VISION-LANGUAGE MODEL CONFIDENCE IS NOT A PROPERTY OF THE ANSWER

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models are increasingly deployed behind a confidence gate: the system reads how confident the model is in its answer and defers when that confidence is low. This makes the confidence signal itself worth attacking. We show that a white-box adversary who perturbs only the input image, within an L_∞ budget of $8/255$ and under the constraint that the model's answer stay byte-identical, can invert the confidence ranking, lowering it on correct answers and raising it on wrong ones until the signal points the wrong way. The answer never changes, and most of the inversion persists even when the whole next-token distribution is held near the clean one, so the answer does not determine the confidence attached to it. Across four vision-language models and three visual question-answering benchmarks, the attack drives the model-internal readouts we test below chance in 83 of 84 readout-by-cell profiles under an adversary that knows which answers are correct, and six of the seven readouts fall below chance in every cell under an adversary that does not. Training a probe on the model's frozen hidden states does not fix this: the robustness such a probe gains against the attack is paid for with the information that made it useful. Nor does reading confidence from a separate verifier model, which holds only until the attacker can reach it. Nor does an adversarially trained vision encoder, which at its own training budget holds against the attacker without correctness labels but not against the one with them. How far an answer-preserving adversary can reach a signal governs where it survives, and whether a robust and informative readout can be built remains open. For deployment, a gate under this attack admits up to 89.8% of the wrong answers it would otherwise catch, 79.0% when the attacker has no correctness labels, and, corrected for how often the model is wrong, the answers it accepts are less accurate than with no gate at all.

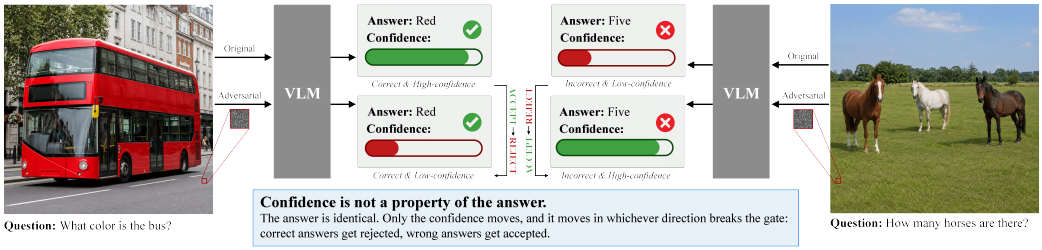

Figure 1: An answer-preserving confidence attack. An image perturbation within an L_∞ budget, constrained to leave the model's answer byte-identical, moves the confidence the model reports in that answer: down on a correct answer until the gate rejects it, and up on a wrong answer until the gate accepts it. The answer never changes, and the gate's decision reverses.

1 INTRODUCTION

A vision-language model returns an answer, but a deployed system rarely takes that answer on faith. It reads off a confidence score and uses it to decide whether to accept the answer, defer to a human, or abstain. This gate is only worth having if the score it reads is hard to fool. We show that it is not. An adversary who perturbs only the input image, within an L_∞ budget of $8/255$, and who is forbidden from changing the model's answer by a single token, can still invert the confidence ranking:

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.

driving it down on answers that happen to be correct, so the gate rejects them, and up on answers that are wrong, so the gate accepts them. Only the model's stated trust in that answer moves, and it moves far enough to reverse the gate's decision (Figure 1).

With white-box gradients one can move any differentiable quantity, so the point is what stays fixed while the score moves. The answer is a hard constraint, fixed exactly across a full generated sequence, yet the confidence attached to it can be pushed down on correct answers and up on wrong ones until the two are anticorrelated. A quantity that can be steered this freely while the behavior it describes stays fixed is a separate signal, read from the same network, that can be corrupted on its own. Prior work established that confidence can be manipulated with the prediction held fixed, for image classifiers (Galil & El-Yaniv, 2021; Obadinma et al., 2024) and, through recoverable and steerable confidence directions, for text language models (Miao & Ungar, 2026; Ji et al., 2025). That a fixed-prediction confidence attack exists is therefore settled. The question those results leave open is which confidence signals survive such an attack and why. In open-ended vision-language generation, with the entire answer preserved instead of a single class label, we show that confidence is separable from the answer it reports, identify the two mechanisms through which an answer-preserving attack corrupts it, and measure how the natural external remedy degrades as the attacker's access grows.

The readouts read from inside the model fail almost without exception. Across four models and three benchmarks, the attack drives the internal readouts below chance in 83 of 84 readout-by-cell profiles under an adversary with correctness labels, and six of the seven readouts fall below chance in every cell under an adversary without them. Training a sturdier signal fails in a way that explains why: what makes a probe more robust to the attack is the loss of the information that made it useful, so robustness and informativeness never arrive together. Moving the check into a separate, independently built model only relocates the problem. Such a verifier survives only until the attacker gains query and then gradient access to it, and with full access it collapses as completely as the internal signals. Reachability, how far an answer-preserving adversary can move a signal, is the axis along which we locate where the separable signal survives.

Two further results fix what the inversion is and what it costs. Holding the answer byte-identical pins only the top token at each position, so the attack might live in the slack the rest of the distribution leaves free; bounding the KL divergence of the full distribution at every answer position to 0.2 nats leaves 62% of the inversion intact. The cost is measured on a gate a deployment would run: a hidden-state gate with its threshold fixed to a 10% false-accept rate on clean data admits up to 89.8% of the wrong answers it had rejected, and up to 79.0% when the attacker has no correctness labels; once each benchmark is reweighted to the model's natural error rate, the answers it accepts are less accurate than accepting every answer in all twelve cells, under either attacker.

