an area under the curve of 0.109 to 0.145 against chance at 0.500, while the affirmative query reaches 0.875 to 0.895. An encoder scored this way is not insensitive to absence but confidently wrong about it, and the fault is the instrument's. The same context prior governs the negated score, which agrees with the affirmative one at 0.93, so that agreement is a consistency check, not an independent replication, and the reversal is not an artifact

of one syntax, since five constructions each stay below chance in every encoder. A checkpoint fine-tuned on negated captions separates the two things this instrument confounds. Its negated score's agreement with the affirmative one falls from 0.914 in its own base to 0.201 and its area under the curve rises from 0.127 to 0.452, so training does make the word act on the query. Yet the value stays at or below chance and the context effect is unmoved, the criterion change differing from its base by $+0.010$ [-0.015, $+0.037$]. Per-encoder results for this section are in Appendices D, E, S, and T.

5 THE CRITERION SHIFT IS THE STABLE COMPONENT ON COCO

Section 4 showed that the two failures rank concepts alike, and this section asks which component of that shared ranking is stable. In the encoders we adjust for confusability, give both classes the same context change, read the representation directly, and intervene on the prior, and Sections 5.1 and 5.2 turn to the decoders and to a test that subtracts the terms the account names.

Confusability as the rival reading. The shared axis admits more than one reading. A concept can sit high on it because context moves its score, which is the bias we are after, or because its target evidence is weak, as for small or visually confusable objects, which is not bias. We therefore adjust for two externally trained measurements of confusability, treated as proxies rather than ground truth, the maximum class confidence of a COCO-trained DETR detector (Carion et al., 2020) and the maximum box-token affinity over the span naming O from Grounding DINO (Liu et al., 2024), which differs in architecture, supervision, and corpus. Over five checkpoints and three specifications the present and absent positions both rise with the prior, significantly in all fifteen fits, and the absent-class coefficient survives a within-concept specification in all ten refits at $+0.19$ to $+0.39$. These are agreement counts over a fixed checkpoint set rather than a search over tests, as Appendix B discusses, and Appendix N tabulates the coefficients.

The prior predicts the score. Equation 3 predicts that absent-object scores rise with the estimated context prior. They do, at significance below 10^{-31} in every checkpoint after adjusting for detector response and the number of context objects, and the slope survives rebuilding the table on Visual Genome, Conceptual Captions, or web alt-text and re-forming the strata from Visual Genome, so the statistic that partitions the images need not be the one that predicts the score (Appendix G).

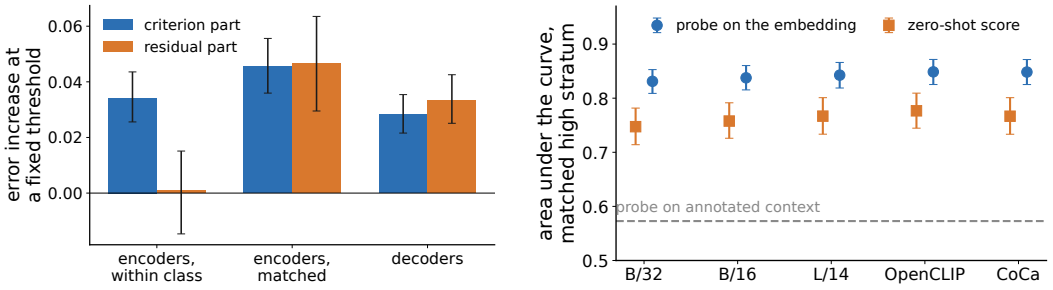
Matching the context change. In the standardized frame of Section 3, not at the answer threshold of Figure 1c, the unadjusted anchor present and absent positions rise by $+0.274$ and $+0.271$ from the low to the high stratum. Within-class strata ask how each class responds across its own range of scenes, however, not how the two respond to one context change. Present images vary more in expectedness than absent ones, so on E_O the present class receives about twice the absent class's change, and near-equal movements cannot show that context moves the classes together. We therefore also match the two classes on the whole annotated corpus. Both share one set of absolute expectedness bins, and absent images are reweighted to the present images' distribution over them, so the change from low to high is one intervention for both (Appendix J). The criterion still rises, by $+0.484$ [$+0.407$, $+0.562$] over the five encoders, with every encoder's interval excluding zero. What the matching changes is the company the shift keeps. At matched context absent scores rise about twice as far as present ones, so the zero-shot score also separates the classes less.

The criterion is stable and sensitivity is not. Table 1 sets the two components side by side, and the comparison needs no confusability proxy. The criterion rises in every row. At a threshold fitted on the low stratum and carried to the high one, the part of the error increase that translating both classes reproduces is likewise positive with an interval excluding zero in every design, as Figure 3a shows, while the residual is indistinguishable from zero only for the encoders within class. The sensitivity change is design dependent. On the within-class strata the encoders' per-concept change carries no systematic sign, with no signed-rank test significant in any of the five, whereas at matched context it falls in all five, as the decoders' does in six of seven on the within-class strata.

The representation keeps the evidence. Where sensitivity falls, what the zero-shot readout loses is still available in the representation it reads, the distinction Section 1 draws between a decision variable and the representation it is read from. A linear probe on the frozen embedding, fitted on

Table 1: The criterion and sensitivity components under three designs. The within-class design cuts each class at its own expectedness terciles, and the matched design gives both classes shared absolute bins. The criterion change is pooled with a 95 percent concept-bootstrap interval. For the per-concept sensitivity change we give the checkpoints whose signed-rank test is significant and the range over checkpoints of the share of concepts beyond 0.15 standardized units below or above zero.

	Criterion	Significant	Share of concepts	
Design	change	tests	below -0.15	above $+0.15$
Encoders, within class	$+0.312$ $[+0.249, +0.380]$	0 of 5	0.338 to 0.416	0.364 to 0.519
Encoders, matched	$+0.484$ $[+0.407, +0.562]$	5 of 5	0.561 to 0.591	0.212 to 0.242
Decoders, within class	$+0.340$ $[+0.281, +0.399]$	6 of 7	0.455 to 0.766	0.143 to 0.429



(a) Error at a carried threshold.

(b) Readout and representation.

Figure 3: The criterion part is stable and the embedding keeps the evidence. Panel (a) splits the error increase at a threshold fitted on the low stratum and carried to the high one into the part that translating both classes reproduces and the residual, with 95 percent concept-bootstrap intervals. Panel (b) shows, on matched strata, the high-stratum area under the curve of an embedding probe fitted in the low stratum and of the zero-shot score, with the annotated-context probe dashed.

low-expectedness scenes and read without refitting on high-expectedness ones, separates the classes at 0.904 area under the curve for the anchor on the within-class strata, against 0.528 for a probe on the annotated context. On the matched strata it reaches 0.831 to 0.849 across the five encoders, while the zero-shot score reaches 0.747 to 0.777 on the same images, a gain of $+0.072$ to $+0.084$ with every interval excluding zero. The embedding therefore carries the target across the context change where context alone does not, as Figure 3b shows. It holds more evidence of the object than its similarity to the object's name reveals, because that similarity also rises with what the scene suggests. Both conditions of Section 1 hold for the encoders, the first by this probe and the second because context nowhere raises present scores detectably more than absent ones.

An intervention on the prior. A causal test raises the prior while holding the target's evidence fixed. Pasting the top co-occurring context object into absent scenes appears to lower sensitivity, but the per-concept change tracks the pasted category's text similarity to t_O, at a rank correlation of -0.797, because high co-occurrence and visual similarity are entangled in natural categories. Manipulating similarity at the instance level decouples the two. In a factorial crossing target presence, a pasted context prior, and the crop's alignment to the target text, raising the prior at low similarity moves the criterion by $+0.052$ $[+0.017, +0.087]$ while the sensitivity effect stays inside the pre-registered equivalence band at -0.061 $[-0.147, +0.025]$. The absent cell carries most of the inflation, so the reading is bias-dominant for the anchor rather than universal (Appendix M).

5.1 The affirmative shift appears in generative decisions

We next measure the decomposition in seven generative checkpoints. The decision variable is $z_O(x) = \ell_{\text{yes}}(O, x) - \ell_{\text{no}}(O, x)$ from one forward pass on a positive presence question, read in the same frame, and because Equation 3 is derived for contrastive scores it applies to answer logits descriptively. The unadjusted equal-error criterion rises in all seven checkpoints, from $+0.146$ for Qwen2-VL to $+0.509$ for InternVL3.5 (Zhu et al., 2025), and the shift is class asymmetric in every one, the absent positions rising by $+0.295$ to $+0.834$ while no present position rises above $+0.183$.

Unlike an encoder, a decoder has a threshold of its own, the forced choice at $z_O(x) = 0$. At that threshold false-affirmation rates rise in all seven checkpoints while hit rates improve, and the absent-position coefficient excludes zero under DETR in all seven and under both proxies in five. Sensitivity again depends on the specification. Unadjusted, the decision variable separates the classes less at high prior in all seven, but after confusability adjustment five of seven coefficients stay significantly negative under DETR against one under Grounding DINO, and a third detector, OWLv2, carries their mean from -0.502 and -0.270 to $+0.129$. The absent-position coefficient stays positive under all three, so affirmative movement is the specification-invariant result here too.

The forced choice manufactures none of this. In one-sentence descriptions of the same images, the absent categories a decoder invents are four to seven times as contextually expected as those it leaves out, at fixed caption length. Appendix P has every checkpoint.

5.2 Subtracting the account's own nuisance terms

Equation 3 says that a presence decision must read $\mathbf{1}[O \in S(x)]$ off a score carrying two nuisance terms the encoder can estimate for itself. That makes the account falsifiable in a direction accuracy cannot reach, since terms the score does not carry could not lower the bias index and raise discriminability at once. Write $\widehat{\pi}_O^{\text{bank}}(x) = \langle f(x), c_O \rangle$ for a companion term whose direction $c_O = \sum_{a \neq O} \omega_a t_a$ weights each prompt by the positive part of the corpus association table of O, and $\bar{s}_{\neg O}(x) = \langle f(x), r_O \rangle$ for the unweighted mean over the same bank. The adjusted score $\tilde{s}_O(x) = s_O(x) - \beta_1 \widehat{\pi}_O^{\text{bank}}(x) - \beta_2 \bar{s}_{\neg O}(x)$ is then the single query direction $t_O - \beta_1 c_O - \beta_2 r_O$, so it needs neither training nor test-image annotation. In words, the rewrite removes from the query what the object's usual companions contribute to its score and what every query in the bank shares.

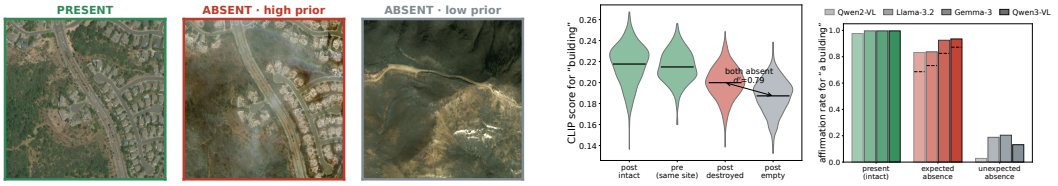
On held-out images the untuned setting $\beta_1 = \beta_2 = 0.5$ raises the area under the curve for present against high-prior absent images by $+0.017$ to $+0.023$ with every interval excluding zero, while the bias index falls by 68 to 83 percent in all five encoders. The reference term raises discriminability, the companion term lowers the bias index, and only the pair moves both. Under one shared threshold the pair halves the false-alarm gap between the two absent strata in all five encoders, from $+0.105$ to $+0.054$, an equalization rather than a reduction, since the pooled false-alarm rate rises in four of the five. A per-concept probe with target labels detects better, at pooled error 0.106 against the adjustment's 0.323, and the scalars do not transfer across domains, so the joint movement is evidence for the decomposition and deployment belongs to a probe (Appendix Q).

Threshold corrections trade the shift against evidence. Subtracting an estimated prior from the score, as in Equation 4, behaves differently. The annotation-based estimate $\widehat{\pi}_O^{\text{ann}}(x)$ averages $+3.84$ on present against -0.19 on absent images inside the high stratum, so the context prior predicts presence, and raising γ lowers the anchor's sensitivity from 2.175 at zero strength to 1.985 at the strength a development split selects, while false alarms rise at a matched miss rate. The same adjustments still identify which bias is at work. On the POPE popular split (Li et al., 2023), whose negatives are sampled by marginal frequency, subtracting a constant log class prior gains 6.0 accuracy points and the conditional adjustment nothing, while on its adversarial split, sampled by co-occurrence with the depicted context, the conditional adjustment gains 2.9 points and the constant one nothing. That double dissociation separates a marginal-frequency bias from a context-prior bias, and Appendix K gives it with intervals and for the decoders.

6 An out-of-distribution sign test on post-disaster imagery

We test the account on xBD post-disaster overhead imagery, a domain none of these models was built for. Querying the concept building, we treat destroyed tiles as expected absence, open terrain as unexpected absence, and each site's pre-disaster image as the present counterpart, as Appendix X details. Figure 4 shows the tiles and the four checkpoints that rarely affirm a building on a blank input, and the anchor encoder separates destroyed from empty tiles by 0.79 sensitivity units.

Here the word bias earns its place, since weighing context is defensible when a surviving street grid raises the odds that a building stands. For a decoder we test the first condition of Section 1, recoverability from the same representation, on this one concept and domain. On tiles at least ninety-five percent destroyed against intact ones from the same events and capture, a probe on the final hid-



(a) The three tile conditions.

(b) Contrastive score and affirmative rate.

Figure 4: Panel (a) shows a block before a wildfire, the same block after it, whose surviving street grid makes the absence expected, and open terrain, where it is not. In panel (b) destroyed tiles, though absent, score far above empty ones, bars for tiles at least seventy percent destroyed and dashes for the strict subset at least ninety-five percent destroyed that the abstract and Figure 1 quote.

den state Qwen2-VL's own answer logits are read from reaches 0.941 area under the curve with whole events held out, against 0.808 for those logits' difference, a gap of $+0.133$ $[+0.026, +0.211]$. Qwen2-VL affirms a building on 0.69 of those tiles against 0.03 of empty ones, and on Qwen3-VL-8B the probe again beats the logits, at 0.969 against 0.930 area under the curve, a gap of $+0.039$ $[+0.014, +0.049]$. Capture date is a confound, since undamaged tiles separate by date alone at 0.735 and 0.788, so the comparison holds the capture fixed.

The language-model head is linear, so the probe and the answer are two linear readouts of one vector, and what fails is the readout the model applies, not the availability of the evidence. The probe is supervised while the head is zero shot, so this is a sign test of the first condition, not a proof that the head misplaced its criterion, and where the head already ranks well the gap closes, to $+0.021$ for Qwen2-VL on COCO and to nothing for Qwen3-VL-8B (Appendix Y). The second condition holds as on COCO, since in no checkpoint does context raise present positions as far as absent ones.

Here a damage score predicts absence, not presence, which flips the sign of Equation 4's trade-off, so subtraction raises raw separation and cuts Qwen2-VL's destroyed-tile false-alarm rate by 0.100 $[0.045, 0.155]$ at matched recall while the open-terrain control holds, though rates stay high.

7 Discussion and Conclusion

The two components call for different repairs. A blind spot needs a distinction the representation lacks, which negated captions are meant to supply, yet our encoders' embeddings already carry the target across the context change, and the negation-trained checkpoint we test leaves the context effect unmoved. A criterion shift needs a different readout. Moving the threshold with context cannot supply one, because in natural scenes context is also evidence of presence, so each context-conditioned threshold correction we test gives up presence signal with the shift. Changing the decision variable does better, as the encoders' query rewrite of Section 5.2 lowers the bias index while raising discriminability, but it only equalizes error across scenes, and a probe on a decoder's own hidden state loses its operating point across the context change as the answer does (Appendix Y). The diagnosis therefore points repair at the readout rather than the representation.

Four qualifications bound what we claim. Sensitivity is not preserved, since its sign depends on how the strata are cut, on the proxy, on the intervention, and on the corpus, and the decoders were measured on within-class strata only. Equation 3 earns its place through its consequences, not its route, so the context term's source stays unidentified. The decoder recoverability of Section 6 rests on one concept in one out-of-domain corpus and nearly vanishes on natural images, so it is a sign test about what a linear readout of the deciding vector could do, not about the answer's calibration. Finally, our five encoders are closely related contrastive checkpoints, our cells and statistics COCO-centric, and the covariates proxies, as Appendix Z details.

Within those bounds, the two failures share a context-dependent axis whose stable component is affirmative movement, present under every specification we try in both families and under matched strata in the encoders, while what else moves is a readout that separates the classes less than the representation it reads. Evaluations of absence should report that decomposition rather than accuracy, and so should interventions that suppress the prior, since a benchmark gain is consistent with either component moving, or both, and only the decomposition says which a method repaired.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and revising the manuscript. Specifically, they were used to improve the clarity, structure, and presentation of the text during the writing and editing process. The authors reviewed and verified all AI-assisted content and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Every result in the paper is inference, linear probing, or statistics on frozen models, with no training anywhere, one fixed seed throughout, and pinned checkpoint revisions recorded alongside the outputs they produced. Appendix B gives the protocol, the cell construction, and the prompt ensembles, and Appendix Q gives the detector's operating-point rules and its prompt banks. The appendices named in each result state the estimator and the inference procedure used, including the concept-clustered and concept-bootstrap choices that follow from images recurring across concepts. Model outputs are cached per instance, so every table and figure is regenerated from those caches by the accompanying scripts without a GPU, and the paired bootstraps reuse one shared resampling draw across the methods being compared. The code will be released.

REFERENCES

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip HS Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29612–29622. IEEE, 2025.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004.

Irving Biederman, Robert J Mezzanotte, and Jan C Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, 14(2):143–177, 1982.

Yuliang Cai, Jesse Thomason, and Mohammad Rostami. Tng-clip: Training-time negation data generation for negation awareness of clip. *arXiv preprint arXiv:2505.18434*, 2025.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4209–4221. IEEE, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Patrick Ebel, Mounia El Baz, Junjue Wang, Weihao Xuan, Heli Qi, Zhuo Zheng, Naoto Yokoya, Junghwan Park, Jaewan Park, Arthur Elskens, et al. Artificial intelligence for earthquake response: Outcomes and insights from a global spaceborne rapid mapping challenge. *IEEE Geoscience and Remote Sensing Magazine*, 2026.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

David Marvin Green, John A Swets, et al. *Signal detection theory and psychophysics*, volume 1. Wiley New York, 1966.

Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xbd: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*, 2019.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427. IEEE, 2024.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

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*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007, 2023.

Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527, 2007.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Pei Peng, Ming-Kun Xie, Hang Hao, Tong Jin, and Sheng-Jun Huang. Representation-level counterfactual calibration for debiased zero-shot recognition. *Advances in Neural Information Processing Systems*, 38:134547–134584, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4035–4045, 2018.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learn "no" to say "yes" better: Improving vision-language models via negations. *arXiv preprint arXiv:2403.20312*, 2024.

Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11067–11075. IEEE, 2020.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Stefano Soatto. Linear spaces of meanings: compositional structures in vision-language models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15349–15358. IEEE, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14635–14645. IEEE, 2025.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Yuhui Zhang, Yuchang Su, Yiming Liu, and Serena Yeung-Levy. Negvqa: Can vision language models understand negation? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3707–3716, 2025.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, volume 2024, pp. 56969–56998, 2024.

Beier Zhu, Kaihua Tang, Qianru Sun, and Hanwang Zhang. Generalized logit adjustment: Calibrating fine-tuned models by removing label bias in foundation models. *Advances in Neural Information Processing Systems*, 36:64663–64680, 2023.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

GUIDE TO THE APPENDICES

The appendices follow the order of the argument. Appendix A extends the related work, Appendix B gives the protocol and the cell construction, and Appendix C derives the first-order model. For the shared axis, Appendices D and E decompose and validate the probe-error index. For the stability of the criterion shift, Appendices F and G report the context-prior regressions and their external-corpus robustness, with the bidirectionality of the prior in Appendix G.1, Appendix H sets a Bayes-optimal baseline for the movement, Appendices I and J carry the fixed-threshold transfer and the matched strata, Appendix K the score adjustments, and Appendices L to O the supervision and covariate checks, the compositional factorial, the within-concept identification, and the proxy-dependence tests. Appendices P to R give the generative checkpoints, the detector, and the decoding-side mitigation in detail. For what the representation keeps, Appendices S and T treat the negated instrument, Appendices U and V the free-form and image-free measurements, Appendix W the Visual Genome replication with the stratification swap of Appendix W.1, and Appendices X and Y the post-disaster replication and the same probe on natural images. Appendix Z collects the extended limitations.

A EXTENDED RELATED WORK

Scene context in human vision. Human vision research has long shown that object recognition depends on relational consistency, contextual associations, and rapid scene-level expectations (Biederman et al., 1982; Bar, 2004; Oliva & Torralba, 2007). This literature distinguishes target evidence from contextual facilitation without assuming that the two are independent. The same dependence appears in recognition models as a bias, since a classifier trained on natural co-occurrence comes to rely on the surroundings and loses accuracy on the images where the usual context is missing, which is why one line of work decorrelates a category's features from the context it co-occurs with (Singh et al., 2020). That work changes the representation by training, whereas we hold the checkpoint fixed and ask which component of a frozen decision the context moves. Our decomposition brings the same distinction to object-absence judgments in vision-language models.