We make three contributions. (1) We show that an image-only perturbation that leaves the generated answer byte-identical drives model-internal confidence below chance in 83 of 84 readout-by-cell profiles, across four vision-language models, three benchmarks, and seven readouts, that an attacker with no correctness labels still drives six of the seven readouts below chance in every cell, and that most of the inversion survives when the full next-token distribution is held near the clean one. (2) We locate where discrimination is lost. Splitting attacked AUROC over same-answer and cross-answer pairs separates two routes, merging items that share an answer into a common score range and moving each item's score on its own, and long, unique answers isolate the second; no probe we train on frozen hidden states, adversarially trained or not, is both informative and robust, and an adversarially trained vision encoder narrows the attack's reach without closing it. (3) We show that the weakness extends beyond the model and into deployment: an independent verifier degrades as the attacker moves from transfer to query to gradient access, and a confidence gate under attack admits up to 89.8% of the wrong answers it had rejected, 79.0% without correctness labels, at the model's natural error rate accepts answers less accurate than the full set it would accept with no gate, and lets through up to 88% of the wrong answers an attacker first induces through the image.

2 Related Work

Confidence as a deployed safety signal. Selective prediction trades coverage for reliability by accepting confident predictions and deferring the rest (El-Yaniv & Wiener, 2010), and failure detection

frames calibration, misclassification detection, and out-of-distribution detection as one problem in which a score must rank correct predictions above failures (Jaeger et al., 2023). This ranking is what a gate depends on, and what we measure with the answers held fixed. For vision-language models, selective question answering has been extended to black-box models whose unreliable answers are flagged by inconsistency across related questions (Khan & Fu, 2024). Prior work asks this on clean, shifted, or adversarially collected inputs; we ask it on inputs perturbed against the confidence signal itself.

Prediction-preserving confidence attacks. The closest predecessor is ACE (Galil & El-Yaniv, 2021), which holds an image classifier's prediction fixed and uses correctness labels to lower confidence on correct items and raise it on incorrect ones. Calibration attacks extend this to bidirectional over- and under-confidence across several uncertainty channels (Obadinma et al., 2024; Kopetzki et al., 2021), verbal confidence in language models has been attacked directly (Obadinma & Zhu, 2025), and the adaptive-attack literature has shown repeatedly that detectors and defenses must be tested against an adversary aimed at the specific mechanism (Carlini & Wagner, 2017; Athalye et al., 2018a; Tramèr et al., 2020). We take both lessons as given: in the core matrix five readouts are attacked through their own gradients, CCPS through a differentiable surrogate, and token probability through the endpoints aimed at the others. We ask which signals survive when the preserved object is a complete generation held byte-identical instead of a class label, and on short answers find none that stays both informative and robust.

Where model confidence is read, and how it has been stressed. The readouts we attack span the ways confidence is estimated in practice: output likelihood and self-evaluation (Kadavath et al., 2022), learned probes over hidden states (Azaria & Mitchell, 2023; Beigi et al., 2024), and stability under internal perturbation (Khanmohammadi et al., 2025). Multi-pass estimators that read uncertainty from the semantic distribution of several sampled responses (Farquhar et al., 2024; Zhang et al., 2024) are a further family we do not attack. A parallel line shows that confidence and truth are recoverable as approximately linear directions that can be steered at inference (Miao & Ungar, 2026; Cho et al., 2026; Ji et al., 2025; Park et al., 2024; Marks & Tegmark, 2024), and recent work has begun to stress language-model uncertainty directly, through a training-time backdoor on multiple-choice distributions (Zeng et al., 2024) and through prompt transformations that degrade estimator ranking (Bakman et al., 2025). None of these holds a complete open-ended generation fixed, perturbs the image against a frozen model at test time, or asks what the estimator families share once the answer is pinned.

Neighboring guarantees and attacks. Certified-uncertainty methods bound worst-case class probability or robustly separate out-of-distribution inputs within a neighborhood (Meinke & Hein, 2020; Bitterwolf et al., 2020; Meinke et al., 2022), and adversarially robust conformal methods guarantee coverage of a prediction set under perturbation (Gendler et al., 2022; Yan et al., 2024). None holds the prediction fixed while certifying that correct and incorrect answers stay separable. The amplitude certificate we test in Appendix C.6 is a sufficient condition for a readout to survive, and no readout we test meets it. Adversarial work on generative VLMs targets the generated content, transferring targeted responses or eliciting unsafe behavior (Zhao et al., 2023; Bailey et al., 2024; Qi et al., 2024); there the answer is what changes, while ours holds it exact. Adversarially fine-tuning the vision encoder defends against such content attacks (Schlarmann et al., 2024); we test whether it also protects the confidence attached to an unchanged answer.

3 Threat Model and Measurement

The answer-preserving image attack. We assume a white-box adversary that perturbs only the input image within an L_∞ budget of $8/255$, and the attack works in valid pixel space: every optimization step is projected onto $[x-\epsilon, x+\epsilon] \cap [0,1]$, and the returned endpoint is serialized as an 8-bit image and verified after the fact. The defining constraint is that the model's answer must not change. During optimization every emitted answer token, including the terminator, must remain the argmax at its position, and the returned image is then checked by regenerating the full answer with greedy decoding and requiring it to be byte-identical to the clean answer; any perturbation that alters a single token is rejected. Byte identity is the strictest form of answer preservation; relaxing it to semantic equivalence could only enlarge the attacker's feasible set. Every item stays in the pool: one for which the search finds no answer-preserving iterate keeps its clean image and clean score and counts against the attack, which happens for at most 1.1% of any objective's endpoints in a cell (Appendix A.2). The full setup, attack algorithm, and verification checks are in Appendix A.

What we measure. For a given readout and item, we run the attack in both directions, once to push the score as low as it will go and once as high, subject to the answer-preservation constraint. The attack is projected gradient descent (PGD) (Madry et al., 2018) on the image, run per item and direction with two restarts, one from the clean image and one from a random start, and a larger search for COUPLED (Table 8); it may pass through iterates that change the answer but returns only the best one that preserves it. The two directions trace out the readout's answer-preserving reachable range, witnessed by the extremal images the attack finds. The quantity of interest is discrimination: whether the score still ranks correct answers above incorrect ones, measured by AUROC over a balanced pool with the answers and their correctness labels held fixed. We report it in two forms. The worst-case form asks how far discrimination can be driven when the adversary knows which answers are correct and pushes each in the harmful direction; it bounds the damage, and an attacker who induced a wrong answer holds exactly this knowledge about the answer it planted. The label-free form asks the same of an adversary with no correctness labels, who chooses the push direction from the model's own clean confidence.

Models, benchmarks, and readouts. We evaluate four vision-language models spanning four architectures and 2B to 12B parameters: InternVL3.5-2B (Wang et al., 2025), Molmo2-4B (Clark et al., 2026), LLaVA-OneVision-7B (Li et al., 2024), and Gemma-3-12B (Team et al., 2025). We use three visual question-answering benchmarks: GQA (Hudson & Manning, 2019), which is in-distribution for the trained probes, and VQAv2 (Goyal et al., 2017) and POPE (Li et al., 2023), which are out-of-distribution. A fourth benchmark, A-OKVQA (Schwenk et al., 2022), is used only for the long-answer experiment of Section 5. Each model and benchmark contributes a canonical balanced pool of 1000 items, split evenly between correct and incorrect answers. We attack seven confidence readouts: answer-span token probability (Liu et al., 2024); two hidden-state probes, $P(\mathrm{IK})$ (Kadavath et al., 2022) and SAPLMA (Azaria & Mitchell, 2023), read at the prompt-final and answer-final tokens; a perturbation-stability estimator, CCPS (Khanmohammadi et al., 2025), with a differentiable variant trained on its features as its attack surrogate; and two further probes on the answer-final hidden state: COUPLED, an answer-coupled probe built as a defense that reads only the projection of that state onto a per-item answer-coupled subspace, and FULL, its unprojected control, an independently fitted probe on the same state SAPLMA reads (Appendix A.1). Verbalized confidence (Tian et al., 2023) and self-consistency (Wang et al., 2023) appear only in the universal-perturbation control (Appendix I.1). Correctness labels come from a vision-capable judge, gpt-5.4-mini, which agrees with blind human adjudication on 99.0% of a stratified 300-item sample. Headline results use one canonical item draw at seed 23, and two further seeds reproduce them (Appendix D.3).

Scope. Because the answer is held byte-identical, an accept-or-defer decision that reverses under attack reverses because the monitor was corrupted. A signal meant for oversight must hold against the worst input it can be shown, so we give the adversary full gradient access to the answering model. The attacker need not be the user who reads the answer: in document and agent pipelines the party that supplies the image is not the one relying on the gate, and an attacker can first induce a wrong answer through the image and then push that answer past the gate, which we demonstrate in Section 4.

4 Answer-Invariant Confidence Inversion

The model-internal readouts invert below chance under the attack. We run the answer-preserving attack against each of the seven readouts, in both directions, across all four models and three benchmarks, and measure the discrimination that survives. Table 1 reports the median over the seven readouts in each cell: clean AUROC ranges from 0.72 to 0.83, and under the worst-case attack the median falls to 0.09 or lower in every one of the twelve cells, with a narrow spread across readouts. An AUROC near zero means the readout now ranks incorrect answers above correct ones almost perfectly: the model reports its highest confidence where it is wrong. All seven readouts fall below chance in eleven of the twelve cells; in the twelfth, Molmo/POPE, six do and the answer-coupled readout does not, which the decomposition below explains. The collapse is stable across three attack seeds (largest across-seed standard deviation 0.046), and it rests neither on the harshest endpoint selection nor on the full budget: scoring each readout only on the endpoint aimed at it leaves the same 83 of 84 profiles below chance, and so does halving the budget to $4/255$. The per-readout results, the multi-seed evidence, and both checks are in Appendix D.

A label-free attacker, with no correctness oracle, inverts the readouts too. An attacker who did not plant the answer has no correctness labels. The label-free column replaces the oracle with

Table 1: The answer-preserving attack inverts the model-internal readouts, and an exact decomposition shows why one profile resists. Left (seven-readout median): discrimination between correct and incorrect answers (AUROC), median over the seven readouts per cell, clean and under attack, with the interquartile range across readouts; the label-free column is the median over all seven readouts. *WC-oracle* chooses each item's push direction from ground-truth correctness; *label-free* uses only the model's own clean confidence. The attacked median is below chance (0.5) in every cell. Right (binding readout): the decomposition of Eq. (1) for each cell's binding readout (COUPLED, except P(IK) on InternVL/GQA and VQAv2), with w the same-answer pair weight and ρ the same-answer range overlap. †Molmo2-4B/POPE is the sole profile above chance ($A_{\text{adv}} = 0.520$), with the highest w and lowest ρ in the matrix.

		Clean	Attacked AUROC			Decomposition				
Model	Dataset	(median)	WC-oracle	IQR	label-free	w	A_{same}	A_{cross}	A_{adv}	ρ
InternVL3.5-2B	GQA	0.797	0.016	0.005	0.307	0.040	0.000	0.022	0.021	1.000
	VQAv2	0.724	0.015	0.011	0.362	0.045	0.000	0.029	0.028	1.000
	POPE	0.779	0.000	0.000	0.321	0.498	0.002	0.002	0.002	0.998
Molmo2-4B	GQA	0.780	0.009	0.003	0.359	0.066	0.450	0.279	0.291	0.441
	VQAv2	0.787	0.039	0.029	0.358	0.051	0.514	0.236	0.250	0.384
	POPE†	0.795	0.000	0.000	0.311	0.525	0.567	0.467	0.520	0.359
LLaVA-OV-7B	GQA	0.828	0.007	0.008	0.306	0.034	0.037	0.049	0.049	0.958
	VQAv2	0.758	0.020	0.015	0.354	0.043	0.056	0.052	0.052	0.931
	POPE	0.745	0.000	0.001	0.333	0.241	0.032	0.038	0.037	0.962
Gemma-3-12B	GQA	0.745	0.060	0.050	0.442	0.070	0.097	0.349	0.331	0.883
	VQAv2	0.774	0.086	0.020	0.433	0.053	0.152	0.346	0.336	0.793
	POPE	0.733	0.008	0.022	0.336	0.465	0.044	0.295	0.178	0.942

the model's own clean confidence: the attacker treats a clean score above the median of a disjoint reference set as a stand-in for a correct answer, and pushes the confident-looking items down and the unconfident-looking ones up, using the same attack endpoints as the oracle and changing only the direction rule. This drives six of the seven readouts below chance in all twelve cells, to between 0.22 and 0.47, and 78 of the 84 profiles overall (Appendix D.2). The exception is the answer-coupled readout, which falls below chance on every InternVL and LLaVA cell and holds between 0.52 and 0.69 on every Molmo and Gemma cell; its strongest cell is Molmo/POPE, the profile that survives the oracle, so the survivor is the same under both adversaries. The oracle inverts the ranking almost completely and the label-free attacker only pushes it past chance, but under either a gate can be made to rank wrong answers above right ones.