Interventions on the language prior. That the language prior contributes to object hallucination is by now an established diagnosis rather than a hypothesis, and recent work acts on it directly. One line edits the weights so that features are projected onto the null space of a subspace carrying hallucinated content, which suppresses the prior the language model contributes (Yang et al., 2025), and another intervenes during the forward pass on the attention heads that encode image-level and object-level visual information, deliberately preserving the prior while strengthening visual evidence (Chen et al., 2025). Our contribution is therefore not the claim that context drives the failure. It is a design that decomposes the failure into a sensitivity and a criterion component and reports them separately, a per-concept axis on which the absence insensitivity of contrastive encoders and the hallucinations of generative decoders are the same ranking, and a separation of what is missing from the representation from what is missing in the decision read out of it, which Section 6 measures inside one checkpoint. The interventions above are evaluated by accuracy on benchmarks that mix the two components, so a gain there is consistent with either a criterion movement or a sensitivity change, which is the distinction our design makes visible.

Model generations and coverage. Open generative vision-language models have advanced rapidly, and our behavioral experiments cover eight checkpoints from six families, namely LLaVA-1.5 (Liu et al., 2023), InstructBLIP (Dai et al., 2023), Qwen2-VL (Wang et al., 2024), both sizes of Qwen3-VL (Bai et al., 2025), InternVL3.5 (Zhu et al., 2025), Gemma-3 (Team et al., 2025) and Llama-3.2-Vision (Grattafiori et al., 2024), with the Qwen3-VL pair supplying the current-generation comparison of Appendix P. Transfer to families we do not evaluate, among them LLaVA-OneVision-2 and Gemma 4 (An et al., 2026; Team et al., 2026), remains an empirical question.

B PROTOCOL DETAILS

We use the 2017 training release of the COCO detection corpus (Lin et al., 2014) with instance annotations. For each category O the expectedness of Equation 1 is computed from the conditional co-occurrence table $\widehat{P}^{\mathrm{COCO}}(O \mid a)$ estimated over the full annotation corpus, images whose present

set is empty or equals $\{O\}$ are excluded for that category because E_O is undefined there, and absent images are split at the terciles of E_O into high and low strata, with present images split by the terciles of E_O among present images. Categories with fewer than 80 images in any required cell are dropped, which removes bear, toaster, and hair drier and retains 77 categories, and each cell is capped at 500 images sampled with a fixed seed of 42. The score uses an ensemble of ten affirmative prompt templates whose text embeddings are averaged and renormalized, and no negation template appears anywhere in the pipeline. Probes for $B(O)$ are logistic regressions with balanced classes and five-fold stratified cross-validation, and no image contributes to both classes of a probe. The table $\widehat{\mathrm{PMI}}^{\mathrm{COCO}}$ uses Laplace smoothing with strength 0.5 and spans values from -5.75 to 3.65.

The signal detection quantities use the pooled standard deviation within a concept, the standardized frame takes the per-concept grand mean of present scores as its origin so that movements of the two class distributions can be tracked independently, and the criterion is the equal-error threshold in that frame.

That choice of origin deserves an argument rather than an assumption, because a reader may suspect that measuring a criterion against a fixed origin makes its movement automatic. It does not, for two reasons. The origin is a single constant per concept subtracted from both class positions, so it cancels from every quantity that compares the two classes with each other, which is to say from sensitivity, which never refers to it, and from the class asymmetry. What it leaves is two separate between-cell contrasts, the change in the mean present score and the change in the mean absent score, and nothing in the construction constrains those two to be equal or even to share a sign. Representational degradation predicts that the second moves while the first does not, which is the pattern the decoders of Section 5.1 show, and this frame would have recorded that pattern as faithfully as the one we observe. The measured content of the criterion result is therefore the two changes themselves, $+0.274$ and $+0.271$ for the anchor encoder on the within-class cells, and the frame records an unequal pair just as faithfully, since the same encoder gives $+0.299$ and $+0.650$ once the two classes are matched (Appendix J).

This paper reports many coefficients, across checkpoints, confusability proxies and covariate specifications, so multiplicity deserves a statement rather than silence. Three features of the design limit its force. The thresholds that decide the primary questions were fixed before the analyses that use them, at a rank correlation of 0.30 for the unpartialled association, 0.20 for the partialled one, and a meaningful-loss margin of -0.15 standardized units for sensitivity, so those tests are not selected after the fact. The claims we rest on are counts of agreement rather than single coefficients, since what carries Section 5 is that all fifteen encoder-by-specification fits move the same way and all seven decoder criteria rise, and a count over a fixed set of checkpoints does not gain from testing more of them. And the significance levels behind those claims are far from any conventional boundary, below 10^{-31} for the prior coefficient on absent-object scores and below 10^{-7} for the shared axis, so no family-wise adjustment we could apply would reverse them. Where multiplicity would bite we do not lean on significance at all. The sensitivity comparison is reported as a distribution of per-concept changes and a sign test rather than as a coefficient that clears a threshold, and the proxy-dependent coefficients are reported as disagreeing rather than as a selected winner. We do not apply a formal family-wise correction, because the checkpoints share concepts and images and their tests are therefore strongly dependent, which makes an independence-based adjustment both conservative and mis-specified. The dependence-preserving alternative is the concept bootstrap with one shared draw across checkpoints, which this paper uses for every difference it claims.

One implementation detail of the encoder library deserves a check rather than an assumption. Recent versions of the library build the ViT-B/32, ViT-B/16 and ViT-L/14 architectures with a standard Gaussian error linear unit, whereas the OpenAI weights loaded into them were trained with the faster sigmoid approximation of it. Because every claim in this paper is a contrast between prior strata within one checkpoint rather than an absolute score, the quantity to check is the contrast. Recomputing the ViT-B/32 decomposition under the activation the weights were trained with moves the sensitivity change from $+0.000$ to -0.007 and the criterion change from $+0.336$ to $+0.347$, with per-concept correlations between the two versions of 0.945 to 0.981 across the four quantities, so the substitution does not affect anything reported here.

C First-order model derivation and algebra

The three assumptions used in Section 3 are stated here.

Assumption 1 (Alignment optimum)**.** *At the population optimum of the contrastive objective, inner products between image and text embeddings recover pointwise mutual information up to an increasing affine transformation, so that $\langle f(x), g(y)\rangle$ is an increasing affine function of $\log p(x, y) - \log p(x) - \log p(y)$ for an image x and a caption y, where g denotes the text embedding map trained jointly with f (Oord et al., 2018).*

Assumption 2 (Additive composition)**.** *Image embeddings compose approximately additively over scene constituents, $f(x) \approx \sum_{a\in S(x)} w_a u_a + \varepsilon$, where u_a denotes a unit direction associated with category a, $w_a \geq 0$ a salience weight, and ε a residual (Trager et al., 2023).*

Assumption 3 (Category transport)**.** *For recurring scene populations dominated by category a, the category cross term increases approximately affinely with the association between observing a and mentioning O, so that $\langle u_a, t_O\rangle \approx \lambda\,\mathrm{PMI}^{\mathrm{pre}}(a, O)+\lambda_0$ with a transport slope $\lambda > 0$ and an intercept λ_0, where $\mathrm{PMI}^{\mathrm{pre}}$ denotes association in the model's pretraining distribution and is estimated below by corpus-specific image co-occurrence or caption co-mention tables.*

Assumption 3 is a category-level analogue of Assumption 1 but is not implied by it.

C.1 Motivation for the score decomposition

Assumption 2 gives $f(x) \approx \sum_{a\in S(x)} w_a u_a + \varepsilon$, so linearity yields $s_O(x) = \langle f(x), t_O\rangle \approx \sum_{a\in S(x)} w_a\langle u_a, t_O\rangle + \langle \varepsilon, t_O\rangle$. Assumption 3 then replaces each category cross term by $\lambda\,\mathrm{PMI}^{\mathrm{pre}}(a, O) + \lambda_0$. Assumption 1 is not needed for this algebra. The pretraining association can be estimated from caption co-mention or image co-occurrence, which lets the corpus comparison ask how much of the same structural term each corpus-specific estimate recovers.

Substitution and separation of $a = O$ gives $s_O(x) \approx w_O\big(\lambda\,\mathrm{PMI}^{\mathrm{pre}}(O, O) + \lambda_0\big)\mathbf{1}[O \in S(x)] + \lambda \sum_{a\in S(x)\setminus\{O\}} w_a\,\mathrm{PMI}^{\mathrm{pre}}(a, O) + \lambda_0 \sum_{a\in S(x)\setminus\{O\}} w_a + \langle \varepsilon, t_O\rangle$. Absorbing the first bracket into v_O, replacing the salience weights in the middle sum by a common scale $\bar{w}$ folded into $\kappa = \lambda\bar{w}$, and collecting the affine context-size term and remaining error into $\varepsilon_O(x)$ gives Equation 3. The context-prior regressions therefore include context size as a covariate, and Appendix L adds it to the position and sensitivity specifications. The 6.5 to 14.7 percent explained-variance gain for $\widehat{\pi}_O^{\mathrm{COCO}}$ and the 0.05 to 0.72 percent gain for $\widehat{\pi}_O^{\mathrm{CC}}$ measure how much score variance each corpus estimate explains, which Appendix G interprets.

C.2 Algebra for common shifts

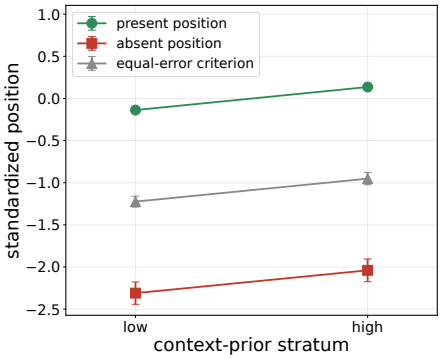


Figure 5: Paired contrasts on CLIP ViT-L/14. The present and absent score distributions and the equal-error criterion move upward together across the two within-class prior strata, which give the two classes different context changes (Appendix J).

Fix a concept O. Within a cell at prior level p, Equation 3 gives the first-order class-conditional means $\mu^{\mathrm{P}}_{O,p} = v_O + \kappa\,\mathbb{E}[\pi_O^{\mathrm{pre}} \mid \mathrm{present}, p] + \mathbb{E}[\varepsilon_O \mid \mathrm{present}, p]$ and $\mu^{\mathrm{A}}_{O,p} = \kappa\,\mathbb{E}[\pi_O^{\mathrm{pre}} \mid \mathrm{absent}, p] +$

$\mathbb{E}[\varepsilon_O \mid \text{absent}, p]$. Treating residual mean changes across prior levels as negligible, define $\Delta_O^{\mathrm{P}} = \mathbb{E}[\pi_O^{\mathrm{pre}} \mid \text{present}, \text{hi}] - \mathbb{E}[\pi_O^{\mathrm{pre}} \mid \text{present}, \text{lo}]$ and Δ_O^{A} analogously. Moving from low to high prior changes raw separation by $\kappa(\Delta_O^{\mathrm{P}} - \Delta_O^{\mathrm{A}}) + o(1)$. Equal increments preserve raw separation to first order and shift both means and the equal-error criterion together. Standardized sensitivity is also preserved only if pooled dispersion is stable. Empirically the paired anchor increments of the present distribution, absent distribution, and criterion are $+0.274$, $+0.271$, and $+0.272$ standardized units on the within-class strata, as Figure 5 shows. Those strata give the two classes different context changes, so equal increments there do not establish a common one, and once the classes are matched the absent increment is about twice the present one (Appendix J). Dispersion and sensitivity are evaluated separately.

C.3 Two further implications

Equation 3 makes the absent-score shift between strata equal to κ times $\mathbb{E}[\pi_O^{\mathrm{pre}} \,|\, \text{absent}, \text{hi}] - \mathbb{E}[\pi_O^{\mathrm{pre}} \,|\, \text{absent}, \text{lo}]$, so two models that encode similar context-prior structure order concepts alike up to their model-specific scales, and a monotone generative readout may inherit that ordering. This implication motivates rank correlations. The weak agreement between $\widehat{\pi}_O^{\mathrm{CC}}$ and $\widehat{\pi}_O^{\mathrm{COCO}}$ shows that different corpus estimators need not recover the same ordering.

For each queried concept O_j, the context-alignment profile $q_{O_j} = (\mathrm{PMI}^{\mathrm{pre}}(a, O_j))_a$ specifies the linear combination of context-category coordinates contributing to its prior term. An intervention removing an r-dimensional shared subspace can eliminate all profiles only if they span at most r dimensions. This rank observation is elementary and does not claim that natural association profiles satisfy the condition.

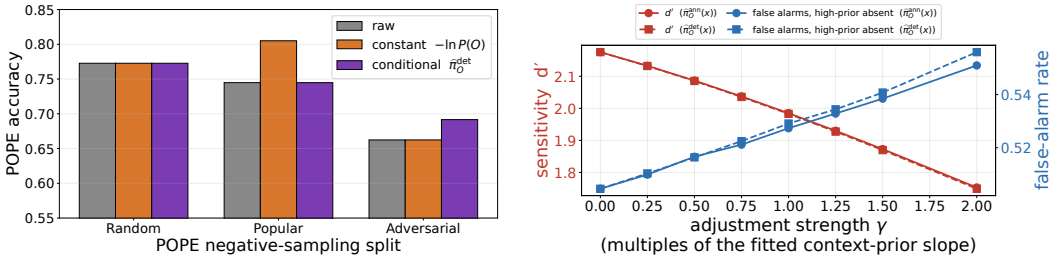
C.4 Correction identity and the shape of its failure

Within a fixed evaluation cell, the corrected score $\tilde{s}_O^e(x) = s_O(x) - \gamma \widehat{\pi}_O^e(x)$ has present-class mean $\mu_O^{\mathrm{P}} - \gamma m_O^{\mathrm{P}}$ and absent-class mean $\mu_O^{\mathrm{A}} - \gamma m_O^{\mathrm{A}}$. Subtracting the latter from the former gives $\tilde{D}_O(\gamma) = D_O(0) - \gamma(m_O^{\mathrm{P}} - m_O^{\mathrm{A}}) = D_O(0) - \gamma\delta_O$. Let $\tilde{\sigma}_O(\gamma)$ denote the pooled standard deviation of the corrected score and let $\tilde{d}'_O(\gamma) = \tilde{D}_O(\gamma)/\tilde{\sigma}_O(\gamma)$. Differentiating at zero gives

$$\left.\frac{\mathrm{d}\tilde{d}'_O(\gamma)}{\mathrm{d}\gamma}\right|_{\gamma=0} = -\frac{\delta_O}{\sigma_O(0)} - \frac{D_O(0)}{\sigma_O(0)^2}\left.\frac{\mathrm{d}\tilde{\sigma}_O(\gamma)}{\mathrm{d}\gamma}\right|_{\gamma=0},$$

where $\sigma_O(0) = \tilde{\sigma}_O(0)$ is the uncorrected pooled dispersion. The first term is the contribution from changing class means, and the second is the contribution from changing dispersion. If dispersion is locally stationary, the second term vanishes and the sensitivity derivative is determined by the class gap. The mean of δ_O across concepts in the high-prior stratum is $3.84 - (-0.19) = 4.03$. A third baseline belongs beside the two the body evaluates, namely a per-concept threshold calibrated on the development split, which is worse than the raw score on all three negative-sampling splits, at 0.754 against 0.773 on the random split, 0.566 against 0.745 on the popular one and 0.620 against 0.662 on the adversarial one. It fails because a per-concept constant cannot represent a within-concept contrast. Figure 6 plots the double dissociation in POPE accuracy beside the resulting near-linear sensitivity decline of Figure 6b, while the generative decomposition in Appendix P illustrates why a nonzero dispersion contribution can offset a change in raw separation.

Remark 1 (Increments versus levels)**.** *The common-shift and correction statements condition on different moments. The former requires the increment of the estimated context prior between expectedness strata to be common to the classes. The paired shifts of* $+0.274$ *and* $+0.271$ *on the within-class strata resemble it, but those strata give the classes different context changes, and on matched strata the anchor's absent position rises by* $+0.650$ *against* $+0.299$ *for the present one (Appendix J), so the condition is not met. Standardized sensitivity additionally requires stable pooled dispersion. Correction instead depends on the prior level within a stratum. The measured class gap of* 4.03 *rejects class balance and therefore predicts a loss of raw separation under positive correction strength. The derivative above separates this mean contribution from any accompanying dispersion change. Both conditions concern natural context statistics rather than fixed properties of the model.*



(a) Double dissociation on POPE.

(b) Sensitivity is not preserved.

Figure 6: Score adjustment as a test between two bias accounts. Panel (a) shows a double dissociation in POPE accuracy for CLIP ViT-L/14, where the constant adjustment changes only the popular split and the conditional adjustment only the adversarial split. Panel (b) shows that a larger adjustment strength lowers sensitivity and raises the false-alarm rate on high-prior absent images at a matched miss rate.

D ORTHOGONAL DECOMPOSITION OF THE HIGH-PRIOR PROBE-ERROR INDEX

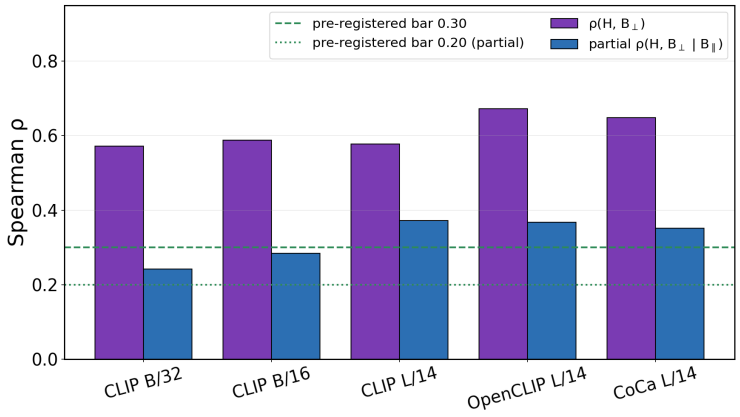


Figure 7: The association between hallucination and the probe-error index remains after removal of the shared one-dimensional channel. For each encoder the left bar shows the rank correlation between H and the off-axis index $B_\perp$ and the right bar shows the partial correlation given the parallel-channel index $B_\parallel$, with the two pre-registered support bars marked, 0.30 for the rank correlation and 0.20 for the partial correlation.

Because $H(O)$ is a functional of the projection onto t_O while $B(O)$ uses the full embedding, a correlation between them could in principle reflect the same one-dimensional separation read twice. Figure 7 shows that the probe-error index measured strictly off that axis still predicts hallucination in all five checkpoints, with $\rho(H, B_\perp)$ between 0.572 and 0.673 against a pre-registered bar of 0.30, and partial correlations given $B_\parallel$ between 0.242 and 0.372 against a pre-registered bar of 0.20, every value significant. The off-axis probe retains dynamic range with a cross-concept standard deviation of its area-under-curve values near 0.03, so the result is not a ceiling artifact.

A second construction check concerns the hallucination index's denominator, which is the per-concept gap between present images and low-prior absent ones and therefore a detectability scale in its own right. Because the probe-error index is also a difficulty measure, part of the association could in principle be definitional, and that worry has a real basis, since the reciprocal denominator on its own correlates with B at $+0.626$ to $+0.660$, as large as the association it would be explaining. It is nevertheless not what produces it. Partialling the denominator out leaves $\rho(H, B)$ at $+0.436$ to $+0.576$ with every bootstrap interval excluding zero, the control set of Section 4 on its own leaves $+0.614$ to $+0.692$ and adding the denominator to it leaves $+0.571$ to $+0.651$, and the same control against the off-axis index leaves $+0.571$ to $+0.639$. The numerator is not separately sufficient either, since the raw absent-position shift alone predicts $B_\perp$ at only $+0.159$ to $+0.253$, significant in one

of the four encoders, and reaches $+0.376$ to $+0.491$ only once the denominator is conditioned on. Both parts of the ratio therefore carry information about the probe-error index that the other does not, which is why the index is written as a ratio, and a reader should credit the association to neither part alone.

A stronger version of that check, and a limit on every correlation the paper reports for these two indices, comes from their reliability. Splitting the images inside every cell in half with the run seed and recomputing both indices independently on each half gives split-half reliabilities, corrected for the halving, of 0.927 to 0.967 for the hallucination index and 0.968 to 0.975 for the probe-error index over the five checkpoints, so neither is appreciably noisy and the probe-error index in particular is a stable measurement despite its narrow range. The disattenuated correlation between them, the observed value divided by the geometric mean of the two reliabilities, is therefore only 0.601 to 0.713 against the observed 0.575 to 0.687. Attenuation accounts for 0.02 to 0.03 of the distance from one and no more, so the honest description is a strong common axis on which roughly half the rank variance is shared, not a single axis on which the two indices are the same measurement.

E Construct validity of the embedding probe

High probe accuracy can arise because the embedding records the target object, because it records context that predicts the target, or because it records both. We test these alternatives with three feature sets. The context feature vector $C_O(x)$ contains the binary presence indicators of the 79 non-target COCO categories together with the number of contextual companions. The embedding feature vector is the previously defined frozen image embedding $f(x)$. The combined feature vector concatenates $C_O(x)$ and $f(x)$. All probes use balanced folds, nested cross-validation, and out-of-fold predictions.

Within the high-prior stratum, context alone is already strong, reaching an area under the receiver operating characteristic curve of 0.961. Relative to it the embedding gains $+0.012$ [$+0.006$, $+0.017$] in area under the curve and reduces log loss by 0.023 [0.005, 0.042]. The combined probe reduces log loss by 0.042 [0.026, 0.059]. These increments show that the embedding adds held-out information beyond the annotated context, but the strength of the context-only baseline prevents the within-stratum result from identifying target evidence by itself.

The cross-prior transfer test is more discriminating. We train each probe to separate present from absent images in the low-prior stratum and evaluate it without refitting in the high-prior stratum. Table 2 shows that context-only prediction collapses to an area under the curve of 0.528, whereas the embedding retains 0.904. The paired embedding-over-context gain is $+0.376$ [$+0.331$, $+0.421$], and the log-loss reduction is 2.48 with a concept-level bootstrap significance below 10^{-4}. Adding context to the embedding does not improve transfer and slightly worsens calibration, which is expected because the companion configuration learned at low prior does not transport to high-prior scenes. The representation therefore carries presence-discriminating information that generalizes across prior context, although it may also encode unannotated visual context.