Attacked discrimination splits exactly into a same-answer and a cross-answer term. Partition the correct-incorrect pairs a readout must rank into those whose two answers are the same string and those whose answers differ, and let w be the fraction that share an answer (Figure 2). AUROC is a mean over these pairs, so it separates without residual into a same-answer part and a cross-answer part,

$$A_{\text{adv}} \;=\; w\,A_{\text{same}} \;+\; (1-w)\,A_{\text{cross}}, \tag{1}$$

where A_{same} and A_{cross} are the attacked discrimination on same-answer and cross-answer pairs. The identity holds for any scoring and carries no assumption (Appendix B); its use is that the two terms fail for different reasons. On same-answer pairs the answer string carries no information, so the readout's discrimination lives entirely in the score, and the attack destroys it by pushing two items that share an answer until their reachable score ranges overlap and no threshold separates them; we measure this overlap per cell as ρ, the fraction of same-answer correct-incorrect pairs whose reachable ranges overlap. On cross-answer pairs the answers differ, and the attack corrupts the ranking by moving each item's score on its own. The weight w decides which term dominates a given cell. Appendix C develops both mechanisms in full.

The same-answer weight sets which term dominates, and explains the one surviving profile. Where answers are rarely shared the collapse is almost entirely cross-answer: on LLaVA/GQA, w is 0.034 and the cross-answer term is 97% of A_{adv}. Where answers are often shared the same-answer term carries real weight, and whether it collapses turns on ρ. Two POPE cells show this at nearly fixed weight. On InternVL/POPE, $w = 0.498$ but the overlap is near-total, $\rho = 0.998$, so the same-

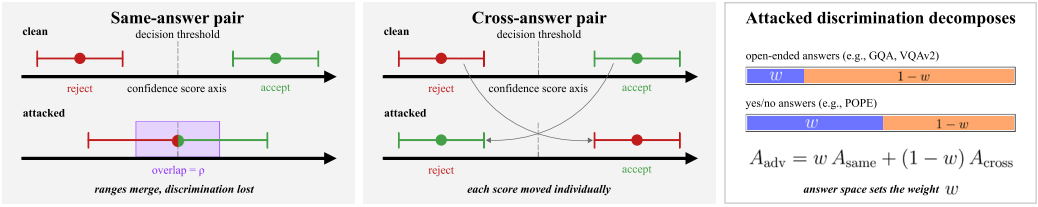

Figure 2: The exact decomposition. A confidence readout must rank correct against incorrect answers, and these pairs split into two kinds: same-answer pairs, whose two items produce the identical answer string, and cross-answer pairs, whose answers differ. Attacked discrimination is the weighted sum of the two, $A_{\text{adv}} = w\,A_{\text{same}} + (1-w)\,A_{\text{cross}}$, with the same-answer weight w setting which mechanism dominates. The one surviving profile, Molmo/POPE, has high w and low ρ: its same-answer term is large and the attack cannot merge it.

answer intervals merge, A_{same} falls to 0.002, and the cell collapses with the rest. On Molmo/POPE, $w = 0.525$ but the overlap is the lowest in the matrix, $\rho = 0.359$: the same-answer intervals do not merge, A_{same} stays at 0.567, and because same-answer pairs make up half the cell, the readout holds at $A_{\text{adv}} = 0.520$, marginally above chance in point estimate, with a bootstrap interval that includes 0.5 on every seed (Appendix D.3). With the weight nearly fixed across these two cells, the overlap alone accounts for the difference. Low overlap alone is not enough: Molmo's other two cells also have low ρ but small weight, and collapse to 0.29 and 0.25 as the cross-answer term dominates. Survival needs both a same-answer term the attack cannot merge and enough weight for it to carry the cell, and only Molmo/POPE has both.

The answer does not determine the confidence, even when the full distribution is constrained. Holding the answer byte-identical pins the top token at every position but leaves the rest of the next-token distribution free, and the attack does move it. We therefore add a bound on the KL divergence of the full next-token distribution at every answer position, enforced by rejecting any step that violates it, and re-run the attack. At a bound of 0.2 nats a mean 62% of the full inversion survives for the SAPLMA readout, which stays below chance in all twelve cells, and 69% for P(IK) (Appendix E). For three of the four models this bound is 35 to 59 times the divergence induced by a single-greylevel pixel dither. Gemma is the decisive case. Its output distribution is so sensitive to input change that a 0.2-nat bound holds its sampled continuations at the clean noise floor, behaviorally indistinguishable from the clean model, and there 83 to 93% of the inversion still survives (Appendix E). Tightening to 0.05 nats is where the picture splits. For InternVL, Molmo, and LLaVA that bound is still nine to fifteen times their benign-dither divergence, and they return toward chance with that slack unused; for Gemma it is below its own dither floor, and Gemma stays inverted. The answer does not determine confidence at any bound we test; confidence depends in part on the distribution's fine structure, and the erosion from 62% to 36% measures how much of it lives there.

A confidence gate under attack accepts worse answers than no gate would. We simulate a gate directly: the FULL hidden-state gate, with its threshold fixed to a 10% false-accept rate on clean data and held fixed under attack, as in deployment. With the answer unchanged, the attack turns previously rejected wrong answers into accepted ones, up to 89.8% in the worst cell, and up to 79.0% when the push direction comes from the label-free rule. Because the relevant comparison depends on how often the model is wrong, we reweight each benchmark to its natural error rate. Under that accounting, the attacked gate accepts answers at lower accuracy than accepting every answer in all 12 cells, whether the attacker knows the labels, does not, or aims at a different probe, P(IK), and lets the gate read the result. On InternVL/POPE, balanced-pool accepted accuracy falls from 0.744 to 0.006. The verdict holds at every threshold we test and over the whole risk-coverage curve (Appendix F).

An attacker who plants a wrong answer can also get it past the gate. We run this combined attack on POPE for LLaVA and Molmo, starting from 200 correctly answered items per model and staying within a single $8/255$ budget around the clean image. A targeted perturbation first flips the answer, which succeeds on 98.5% and 100% of items. On its own the flip is caught: the gate's confidence in the induced answer drops to a median of 0.51 and 0.64, and the gate accepts only 3.5% and 1.0% of the wrong answers. The answer-preserving attack then raises that confidence while holding the induced answer byte-identical, and the gate accepts it for 77% of the originally correct items on LLaVA and 88% on Molmo (Appendix F.3). Inducing the error and hiding it from the gate draw on the same budget, and the attacker needs no correctness labels, because it knows the answer it planted is wrong.

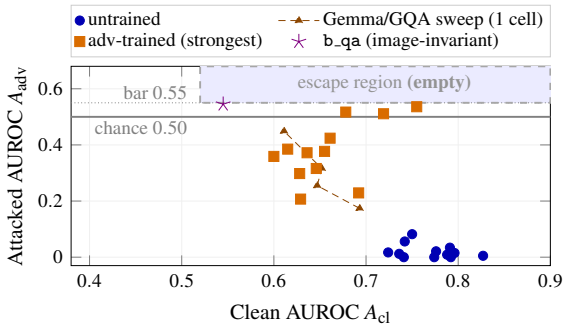

Figure 3: No trained readout is both informative and robust. Each point is a readout for one of the twelve cells, by clean (x) and attacked (y) discrimination. Blue: untrained, informative but collapsing. Orange: adversarially trained at each cell's strongest setting; the dashed line traces the full weight sweep on Gemma/GQA. The shaded escape region, clean above the answer-string prior and attacked above 0.55, is empty. `b_qa`, the pixel-free control, cannot be moved but sits below the prior: unreachable yet too weak to gate.

5 No Internal Readout Escapes on Short Answers

Adversarial training gains robustness only by giving up informativeness. On each of the twelve cells we train a family of confidence probes on the same hidden states with adversarial training that rewards invariance to the answer-preserving attack, sweeping the robustness weight over $\{0.3, 1, 3, 10\}$ with three seeds at the strongest setting, and then attack each trained probe with the same answer-preserving attack and L_∞ budget of $8/255$ used throughout. Across all twelve cells and every weight, and for a separate gradient-penalty variant, no probe is both informative on clean inputs and robust under attack. The two properties trade off against each other (Figure 3). On Gemma/POPE, the lightest-penalty probe has clean AUROC 0.865 and attacked AUROC 0.061; raising the penalty until attacked AUROC reaches its maximum of 0.385 drags clean AUROC down to 0.629. Robustness is bought with the discrimination the probe was meant to provide. Figure 3 traces the same trade-off across the full weight sweep on Gemma/GQA, and its escape region is empty in all twelve cells. This trade-off is a property of post-hoc probes over frozen representations: a head cannot move the representation it reads, so invariance can only be bought by ignoring the directions the attack moves. A robust vision encoder, which does change the representation, is tested below. The collapse is also stable across attack configurations and under an answer-margin constraint (Appendix G.2).

The closest probe misses a pre-specified escape bar. We fixed the criterion for an escape before running the sweep: at the same operating point, clean AUROC above the answer-string prior and attacked AUROC above 0.55. One cell, LLaVA/POPE, comes closest. Its best probe reaches clean AUROC 0.755 and attacked AUROC 0.536 as three-seed means (the seed-23 values are 0.768 and 0.544), but 0.536 sits below the bar, and the three-seed standard deviation of 0.008 places it there reliably. None of the defenses we trained escaped (Appendix G).

A robust vision encoder narrows the attack's reach but does not close it. A probe cannot change the representation it reads, so we also test a defense that does. On LLaVA-1.5-7B we replace only the vision tower, swapping the standard CLIP ViT-L/14 for FARE-4, the same encoder adversarially fine-tuned at $4/255$ (Schlarmann et al., 2024), and hold the language model, projector, probe recipe, and attack fixed, at both LLaVA-1.5's 336-pixel input and FARE's native 224. Both towers yield informative readouts, and neither yields a robust one. At $4/255$, FARE's own training budget, the worst-case attack drives P(IK), SAPLMA, and token probability below chance with either tower, to between 0.00 and 0.08 on the standard tower and between 0.10 and 0.44 on FARE (Appendix G.3). Against an attacker without correctness labels the robust encoder does hold at this budget, keeping every readout between 0.53 and 0.64 while the standard tower's fall below chance. What FARE changes is how far the attack reaches. On GQA the reachable range falls from about twice the clean gap between correct and incorrect scores to between 0.95 and 1.4 times that gap, and at $2/255$ below it, where FARE's readouts move to about chance while the standard tower stays inverted. The closest case, FARE's SAPLMA at native resolution and $2/255$, reaches 0.564, but a pre-registered rerun with a stronger attack places it below the escape bar. White-box access alone therefore does not guarantee the inversion; how deep it goes tracks how far an answer-preserving perturbation can move a score relative to the separation the readout was trained to produce.

Long, unique answers switch off same-answer merging and isolate the cross-answer term. On short answers the two terms of Eq. (1) run together, because answers are often shared and both

are active at once. When every answer is distinct there are no same-answer pairs, w goes to zero, and only the cross-answer term remains, so long, unique answers let us ask whether that term is attackable on its own. Because the three core benchmarks do not elicit long answers, this experiment uses A-OKVQA, whose questions call for reasoning: on 1000 items we induce long answers with a reasoning prompt, giving median answers of 77 to 97 tokens on all four models, and Molmo is the one model whose answers are long even without it (25.5-token median under a neutral prompt); P(IK) and SAPLMA are retrained on an image-disjoint 600-item A-OKVQA split and evaluated on balanced pools of 180 to 298 items. Holding a roughly ninety-token answer byte-identical is a severe constraint: for the median item only 2 of the 62 attack iterates preserve it, and between 8 and 16 candidate perturbations per item are rejected for breaking it, so any collapse here is not an artifact of preservation being easy on short answers.