Table 2: Probe construct-validity results. Within-stratum values use out-of-fold predictions on high-prior scenes. Transfer values train on low-prior scenes and test on high-prior scenes. Lower log loss is better.

Feature set	High-prior area under curve	High-prior log loss	Transfer area under curve	Transfer log loss
Annotated context	0.961	0.198	0.528	3.600
Image embedding	0.973	0.175	0.904	1.117
Context and embedding	0.978	0.156	0.900	1.433

F Per-model pointwise mutual information regressions

Figure 8 shows the per-model regressions underlying Section 5. Absent images are sampled uniformly across the range of $\widehat{\pi}_O^{\mathrm{COCO}}$ rather than from the tercile cells, so the slope is not induced by the cell construction, and a fresh detector pass provides the confusability covariate on the sampled images. Raw slopes range from 0.0005 to 0.0014 in score units per unit of $\widehat{\mathrm{PMI}}^{\mathrm{COCO}}$, which correspond to the standardized mixed-model coefficients between 0.0106 and 0.0226 in Table 3. The

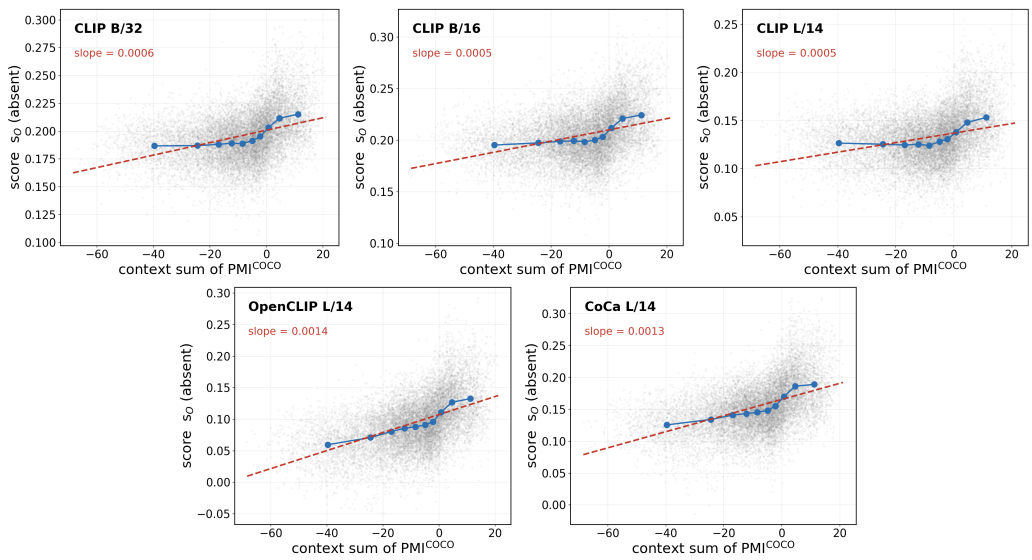


Figure 8: Scores on absent images track $\widehat{\pi}_O^{\mathrm{COCO}}$ in all five checkpoints. Each panel plots the raw score of the queried category on images where it is absent against the context sum of $\widehat{\mathrm{PMI}}^{\mathrm{COCO}}$, with binned means and the mixed-model fit overlaid, for the three OpenAI CLIP models, OpenCLIP trained on LAION-2B, and CoCa.

increase in explained variance from adding $\widehat{\pi}_O^{\mathrm{COCO}}$ beyond the confusability measurement ranges from 6.5 to 14.7 percent.

G External-corpus robustness of the context prior

Appendix F gives the per-checkpoint fits of the COCO estimator that this comparison is measured against. We repeated the primary regression of Section 5 with $\widehat{\pi}_O^{\mathrm{CC}}$ rather than $\widehat{\pi}_O^{\mathrm{COCO}}$, processing 3,318,333 Conceptual Captions (Sharma et al., 2018) captions with a lexical matcher over the 80 category names and a fixed synonym table logged before results were inspected. Table 3 reports the outcome. The $\widehat{\pi}_O^{\mathrm{CC}}$ slope is positive at significance below 10^{-15} in all five checkpoints and is numerically similar to the $\widehat{\pi}_O^{\mathrm{COCO}}$ slope. Its incremental explanatory value is 0.0005 to 0.0072, compared with 0.065 to 0.147 for $\widehat{\pi}_O^{\mathrm{COCO}}$. The caption co-mention analogue of $\widehat{\mathrm{PMI}}^{\mathrm{COCO}}$ agrees only weakly with the image co-occurrence table across off-diagonal entries, with a Spearman correlation of 0.115 over 6,320 pairs. Conceptual Captions preprocessing also contributes measurement noise because named entities are replaced with a literal person token and five categories occur fewer than one hundred times. This noise alone does not explain why the external estimate adds almost no variance for several checkpoints.

This result weakens the claim that $\widehat{\pi}_O^{\mathrm{COCO}}$ directly recovers pretraining caption co-occurrence. A viable alternative is that COCO-style object annotations summarize recurring visual scene types that the encoders represent, an interpretation supported by the 0.860 agreement between $\widehat{\mathrm{PMI}}^{\mathrm{COCO}}$ and $\widehat{\mathrm{PMI}}^{\mathrm{VG}}$. Another possibility is evaluation-specific alignment between the COCO cell construction and $\widehat{\pi}_O^{\mathrm{COCO}}$. The positive $\widehat{\pi}_O^{\mathrm{CC}}$ slope provides sign-level evidence that caption co-mention may contribute, but neither its magnitude nor significance identifies it as the dominant or unique source.

Distinguishing these accounts calls for a comparison that holds the corpus fixed and varies the model, and the five encoders permit one, because two of them were pretrained on LAION-2B web alt-text and three were not. If pretraining co-mention drives the effect, a web alt-text table should account for relatively more of those two encoders' absent-object scores. LAION-2B is gated for this account, so we approximate its distribution rather than claim it, counting co-mentions over 12,800,000 DataComp CommonPool (Gadre et al., 2023) alt-texts, which are Common Crawl text collected the same way, and separately over the 1,874,440 of them at or above the CLIP ViT-L/14

image-text similarity of 0.28 that LAION-2B-en was filtered at. The same lexical matcher and synonym table are used, so a mention means here what it means above. The unfiltered pool agrees with the image co-occurrence table at a Spearman correlation of 0.216 over the off-diagonal entries, about twice the Conceptual Captions figure, so part of that weak agreement was a property of that corpus rather than of caption co-mention in general, and it remains weak in absolute terms.

The model-specific prediction fails. Measured as explained within-concept variance, the web table falls short of the image co-occurrence table by 0.020 for the three encoders that never saw web alt-text and by 0.027 for the two that were trained on it, so the shortfall does not favor the encoders whose pretraining distribution the stand-in approximates. The same holds when absent images are drawn uniformly across the COCO prior rather than at random, which is the version that matters least because stratifying on one table inflates the variance of that table's own regressor. All four tables give a positive slope at significance below 10^{-8} in all five encoders, so what is at issue is attribution and not the additive form. We therefore read the measured prior as tracking recurring visual scene structure more closely than caption co-mention, and we keep two limits on that reading, since the stand-in is not LAION and a curated annotation table may be a less noisy measurement of a regularity that web images also carry, which no comparison of explained variance can separate.

Table 3: The context-prior regression on absent images under $\widehat{\pi}_O^{\mathrm{CC}}$ and $\widehat{\pi}_O^{\mathrm{COCO}}$, adjusted for the confusability measurement and context size and fitted with concept random slopes. Slopes are standardized mixed-model coefficients, and the explained-variance gain is the increase from adding the corresponding corpus estimator beyond the confusability measurement.

Model	Conceptual Captions slope	Explained-variance gain	COCO slope	Explained-variance gain
CLIP ViT-B/32	+0.0104	0.0072	+0.0114	0.094
CLIP ViT-B/16	+0.0090	0.0061	+0.0106	0.086
CLIP ViT-L/14	+0.0103	0.0053	+0.0112	0.065
OpenCLIP ViT-L/14	+0.0198	0.0005	+0.0226	0.147
CoCa ViT-L/14	+0.0187	0.0005	+0.0224	0.142

G.1 Bidirectionality of the context prior

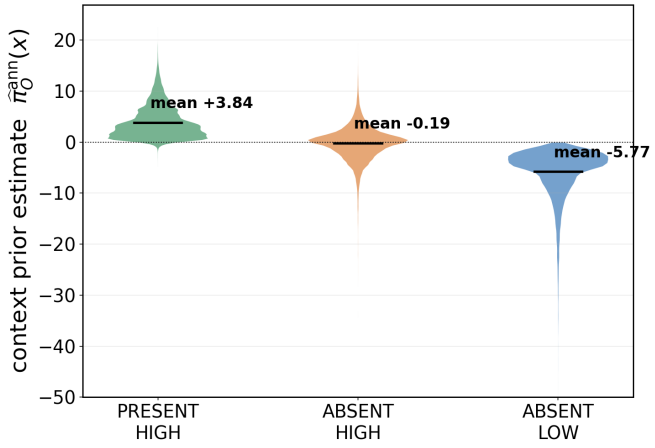


Figure 9: The annotation-based context-prior estimate is strongly presence-predictive. Violin plots of $\widehat{\pi}_O^{\mathrm{ann}}(x)$ computed from annotated context objects show a mean of $+3.84$ on present images in the high-expectedness stratum, -0.19 on absent images in the same stratum, and -5.77 on absent images in the low stratum, so the quantity subtracted by the conditional correction carries genuine evidence of presence.

Figure 9 documents the failure of the class-balance condition in Equation 4. The separation between the present and absent distributions of $\widehat{\pi}_O^{\mathrm{ann}}$ within the same expectedness stratum ensures that subtraction lowers raw class separation because present scores decrease more than absent scores. Appendix C.4 separates this mean contribution from the dispersion contribution to standardized sensitivity, and Figure 6b shows that their combined effect is a near-linear sensitivity decline over the tested correction strengths.

H A BAYES-OPTIMAL BASELINE FOR THE CRITERION MOVEMENT

An ideal observer moves its criterion when the base rate of presence changes, so the movement reported in Section 5 raises a calibration question that the two qualitative conditions of Section 1 leave open. Here we answer it with a number.

For concept O and expectedness stratum p, let π_p denote the corpus probability that O is present, measured by applying that concept's stored expectedness thresholds to the whole COCO pool and counting how often O actually occurs. Under equal-variance Gaussian signal detection with equal costs, the optimal threshold lies $\ln[(1-\pi_p)/\pi_p]/\bar{d}'_O$ standardized units above the midpoint of the present and absent distributions, where $\bar{d}'_O$ is the concept's mean sensitivity over the two strata and is held common across them so that the baseline reflects the change in base rate alone. A higher base rate lowers that threshold, so the ideal observer becomes more liberal by the difference of the two values. Its measured counterpart is the change in the equal-error criterion, which is what a reader applying one threshold across scenes experiences as liberalization. The design's own cells cannot supply π_p, because they split the two classes at class-specific thresholds in order to balance them.

Averaged over concepts the corpus presence rate rises from 0.003 in the bottom expectedness bin to 0.082 in the top, so the ideal liberalization is large, between 2.57 and 2.99 standardized units for the encoders and between 0.94 and 1.42 for the decoders, the two families differing because the conversion divides by sensitivity. The observed movement is much smaller, at 0.272 to 0.350 and at 0.146 to 0.509. The equal-weight ratio of observed to ideal movement is 0.112 $[0.086, 0.142]$ over the five encoders and 0.302 $[0.239, 0.330]$ over the seven decoders, from ten thousand bootstrap draws that resample one set of concepts for every checkpoint. No encoder concept, and at most one of the 77 concepts in any decoder, moves further than the ideal observer would. Reading the base rate off the expectedness values instead of counting occurrences raises the per-checkpoint ratios to between 0.168 and 0.241 and between 0.200 and 0.822, which changes no conclusion.

The ideal shift divides a log-odds gap by a sensitivity, so it inherits the small-denominator tail that makes $H(O)$ a rank statistic in Section 4, and the equal-weight mean of the ideal shift is accordingly pulled up by concepts the encoders barely discriminate. Among the five encoders the mean ideal shift runs from 2.57 to 2.99 against a median of 1.99 to 2.38, and the concept with the smallest sensitivity in the anchor contributes an ideal shift of 8.6 against a median of 2.2. We therefore report the ratio in four further summaries that the tail cannot drive. Over the encoders the ratio of medians is 0.108 to 0.159, the median of the per-concept ratios is 0.105 to 0.156, the ratio of ten percent trimmed means is 0.105 to 0.153, and dropping the fifth of concepts with the lowest sensitivity gives 0.119 to 0.176. Over the decoders the same four give 0.091 to 0.477, 0.101 to 0.510, 0.111 to 0.473 and 0.128 to 0.458. The equal-weight ratio of medians is 0.135 $[0.094, 0.178]$ for the encoders and 0.311 $[0.240, 0.358]$ for the decoders under the same shared concept draw. Every summary stays far below one, so the tail moves the second digit and not the conclusion.

Two readings follow and the paper needs both. Against an account on which these models over-trust context, the movement is not excess by this standard but a fraction of what the corpus would license, so we do not describe the effect as over-reliance on priors, and a mitigation that merely damped context following would be pushing an already conservative quantity further from the corpus value. Against an account on which the movement is simply correct Bayesian updating, the frame in which the failure is measured is balanced by construction, with equal numbers of present and absent images in every cell, so the optimal shift there is exactly zero and the whole of the observed movement is excess. Magnitude therefore does not settle the normative question in either direction, which is why Section 1 rests the term on recoverability and on common translation rather than on how far the criterion travels.

I ONE THRESHOLD, CARRIED ACROSS THE CONTEXT CHANGE

Two objections to Section 5 share an answer. The first is that a contrastive encoder applies no threshold of its own, so the equal-error criterion is a statistic we compute from the score distribution rather than a rule the model follows. The second is that describing the failure as criterion dominant asserts a share of the error without reporting one. Both are addressed by removing the construction. For each concept we fit one threshold on the low-expectedness stratum alone, at the point where its miss and false-alarm rates meet, and then read the balanced error on the high stratum with that

threshold unchanged, so the operating point is chosen before the context changes and is never refitted to it.

The resulting error increase then splits exactly and without a distributional assumption. Translating both high-stratum class distributions down by the measured criterion movement returns the criterion to where the low stratum placed it and leaves sensitivity untouched, since a common translation changes neither class separation nor pooled dispersion. Evaluating the same threshold on the translated scores gives a third error rate, and the difference between the observed high-stratum error and that rate is the part the translation produced, while the difference between that rate and the low-stratum error is the part the changes in separation and dispersion produced. The two sum to the total by construction, which is what makes a share of the error well defined.

Table 4 gives both components on the within-class cells. For the encoders the whole of the increase is the translation there. Its component is positive with an interval excluding zero in all five checkpoints while the residual component's interval contains zero in all five, at $+0.034$ $[+0.026, +0.044]$ against $+0.001$ $[-0.015, +0.015]$ on average, and a threshold fitted where the context is unexpected multiplies its false-alarm rate by 1.7 to 2.0 once the context becomes expected, carrying it from between 0.163 and 0.182 to between 0.314 and 0.350. The picture is unchanged at a second operating point admitting ninety percent of present images, where the two components are $+0.044$ $[+0.032, +0.058]$ and -0.006 $[-0.020, +0.007]$. We give the two components rather than their ratio here, because the encoders' total is itself small and a ratio of small signed quantities inherits whatever interval its denominator has, in this case 1.019 $[0.675, 2.001]$. It holds at a threshold placed without the high stratum in view, and it is a property of cells cut inside each class, since once the two classes are matched on shared expectedness bins the residual grows to the size of the criterion part (Appendix J).

For the decoders the same measurement returns a shared effect rather than a dominant one. Both components are positive with intervals excluding zero in all seven checkpoints, and the translation accounts for 0.444 $[0.367, 0.528]$ of the increase, an interval that excludes one as clearly as it excludes zero. At the ninety percent recall operating point the translation accounts for almost none of it, at $+0.004$ $[-0.001, +0.010]$ against a residual of $+0.027$ $[+0.018, +0.036]$, because a threshold in the tail of the present distribution is more exposed to the dispersion growth of Table 18 than to a movement of both classes together. That table also carries the residual's own division between class separation and pooled dispersion, which we measure on the sensitivity change directly rather than on the error, since apportioning an error difference between the two would require an ordering of the two counterfactuals that neither quantity supplies. The criterion component of a decoder's absence error is therefore real and present in every checkpoint while being neither the larger part of that error nor stable across operating points, which is the same asymmetry Section 5.1 reports on the ranking axis.

One reading the table does not support deserves stating. That the encoders' mean residual is indistinguishable from zero does not make their per-concept sensitivity changes small. Averaged over concepts the absolute residual component is the larger of the two, since the criterion accounts for a sign-free share of only 0.417 $[0.350, 0.484]$ of the total unsigned movement for the encoders and 0.426 $[0.382, 0.470]$ for the decoders, so what separates the families here is that only the encoders' sensitivity changes cancel. That is the fact Section 5 states as an absence of systematic sign, seen through its cost at a fixed threshold.

J Matching the context change the two classes receive

The within-class cells of Section 4 cut each class at its own expectedness terciles, which balances the classes in every cell but does not give them the same change in context. On E_O the present class's change from the low to the high cell is twice the absent class's, at a median ratio over concepts of 2.01 and larger in 69 of the 77 concepts. We therefore repeat the decomposition on strata that give both classes one change. The design uses the whole annotated corpus, 117,266 images for which E_O is defined, embedded with the same encoders and checked against the cached embeddings of the cell images. For each concept, 21 bin edges are placed at the quantiles of E_O over its present images and shared by both classes, bins holding fewer than ten images of either class are dropped, and the bottom and top seven bins form the low and high strata. Absent images are then reweighted bin by bin to the present images' distribution over the kept bins, so the two classes have one distribution

Table 4: The context-induced error at one threshold fitted on the low-expectedness stratum and carried unchanged to the high one. Entries are balanced error rates averaged over 77 concepts, with 95 percent intervals from a concept bootstrap using one shared draw across every checkpoint. The criterion and residual columns are the exact additive split of the total into the part produced by translating both class distributions and the part produced by changes in their separation and dispersion. The last column gives the false-alarm rate on absent images at that threshold in the two strata.

Checkpoint	Total error change	Criterion component	Residual component	False alarm, low to high
Contrastive encoders				
CLIP ViT-B/32	$+0.041$	$+0.034$ $[+0.026, +0.044]$	$+0.006$ $[-0.010, +0.021]$	0.181 to 0.350
CLIP ViT-B/16	$+0.038$	$+0.033$ $[+0.024, +0.043]$	$+0.005$ $[-0.009, +0.019]$	0.176 to 0.329
CLIP ViT-L/14	$+0.021$	$+0.031$ $[+0.023, +0.040]$	-0.010 $[-0.026, +0.006]$	0.182 to 0.314
OpenCLIP ViT-L/14	$+0.037$	$+0.034$ $[+0.025, +0.044]$	$+0.003$ $[-0.013, +0.019]$	0.163 to 0.323
CoCa ViT-L/14	$+0.037$	$+0.037$ $[+0.028, +0.048]$	-0.000 $[-0.017, +0.016]$	0.169 to 0.339
five-encoder mean	$+0.035$ $[+0.022, +0.048]$	$+0.034$ $[+0.026, +0.044]$	$+0.001$ $[-0.015, +0.015]$	
Generative decoders				
Qwen2-VL-7B	$+0.051$	$+0.015$ $[+0.007, +0.024]$	$+0.036$ $[+0.021, +0.051]$	0.055 to 0.156
InstructBLIP-7B	$+0.080$	$+0.040$ $[+0.026, +0.055]$	$+0.040$ $[+0.025, +0.057]$	0.058 to 0.240
Qwen3-VL-4B	$+0.065$	$+0.030$ $[+0.022, +0.038]$	$+0.036$ $[+0.028, +0.044]$	0.023 to 0.166
Qwen3-VL-8B	$+0.065$	$+0.032$ $[+0.024, +0.040]$	$+0.033$ $[+0.025, +0.041]$	0.024 to 0.166
InternVL3.5-8B	$+0.072$	$+0.044$ $[+0.035, +0.053]$	$+0.029$ $[+0.021, +0.037]$	0.033 to 0.193
Gemma-3-12B	$+0.048$	$+0.013$ $[+0.008, +0.018]$	$+0.035$ $[+0.022, +0.050]$	0.075 to 0.180
Llama-3.2-11B	$+0.050$	$+0.025$ $[+0.019, +0.031]$	$+0.026$ $[+0.017, +0.035]$	0.047 to 0.162
seven-decoder mean	$+0.062$ $[+0.050, +0.073]$	$+0.028$ $[+0.022, +0.035]$	$+0.033$ $[+0.025, +0.043]$	

over the same absolute bins in each stratum and the change from low to high is one intervention for both. The present distribution is the target because the reverse direction starves the present class at low expectedness, which is why those cells are cut inside each class. Sixty-six concepts keep enough bins. Every quantity is computed in the frame of Section 3 with weighted moments, the fixed-threshold split follows Appendix I, and intervals come from the 77-concept bootstrap with one shared draw across encoders and designs.

Table 5 gives the result and the checks on it. The criterion rises in every row, with every encoder's interval excluding zero, while the matching makes absent scores rise about twice as far as present ones and so costs the zero-shot score sensitivity. Restricted to the same 66 concepts the within-class cells give a sensitivity change of $+0.049$, so the difference is the matching and not the concept set. Three checks address what matching could itself introduce. The matched high stratum draws absent images from scenes arranged as if the object were there, where an unannotated instance is likelier, and removing the absent images a COCO-trained DETR detector calls present leaves the result in place. Removing the linear detector-confidence component from absent scores shrinks the sensitivity loss without removing it, and since that detector itself reads context, this adjustment also removes part of the effect it is meant to isolate. Matching the number of context objects as well, and removing log target area from present scores, leave the criterion where it was and enlarge the loss. Matching on the summed prior of Equation 3 instead of E_O gives the same pattern. Matching to the absent class's distribution keeps only 21 concepts, since present images are rare at low expectedness, and there the criterion still rises while the sensitivity change is not resolved from zero, so the sensitivity loss concentrates at the expectedness levels typical of present images.