With merging gone, the attack still inverts internal readouts on two of four models. On Gemma, P(IK) and SAPLMA fall from clean 0.67 and 0.60 to 0.38 and 0.40, and on Molmo from 0.68 and 0.62 to 0.43 and 0.43. The token-probability readout collapses on all four models, as low as 0.08 on InternVL. The attack can therefore move a single fixed answer's confidence far enough to corrupt the ranking with no shared-answer structure to exploit, which is the second mechanism shown in isolation. On InternVL the trained probes resist, P(IK) and SAPLMA both holding at 0.62 with bootstrap intervals above 0.55; on LLaVA they stay at 0.58 to 0.62 in point estimate, but that cell has only 65 to 69 usable items and its intervals include 0.5, so LLaVA's resistance is suggestive rather than established. None of reasoning training, native answer length, backbone, size, the strength of the answer-final correctness signal, or input-to-logit sensitivity predicts the split. What separates the groups is how far a model's readout moves under an answer-preserving perturbation, which this analysis measures but does not explain (Appendix C.4).

6 Outside the Model: The Reachability Ladder

An independent verifier resists the transferred attack only while the attacker has no access to it. If every readout inside the answering model is reachable, one can instead hand the image, question, and answer to a second, independently trained model and ask it to judge correctness. The attacker is white-box on the answering model but, at first, has no access to this verifier. In that setting the verifier survives: across the twelve base-verifier pairs its discrimination under the transferred attack stays around 0.57 on average, while the answering model's own internal readout on the same attacked images sits near 0.05. This looks like a defense; the rest of the section shows it relocates the attack.

Verifier discrimination falls as the attacker gains optimization power over the verifier. We grant progressively more verifier access and measure what survives (Table 2). With score-only access, the attacker queries the continuous score and runs gradient-free search; mean discrimination falls to 0.48, already below chance. With white-box access to the verifier's gradient, it collapses to 0.06–0.35 across the nine pairs with an exact differentiable verifier path (Appendix H.3), a mean of 0.223, matching an internal readout, and white-box is the strongest attack in ten of twelve pairs. Throughout, the base model's answer is held byte-identical under the internal attack's constraint: a perturbation is kept only when a fresh greedy decode of the base model reproduces the committed answer, so the verifier's collapse is measured on images the base model still answers identically. The two intermediate rungs, transfer and decision-only, sit close together at 0.57–0.60 and are not separably ordered: a label-aware transfer oracle is already a strong baseline that hard-label boundary search does not reliably beat. The movement that matters is across rungs that add real optimization power, from clean through score-only to white-box.

The verifier's survivors are saturated items, already decided on the clean image. The items a query attacker cannot flip are the ones the verifier was already most certain about, whose scores sit too far from the decision boundary for a bounded perturbation to cross. Across the twelve pairs about half of the items are decision-flippable under hard-label access alone, so the surviving discrimination rests on the unflipped remainder, the items that began farthest from the boundary. On the strongest pair the mechanism is explicit: the unflippable items sit at mean distance 0.335 from the decision boundary against 0.047 for the flipped ones, and the attack needs a median of 132 queries to reach its first flip, against 52 to 63 on weaker verifiers. A strong verifier is harder to attack because more of its answers begin beyond what a bounded perturbation can move, and survival tracks clean quality.

Table 2: An independent verifier only relocates the attack: its reliability falls as the attacker gains optimization power over it. Mean verifier AUROC over the twelve base→verifier pairs at increasing levels of attacker access. *Clean* is the verifier's unattacked AUROC and *no access* is its transfer AUROC against the answer-preserving attack aimed at the base model; *decision-only* (HopSkipJump, Chen et al., 2020) and *score-only* (SPSA, Uesato et al., 2018) query the verifier under a fixed query budget, and *white-box* attacks its gradients directly. Discrimination falls from clean through score-only to white-box; the two intermediate rungs are close. The three rows report all twelve pairs, the eleven excluding the under-powered white-box pair, and the nine whose verifier admits an exact differentiable path. Per-pair values, confidence intervals, and caveats are in Appendix H.

	clean	no access	decision-only	score-only	white-box
All 12 pairs	0.655	0.571	0.601	0.480	0.287
Excluding under-powered cell	0.646	0.558	0.590	0.468	0.246
Bit-exact verifiers (9)	0.688	0.589	0.620	0.496	0.223

An external check relocates the phenomenon. Reachability, how far an answer-preserving adversary can move the signal, connects the verifier's survival to the internal results. Inside the model, a readout collapses because the attack reaches it directly, and training cannot move it out of reach without emptying it of the information that made it useful. Outside the model, an independent verifier collapses once the attacker can reach it, by query or by gradient, and survives only where reach fails: at no access, or where an item sits so far from the verifier's decision boundary that the budget cannot cross it.

7 Conclusion

Summary. Confidence increasingly decides whether a model's answer is trusted, deferred, or shown to a user, and when it is read from inside the model, an adversary who leaves the answer untouched can turn it against its purpose, or plant a wrong answer and carry it past the gate. Neither reading the check from a separate model nor hardening the vision encoder removes this: the verifier holds only until the attacker can reach it, and the robust encoder shrinks the attack's reach without stopping a worst-case adversary at its own training budget. Under that adversary, on the short-answer benchmarks of the core matrix, we find no readout, inside the model or outside it, that stays both informative and reliable. An adversary without correctness labels is weaker, and the answer-coupled readout on two models and every readout behind the robust encoder at its training budget stay above chance against it. On long answers one model holds above chance and a second only in point estimate, which is the open question below. A confidence signal that can be steered while the answer it describes is held exact is a channel of its own, corruptible on its own.

Limitations. The perturbation is small but not invisible: at $8/255$ it is comparable on LPIPS to a two-pixel crop-and-resize, and on a minority of items is perceptible on close inspection (Appendix A.4). The attack is per-item and white-box; it does not transfer across models, and no universal perturbation carries it (Appendix I.1). The threat model is therefore an adversary with gradient access to the answering model and control of the input image, as in open-weight deployment; a server-side re-encode outside the attacker's control is partly covered by the purifier results in Appendix I.2, where JPEG at quality 50 removes between a quarter and two thirds of the manipulable range. The label-free attack is weaker than the oracle, pushing the readouts past chance rather than inverting them; the answer-coupled readout resists it on Molmo and Gemma, and so does every readout behind the robust encoder at $4/255$. The long-answer experiment is a single benchmark, A-OKVQA, with pools of 180 to 298 items and three readouts; input purification is evaluated only nonadaptively and on three models (Appendix I.2). The defense trade-off is measured on post-hoc probes over frozen hidden states and on one adversarially trained vision encoder, not on end-to-end adversarial fine-tuning of the full model, and the attacked readouts are single-pass; multi-pass estimators are not evaluated.

Open question. Whether a robust and informative readout can be built remains open, and the clearest lead comes from the one place the pattern is not uniform. On long answers, whether an internal readout resists the attack is decided by how movable it is item by item, and that movability varies across models in a way none of the properties we examined predicts. Understanding what makes one model's confidence reachable and another's not is where a full account of this phenomenon, and any principled defense against it, will have to begin.

AI USE STATEMENT

In this work, we used generative AI tools to aid with editing, language refinement, and polishing of the main manuscript, and to assist in drafting selected portions of the appendix. All AI-assisted manuscript content was manually reviewed, revised where necessary, and checked by the authors for technical accuracy and consistency with the underlying experiments, analyses, and reported results. During research execution, generative AI assistants were used to support programming, implementation, and the running of experiments, and as discussion partners when planning experiments, checking analyses, and reviewing drafts of the paper. The research question, the threat model, the experimental design, and the interpretation of results were decided by the authors, who evaluated every AI suggestion before adopting it. Any AI-assisted code incorporated into the experimental workflow was manually reviewed, executed, and tested by the authors before use, and the resulting analyses and reported numerical results were checked against the experimental outputs. A generative AI model was also used as an automated correctness judge for model answers, as described in Section 3 and Appendix A.5; its labels were validated against blind human adjudication. Generative AI tools were not used for retrieval or discovery of related work, and we have not used generative AI tools to generate synthetic data sets; translation and qualitative or thematic data analysis are not applicable to this work. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, code, analyses, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies the robustness of confidence mechanisms used to decide whether vision-language model outputs should be trusted, deferred, or rejected. The adversarial techniques developed for this purpose are dual-use: although they are intended to expose weaknesses in confidence-based safety mechanisms, similar methods could be used to manipulate such mechanisms in deployed systems. We therefore frame the attacks as robustness and integrity evaluations and explicitly characterize the assumptions, scope, and limitations of the threat models considered. We also evaluate defensive approaches and report negative results alongside successful attacks in order to distinguish the demonstrated vulnerabilities from broader claims not supported by our experiments. Our goal is to support the development and evaluation of confidence mechanisms that remain reliable under adversarial manipulation.

REPRODUCIBILITY STATEMENT

The supplementary material contains the full code release together with a representative data sample and the artifact stores used to regenerate the reported results. The release reproduces all tables and the data-derived figures in the paper from the provided artifacts and includes a manifest linking each result to the corresponding script and inputs. The codebase contains the complete experimental pipeline, including dataset preparation and the committed seed-23 item identifiers, model inference and hidden-state extraction, the answer-preserving attack, probe training, the verifier experiments, correctness evaluation, and result aggregation. The exact model checkpoints, software environment, hyperparameters, sweep ranges, and selection criteria are documented in Appendix A and Table 8; the decomposition and supporting derivations are given in Appendix B; and the principal verification checks on answer preservation, pixel validity, perturbation magnitude, and correctness-label agreement are reported in Appendices A.2–A.5. The accompanying data sample contains 300 items stratified across all four models and three benchmarks, with the corresponding clean and adversarial images, questions, model answers, correctness labels, and clean and attacked scores for all seven confidence readouts. Together, the released code, artifacts, configuration, and documented evaluation protocol are intended to support reproduction of both the main results and the analyses reported in the appendices.

REFERENCES

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 274–283. PMLR, 10–15 Jul 2018a. URL https://proceedings.mlr.press/v80/athalye18a.html.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples, 2018b. URL https://openreview.net/forum?id=BJDH5M-AW.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: adversarial images can control generative models at runtime. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Yavuz Faruk Bakman, Duygu Nur Yaldiz, Sungmin Kang, Tuo Zhang, Baturalp Buyukates, Salman Avestimehr, and Sai Praneeth Karimireddy. Reconsidering LLM uncertainty estimation methods in the wild. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29531–29556, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1429. URL https://aclanthology.org/2025.acl-long.1429/.

Mohammad Beigi, Ying Shen, Runing Yang, Zihao Lin, Qifan Wang, Ankith Mohan, Jianfeng He, Ming Jin, Chang-Tien Lu, and Lifu Huang. InternalInspector i^2: Robust confidence estimation in LLMs through internal states. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12847–12865, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.751. URL https://aclanthology.org/2024.findings-emnlp.751/.

Julian Bitterwolf, Alexander Meinke, and Matthias Hein. Certifiably adversarially robust detection of out-of-distribution data. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, pp. 3–14, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450352024. doi: 10.1145/3128572.3140444. URL https://doi.org/10.1145/3128572.3140444.

Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. Hopskipjumpattack: A query-efficient decision-based attack, 2020. URL https://arxiv.org/abs/1904.02144.

Seonglae Cho, Zekun Wu, Kleyton Da Costa, and Adriano Koshiyama. The confidence manifold: Geometric structure of correctness representations in language models, 2026. URL https://arxiv.org/abs/2602.08159.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding, 2026. URL https://arxiv.org/abs/2601.10611.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/cohen19c.html.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL http://jmlr.org/papers/v11/el-yaniv10a.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Ido Galil and Ran El-Yaniv. Disrupting deep uncertainty estimation without harming accuracy. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9L1BsI4wP1H.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2017.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Paul F Jaeger, Carsten Tim Lüth, Lukas Klein, and Till J. Bungert. A call to reflect on evaluation practices for failure detection in image classification. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=YnkGMIh0gvX.

Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating verbal uncertainty as a linear feature to reduce hallucinations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3769–3793, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.187. URL https://aclanthology.org/2025.emnlp-main.187/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Zaid Khan and Yun Fu. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10854–10863, 2024. doi: 10.1109/CVPR52733.2024.01032.