Table 5: The decomposition on matched strata, pooled over the five encoders, in standardized units with 95 percent concept-bootstrap intervals. The present and absent columns give each class's change from the low to the high stratum, and the last column is the residual part of the fixed-threshold error increase of Appendix I.

Design	Concepts	Present	Absent	Criterion	Sensitivity	Residual error
Within-class cells, each class cut at its own terciles	77	$+0.273$	$+0.351$	$+0.312$ $[+0.249, +0.380]$	$+0.007$ $[-0.135, +0.141]$	$+0.001$ $[-0.015, +0.015]$
Matched on expectedness	66	$+0.299$	$+0.669$	$+0.484$ $[+0.407, +0.562]$	-0.233 $[-0.363, -0.099]$	$+0.047$ $[+0.030, +0.064]$
Also matched on the number of context objects	65	$+0.276$	$+0.633$	$+0.455$ $[+0.376, +0.539]$	-0.233 $[-0.372, -0.088]$	$+0.048$ $[+0.030, +0.066]$
Absent images detected as present removed	66	$+0.299$	$+0.654$	$+0.477$ $[+0.401, +0.554]$	-0.205 $[-0.339, -0.067]$	$+0.044$ $[+0.027, +0.061]$
Absent scores adjusted for detector confidence	66	$+0.301$	$+0.586$	$+0.443$ $[+0.370, +0.518]$	-0.137 $[-0.272, -0.000]$	$+0.031$ $[+0.014, +0.049]$
All of the above, with target area adjusted	65	$+0.203$	$+0.665$	$+0.434$ $[+0.347, +0.523]$	-0.370 $[-0.515, -0.222]$	$+0.058$ $[+0.041, +0.077]$
Matched on the summed prior, target area adjusted	73	$+0.094$	$+0.534$	$+0.314$ $[+0.229, +0.398]$	-0.299 $[-0.462, -0.133]$	$+0.046$ $[+0.032, +0.062]$
Matched to the absent class's distribution	21	$+0.151$	$+0.325$	$+0.238$ $[+0.132, +0.335]$	-0.072 $[-0.261, +0.152]$	$+0.012$ $[-0.019, +0.040]$

What the zero-shot readout loses is still in the embedding. For each encoder we fit a logistic probe on the frozen embedding in the matched low stratum, drawing each cell with probability proportional to its matching weight and balancing the classes as in Appendix E, and read it without refitting in the matched high stratum. Table 6 shows that the probe separates the high-stratum classes at 0.831

to 0.849 area under the curve, well above the 0.573 of a probe on the annotated context and above the zero-shot score on the same images in every encoder. The same code reproduces the anchor's transfer of 0.904 against 0.528 on the within-class cells.

Table 6: Transfer from the matched low stratum to the matched high stratum, area under the curve for present against absent images averaged over 66 concepts, with 95 percent concept-bootstrap intervals. The last column is the embedding probe minus the zero-shot score on the same images.

Encoder	Embedding probe	Context probe	Zero-shot score	Probe minus score
CLIP ViT-B/32	0.831 [0.809, 0.853]	0.573 [0.553, 0.592]	0.747 [0.714, 0.782]	+0.084 [+0.066, +0.103]
CLIP ViT-B/16	0.838 [0.815, 0.860]	0.573 [0.553, 0.592]	0.758 [0.726, 0.791]	+0.080 [+0.064, +0.097]
CLIP ViT-L/14	0.843 [0.819, 0.866]	0.573 [0.553, 0.592]	0.767 [0.733, 0.801]	+0.076 [+0.061, +0.093]
OpenCLIP ViT-L/14	0.849 [0.825, 0.872]	0.573 [0.553, 0.592]	0.777 [0.745, 0.809]	+0.072 [+0.058, +0.086]
CoCa ViT-L/14	0.848 [0.825, 0.871]	0.573 [0.553, 0.592]	0.767 [0.733, 0.801]	+0.082 [+0.065, +0.098]

K Score adjustments as a test between two bias accounts

Section 5.2 summarizes this test, and we give it in full here. Two bias accounts are in play. A marginal-frequency account says a model over-affirms objects that are common, whatever the scene shows, and a context-prior account says it over-affirms objects that the depicted scene predicts. Conditional score subtraction is here not a mitigation but an intervention whose selective effects distinguish the two. A constant adjustment subtracts the same amount for a concept in every image, so it can correct only how often that concept is affirmed overall, while a conditional one moves with the scene. Equation 4 shows that class-balanced prior estimates preserve raw separation, a condition natural scenes violate. The annotation-based estimate $\widehat{\pi}_O^{\mathrm{ann}}(x)$ averages $+3.84$ on present against -0.19 on absent images in the high stratum, so the context prior predicts presence and what is subtracted as bias is evidence.

We evaluate the adjusted score $\tilde{s}_O^e(x) = s_O(x) - \gamma\, \widehat{\pi}_O^e(x)$ against the raw score, and against a constant adjustment that subtracts the log marginal class prior (Menon et al., 2020; Zhu et al., 2023). The strength γ is selected on a development split and is zero whenever no adjustment helps there, and $e \in \{\mathrm{ann}, \mathrm{det}\}$ names the estimator. The popular split samples negative objects by marginal frequency, which the constant adjustment removes, so it gains 6.0 points [4.3, 7.8] over a raw accuracy of 0.745 while the conditional adjustment gains nothing. The adversarial split samples them by co-occurrence with the depicted context, which a constant offset cannot represent, so the conditional adjustment gains 2.9 points over both alternatives [1.7, 4.2] on a raw accuracy of 0.662 while the constant one gains nothing. The two estimates agree, so the effect follows the prior, not annotation access, and the two splits form the double dissociation the two accounts predict.

The same evaluation forbids reading it as a general improvement, since increasing γ lowers sensitivity from 2.175 at zero strength to 1.985 at the selected one, and false alarms rise at a matched miss rate. Constant adjustment therefore addresses marginal-frequency bias and conditional adjustment context-prior bias, at the cost of genuine contextual evidence. The generative route is weaker, since in POPE's own units a class-constant offset gains 0.3 to 3.1 accuracy points on the frequency split in all seven decoders and nothing on the co-occurrence split, as Appendix P tabulates.

L Supervision and covariate robustness of the identification

Figure 10 displays the five DETR-adjusted coefficients against the null line. Every interval covers zero at comparable baseline sensitivity. Under the Grounding-DINO specification the coefficients shift uniformly in the positive direction and two intervals exclude zero. Across the resulting 10 model-by-proxy fits, one coefficient is negative, CLIP ViT-B/16 at -0.002 under DETR, and it is not significantly below zero. Captioning supervision, which exposes CoCa to fluent descriptions rather than only paired contrast, therefore does not create the negative sensitivity effect required by the representational account.

The directional identification is also unchanged when the number of annotated context objects is added as a covariate. The prior coefficient on sensitivity remains non-negative for the anchor, the coefficient on the absent-distribution position remains positive, and the present-distribution coefficient

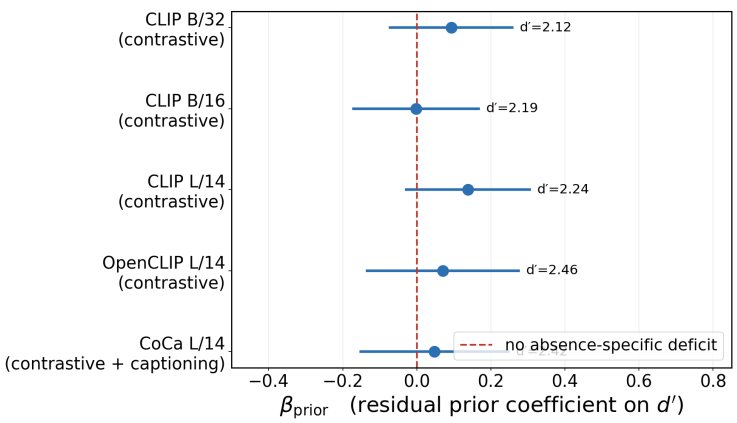


Figure 10: No supervision variant exhibits a significant absence-specific sensitivity deficit under the DETR specification. The coefficient of the prior level on sensitivity is shown with 95 percent confidence intervals for five checkpoints spanning contrastive training on two corpora and contrastive training augmented with captioning, with each checkpoint's baseline sensitivity annotated.

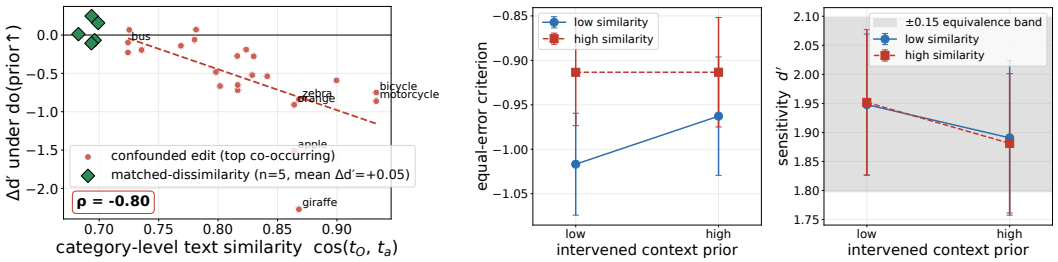
keeps its sign with weaker significance. This covariate adjustment matters because the derivation in Appendix C.1 produces an affine offset that grows with context size and could otherwise masquerade as a prior effect, while the count is strongly collinear with the prior level by construction and therefore inflates the variance of the prior coefficient, which is why the primary specifications exclude it.

M A CONTROLLED COMPOSITIONAL FACTORIAL

The factorial crosses target presence with a pasted context prior and with the pasted object's visual similarity to the target, so that the confusability channel exposed in Section 5 becomes a controlled factor rather than a confound. The design as originally pre-registered measured similarity at the category level, as the text cosine between the context category and the target. That design is structurally infeasible on COCO. Category-level similarity and pointwise mutual information correlate at a Spearman coefficient of 0.863 across the 80-category pool, and none of the 98 category pairs with pointwise mutual information at or above 2.0 falls below a text cosine of 0.70, so the decisive high-prior low-similarity context category does not exist in the corpus.

We therefore amended the design before any outcome was inspected, moving the similarity manipulation to the instance level and defining it as the pasted crop's image-embedding alignment to the target text $\cos(f(\text{crop}), t_O)$, which is the exact quantity that an additive composite leaks into the score. Data-driven terciles at 0.158 and 0.184 replace the text-cosine thresholds, and the realized crop similarity additionally enters as a continuous covariate. After building the crop bank the census admits the decisive low-similarity presence-by-prior square with at least five crops per cell in 58 of 77 concepts, and the realized factorial contains 276 cells, of which 60 concepts supply the paired low-similarity contrast quoted in the main text, confirming that instance-level similarity decouples what category-level statistics couple.

Figure 11 shows the category-level hazard, and Table 7 reports the main effects of the anchor mixed model, while the contrasts quoted in the main text are paired simple effects within the decisive low-similarity square, so the two sets of estimates are close but not interchangeable. In the mixed model the prior raises the criterion and the absent-distribution position while its effect on sensitivity stays inside the equivalence band of Appendix O, and in the low-similarity square the paired do-prior contrast moves the criterion by $+0.052$ $[+0.017, +0.087]$ and the absent position by $+0.088$ $[+0.044, +0.132]$ with the sensitivity change at -0.061 $[-0.147, +0.025]$. The confusability channel is not resolved for the anchor at -0.087 $[-0.520, +0.345]$, where the realized similarity range is compressed. Two qualifications are reported with equal prominence. The interventional inflation is carried mostly by the absent cell, with the present-side position rising by only $+0.015$ and not significantly. We attribute this pattern to score saturation on present images where the object's own evidence dominates and treat it as a boundary of the additive approximation rather than as support



(a) The insertion confound.

(b) The factorial resolution.

Figure 11: Panel (a) shows that the apparent sensitivity drop from pasting co-occurring objects is predicted by the category-level text cosine to the target and is absent for the dissimilar pastes drawn as diamonds. In panel (b) the similarity factor is instead the instance-level alignment of the pasted crop to the target text, and the intervened prior raises the criterion at low similarity while sensitivity stays inside the equivalence band.

for representational degradation, which would lower sensitivity and is not observed for the anchor. The factorial therefore supports a bias-dominant response for the anchor.

A complementary removal-like edit deletes a high-prior low-similarity context object by diffusion inpainting. The result is inconclusive at three concepts, because the mean score change of $+0.0054$ runs opposite to the predicted decrease and is of the same order as the $+0.0024$ produced by a no-operation inpainting control, with per-concept signs that disagree, so we do not cite it as independent confirmation.

Table 7: Mixed-effects main-effect coefficients of the compositional factorial on the anchor model, with concept random intercepts, the realized crop similarity as a continuous covariate, and 90 percent confidence intervals. Positions are measured in the prior-independent per-concept standardized frame. The paired simple effects within the low-similarity square, quoted in the text, are estimated separately.

Term	Sensitivity d'	Criterion	Absent-distribution position
Prior	-0.058 $[-0.146, +0.031]$	$+0.054$ $[+0.019, +0.088]$	$+0.093$ $[+0.048, +0.138]$
Similarity	-0.087 $[-0.520, +0.345]$	$+0.101$ $[-0.068, +0.270]$	$+0.116$ $[-0.103, +0.335]$
Prior by similarity interaction	-0.032 $[-0.160, +0.096]$	-0.050 $[-0.099, +0.000]$	-0.061 $[-0.126, +0.003]$

N WHERE THE CONFUSABILITY ADJUSTMENT IS IDENTIFIED

The two proxies of Section 5 correlate at 0.907 with each other and, averaged over the encoders, at 0.468 and 0.583 with each encoder's absent-image score, and Grounding DINO covers 53 to 57 percent of absent instances, so every comparison between them uses its covered images. Table 8 gives the anchor's regressions under both proxies, together with the five-checkpoint sensitivity coefficients under a concept random intercept and under concept fixed effects. The position coefficients agree across proxies and specifications, while the sensitivity coefficients agree only under fixed effects, for the reason that follows.

Table 8: Regressions across 77 concepts under two proxies, CLIP ViT-L/14 as the anchor. Entries give the prior coefficient with a 95 percent interval or a significance level, and the last two columns average the five checkpoints under a concept random intercept and under concept fixed effects.

Proxy specification	Anchor sensitivity d'	Absent position	Present position	Five-model d'	Five-model within-concept d'
DETR	$+0.138$ $[-0.028, +0.304]$	$+0.234$ ($p = 2.0\times10^{-14}$)	$+0.215$ ($p = 0.0005$)	$+0.069$	-0.004
Grounding DINO	$+0.293$ $[+0.047, +0.538]$	$+0.165$ ($p = 9.7\times10^{-5}$)	$+0.174$ ($p = 0.007$)	$+0.250$	-0.002

The prior level is a within-concept factor, because every concept contributes one cell at each level, so the confound a confusability covariate exists to remove is whether a concept's high-prior cells hold more detectable instances than that same concept's low-prior cells. A concept random intercept

nevertheless fits one covariate slope from both within- and between-concept variation, and between-concept variation dominates, at two to three times the within-concept variance of the proxy. The two sources disagree in sign. Between concepts the proxy predicts sensitivity at -0.97 to -1.53 standardized units for the encoders and -0.79 to -2.19 for the decoders, because the categories a detector is confident about are frequent, large ones that a contrastive encoder separates worst, while within a concept the correlation is -0.05 to $+0.09$ for the encoders and -0.34 to $+0.05$ for the decoders. Applying a negative between-concept slope to a positive within-concept confusability increment raises the prior coefficient, and raises it further for the proxy with the stronger between-concept correlation, which is the published offset.

We therefore re-estimate with concept fixed effects, which for a two-level within factor is the paired regression of the high-minus-low change in the outcome on the changes in the covariates, with the intercept as the prior effect at a zero covariate change and log frequency dropping out because it does not vary within a concept. A Mundlak specification that enters the concept mean and the within-concept deviation of each covariate as separate regressors reproduces that intercept to three decimals in all ten contrastive fits, so the change of frame is the standard within-between decomposition rather than an artifact of the paired arithmetic. Table 9 reports the outcome.

Table 9: The prior coefficient on sensitivity and on the absent-class position, averaged over checkpoints, under the published concept random intercept and under concept fixed effects. The gap is between the two proxy specifications, which is the quantity the paper previously reported as leaving sensitivity unidentified.

Family	Frame	Grounding DETR	Grounding DINO	Proxy gap per checkpoint	Significantly negative
Five encoders	concept random intercept	$+0.069$	$+0.250$	0.180	0 of 5 under either
Five encoders	concept fixed effects	-0.004	-0.002	0.024	0 of 5 under either
Seven decoders	concept fixed effects	-0.502	-0.270	0.232	5 of 7 under DETR, 1 under Grounding DINO

For the encoders this answers which proxy to believe. The two proxies agree to 0.024 per checkpoint instead of 0.180, no interval excludes zero under either, the range is -0.094 to $+0.089$, no checkpoint places more than half its bootstrap mass beyond the -0.15 margin, and the absent-position coefficient survives the change of frame in all ten fits at $+0.194$ to $+0.393$. For the decoders it does not. Their proxies disagree within concepts as well, the gap falling only from 0.63 to 0.232, and five of the seven coefficients stay significantly negative under DETR while one does under Grounding DINO, so the decoders keep a sensitivity cost under the weaker control that the encoders do not. Their absent-position coefficient survives in thirteen of fourteen fits, the exception being Qwen2-VL under Grounding DINO, which is the same checkpoint and control that Appendix O already isolates. The reading we take from this is that the criterion result is invariant to the frame while the sensitivity result is not, and that the encoders and the decoders should be reported separately rather than under one hedge.

Agreement between the proxies is not the same as precision, and the change of frame costs the second. Concept fixed effects discard between-concept variation, which was tightening the published interval at the price of identifying the covariate slope in the wrong place, so the within-concept estimates carry wider intervals. Redoing the paper's non-inferiority analysis in the new frame, with the same dependence-preserving bootstrap that resamples one set of concepts for every checkpoint before averaging with equal weight, gives the entries in Table 11, which separate the three quantities the paired frame offers. The Grounding-DINO specification that passed in the superseded frame, at a lower bound of -0.043, we withdraw as an artifact of the same misplaced identification that produced its point estimate.

Two proxies cannot say which of them is the outlier, so we add a third that differs from both on every axis we can vary. OWLv2 (Minderer et al., 2023) has a contrastively pretrained image tower rather than a ResNet-50 or a Swin transformer, its text queries enter as embeddings compared against per-patch predictions rather than through a fused cross-attention decoder, and its supervision is web self-training with pseudo-boxes rather than COCO or Objects365. Over the cached images it separates annotated presence from annotated absence at a pooled area under the curve of 0.988, and at 0.993 averaged over the eighty categories, so it measures target evidence at least as well as the other two, and it is the least redundant of the three, agreeing with DETR at a rank correlation of 0.825 where

DETR and Grounding DINO agree at 0.914. Table 10 reports every within-concept coefficient under all three.

For the encoders the agreement was not an accident of having two. The three specifications give per-checkpoint means of -0.004, -0.002 and $+0.014$ with a mean spread across the three of 0.032, and the absent-class coefficient stays positive in all five checkpoints under all three, between $+0.153$ and $+0.393$. For the decoders the third proxy makes the non-identification explicit instead of resolving it, since the per-checkpoint means run -0.502, -0.270 and $+0.129$, so the residual changes sign across three specifications that all measure detectability well. One reading is that a proxy sharing more of the encoder's representation absorbs more of the effect, and the three linear correlations with the anchor's absent-image score do run in the same order at 0.45, 0.52 and 0.54, but the corresponding rank correlations at 0.56, 0.53 and 0.58 do not, so the ordering is not stable under the choice of correlation measure and we rest nothing on it. What survives every specification is the criterion result, since the decoders' absent-class coefficient stays positive under all three proxies in six of the seven, Qwen2-VL being the exception under both open-vocabulary ones.

Table 10: The within-concept prior coefficient under three confusability proxies, on sensitivity and on the absent-class position. The proxies are a COCO-trained DETR, an open-vocabulary Grounding DINO, and an OWLv2 detector with a contrastively pretrained image tower and web self-training. Each entry is the intercept of the paired regression described above, so it is the effect at a zero change in the covariates.

	Sensitivity			Absent-class position		
Model	DETR	Grounding DINO	OWLv2	DETR	Grounding DINO	OWLv2
CLIP ViT-B/32	$+0.023$	$+0.089$	$+0.076$	$+0.355$	$+0.300$	$+0.221$
CLIP ViT-B/16	-0.076	-0.094	-0.053	$+0.325$	$+0.299$	$+0.223$
CLIP ViT-L/14	$+0.069$	$+0.070$	$+0.072$	$+0.244$	$+0.194$	$+0.153$
OpenCLIP ViT-L/14	-0.010	-0.022	-0.015	$+0.321$	$+0.292$	$+0.207$
CoCa ViT-L/14	-0.026	-0.052	-0.010	$+0.393$	$+0.386$	$+0.232$
Qwen2-VL-7B	-0.304	-0.190	$+0.243$	$+0.198$	$+0.058$	-0.017
InstructBLIP-7B	-0.049	$+0.086$	$+0.491$	$+0.319$	$+0.292$	$+0.265$
Qwen3-VL-4B	-0.722	-0.425	$+0.073$	$+0.488$	$+0.276$	$+0.243$
Qwen3-VL-8B	-0.804	-0.528	$+0.148$	$+0.538$	$+0.339$	$+0.321$
InternVL3.5-8B	-0.354	-0.043	$+0.136$	$+0.701$	$+0.463$	$+0.430$
Gemma-3-12B	-0.569	-0.355	-0.020	$+0.311$	$+0.121$	$+0.108$
Llama-3.2-11B-V	-0.713	-0.435	-0.167	$+0.489$	$+0.269$	$+0.143$

The intercept, which the second and third rows of each family report, is the effect at a zero change in the covariates, and the observed mean change in confusability is 0.92 standardized units under DETR and 1.26 under Grounding DINO, so that row extrapolates almost a full standard deviation and inherits the variance of doing so. The total effect is the mean high minus low change with no covariate. It is what the prior does as the data realize it, it is identical under both proxies because a covariate cannot alter a difference of means, and for the encoders it is $+0.007$ with a lower bound of -0.108 that clears the margin and survives dropping any one checkpoint. The middle rows say why the proxies once looked decisive and no longer do, since the part of the total that the confusability change can account for is -0.001 under DETR and -0.003 under Grounding DINO, and an adjustment worth less than 0.01 cannot be what separates two specifications by 0.18. For the decoders the same decomposition reads differently. The total is -0.563 with a lower bound of -0.759, and Grounding DINO attributes -0.307 of it to confusability against -0.068 for DETR, which is where their remaining disagreement lives.

Because the encoders' bound is marginal, the claim we rest on is one that needs neither a proxy nor a margin. It concerns the per-concept changes themselves, and it starts from the observation that their spread is not a sample-size problem. The standard deviation of the change across concepts is 0.57 to 0.72 for the encoders, of which the analytic sampling variance of d' in the two cells accounts for 10 to 14 percent, so 86 to 90 percent of it is variation of the effect from one concept to the next, and no number of images per cell would reduce it. Given that spread, an equal-weight mean near zero could hide a loss on some concepts and a gain on others, so we ask how the changes are distributed. For the encoders they carry no systematic sign, since 0.338 to 0.416 of concepts lie below -0.15 against 0.364 to 0.519 above $+0.15$, and a signed-rank test on the per-concept changes is significant in none of the five, the smallest probability being 0.09 for CLIP L/14 in the positive direction. For the decoders they are asymmetric, since 0.455 to 0.766 lie below the negative margin

Table 11: One-sided non-inferiority of the equal-weight mean within-concept prior effect on d' against the meaningful-loss margin of -0.15, from ten thousand bootstrap draws that resample the same 77 concepts for every checkpoint. The total effect is the mean high minus low change and is the same under both proxies, the attributed part is the fitted covariate slope times the observed mean covariate change, and the residual is the intercept at a zero covariate change. The last column drops one checkpoint at a time and reports the smallest bound.

Family	Estimand	Proxy specification	Estimate	Lower bound	Probability at or below margin	Worst leave-one-out bound
Five encoders	total effect	either	$+0.007$	-0.108	0.013	-0.134
Five encoders	attributed to confusability	DETR	-0.001	-0.085	0.003	-0.092
Five encoders	attributed to confusability	Grounding DINO	-0.003	-0.145	0.046	-0.157
Five encoders	residual at zero change	DETR	-0.004	-0.181	0.090	-0.205
Five encoders	residual at zero change	Grounding DINO	-0.002	-0.243	0.166	-0.273
Seven decoders	total effect	either	-0.563	-0.759	0.999	-0.848
Seven decoders	attributed to confusability	DETR	-0.068	-0.227	0.187	-0.242
Seven decoders	attributed to confusability	Grounding DINO	-0.307	-0.699	0.741	-0.747
Seven decoders	residual at zero change	DETR	-0.502	-0.798	0.964	-0.862
Seven decoders	residual at zero change	Grounding DINO	-0.270	-0.749	0.673	-0.830

against 0.143 to 0.429 above the positive one, and the signed-rank test is significant in six of the seven, the exception being InstructBLIP, whose median change is -0.040. This is the sharpest form of the separation between the two families, since it uses no confusability proxy, no covariate, and no equivalence margin, and it says that contextual expectedness changes an encoder's discriminability in both directions about equally often while it lowers a decoder's.

O Confusability-proxy dependence and sensitivity tests

The second confusability measurement tests how strongly results depend on one COCO-trained detector. DETR and Grounding DINO agree at a mean correlation of 0.907 across model-specific analysis sets, while their mean correlations with the encoders' scores are 0.468 and 0.583. A preregistered threshold classified these correlations as nonredundant, but it does not validate either detector as ground-truth target evidence. Table 12 reports every sensitivity coefficient. Swapping DETR for Grounding DINO adds between 0.15 and 0.23 to the coefficient in each checkpoint, on average 0.18. No Grounding-DINO fit is negative, under DETR only CLIP ViT-B/16 at -0.002 falls below zero, and under the joint specification none stays negative. The position coefficients are positive in all fifteen checkpoint-by-specification fits, significantly so in all fifteen for both classes. We interpret this as robustness of upward position movement together with material proxy dependence of sensitivity. The equal-weight mean of the five joint-column coefficients is $+0.144$, between the DETR and Grounding DINO means.

Table 12: Prior coefficients on sensitivity for five contrastive checkpoints under DETR, Grounding DINO, and joint proxy specifications. A negative coefficient is the direction predicted by prior-induced representational degradation. The joint column reports point estimates only, and Table 13 does not evaluate joint non-inferiority.

Model	DETR	Grounding DINO	Joint
CLIP ViT-B/16	-0.002	$+0.154$	$+0.063$
CLIP ViT-B/32	$+0.093$	$+0.327$	$+0.191$
CLIP ViT-L/14	$+0.138$	$+0.293$	$+0.201$
CoCa ViT-L/14	$+0.048$	$+0.211$	$+0.114$
OpenCLIP ViT-L/14	$+0.071$	$+0.263$	$+0.151$

The Grounding-DINO cache covers 57 percent of low-prior and 53 percent of high-prior absent instances, with all 77 concepts retaining at least 20 images per cell. On this identical covered subset, the Qwen2-VL behavioral prior log-odds is $+0.737$ without a proxy, $+0.570$ with DETR, and $+0.174$ with Grounding DINO. The loss of significance is therefore associated with proxy specification rather than coverage. It remains ambiguous whether Grounding DINO removes nuisance confusability, the contextual effect of interest, genuine target evidence, or a mixture. Two-way clustering by concept and image gives the same intervals to three decimal places, so image reuse does not alter this conclusion.

The non-inferiority analysis asks whether the proxy-adjusted contrastive sensitivity association exceeds the meaningful-loss margin of -0.15 standardized units. Because all five checkpoints share concepts and images, the primary bootstrap resamples the same 77 concepts for every model, refits each mixed model with a concept random intercept, and averages checkpoints with equal weight. Table 13 reports the results in that frame, where Grounding DINO passes including every leave-one-model-out analysis and DETR is unresolved with a one-sided 95 percent lower bound of -0.155, a bootstrap probability 0.053 at or below the margin, and a worst leave-one-model-out bound of -0.183. We report the table because it is the analysis the literature would expect, but Appendix N supersedes it, because the covariate slope here is identified from between-concept variation, which both biases the point estimate upward and narrows the interval, so the Grounding-DINO pass is withdrawn and Table 11 carries the version of this test we stand behind.

Table 13: Dependence-preserving one-sided non-inferiority analysis for the prior coefficient on sensitivity under the two single-proxy specifications. The lower bound is compared with the meaningful-loss margin of -0.15. The worst leave-one-model-out value is the smallest lower bound after dropping one checkpoint.

Proxy specification	Equal-weight coefficient	Lower bound	Probability at or below margin	Worst leave-one-model-out bound
DETR	$+0.038$	-0.155	0.053	-0.183
Grounding DINO	$+0.217$	-0.043	0.009	-0.076

These tests do not support a general invariance claim. The DETR specification leaves the non-inferiority question unresolved, whereas Grounding DINO passes under its proxy-specific estimand. The low-similarity intervention separately places the anchor change at -0.061 $[-0.147, +0.025]$. The robust result across this appendix is upward score-position movement, not a common sensitivity value.

P THE GENERATIVE CHECKPOINTS IN DETAIL

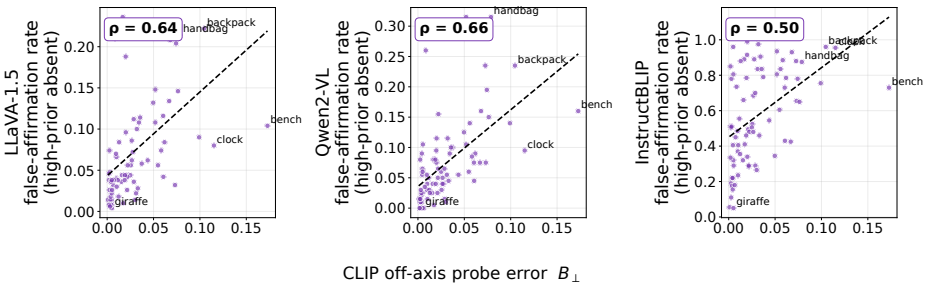


Figure 12: The shared axis transfers across architectures and tasks. The per-concept false-affirmation rate on high-expectedness absent images predicts the off-axis probe-error index $B_\perp$ of CLIP ViT-L/14 at Spearman correlations of 0.637 for LLaVA-1.5, 0.661 for Qwen2-VL, and 0.499 for InstructBLIP.

We repeated the two cross-architecture analyses on every checkpoint that the protocol can be run on, reading the decision variable as the affirmative minus the negative answer logit in a single forward pass and pinning model revisions for reproducibility. The per-concept hallucination rate ranges from 0.004 to 0.236 with mean 0.075 for LLaVA and tracks the anchor's off-axis probe-error index $B_\perp$ at Spearman correlations between 0.499 and 0.661 over the seven decoders of Table 14 and at 0.637 for LLaVA, each over 77 concepts and each with significance below 10^{-5}, with three of those scatter plots in Figure 12. This is the same association as Figure 13 measured on the other literature's instrument rather than on ours.

Table 14 states the paper's claim in that protocol's own units. Accuracy falls from negatives sampled at random to negatives sampled by frequency to negatives sampled by co-occurrence with the depicted scene in all seven checkpoints, and the false-alarm rate is higher on the co-occurrence split than on the random split in all seven, by margins running from 0.0003 for InstructBLIP, which is why its two entries agree to three decimals, to 0.086 for Llama-3.2-Vision, and rising monotonically

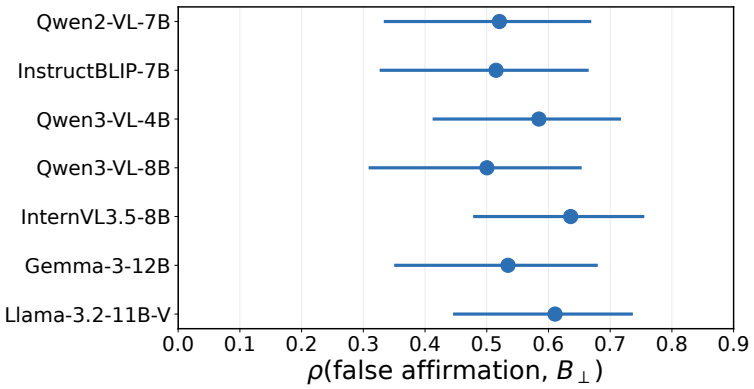


Figure 13: The same association on every checkpoint of the signal-detection design, measured with one instrument throughout. Each entry is the Spearman correlation across the 77 concepts between a checkpoint's rate of false affirmative answers on high-expectedness absent images at the forced-choice threshold and the off-axis probe-error index $B_{\perp}$ of CLIP ViT-L/14, with a 95 percent interval from the Fisher transformation. LLaVA-1.5 is absent because its rate comes from the POPE protocol and appears in Figure 12.

across the three in four of them. The absolute levels are not comparable with published POPE numbers, because the negatives are resampled here from the annotations of this study's corpus rather than taken from the released question files, so the comparison to read is the one across splits within a row.

Table 15 reports the adjustment dissociation on the same runs, and it separates cleanly on one side only. A class-constant offset helps exactly where a marginal-frequency account predicts, on the frequency-sampled split, with an interval excluding zero in five of the seven, and does nothing at all on the co-occurrence-sampled split in every one. The context-conditional estimate is the weaker half, resolving on two checkpoints, touching zero on a third, and leaving the remaining four unsettled. On InstructBLIP it moves the adversarial criterion in the predicted direction, with the false-alarm rate falling from 0.081 to 0.068 while the miss rate rises from 0.198 to 0.217, without a net accuracy gain because that model is already conservative on that split. We therefore scope the accuracy form of the conditional dissociation to the checkpoints where it resolves, and note that the anchor encoder of Appendix K, where it gains 2.9 points, remains the case the argument rests on. What holds across every decoder is the direction of the constant offset and the correlation between hallucination and the operational probe-error index.

The adjustment strength is selected on the same split that evaluates it, so both the estimate and its interval inherit the split's size, and a larger split is a check on the selection and not only on the interval. Recomputing all seven splits with two and a half times as many images changes the picture in ways that a narrower interval alone would not. InstructBLIP's null becomes a gain of 1.38 points $[+0.84, +1.92]$, Qwen2-VL's apparent 1.11 points becomes exactly zero because no strength improves the larger development split, InternVL3.5 and Llama-3.2-Vision stay positive at 0.47 $[+0.04, +0.90]$ and 0.93 points $[+0.37, +1.49]$, and the remaining three stay unresolved with intervals narrower by 14 to 32 percent and with every point estimate above zero. Both arms sharpen. The class-constant gain on the frequency-sampled split is now resolved in all seven, between 0.45 and 3.92 points, where the smaller split left two of them touching zero, and the same offset becomes significantly harmful on the co-occurrence-sampled split in two checkpoints, at -0.24 and -0.22 points, where the smaller split had shown exactly zero. Two readings follow. The direction of the dissociation is supported on both arms, since a class-constant offset helps only where negatives are sampled by frequency and hurts where they are sampled by context, while the context-conditional estimate helps only on the latter, in three of the seven with an interval excluding zero and nowhere with a negative point estimate. And the per-checkpoint magnitude on the conditional arm remains the weaker quantity, so we report the direction as the result and treat those magnitudes as provisional.

We additionally apply the continuous signal detection analysis to $z_O(x)$, the affirmative answer logit minus the negative answer logit, on seven checkpoints that span five connector designs. These are the projection of image tokens into the text sequence for the three Qwen models, a Q-Former

Table 14: The standard negative-sampling protocol on the raw decision variable, with negatives resampled from this study's corpus. The three groups are the rules by which negatives are sampled, at random, by marginal frequency, and by co-occurrence with the depicted scene, and within each group the first column is accuracy and the second is the false-alarm rate on the negatives, and the last column repeats the rank correlation between the per-concept false-affirmation rate on high-expectedness absent images and the anchor encoder's off-axis probe-error index.

	Random		Frequency		Co-occurrence		
Model	Accuracy	False alarms	Accuracy	False alarms	Accuracy	False alarms	Correlation with $B_\perp$
Qwen2-VL-7B	0.950	0.052	0.930	0.086	0.896	0.127	$+0.661$
InstructBLIP-7B	0.916	0.081	0.891	0.120	0.865	0.081	$+0.499$
Qwen3-VL-4B	0.957	0.033	0.925	0.092	0.918	0.072	$+0.585$
Qwen3-VL-8B	0.961	0.032	0.919	0.105	0.913	0.100	$+0.642$
InternVL3.5-8B	0.953	0.027	0.927	0.063	0.900	0.083	$+0.651$
Gemma-3-12B	0.915	0.030	0.885	0.088	0.879	0.105	$+0.573$
Llama-3.2-11B-V	0.939	0.052	0.919	0.057	0.868	0.138	$+0.641$

Table 15: The score-adjustment dissociation across seven generative checkpoints and LLaVA-1.5, reported as accuracy differences with 95 percent paired-bootstrap intervals over ten thousand shared resamples. A zero entry means the development split selected no adjustment, so the method reproduces the raw score exactly.

Model	Adversarial, conditional minus raw	Popular, constant minus raw	Adversarial, constant minus raw
LLaVA-1.5	$+0.004$ $[-0.009, +0.017]$	$+0.017$ $[+0.005, +0.029]$	0.000
Qwen2-VL	$+0.011$ $[+0.000, +0.022]$	$+0.011$ $[+0.003, +0.020]$	0.000
InstructBLIP	-0.001 $[-0.012, +0.009]$	$+0.031$ $[+0.019, +0.044]$	0.000
Qwen3-VL-4B	$+0.001$ $[-0.002, +0.003]$	$+0.014$ $[+0.004, +0.023]$	0.000
Qwen3-VL-8B	$+0.007$ $[-0.001, +0.015]$	$+0.018$ $[+0.011, +0.025]$	-0.001 $[-0.004, +0.003]$
InternVL3.5	$+0.008$ $[+0.003, +0.015]$	$+0.011$ $[+0.004, +0.018]$	0.000
Gemma-3	$+0.001$ $[-0.006, +0.007]$	$+0.007$ $[+0.000, +0.013]$	0.000
Llama-3.2-Vision	$+0.012$ $[+0.003, +0.021]$	$+0.003$ $[-0.002, +0.009]$	-0.001 $[-0.005, +0.004]$

for InstructBLIP, InternViT with a pixel-shuffle projector for InternVL3.5, a SigLIP-2 tower with a projector for Gemma-3, and gated cross-attention for Llama-3.2-Vision, the only one of the five that does not concatenate image tokens into the text sequence. InternVL is also the family an earlier version of this study had to abandon because its released inference code returned invalid logits under our library version, so the official port is what makes it available here. Table 16 reports the proxy-adjusted coefficients and Table 17 the criterion, position and behavioral quantities.

The criterion rises in all seven, from $+0.146$ to $+0.509$, with InternVL3.5 and Qwen3-VL-8B giving the largest values, and the rise is class asymmetric in all seven, since the absent position moves more than the present position by between $+0.209$ and $+0.650$. The unadjusted sensitivity change is negative in all seven and tail-robust, with per-concept medians between -0.040 and -1.062, signed-rank significance below 10^{-5} in five of them and 0.01 in Qwen2-VL, and no resolved change in InstructBLIP. It nevertheless does not identify a sensitivity cost, because the two proxies disagree systematically. Four coefficients are significantly negative under DETR, none is under Grounding DINO, two turn significantly positive there, and the sign reverses between the two specifications in five of the seven, InternVL3.5 being the starkest case because its reversal reaches significance under both proxies. The absent-position coefficients are the stable quantity, positive under both proxies in every checkpoint and significantly so under both in five of the seven, the two exceptions being the Grounding-DINO intervals of Qwen2-VL and Gemma-3, while the behavioral prior coefficient on absent images survives both proxies in six of the seven. We describe this pattern as class-asymmetric score movement rather than a decoder-localized component, because dispersion compression and score saturation are observationally consistent with the same result.

The within-family scale pair separates model size from the effect. Qwen3-VL at four billion and at eight billion parameters agree within their intervals on the criterion change, the absent-position change, the unadjusted sensitivity change and the proxy-adjusted sensitivity coefficient, so the smaller checkpoint is not the more biased one. Scale within a family therefore neither creates nor removes the signature at the sizes we can run.

Table 18 decomposes the unadjusted sensitivity change into contributions from the concept-level raw separation $D_{O,p}$ and pooled dispersion $\sigma_{O,p}$ defined in Section 3. The reported model-level val-

Table 16: Generative signal detection and behavior over the 77-concept design, for seven checkpoints spanning five connector designs and a within-family scale pair. Unadjusted columns give the low-to-high change in standardized units. Adjusted columns give the prior coefficient from the identification regression of Section 5 under each detector measurement, together with mean present target area and log frequency, under concept-clustered inference. The last column gives the affirmation rate on absent images at the fixed threshold $z_O(x) = 0$ at the two prior levels.

	Absent position			Sensitivity d'			False affirmation
Model	Unadjusted	DETR	Grounding DINO	Unadjusted	DETR	Grounding DINO	at $z_O(x) = 0$
Qwen2-VL-7B	$+0.295$	$+0.181$ $[+0.073, +0.288]$	$+0.010$ $[-0.135, +0.154]$	-0.285	-0.180 $[-0.512, +0.153]$	$+0.220$ $[-0.257, +0.698]$	0.044→0.096
InstructBLIP-7B	$+0.372$	$+0.315$ $[+0.260, +0.370]$	$+0.279$ $[+0.201, +0.357]$	-0.007	$+0.147$ $[-0.296, +0.590]$	$+0.743$ $[+0.089, +1.398]$	0.324→0.569
Qwen3-VL-4B	$+0.667$	$+0.481$ $[+0.409, +0.553]$	$+0.260$ $[+0.178, +0.343]$	-0.893	-0.539 $[-0.963, -0.115]$	$+0.265$ $[-0.366, +0.895]$	0.004→0.030
Qwen3-VL-8B	$+0.695$	$+0.532$ $[+0.455, +0.608]$	$+0.324$ $[+0.233, +0.414]$	-0.907	-0.613 $[-1.098, -0.129]$	$+0.174$ $[-0.536, +0.884]$	0.013→0.064
InternVL3.5-8B	$+0.834$	$+0.671$ $[+0.564, +0.779]$	$+0.389$ $[+0.258, +0.521]$	-0.502	-0.215 $[-0.475, +0.044]$	$+0.492$ $[+0.094, +0.890]$	0.017→0.108
Gemma-3-12B	$+0.416$	$+0.299$ $[+0.216, +0.382]$	$+0.095$ $[-0.004, +0.193]$	-0.580	-0.400 $[-0.731, -0.070]$	$+0.143$ $[-0.329, +0.614]$	0.066→0.142
Llama-3.2-11B-V	$+0.620$	$+0.464$ $[+0.357, +0.572]$	$+0.207$ $[+0.072, +0.342]$	-0.764	-0.579 $[-0.855, -0.303]$	-0.015 $[-0.409, +0.379]$	0.030→0.104

Table 17: The generative decomposition over seven checkpoints, 77 concepts, high minus low prior and paired across concepts. Median and signed-rank columns corroborate the sensitivity change without assuming symmetry, because generative d' is heavy tailed. The last two columns give the prior coefficient on false affirmation over absent instances in log-odds, under each detector measurement with concept-clustered inference. A concept bootstrap over the same 77 concepts agrees with the clustered interval on whether zero is excluded in every one of those fourteen cells.

	Standardized position changes				Sensitivity, tail-robust		Behavioral prior, absent	
Model	Criterion	Absent	Present	Asymmetry	Median $\Delta d'$	Signed rank	DETR	Grounding DINO
Qwen2-VL-7B	$+0.146$ $[+0.067, +0.224]$	$+0.295$	-0.002	$+0.297$	-0.353	0.01	$+0.69$ $[+0.47, +0.90]$	$+0.17$ $[-0.09, +0.43]$
InstructBLIP-7B	$+0.267$ $[+0.197, +0.337]$	$+0.372$	$+0.163$	$+0.209$	-0.040	0.5	$+0.97$ $[+0.82, +1.12]$	$+0.72$ $[+0.52, +0.91]$
Qwen3-VL-4B	$+0.392$ $[+0.325, +0.459]$	$+0.667$	$+0.117$	$+0.550$	-1.062	1e-07	$+1.73$ $[+1.30, +2.17]$	$+0.98$ $[+0.38, +1.59]$
Qwen3-VL-8B	$+0.431$ $[+0.367, +0.494]$	$+0.695$	$+0.166$	$+0.529$	-1.019	5e-08	$+1.40$ $[+1.10, +1.69]$	$+0.65$ $[+0.25, +1.05]$
InternVL3.5-8B	$+0.509$ $[+0.439, +0.577]$	$+0.834$	$+0.183$	$+0.650$	-0.592	7e-06	$+1.79$ $[+1.52, +2.07]$	$+1.24$ $[+0.93, +1.54]$
Gemma-3-12B	$+0.250$ $[+0.192, +0.306]$	$+0.416$	$+0.083$	$+0.332$	-0.600	5e-06	$+0.77$ $[+0.60, +0.94]$	$+0.34$ $[+0.15, +0.54]$
Llama-3.2-11B-V	$+0.388$ $[+0.326, +0.446]$	$+0.620$	$+0.155$	$+0.466$	-0.867	2e-08	$+1.19$ $[+0.86, +1.52]$	$+0.63$ $[+0.31, +0.95]$

ues aggregate the corresponding concept-level changes. The decomposition averages the two orders in which separation and dispersion can be changed, giving a two-factor Shapley allocation whose contributions sum to the observed sensitivity change. Every model loses raw mean separation at high prior, and what differs is the dispersion. Qwen2-VL holds it fixed and therefore exposes the separation loss in standardized sensitivity, InstructBLIP and InternVL3.5 contract it so that a positive dispersion contribution offsets part of the loss, and the remaining four expand it so that both contributions push the same way, which is why the two Qwen3-VL checkpoints show the largest unadjusted changes. The InstructBLIP row is the informative one, because a separation contribution of -0.213 is almost cancelled by a dispersion contribution of $+0.206$, so a flat standardized change there records two offsetting movements rather than an absence of movement. This matters for how much of the standardized loss counts as lost information, because growth in pooled dispersion lowers d' while leaving class separation untouched, and in the four checkpoints where dispersion expands it accounts for between a third and two fifths of the change, at 37, 34, 39 and 34 percent for Qwen3-VL-8B, Qwen3-VL-4B, Gemma-3 and Llama-3.2-Vision. This is the same arithmetic that Appendix R uses to show that a standardized bias index can be moved by adding noise to a decision variable. Raw separation does fall in every checkpoint, so the observation rescales the sensitivity result rather than removing it.

Table 18: Decomposition of the unadjusted generative sensitivity change. The separation contribution holds dispersion fixed in both possible orders, and the dispersion contribution is defined analogously.

Model	Mean separation, low to high	Pooled dispersion, low to high	Sensitivity change	Separation contribution	Dispersion contribution
Qwen2-VL-7B	$4.80 \to 4.41$	$1.21 \to 1.21$	-0.285	-0.312	$+0.027$
InstructBLIP-7B	$5.67 \to 5.38$	$1.52 \to 1.47$	-0.007	-0.213	$+0.206$
Qwen3-VL-4B	$28.10 \to 24.99$	$5.13 \to 5.40$	-0.893	-0.591	-0.302
Qwen3-VL-8B	$16.85 \to 15.06$	$3.03 \to 3.24$	-0.907	-0.569	-0.338
InternVL3.5-8B	$13.43 \to 11.32$	$3.02 \to 2.88$	-0.502	-0.715	$+0.213$
Gemma-3-12B	$24.46 \to 22.09$	$6.26 \to 6.71$	-0.580	-0.354	-0.227
Llama-3.2-11B-V	$9.74 \to 8.58$	$2.21 \to 2.36$	-0.764	-0.501	-0.263

At the behavioral rule $z_O(x) > 0$, Qwen2-VL's absent false-affirmation rate rises from 0.044 to 0.096 and InstructBLIP's rises from 0.324 to 0.569, while both hit rates increase by only 0.008 from already high baselines. Concept bootstrap intervals for the false-affirmation changes are [0.039, 0.066] and [0.210, 0.280]. Logistic regressions with a presence-by-prior interaction, detector confidence, mean target area, and log frequency confirm the class asymmetry. The InstructBLIP absent-class prior log-odds remain positive with confidence intervals excluding zero under DETR and Grounding DINO. The Qwen2-VL coefficient is positive under DETR but is not statistically distinguishable from zero after conditioning on Grounding DINO. The same-subset analysis in Appendix O attributes this difference to the covariate specification rather than coverage without identifying which signal the measurement removes. Conventional binary sensitivity decreases when false affirmations rise against ceiling hit rates, so continuous $z_O(x)$ remains the primary signal detection variable.

The newest checkpoint, Qwen3-VL-8B-Instruct at a pinned revision, reproduces the affirmative signature and sharpens the sensitivity question at once. Its criterion change has a median of $+0.433$ with 71 of 77 concepts positive, its false-affirmation rate on absent images rises while the hit rate moves only from 0.933 to 0.956, and its behavioral prior coefficient survives both proxies as InstructBLIP's does and Qwen2-VL's does not. Its unadjusted sensitivity change of -0.907 [-1.274, -0.505] is also the largest we measure, and it is the clearest case of the disagreement, since the DETR specification resolves that change as a loss and Grounding DINO does not resolve it at all.

Q THE ABSENCE DETECTOR

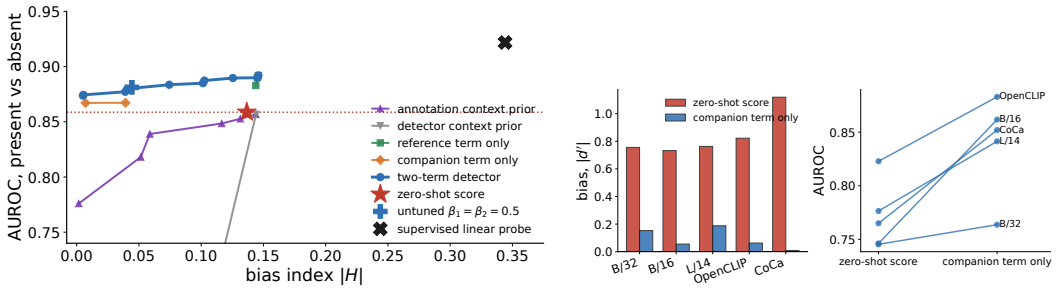
The adjustment of Section 5.2 is $\tilde{s}_O(x) = s_O(x) - \beta_1 \widehat{\pi}_O^{\text{bank}}(x) - \beta_2 \bar{s}_{\neg O}(x)$ with $\widehat{\pi}_O^{\text{bank}}(x) = \sum_{a \neq O} \omega_a \langle f(x), t_a \rangle$, weights ω_a proportional to the positive part of $\widehat{\text{PMI}}^{\text{COCO}}(a, O)$ and normalized over the companions of O, and $\bar{s}_{\neg O}(x)$ the mean score of the same bank with O removed. Because both terms are inner products with fixed text directions, the adjustment is $\langle f(x), t_O - \beta_1 c_O - \beta_2 r_O \rangle$ with $c_O = \sum_{a \neq O} \omega_a t_a$ and r_O the unweighted mean of the same directions, so it costs one extra matrix product against the bank at inference and nothing else. Images are split in half within every cell, all scalars and thresholds are chosen on the development half, and every number in Table 19 is measured on the other half with a paired 77-concept bootstrap. Nothing here is trained, and the probe rows are linear probes on frozen embeddings.

Two decision rules are reported. The target-bias rule takes each method to the highest area under the curve among settings whose development mean bias index is at most 0.05, and the no-cost rule to the smallest development bias among settings that do not lower the area under the curve. A method that cannot reach the target without losing discriminability is reported as infeasible rather than pushed to a degenerate end of its grid, which is what happens to the two subtraction baselines, where at the weight that removes their bias, the annotation prior loses 5 to 10 points and the detector prior 51 to 86, because a detector confidence in the unit interval must be scaled far above the cosine it is subtracted from. Figure 14 draws the resulting frontier for the anchor and the satellite transfer.

Table 19: The two-term detector on held-out images. The untuned columns use $\beta_1 = \beta_2 = 0.5$ with no fitting, the fitted columns use the target-bias rule, and the last pair applies the COCO scalars unchanged to Visual Genome. H is the bias index of Equation 2, and AUROC abbreviates the area under the receiver operating characteristic curve for present against high-prior absent images.

	Zero-shot score		Untuned $\beta_1 = \beta_2 = 0.5$		Fitted			Visual Genome	
Encoder	AUROC	H	AUROC	Change [95%]	H	AUROC	H	AUROC	H
CLIP ViT-B/32	0.845	+0.188	0.864	+0.018 [+0.011, +0.026]	+0.032	0.864	+0.032	0.859	+0.055
CLIP ViT-B/16	0.854	+0.170	0.871	+0.017 [+0.010, +0.025]	+0.032	0.874	+0.067	0.866	+0.064
CLIP ViT-L/14	0.859	+0.137	0.881	+0.023 [+0.014, +0.032]	+0.044	0.881	+0.044	0.877	+0.075
OpenCLIP ViT-L/14	0.869	+0.172	0.886	+0.018 [+0.011, +0.025]	+0.044	0.886	+0.044	0.883	+0.075
CoCa ViT-L/14	0.861	+0.194	0.880	+0.019 [+0.011, +0.027]	+0.049	0.880	+0.049	0.878	+0.080

The prompt banks for the satellite instantiation are written out rather than derived, because no association table exists for that domain. The companion bank names damage, and the reference bank names overhead scene types. A ten-prompt companion bank that also named urban layout worked less well than the five damage prompts in four of the five encoders, which gives the design rule that a companion bank must exclude cues the target itself produces. When companions are other anno-



(a) Attainable pairs on COCO, anchor encoder.

(b) Satellite tiles, bias and discriminability.

Figure 14: The detector on both axes. Panel (a) shows the undominated settings of each method for CLIP ViT-L/14, with the area under the curve for present against high-prior absent images against the absolute bias index. The dotted line is the raw score, and the supervised probe sits high and far to the right. Panel (b) shows, per encoder, the destroyed-against-empty separation and the intact-against-destroyed area under the curve before and after the companion term.

tated categories, as on COCO, that exclusion is automatic. Table 20 reports the satellite numbers, including the supervised probes.

Table 20: Post-disaster tiles. Bias is d' between destroyed and empty tiles, both of which lack a building, so zero is the target. False-alarm rates are on destroyed tiles at a development-fitted threshold giving 0.90 recall on intact tiles. AUROC abbreviates the area under the receiver operating characteristic curve, here for intact against destroyed tiles. The last three columns are linear probes trained on the same development half, first against all absent tiles and then against destroyed tiles only.

	Zero-shot score		With companion term		False alarms		Probe		Probe, hard
Encoder	AUROC	Bias d'	AUROC	Bias d'	Before	After	AUROC	Bias d'	Bias d'
CLIP ViT-B/32	0.745	+0.756	0.764	+0.153	0.610	0.675	0.925	+0.634	−0.281
CLIP ViT-B/16	0.746	+0.732	0.862	+0.055	0.730	0.500	0.920	+0.903	−0.409
CLIP ViT-L/14	0.776	+0.763	0.841	+0.188	0.725	0.565	0.961	+0.759	−0.507
OpenCLIP ViT-L/14	0.823	+0.822	0.883	+0.063	0.610	0.400	0.958	+0.822	−0.731
CoCa ViT-L/14	0.765	+1.119	0.852	+0.008	0.750	0.555	0.953	+1.008	−0.612

Table 21: One threshold shared by all concepts, fitted on development images for 0.90 present recall and applied to held-out images, averaged over the five encoders. The strata gap is the difference between the two absent false-alarm rates, which is what a single threshold cannot absorb. The change against the raw score carries a paired 77-concept bootstrap.

Method	False alarms, high prior	False alarms, low prior	Strata gap	Change in the gap	Pooled rate
Zero-shot score	0.369	0.264	+0.105	—	0.317
Two terms, untuned	0.356	0.302	+0.054	−0.051, 5 of 5 exclude zero	0.329
Supervised probe	0.205	0.011	+0.194	+0.089, 0 of 5 improve	0.108

Table 21 separates the two axes on which a detector can be judged. The supervised probe reaches a pooled error rate of 0.106 against 0.307 for the raw score, so it is the better detector by total error and we say so throughout, but it reaches that rate by treating context as evidence, and under a single threshold it false-alarms on 0.202 of high-prior absent images against 0.011 of low-prior ones. Its strata gap is therefore wider than the raw score's in all five encoders while the untuned two-term adjustment halves the raw gap in all five with every bootstrap interval excluding zero. A threshold fitted where context is sparse keeps its meaning where context is rich under the adjustment and does not under the probe, which is the property a deployed presence test needs and the one the method owns.

What the probe's advantage costs in supervision, and where the adjustment is not optimal. The concession that a probe wins wherever labels exist needs the requirement spelled out, because the probe of Table 21 is fitted per concept and a new target therefore needs annotated images of that target. Table 22 measures the ladder on the held-out image half, adding the adjustment with its

scalars fitted leaving the evaluated concept out and the probe denied labels for that concept in two forms. Three things follow, one of them against the method. The scalars transfer across concepts, since fitting them without the evaluated concept gives the same mean gain as fitting them on all of them, positive in all five encoders. Reusing annotation across concepts in the obvious way fails outright, since a probe trained on the other seventy-six concepts' embeddings ranks below chance at 0.437, presence of different targets not being a shared direction. But a logistic fit on the same three score terms, also trained on the other concepts, ranks better than the fixed-scalar rule at 0.882 against 0.876 while carrying about five times its bias at 0.269 against 0.059. Any fit that maximizes ranking re-uses context as evidence, which is what Equation 4 predicts and what the target-bias rule exists to prevent, so the adjustment is not the best available combination of its own terms but the combination that keeps the bias low. A reader who wants ranking alone should fit the terms and accept the bias.

Table 22: The detector against probes denied labels for the evaluated concept, on the held-out image half, averaged over the five encoders. The area under the curve is for present against high-prior absent images and the bias column is the absolute mean of the bias index. The fraction closed is of the gap between the raw score and the per-concept probe that does have the target's labels.

Score	Labels for the target concept	Area under the curve	Change against raw	Fraction of the gap closed	Absolute mean bias
Raw cosine	no	0.858	—	0.00	0.172
Two-term, untuned scalars	no	0.876	$+0.019$ $[+0.011, +0.027]$	0.30	0.040
Two-term, scalars from all concepts	no	0.876	$+0.019$ $[+0.011, +0.027]$	0.30	0.040
Two-term, scalars leaving one concept out	no	0.876	$+0.019$ $[+0.011, +0.027]$	0.30	0.059
Logistic on the three terms, other concepts	no	0.882	$+0.025$ $[+0.017, +0.032]$	0.39	0.269
Logistic on embeddings, other concepts	no	0.437	-0.421 $[-0.478, -0.364]$	-6.71	0.650
Logistic on embeddings, this concept	yes	0.921	$+0.063$ $[+0.049, +0.078]$	1.00	0.349

Under one threshold shared by all concepts the adjustment moves the operating point rather than uniformly improving it, and this is the weakest part of the result. On COCO at matched development recall the false-alarm gap between the high-prior and low-prior absent strata narrows in all five encoders, on average from $+0.105$ to $+0.054$ and for the anchor from $+0.082$ to $+0.043$, but the pooled rate over both strata rises in four of them by up to 0.020 and only the anchor is flat at 0.321 against 0.318. The reason is visible in the same numbers, since the raw score was accidentally right on unexpected absences for the same reason it was wrong on expected ones, so removing the context term costs on the stratum where the context was already low. On the satellite tiles the pooled rate improves in three of the five encoders and the false-alarm rate on destroyed tiles rises for CLIP ViT-B/32 from 0.610 to 0.675. We therefore claim equalization across strata and a better ranking, not a uniformly lower error rate, and a deployment that cares about the pooled rate under one threshold should fit the scalars against that objective instead.

The generative form fails, and the measurement says why. A decoder exposes no query direction, only the decision variable $z_O(x)$ and a threshold, so the transplantable version of the method lets the threshold follow the context, answering present when $z_O(x) > \theta_O + \beta\widehat{\pi}_O^{\mathrm{bank}}(x)$ with θ_O the concept's development-fitted threshold, β a single slope and $\widehat{\pi}_O^{\mathrm{bank}}$ the same companion term computed by the anchor encoder. Across all seven checkpoints this changes nothing, since the development-chosen β is zero in four of them and 0.1 in the rest, and every false-affirmation change is below 0.004 in the global form and in a per-concept form that centers the context estimate inside each concept. One of those changes has a bootstrap interval excluding zero, Qwen3-VL-4B in the per-concept form at $+0.0033$ $[+0.0012, +0.0056]$, and it raises the rate rather than lowering it. Two measurements explain the failure rather than leaving it as a null. Within a concept, on absent images, the correlation between the decoder's decision variable and the encoder's context estimate runs from $+0.116$ to $+0.221$, positive in 79 to 96 percent of concepts but worth a few percent of the variance, while between concepts the same correlation runs the other way in every checkpoint, from -0.132 to -0.258. A single offset therefore trades the within-concept signal against the between-concept one, and the rewriting of the query direction works in the contrastive case because it acts inside the model whose score it corrects. This is a different result from the conditional adjustment of Appendix K, which uses a detector-based estimate on a decoder's own answer logits and buys a small accuracy gain on the adversarial split. What fails here is transplanting one model's context estimate into another model's threshold.

The decoder's own context estimate removes that objection and still buys nothing. The natural repair is to stop borrowing and ask the decoder itself, replacing the encoder's companion term with $\widehat{\pi}_O^{\text{self}}(x) = \sum_{a \neq O} \omega_a z_a(x)$, the same association-weighted sum evaluated on the model's own answers about the companions of the target, which costs k extra forward passes and no annotation. As a measurement it does what it should, since the within-concept correlation with the decision variable rises from about 0.16 to $+0.564$ for InstructBLIP and $+0.606$ for Qwen3-VL, positive in all 77 concepts, and the between-concept correlation that ran negative for the transplanted estimate turns to $+0.220$ for InstructBLIP. The development-chosen weight is nevertheless zero for InstructBLIP and without measurable effect for Qwen3-VL, and Equation 4 accounts for that exactly, because the decoder's own estimate averages $+0.544$ standardized units on present images against -0.544 on absent ones, a class gap of $\delta_O = 1.087$ that Qwen3-VL reproduces at 1.144, so every unit of weight removes bias and surrenders that much class separation, and at matched recall the two cancel. The identity that defeats the mitigation is the one that fixes the exchange rate, since context predicts presence whenever co-occurrence is bidirectional.

The reference term does not carry over, which bounds criterion-side repair. In the contrastive detector the discriminability gain came almost entirely from the reference term, and the reason it is safe to subtract is that a cosine score carries an image-level offset which the prompt-bank mean captures without carrying target evidence. A decoder's variable is a difference of two answer logits read at one position of one forward pass, so any influence that raises both cancels before a threshold is applied, and the analogous quantity should therefore have little left to remove, which is what the measurement shows. The within-concept correlation between the decision variable and the model's response to non-target concepts on the same absent images gives $+0.630$ to $+0.722$ across the five encoders, worth 41 to 53 percent of the variance, against $+0.203$ and 6.5 percent for InstructBLIP, whose own reference term has a class gap of -0.320 and whose best available weight moves the area under the curve by $+0.0004$. The second family behaves the same way, since Qwen3-VL's reference term has a class gap of -0.379 and leaves its area under the curve at 0.9868 unchanged, with a false-affirmation change of -0.0016 $[-0.0034, +0.0003]$ whose interval contains zero. Taken together the three attempts bound what a threshold can do. Any context estimate with enough purchase on the decision variable to move it is presence-informative, and Equation 4 then fixes the exchange rate, while the one channel that is not presence-informative has no offset left to remove. The bound concerns criterion-side repair rather than the decoders, which are already the stronger rankers on this task, since InstructBLIP separates present from high-prior absent images at 0.959 and Qwen3-VL at 0.987, where the best adjusted encoder score reaches 0.888 and a supervised probe on the same embeddings reaches 0.929. On this particular comparison the ranking is already strong, so the residual context-dependent error is primarily one of operating point, with false affirmation of 0.108 on high-prior absent images against 0.030 on low-prior ones, while the sensitivity change of Section 5.1 lives in the contrast between prior strata rather than here. Repair for a decoder has to change the decision variable, through the input or the decoding, rather than where it is cut.

R A DECODING-SIDE MITIGATION

The bound of Appendix Q concerns rules that move a threshold, and it says nothing about changing the decision variable, which is what decoding-side mitigations do. We therefore run visual contrastive decoding (Leng et al., 2024) through the same cells, forming $z_O^{\text{vcd}}(x) = (1+\alpha) z_O(x) - \alpha z_O(x_t)$ with $\alpha \geq 0$ the contrast strength and x_t the image corrupted by the diffusion forward process at the step that method prescribes for it. The rule is defined on log probabilities, and for this decision variable the two forms are the same quantity, because the log-sum-exp normalizer is shared by the yes and no entries of one forward pass and cancels in their difference, so nothing is approximated. Its plausibility constraint keeps vocabulary entries within a factor of the most probable one, which on a two-alternative presence question binds only when one answer is already implausible and the contrast cannot change the decision, so the unconstrained variable is used. Both passes read the same images in the same order, and one 77-concept bootstrap draw is shared across contrast strengths. A second family of decoding-side mitigations penalizes over-trust over a generated sequence and reallocates attention retrospectively (Huang et al., 2024). This design cannot evaluate it, because that penalty is defined over the tokens of a continuation while the decision variable here is read at a single answer position, so importing it would require a different instrument rather than a different setting of the same one, and we state the boundary rather than approximate the method.

Table 23: Visual contrastive decoding on InstructBLIP over the 77-concept design. The criterion column is the standardized high minus low prior stratum contrast and the raw criterion column is the same contrast before per-concept standardization, dispersion is the pooled within-concept standard deviation, raw separation is the present minus absent difference in answer logits, and the pooled false-alarm rate is measured at a hit rate of 0.90 on present images under one threshold shared by all concepts. A contrast strength of zero is the unmodified decision variable.

α	Criterion	d'	Raw criterion	Raw separation	Dispersion	Offset over evidence	Pooled false alarm
0.0	+0.267	3.896	0.421	5.524	1.551	0.0763	0.068
0.5	+0.181	3.600	0.420	7.655	2.294	0.0549	0.079
1.0	+0.134	3.311	0.419	9.787	3.127	0.0429	0.095
2.0	+0.088	2.993	0.418	14.050	4.874	0.0297	0.122

The standardized criterion falls because dispersion grows. Table 23 shows the standardized criterion contrast dropping from $+0.267$ to $+0.088$ monotonically in the contrast strength, and read alone that column says the mitigation removes the prior-driven shift. The raw columns say otherwise. The same contrast measured before per-concept standardization is flat, moving only from 0.421 to 0.418, while pooled dispersion grows from 1.551 to 4.874, so the standardized quantity falls because its denominator grows and not because the offset was removed. Any evaluation that reports a standardized bias index without the dispersion that scales it can be moved by adding noise to the decision variable.

What the contrast does change, and what it costs. Raw separation between present and absent images grows from 5.524 to 14.050 while the prior-driven offset does not, so the offset measured against target evidence falls from 0.0763 to 0.0297. Suppressing the prior relative to the evidence is a real effect and is plausibly the intended one. It is bought with independent noise, since the corrupted copy's response correlates with the original at only $+0.154$ within a cell and carries a class gap of $+0.763$ standardized units, so dispersion grows by more than evidence does. Sensitivity pooled over strata therefore falls from 3.896 to 2.993, the pooled false-alarm rate at a hit rate of 0.90 rises from 0.068 to 0.122 under a shared threshold and from 0.087 to 0.173 under one threshold per concept, and the gap between the two absent strata narrows only from $+0.078$ to $+0.070$. The decomposition is what separates these outcomes, since accuracy at a tuned threshold would show none of them.

The corruption strength trades the two requirements against each other. For the contrast to lower the offset relative to evidence, the response it subtracts has to track the context. For it to cost nothing, that response must not predict presence. At the prescribed step the corrupted copy satisfies the first at the expense of the second. Repeating the run at the end of the diffusion schedule, where the retained signal fraction is 0.00004, inverts that trade. The subtracted response then has a class gap of -0.001, tracks the prior stratum by -0.007 and correlates with the original at -0.004, which is the class-balanced ideal Equation 4 asks for. It accomplishes nothing, since every raw quantity scales by exactly one plus the contrast strength, the offset measured against evidence stays at 0.0763 at all settings, sensitivity is preserved at 3.862 against 3.896, and the criterion contrast moves by -0.001 $[-0.003, +0.001]$. The contrast has become a rescaling of the decision variable, and a rescaling cannot change a decision. No setting on this axis satisfies both requirements, because in an image the context and the target evidence are carried by the same signal, so a corruption that preserves one preserves the other. This is the non-separability that prevents the adjustment of Appendix K from preserving sensitivity and that gives a decoder's own companion estimate its class gap, now visible along the design axis of a published mitigation.

What the contrast can do is bounded by how much the model affirms without evidence. The mitigation removes the affirmation a model produces from a corrupted image, so a model that already refuses such an image has little for it to remove, and Table 24 shows that this sets the outcome across three families. InstructBLIP affirms on 0.955 of the corrupted copies at a mean decision variable of $+2.29$ and its offset measured against evidence falls by 61 percent, whereas both Qwen checkpoints refuse them, at 0.044 and 0.178, and theirs fall by 9 and 7 percent, with criterion contrasts falling by 13 percent against 67. The cost is paid in every case, since sensitivity falls and the pooled false-alarm rate at a matched hit rate rises in all three. The predictor is available before any contrast is computed, because the image-free levels of Appendix V already say which checkpoints affirm

Table 24: Visual contrastive decoding across three decoder families, at a contrast strength of zero against two. The affirmation rate and mean decision variable are measured on the corrupted copies, which is the response the contrast removes. The offset over evidence is the prior-driven criterion offset divided by the present minus absent separation, both before standardization. Sensitivity is pooled over strata and the pooled false-alarm rate is measured at a hit rate of 0.90 under one shared threshold.

	Corrupted copy		Offset over evidence		Criterion		d'		Pooled false alarm	
Model	Affirms	Mean z	$\alpha = 0$	$\alpha = 2$	$\alpha = 0$	$\alpha = 2$	$\alpha = 0$	$\alpha = 2$	$\alpha = 0$	$\alpha = 2$
InstructBLIP	0.955	$+2.29$	0.0763	0.0297	$+0.267$	$+0.088$	3.896	2.993	0.068	0.122
Qwen2-VL	0.044	-1.24	0.0430	0.0390	$+0.146$	$+0.127$	3.918	3.743	0.055	0.060
Qwen3-VL-8B	0.178	-2.07	0.0943	0.0876	$+0.431$	$+0.373$	5.198	4.693	0.015	0.019

without evidence, and a model that does not is one for which this family of mitigation can only rescale its decision variable while adding noise to it.

S NEGATED QUERIES

The absence-insensitivity literature measures its symptom with negated text, which this paper does not use to score anything, so the connection drawn in Section 4 needs its own measurement. The symptom is reported on both sides of the model families we study, in retrieval and multiple choice for the contrastive encoders (Alhamoud et al., 2025) and in question answering for the generative decoders (Zhang et al., 2025), which is why the instrument rather than either family is what this appendix isolates. We built a six-template negated ensemble, ensembled and normalized exactly like the affirmative one, and applied it to the cached image embeddings of all five checkpoints over the same 77 concepts. The ensemble is confined to this appendix and its own output files, and no quantity derived from it enters any index, regression, adjustment, or table reported elsewhere.

Table 25 gives the result. The negated query ranks absent images above present ones at an area under the curve of 0.109 to 0.145 against a chance value of 0.500, whereas the affirmative query ranks present images above absent ones at 0.875 to 0.895, so a query that treated the negation as an operator would reach a comparable value in its own direction rather than the reverse of it. The negated query exceeds chance on at most one of the 77 concepts in any checkpoint. The two scores agree at 0.925 to 0.947, so the negation changes the ranking hardly at all. The same axis then governs the negated score, because the index of Equation 2 recomputed on it stays positive in 87 to 99 percent of concepts and correlates with the affirmative index at 0.816 to 0.902 and with the off-axis probe-error index at 0.447 to 0.610, every correlation significant below 5×10^{-5}. At the instance level the context prior raises negated absent-image scores by 0.29 to 0.54 score standard deviations after adjustment for detector response, target area and log frequency, with concept-clustered significance below 10^{-21} in all five checkpoints and a paired concept-level contrast whose bootstrap interval excludes zero in all five.

Two scoping remarks belong with this. The negated and affirmative scores correlate at about 0.93, so the correlation between the two indices is a consistency check rather than an independent replication, and the informative quantities are the level of the negated index, the correlation with the orthogonally constructed probe-error index, and the instance-level prior coefficient. The reversal is also not a property of one way of writing negation. Scoring five syntactic constructions separately, an existential form, a prepositional form built on without, a verbal form built on does not contain, a participial form built on lacking, and a nominal form asserting that the object is absent, each on its own three-template ensemble, leaves the negated query below chance in all twenty-five family-by-encoder pairs, at mean areas under the curve of 0.093 to 0.163 against the affirmative query's 0.875 to 0.895. At most one of the 77 concepts exceeds chance in any pair, agreement with the affirmative score stays between 0.86 and 0.97, and the recomputed index remains positive in 71 to 100 percent of concepts, so every construction inherits the same context prior. What this establishes is that these encoders do not act on any of the forms we can write rather than that no formulation could reach them.

Table 25: Negated queries over 77 concepts. N is the area under the curve with which the negated score ranks absent images above present ones, the affirmative column is the same quantity for the affirmative query in its own direction, and agreement is the correlation between the two scores. H_{neg} applies Equation 2 to the negated score, and the last column is the prior coefficient on negated absent-image scores in score standard deviations with concept-clustered inference.

Encoder	N	Affirmative	Agreement	H_{neg}	Fraction > 0	ρ with H	ρ with $B_{\perp}$	Prior coefficient
CLIP ViT-B/32	0.134	0.875	0.930	$+0.235$	0.99	$+0.851$	$+0.568$	$+0.452$
CLIP ViT-B/16	0.117	0.880	0.947	$+0.211$	0.96	$+0.902$	$+0.504$	$+0.458$
CLIP ViT-L/14	0.145	0.879	0.925	$+0.163$	0.87	$+0.816$	$+0.447$	$+0.290$
OpenCLIP ViT-L/14	0.109	0.895	0.928	$+0.204$	0.97	$+0.886$	$+0.587$	$+0.497$
CoCa ViT-L/14	0.113	0.889	0.934	$+0.237$	0.99	$+0.895$	$+0.610$	$+0.541$

T A NEGATION-TRAINED ENCODER IN THE SAME DECOMPOSITION

The remedies proposed in the absence literature presuppose the degradation component, so the direct test of that reading is to put a checkpoint from that family through this decomposition and ask which component it moves. CoN-CLIP (Singh et al., 2024) is the case that admits the cleanest comparison. It fine-tunes OpenAI CLIP ViT-B/32 on captions carrying negation, which is one of our own checkpoints, and it freezes the image tower, which we verify by comparing every visual parameter with its base rather than taking it on trust. Both arms therefore share one set of image embeddings and differ only in the text tower, so nothing here is a cross-family comparison. Both are also built with the activation the OpenAI weights were trained with, so the base entries below are the recomputation of Appendix B and differ slightly from the corresponding row of Table 25.

The context effect does not move. Between the low and high expectedness strata the base model changes sensitivity by -0.007 and the equal-error criterion by $+0.347$, while CoN-CLIP changes them by $+0.020$ and $+0.357$. A paired bootstrap over the 77 concepts places the differences at $+0.027$ $[-0.012, +0.066]$ for sensitivity, $+0.010$ $[-0.015, +0.037]$ for the criterion, $+0.013$ $[-0.024, +0.051]$ for the present position and $+0.008$ $[-0.013, +0.029]$ for the absent position, so all four intervals contain zero. Training on negated captions leaves the contextual criterion movement of Section 5 where it was.

The instrument does move, and the direction is informative. On the negated ensemble of Appendix S, scored in the same isolated namespace, the base model ranks absent images above present ones at 0.127 while its negated and affirmative scores still correlate at 0.914. CoN-CLIP reaches 0.452 with the two scores correlating at only 0.201, and the number of concepts above chance rises from one to sixteen of 77. Negation training therefore does what the diagnosis in Section 4 says is missing, in that it decouples the negated query from its affirmative counterpart and stops the query from being a paraphrase of the affirmative one. What it does not do is produce a query that ranks absence, since 0.452 is still at or below chance, and the recomputed index stays positive on 82 percent of concepts against 99 percent for the base, so contextual expectedness continues to inflate absent-image scores under the negated query. One checkpoint from one remedy family does not settle what the family can achieve, and CoN-CLIP addresses negation rather than the hard negatives of the other branch, so we report this as a single direct measurement rather than as a general limit. Later work in the same branch generates negated pairs during training rather than fine-tuning on a fixed negated set (Cai et al., 2025), and whether that changes the criterion component is the measurement this appendix would need repeating on.

U FREE-FORM DESCRIPTIONS

Every generative measurement in this paper reads one token, the difference between the answer logits for yes and no on a positive presence question. That leaves open whether the affirmative shift is a property of the probe rather than of what a model says about a scene. We therefore repeat the contrast with an instrument that never names the target, never asks about it and never mentions absence, which also leaves the study's positive-only convention intact.

Seven decoders describe each of 2,500 images from the cell pool in one sentence under greedy decoding, and each caption is parsed for the 80 categories with the surface forms of the lexical matcher used elsewhere. A mentioned category that the annotation does not record is a free-form

invention. Categories whose names carry a common non-object sense, among them orange and remote, are excluded from the numbers below. Because a one-sentence caption does not mention every annotated object, the rate at which captions name present objects runs from 0.455 to 0.558, so these rates are not detection rates and only contrasts computed within an image or within a concept are interpretable.

Table 26: One-sentence descriptions of 2,500 images per decoder. The first three columns give the mean contextual expectedness of the absent categories a caption invents, the mean for the absent categories it does not name, and their paired difference over images carrying at least one invention, with a 95 percent bootstrap interval. The next column is the fraction of those images for which the difference is positive. The two mention-rate columns are the per-concept rate at which absent categories are named in the low and the high expectedness stratum. The coefficient is the logistic slope of mention on standardized expectedness with caption length and the number of named categories held fixed, using concept cluster-robust standard errors. The last column correlates the per-concept free-form difference with the single-token criterion shift of Table 16, and the corresponding correlations with the absent-class position and with the sensitivity change are given in the text.

Model	Invented	Not named	Paired difference	Images positive	Mention rate, low	Mention rate, high	Coefficient	Correlation with the criterion shift
Qwen2-VL-7B	0.167	0.032	+0.135 [+0.119, +0.151]	0.800	0.0006	0.0040	+0.477	+0.074 [−0.178, +0.318]
InstructBLIP-7B	0.209	0.032	+0.177 [+0.163, +0.191]	0.867	0.0008	0.0084	+0.544	+0.350 [+0.118, +0.545]
Qwen3-VL-4B	0.146	0.031	+0.115 [+0.100, +0.129]	0.732	0.0010	0.0043	+0.449	+0.223 [−0.013, +0.444]
Qwen3-VL-8B	0.152	0.031	+0.120 [+0.105, +0.136]	0.754	0.0008	0.0042	+0.452	+0.213 [+0.011, +0.430]
InternVL3.5-8B	0.163	0.031	+0.133 [+0.117, +0.148]	0.792	0.0006	0.0044	+0.478	+0.236 [+0.009, +0.430]
Gemma-3-12B	0.142	0.032	+0.110 [+0.096, +0.125]	0.736	0.0010	0.0048	+0.443	+0.312 [+0.069, +0.522]
Llama-3.2-11B-V	0.160	0.031	+0.129 [+0.115, +0.145]	0.777	0.0008	0.0043	+0.475	+0.183 [−0.043, +0.434]

The primary test is within an image, and Table 26 reports it. Among the categories absent from an image, the ones a caption invents carry between four and a half and seven times the contextual expectedness of the ones it leaves out, in all seven decoders, with every paired interval excluding zero and with the difference positive for between 0.732 and 0.867 of the images that carry an invention. The paper's own stratum contrast appears as well, since the per-concept mention rate on absent images rises from the low to the high expectedness stratum in all seven. Of 71 concepts that difference is positive for between 0.465 and 0.845, where the remainder is mostly ties at zero rather than reversals, since among concepts with any mention in either stratum it is positive for between 0.809 and 0.952.

The control this design needs concerns verbosity, because expected-context scenes are busier and a longer caption mentions more of everything. InstructBLIP's captions are 31.3 against 32.0 words and name 1.83 against 2.15 categories between the two strata, while in the other six decoders those differences stay under two words and under 0.3 categories. The logistic coefficient nevertheless rises under the controls in every decoder, from a range of $+0.409$ to $+0.505$ unadjusted to a range of $+0.443$ to $+0.544$ with caption length and the number of named categories held fixed, so the effect is which absent categories are named and not how many.

Finally the two instruments agree, and they agree on the component the paper identifies. The per-concept free-form difference correlates positively with the single-token criterion shift in all seven decoders and significantly in four of them, and with the absent-class position at between $+0.044$ and $+0.386$, significantly in five. Its correlation with the sensitivity change contains zero in six of the seven, the exception being InternVL3.5 at -0.272 $[-0.485, -0.025]$, which runs in the direction opposite to a discriminability cost that tracks the invention effect. The decomposition of Section 5.1 therefore reproduces across instruments on its criterion component and not on its sensitivity component. Qwen2-VL is the weakest case, its criterion and absent-position correlations both containing zero, which is the checkpoint that Appendices O and N already isolate under one of the two confusability proxies.

For readers who want these captions in the units of the caption literature, the same parse gives a per-mention invention rate between 0.075 and 0.124 and a per-caption rate between 0.114 and 0.196, highest for InstructBLIP and lowest for Qwen2-VL in both, with the per-model values in the released table. Those values are not comparable with published figures for the same quantities, because the prompt asks for one sentence rather than a paragraph, the images are this study's cell pool rather than a standard split, and presence comes from detection annotations, so they place the instrument in familiar units rather than ranking these checkpoints against the literature.

V IMAGE-FREE BASELINES

Section 5.1 measures how contextual expectedness moves a generative decision variable, which leaves open how much affirmation the question alone produces. We therefore repeat the same positive-only question over the same 77 concepts with four inputs that carry no target evidence, a uniform gray canvas, a uniform black canvas, uniform random noise, and, where the processor accepts text without pixels, no image at all. Each condition yields one decision value per concept, and Table 27 places the resulting affirmation rates beside the natural-image cells for the seven checkpoints of Table 17. Decision values are not comparable across models because the checkpoints differ in logit scale, so every comparison below is within a model.

Absolute affirmation levels are partly a property of the question. InstructBLIP affirms on 0.896 of concepts with a gray canvas and on 0.610 under noise, so its high false-affirmation rate on absent images rests on an affirmative default that no image supports, whereas Llama-3.2-Vision refuses every image-free input and both Qwen3-VL checkpoints refuse every one that carries pixels. These defaults are input dependent rather than one constant per model. Qwen2-VL affirms on 0.558 of concepts only when the canvas is black, and Qwen3-VL-8B refuses on all three canvases yet affirms on 0.844 when the question arrives with no image at all, its 4B counterpart on 0.247 in the same condition. Gemma-3 affirms on 0.584 with no image against 0.221 with a black canvas and 0.013 under noise, and InternVL3.5 affirms on 0.857 with a black canvas against 0.338 with no image and 0.052 under noise. One yes-bias number therefore does not characterize a model, let alone the model class, and when an input carrying pixels draws an affirmation at all it is most often the black canvas.

The prior effect is not that default. In every checkpoint a natural image without the target drives the decision value below the image-free level, and the high-minus-low context increment sits on top of that reduced level, so contextual affirmation is measured against a baseline the image has already lowered. Because an image-free condition supplies one constant per concept while the prior contrast is defined within a concept, conditioning on it cannot change that contrast, and the estimated prior coefficient is unchanged to three decimals in all seven. What the baseline can do is separate a marginal channel from a contextual one, and there the models differ. In Qwen2-VL the uniform-gray value tracks corpus frequency at a rank correlation of 0.428 while its correlation with the per-concept criterion shift is 0.006, and in InstructBLIP, Gemma-3 and InternVL3.5 the criterion correlation is not resolved from zero. Three checkpoints are the exception, since Qwen3-VL-8B correlates at 0.496 with frequency and at 0.355 with the criterion shift, its 4B counterpart at 0.392 and 0.357, and Llama-3.2-Vision at 0.326 and 0.329, so there concepts with a stronger marginal default also receive a larger contextual push. Because the pattern appears at both Qwen3-VL parameter counts it is a family property in that case rather than a property of one checkpoint, but it is not confined to one family, and the separation of channels that Appendix K exploits is clean in the remaining four. The count depends on which image-free input measures the default, since Gemma-3 entangles the two channels when the question arrives with no image at all, at 0.618 with frequency and 0.394 with the criterion shift, so the separation is a property of a checkpoint paired with an input rather than of a checkpoint alone.

Table 27: Affirmation rates on image-free inputs beside the natural-image cells, over the same 77 concepts and the same positive-only question. For the image-free columns the rate is the fraction of concepts whose decision value is affirmative, and for the natural-image columns it is the mean per-concept rate over images. Decision values themselves are not comparable across models and are reported in the released tables. InstructBLIP has no text-only entry because its processor requires pixel inputs.

	Image-free inputs				Natural images		
Model	No image	Gray	Black	Noise	Absent, low	Absent, high	Present
Qwen2-VL-7B	0.000	0.000	0.558	0.000	0.044	0.096	0.911
InstructBLIP-7B	—	0.896	0.935	0.610	0.324	0.569	0.988
Qwen3-VL-4B	0.247	0.000	0.000	0.000	0.004	0.030	0.917
Qwen3-VL-8B	0.844	0.000	0.000	0.000	0.013	0.064	0.944
InternVL3.5-8B	0.338	0.416	0.857	0.052	0.017	0.108	0.949
Gemma-3-12B	0.584	0.052	0.221	0.013	0.066	0.142	0.909
Llama-3.2-11B-V	0.000	0.000	0.000	0.052	0.030	0.104	0.938

W VISUAL GENOME REPLICATION

We replicated the core pipeline on 108,077 Visual Genome (Krishna et al., 2017) images with a fixed name mapping under which 74 of the 77 concepts yield full cells. The descriptive score movement replicates. The two indices of Equation 2 correlate positively in all five checkpoints, between 0.43 and 0.53 with every significance level below 10^{-3}, the absent-distribution position rises with the context prior in all five checkpoints, and the absent-image score tracks $\widehat{\pi}_O^{\mathrm{VG}}$ with a positive slope at significance below 10^{-32} in all five checkpoints. The cross-corpus orderings are strong, with per-concept hallucination on COCO and Visual Genome agreeing at a Spearman correlation of 0.723 and $\widehat{\mathrm{PMI}}^{\mathrm{COCO}}$ agreeing with $\widehat{\mathrm{PMI}}^{\mathrm{VG}}$ at 0.860, both shown in Figure 15. Their strong agreement, contrasted with the 0.115 agreement between $\widehat{\mathrm{PMI}}^{\mathrm{COCO}}$ and $\widehat{\mathrm{PMI}}^{\mathrm{CC}}$, makes stable visual scene-type or annotation structure a viable explanation alongside pretraining caption statistics.

Sensitivity does not reproduce a null. The prior raises it by $+0.376$ [$+0.326$, $+0.427$], and the class decomposition locates that difference in the present class rather than the absent one. On COCO the absent distribution moves further than the present distribution at high prior in four of the five encoders, which is why sensitivity is flat or falls, whereas on Visual Genome the present distribution moves further in all five, which is why it rises, and the absent-class movement itself is the same on both corpora at $+0.26$ to $+0.37$ against $+0.27$ to $+0.40$. That pattern is what salience-biased annotation predicts, since extra evidence on high-expectedness present images raises the present class alone. Two limits keep it short of an identification, because only one encoder resolves the reversal on its own and the change in mean target area is comparable across the corpora, so area is not the differentiating variable and we do not measure the centrality or occlusion that would be. We therefore claim replication of the absent-class movement, which is the paper's own claim, and attribute the sensitivity difference to the present class without isolating its cause.

W.1 STRATIFYING COCO WITH A VISUAL GENOME PRIOR TABLE

Appendix G replaces the estimator that predicts the score. The complementary objection concerns the partition instead. The expectedness statistic E_O that forms the cells and the estimate $\widehat{\pi}_O^{\mathrm{COCO}}$ that predicts the score are two summaries of one COCO table, so their association could be a property of that partition rather than of the model. We therefore re-form the strata from a conditional table estimated on Visual Genome and applied to COCO scene composition, holding the images, the presence labels and the scores fixed, which requires no new images and no new forward passes.

The first thing this control reports is how much room it has. Per concept the two expectedness statistics agree on the absent pool at a median Spearman correlation of 0.935, from 0.700 to 0.985, and re-forming the sides moves 3.4 percent of absent images across on average and at most 27.4 percent for a single concept. Two independently estimated image co-occurrence tables therefore almost agree about which absences are expected, which bounds how far this design can separate them. That bound is itself part of the answer, since a partition an independently estimated table reproduces is unlikely to be a COCO-specific artifact, although both corpora annotate natural scenes under similar conventions and a broader annotation-structure account survives this control.

Within that bound the effect replicates. Re-forming both classes by Visual Genome terciles inside each concept's cached pool raises the absent-class position by $+0.345$ to $+0.495$ standardized units in all five encoders with every interval excluding zero, the present-class position by $+0.297$ to $+0.382$, and the equal-error criterion by $+0.348$ to $+0.438$, while the sensitivity change runs from -0.060 to $+0.113$ with no interval excluding zero. That design still draws from a pool the COCO terciles selected. We therefore also split each COCO stratum separately at its own Visual Genome median, a contrast the cell construction did not impose, and the absent-class position rises there by $+0.141$ to $+0.178$ in all five encoders with every interval excluding zero. Both designs use the 77-concept bootstrap with one shared draw across encoders.

X THE POST-DISASTER REPLICATION IN DETAIL

Tiles from all four xBD splits (Gupta et al., 2019; Ebel et al., 2026) are categorized by their damage annotations, with 2,795 post-disaster tiles containing at least five buildings of which at least 90 per-

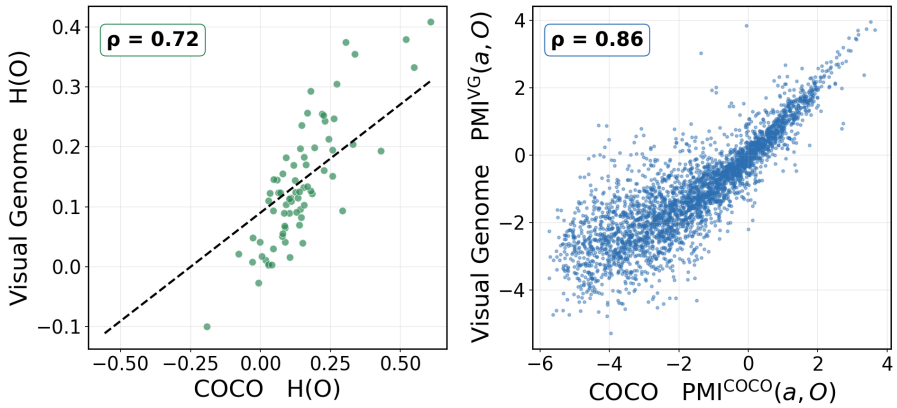


Figure 15: Cross-corpus agreement between COCO and Visual Genome. The per-concept hallucination index agrees at a Spearman correlation of 0.72 and the off-diagonal entries of $\widehat{\mathrm{PMI}}^{\mathrm{COCO}}$ and $\widehat{\mathrm{PMI}}^{\mathrm{VG}}$ agree at 0.86, in contrast with the 0.115 agreement between $\widehat{\mathrm{PMI}}^{\mathrm{COCO}}$ and the caption co-mention table $\widehat{\mathrm{PMI}}^{\mathrm{CC}}$ of Appendix G.

cent are undamaged serving as the present set, 399 tiles with at least 70 percent destroyed buildings serving as expected absence, and 3,858 tiles containing no buildings serving as unexpected absence. The pre-disaster image of every destroyed tile provides a paired present counterpart at identical geography. The score is the mean cosine to a six-prompt affirmative building ensemble with no negation anywhere in the pipeline, the generative decision variable is the difference between the affirmative and negative answer logits of Qwen2-VL in a single forward pass with inputs bounded at 512 pixels, and the encoder loader matches the main study exactly. The anchor encoder separates present tiles from expected absence at 1.02 sensitivity units against 1.74 from unexpected absence. Destroyed urban tiles and empty-terrain tiles are both annotated as lacking the queried building, but they also differ in preserved urban context, destruction cues, terrain appearance, and possible residual structure. Because a tile at the 70 percent threshold can retain standing buildings, we recompute the headline numbers on the strict subset of 86 tiles with at least 95 percent destruction and no undamaged building. On this subset the affirmative rate falls from 0.832 to 0.686 against 0.028 on empty terrain, and the encoder separation of destroyed from empty tiles persists at 0.65 sensitivity units. Residual structures therefore explain part of the raw rate but do not account for the entire association.

The recoverability claim of Section 6 requires the evidence and the decision to be measured in one model, so we probe the deciding checkpoint itself rather than a separately loaded encoder. In the same forward pass that produces the answer logits we retain two representations, the final-layer hidden state at the last position, which is the exact vector the language-model head multiplies to form those logits, and the layer-zero hidden states at the image-token positions averaged over the image, which are the projected visual features as they enter the language model. A logistic probe is fit on each and scored out of fold under folds grouped by disaster event, so no event contributes to both fitting and scoring and a site's two dates never separate. The inverse penalty is chosen by nested cross-validation inside each training fold, the null is a hundred relabeled repeats of the identical pipeline, and the interval is a bootstrap over events rather than over tiles, with the same event draw used for the probe and for the decision when the two are differenced.

Because separability alone does not say what a probe reads, Table 28 reports the comparison in three forms. The published one draws intact tiles without regard to where the destroyed ones come from. The matched one draws them from the same disaster events and the same post-disaster capture, and the strict one restricts the destroyed side to tiles at least ninety-five percent destroyed with no undamaged building left. The permuted null sits at 0.494 to 0.506 throughout and no observed value is reached by any of the hundred repeats.

Two controls decide the reading. Tiles whose buildings stand undamaged in both captures separate by capture date at 0.735 and 0.788, well above chance but far below the paired destroyed contrast at 0.954 and 0.927, and the acquisition direction fitted on those undamaged tiles transfers to the

paired destroyed contrast at 0.888 and 0.852. The paired contrast is therefore substantially a date classifier and we do not rest the claim on it. The damage direction fitted on the matched contrast transfers to the acquisition control at only 0.614 and 0.615, so the two directions are not the same one, and the capture-constant matched and strict contrasts therefore carry the claim, where on both checkpoints the readout probe stands above the decision value with every difference excluding zero. The models' own answers are at 0.481 and 0.541 on the acquisition control, which is the correct response for a question about buildings that stand in both images. The projected visual features carry the distinction at least as well on five of the six comparisons, so the evidence is present when the image enters the language model and still present at the site the answer is taken from. For completeness the anchor contrastive encoder, which does not make the decision, separates intact from destroyed at 0.956 and present from absent tiles at 0.960, which is the cross-model version of the observation and the weaker one. A probe establishes linear availability rather than use, which is why the claim is stated as one about the functional the model applies, a linear head in every case here, and not about the calibration of its answer. That a probe can report the difficulty of its own task rather than the content of a representation is the standing objection to probing, answered there by contrasting a probe against one fitted to a control task (Hewitt & Liang, 2019). The controls above play that role for this design, since the acquisition control holds the probe and the site fixed and changes only what the tiles differ in.

Table 28: Linear separability at the site the decoder's own answer is read from, against the answer itself, on xBD. Every value is an area under the receiver operating characteristic curve, scored out of fold with whole disaster events held out, with intervals bootstrapped over events. Rows are ordered from the loosest comparison to the strictest, and the last row is the control in which buildings are undamaged in both captures.

Comparison	Model	Tiles	Answer	Readout probe	Difference	Projector probe
Intact vs. destroyed, any event	Qwen2-VL-7B	999	0.806	0.951 [0.923, 0.977]	+0.144 [+0.084, +0.219]	0.968
	Qwen3-VL-8B	999	0.872	0.929 [0.887, 0.976]	+0.057 [+0.016, +0.097]	0.961
Same event and capture	Qwen2-VL-7B	714	0.743	0.942 [0.906, 0.970]	+0.199 [+0.093, +0.278]	0.960
	Qwen3-VL-8B	714	0.892	0.946 [0.916, 0.971]	+0.055 [+0.032, +0.070]	0.970
Same event, destroyed above 95%	Qwen2-VL-7B	401	0.808	0.941 [0.911, 0.965]	+0.133 [+0.026, +0.211]	0.885
	Qwen3-VL-8B	401	0.930	0.969 [0.954, 0.990]	+0.039 [+0.014, +0.049]	0.963
Same site before vs. after	Qwen2-VL-7B	798	0.702	0.954 [0.907, 0.983]	+0.253 [+0.193, +0.320]	0.970
	Qwen3-VL-8B	798	0.764	0.927 [0.789, 0.986]	+0.163 [+0.035, +0.233]	0.989
Capture control, buildings undamaged	Qwen2-VL-7B	630	0.481	0.735 [0.608, 0.805]	+0.255 [+0.099, +0.338]	0.762
	Qwen3-VL-8B	630	0.541	0.788 [0.691, 0.835]	+0.247 [+0.160, +0.293]	0.754

The same measurement on five further checkpoints moves in the wrong direction, which is this study's strongest signal about model progress and also its most heavily qualified one, and an image-free control decides which of them may be read that way. Asked the identical question with no evidence in the image, Qwen2-VL, Llama-3.2-Vision, Qwen3-VL, and Gemma-3 affirm a building on none of sixteen noise images and on neither uniform canvas, whereas InternVL3.5 and InstructBLIP affirm on a uniform canvas and InternVL3.5 on 81 percent of the noise images as well. We therefore read the four controlled checkpoints and exclude the other two, whose rates of 0.732 and 0.596 on open terrain are largely defaults rather than terrain context. Among the four the split is by generation and by nothing else. The two 2024 checkpoints, one fusing image tokens into the text sequence and one using gated cross-attention, affirm on 0.832 and 0.837 of destroyed tiles and lose 12.3 and 11.8 points when the buildings are actually destroyed, while the two 2025 checkpoints, from different vendors again, affirm on 0.925 and 0.935 and lose only 6.0 and 5.5. On the strict subset the same ordering holds, from 0.686 and 0.733 to 0.826 and 0.872. Sensitivity to the object actually being gone therefore roughly halved between the two generations, across two vendors and two fusion designs within each generation, and it is not attributable to an image-free default in any of the four. With one checkpoint per vendor per generation the contrast rests on four models and cannot separate generation from vendor-specific training choices. Because the corpus is out of domain for every model, we read these numbers as a within-domain generational contrast under one protocol and not as deployment estimates.

Three additional caveats accompany the result. The destroyed signal comes from tsunami, wildfire, and tornado events because the earthquake subset contains no tile above the destruction threshold. Overhead imagery is far from the natural-photograph distribution the rest of the study uses, so only

within-domain contrasts that share the shift are interpreted. Absolute scores and affirmative rates are not comparable to natural-image benchmarks.

The score-adjustment analysis uses a context estimate computed without annotations, a damage score built from prompts about rubble, collapse, and demolition, which orders the sets at 0.230 on destroyed, 0.220 on intact, and 0.209 on empty tiles. Among urban tiles this score predicts absence and therefore instantiates the sign-flipped case of Equation 4. A pilot that selected the adjustment strength to minimize destroyed false alarms alone reduced them from 0.735 to 0.520 on the anchor but raised empty false alarms from 0.450 to 0.695. We report this one-sided tuning artifact because it shows that a bias adjustment relocates error mass when its objective ignores part of the distribution. Selecting the strength for the pooled absent false-alarm rate and adding a second terrain axis built from prompts about bare ground, fields, forest, and water resolves the artifact. The damage axis reduces destroyed false alarms, while the terrain axis reduces empty-terrain errors, as summarized in Table 29 for the five-model mean at a matched present recall near 0.90. Strengths are selected on a development half and evaluated on a disjoint test half. The per-model paired-bootstrap drops of the pooled rate exclude zero in all five checkpoints and range from 0.03 on CLIP ViT-B/32 to 0.20 on the anchor. OpenCLIP ViT-L/14 is the strongest base detector on this imagery, while the least calibrated checkpoints show the largest operating-point changes.

Table 29: Two-axis conditional calibration on xBD, reported as the mean over five contrastive checkpoints at a matched present recall near 0.90, with strengths tuned on a development half and evaluated on a disjoint test half.

Method	False-alarm rate on destroyed tiles	False-alarm rate on empty tiles	Pooled absent false-alarm rate
Global threshold	0.711	0.347	0.468
Damage prior	0.625	0.346	0.439
Damage and terrain priors	0.621	0.238	0.366

Y THE SAME PROBE ON NATURAL IMAGES

The recoverability result of Section 6 rests on one concept in one corpus these models were not built for, so a reader may reasonably ask whether it describes overhead imagery and the word building rather than object-absence judgments. We therefore repeat it on the 77 COCO concepts of the main design, in the same forward pass that produces the decision, at the two sites Appendix X uses. The readout site is the final-layer hidden state at the last position, the vector the language-model head multiplies to form the two answer logits, and the projector site is the layer-zero hidden states at the image-token positions, mean pooled. Probes are fitted per concept with five-fold cross-validation and an inverse penalty chosen inside the training folds, so no image is scored by a probe that saw it, and the image subsample replays the generator of the published decision-variable cache in its own loop order, which the script checks by recomputing that variable and comparing it to the cache. The answer has three parts and we report all three. The direction survives on one of the two checkpoints, its size survives on neither, and a further measurement runs against a reading a reader might take from Section 6. The second part is what Appendix Q already implies rather than a failure to replicate, since a decision variable that separates present from high-expectedness absent images at 0.987 there leaves a readout almost nothing to recover, so what follows fixes the scope of Section 6 instead of contradicting it.

On the ranking axis the readout probe beats the decision variable on Qwen2-VL, the checkpoint Section 6 rests on, carrying the area under the curve for present against high-expectedness absent images from 0.969 to 0.989, a paired gain of $+0.021$ $[+0.013, +0.032]$ that is positive in 62 of the 77 concepts. On Qwen3-VL-8B there is no gain at all, at $+0.000$ $[-0.001, +0.002]$ and 29 of 77. The reason is visible in the levels rather than in the probes, since that checkpoint's answer logits already rank these images at 0.990, leaving one hundredth of headroom against the 0.192 available on the satellite tiles. The gap of Section 6 is therefore a property of the out-of-domain regime, where the decision variable is a poor ranker, and not of absence judgments in general. What does transfer across both domains is which site carries the evidence, since the projector probe reaches 0.969 and 0.971 and so adds nothing over the answer logits, with a gain of -0.000 $[-0.009, +0.012]$ on Qwen2-VL and -0.019 $[-0.024, -0.014]$ on Qwen3-VL-8B. Table 30 carries the levels behind these contrasts.

The second measurement is the fixed-operating-point contrast of Appendix I, and it goes the other way. Fitting the probe on the low-expectedness stratum alone and taking its equal-error threshold from held-out images of that stratum, so that the high-stratum images sit in no training fold, its balanced error rises from 0.017 to 0.131 on Qwen2-VL while the decision variable's rises from 0.055 to 0.105, and from 0.016 to 0.109 against 0.024 to 0.090 on Qwen3-VL-8B. The probe's error therefore grows by more, not less, by $+0.063$ $[+0.045, +0.082]$ and $+0.028$ $[+0.019, +0.037]$, and the projector probe is worse again at $+0.237$ and $+0.219$. One qualification belongs with this, since a probe whose in-stratum error is a third of the decision variable's places its threshold where the two low-stratum classes barely overlap, and a sharper boundary transfers worse for reasons that are not about context. Even allowing for that, nothing here shows a readout that absorbs the context shift.

Taken together the two measurements say that linear recoverability buys ranking rather than the operating point, which is what Appendix Q concludes from three attempts at threshold-side repair and what this adds a fourth to, on the strongest terms available, namely target labels and the deciding vector itself. It also fixes the scope of Section 6 more tightly than that section could on its own. The probe is supervised and per concept while the decision is zero shot, so the licensed claim remains that the evidence is linearly available in the deciding vector, not that the two are comparable detectors.

Table 30: Probes on the decoder's own representation over the 77 COCO concepts. Areas under the curve are for present against high-expectedness absent images, averaged over concepts with 95 percent concept-bootstrap intervals under one shared draw. The error columns give the balanced error at one equal-error threshold taken from the low-expectedness stratum and carried to the high one, with the probe fitted on that stratum alone.

Checkpoint	Answer logits	Readout probe	Projector probe	Answer-logit error	Readout-probe error
Qwen2-VL-7B	0.969	0.989 [0.987, 0.992]	0.969 [0.962, 0.975]	0.055 to 0.105	0.017 to 0.131
Qwen3-VL-8B	0.990	0.990 [0.988, 0.992]	0.971 [0.964, 0.977]	0.024 to 0.090	0.016 to 0.109

Z Extended limitations

Three qualifications extend the summary in Section 7. First, the interventional evidence is heterogeneous across designs. The compositional factorial establishes an affirmative criterion effect for the anchor encoder, the Visual Genome replication of Appendix W shows a positive sensitivity change, and Qwen3-VL shows a negative DETR-controlled coefficient that Grounding DINO reverses, so no single design licenses a general invariance statement. The magnitude of the decoder reduction is also partly a dispersion effect rather than a loss of separable information, as Appendix P quantifies. A third confusability proxy sharpens the same point, since the decoder residual averages -0.502, -0.270 and $+0.129$ across three specifications that all measure detectability well, so its sign is a property of the specification rather than of the data. The complementary context-removal intervention, which uses diffusion inpainting to delete companion objects, is inconclusive at three concepts, because deleting one peripheral companion moves the score by $+0.0054$ on average, opposite to the predicted decrease and at the same order as the $+0.0024$ produced by a no-operation inpainting control, with per-concept signs that disagree. The interventional score inflation is also carried mostly by the absent cell, which is consistent with score saturation on present images and marks a boundary of the additive approximation rather than a decoder-localized absent component.

Second, category transport is evaluated through its structural predictions rather than through direct access to model-specific pretraining pairs, so the corpus comparison identifies which estimated association table predicts model behavior without establishing which distribution the models actually learned.

Third, the image-free baselines of Appendix V bound one component and leave the other open. They fix what the question contributes without target evidence and show that this contribution is a per-concept constant which cannot generate the within-concept prior contrast, but they say nothing about what unannotated context contributes inside a natural image. Isolating that would require deleting companion objects while preserving the rest of the scene, which is the inconclusive diffusion intervention reported above. The baselines also show that the image-free default is input dependent, affirmative for InstructBLIP on every canvas, affirmative for Qwen3-VL-8B only when no image is supplied at all, and absent from Qwen2-VL except on a black canvas, so we read these levels as diagnostics of a model's default rather than as one language-only quantity to subtract.