Reza Khanmohammadi, Erfan Miahi, Mehrsa Mardikoraem, Simerjot Kaur, Ivan Brugere, Charese Smiley, Kundan S Thind, and Mohammad M. Ghassemi. Calibrating LLM confidence by probing perturbed representation stability. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10448–10514, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.530. URL https://aclanthology.org/2025.emnlp-main.530/.

Anna-Kathrin Kopetzki, Bertrand Charpentier, Daniel Zügner, Sandhya Giri, and Stephan Günnemann. Evaluating robustness of predictive uncertainty estimation: Are dirichlet-based models reliable? In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5707–5718. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kopetzki21a.html.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. URL https://arxiv.org/abs/2408.03326.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Xin Liu, Muhammad Khalifa, and Lu Wang. Litcab: Lightweight language model calibration over short- and long-form responses. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jH67LHVOIO.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Alexander Meinke and Matthias Hein. Towards neural networks that provably know when they don't know. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ByxGkySKwH.

Alexander Meinke, Julian Bitterwolf, and Matthias Hein. Provably adversarially robust detection of out-of-distribution data (almost) for free. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30167–30180. Curran Associates, Inc., 2022. doi: 10.52202/068431-2187. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c2c62117283dda155db754e54dbe8d71-Paper-Conference.pdf.

Miranda Muqing Miao and Lyle Ungar. Closing the confidence-faithfulness gap in large language models, 2026. URL https://arxiv.org/abs/2603.25052.

Stephen Obadinma and Xiaodan Zhu. On the robustness of verbal confidence of LLMs in adversarial attacks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=pAfdjWD9pN.

Stephen Obadinma, Xiaodan Zhu, and Hongyu Guo. Calibration attacks: A comprehensive study of adversarial attacks on model confidence. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=TXzz9xwdpv.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i19.30150. URL https://doi.org/10.1609/aaai.v38i19.30150.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pp. 146–162, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20073-1. doi: 10.1007/978-3-031-20074-8_9. URL https://doi.org/10.1007/978-3-031-20074-8_9.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivi'ere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Jonathan Uesato, Brendan O'Donoghue, Pushmeet Kohli, and Aaron van den Oord. Adversarial risk and the dangers of evaluating against weak attacks. In Jennifer Dy and Andreas Krause

(eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5025–5034. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/uesato18a.html.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Ge Yan, Yaniv Romano, and Tsui-Wei Weng. Provably robust conformal prediction with improved efficiency. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=BWAhEjXjeG.

Qingcheng Zeng, Mingyu Jin, Qinkai Yu, Zhenting Wang, Wenyue Hua, Zihao Zhou, Guangyan Sun, Yanda Meng, Shiqing Ma, Qifan Wang, Felix Juefei-Xu, Kaize Ding, Fan Yang, Ruixiang Tang, and Yongfeng Zhang. Uncertainty is fragile: Manipulating uncertainty in large language models, 2024. URL https://arxiv.org/abs/2407.11282.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/zhang19p.html.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric . In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, Los Alamitos, CA, USA, June 2018. IEEE Computer Society. doi: 10.1109/CVPR.2018.00068. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00068.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation, 2024. URL https://arxiv.org/abs/2411.11919.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Appendix Contents

A Setup and Verification **17**

A.1 Models, environment, and the implemented attack 17

A.2 Answers are preserved byte for byte 19

A.3 The attack remains in valid pixel space 19

A.4 The perturbation is small 19

A.5 The correctness judge agrees with human adjudication 21

A.6 One canonical draw, and nothing tuned against the result 22

B The Exact Decomposition **25**

B.1 The identity 25

B.2 The stored terms reconstruct all eighty-four profiles 25

B.3 The movement is separable from the answer 25

C The Two Mechanisms **28**

C.1 The reachable ranges and the endpoint adversary 28

C.2 The same-answer term: within-answer information, destroyed by merging 28

C.3 The same-answer term admits a measured ceiling 29

C.4 The cross-answer term: isolated by long, unique answers 29

C.5 A causal check on the movement 30

C.6 The uniform certificate, and why no readout meets it 31

D The Phenomenon in Full **34**

D.1 Every readout in every cell 34

D.2 The inversion without a correctness oracle 36

D.3 The collapse is stable across seeds 36

D.4 The collapse does not rest on the harshest endpoint selection 37

D.5 The collapse does not need the full budget 40

E The Distribution Constraint **41**

F Deployment: The Gate Accepts Worse Answers Than No Gate **45**

F.1 The gate harm, cell by cell 45

F.2 The harm does not depend on the threshold 45

F.3 The combined attack: induce a wrong answer, then push it past the gate 45

G No Internal Readout Escapes, in Full **49**

G.1 The full adversarial-training sweep 49

G.2 The collapse is not an artifact of the attack settings 49

G.3 A robust vision encoder 50

H The Reachability Ladder in Full **52**

H.1 The twelve-cell ladder 53

H.2 The intermediate rungs are not separably ordered 53

H.3 The Molmo verifier's differentiable path 53

H.4 Disclosures 54

I Scope and Negative Results **55**

I.1 The attack is per-item and white-box, not universal or transferable 55

I.2 Defenses we did not fully evaluate 55

I.3 An earlier construction reached the same conclusion 56

A Setup and Verification

This appendix records the experimental setup and the checks that make the main results interpretable: the implemented attack, verification that answers are preserved byte for byte, that the perturbed images are valid, that the perturbation is small, that the automated correctness judge agrees with human adjudication, and that no result was obtained by tuning against it. Every table is regenerated from the released stores.

A.1 Models, environment, and the implemented attack

All results are tied to exact checkpoints and a fixed environment. The four vision-language models are InternVL3.5-2B, Molmo2-4B, LLaVA-OneVision-7B, and Gemma-3-12B, obtained from the repositories listed in the released artifact manifest; model-family names do not identify weights uniquely, so the repository identifiers define the systems evaluated here. Attacks, probe training, and hidden-state extraction ran on NVIDIA H200 GPUs under a pinned software stack, and every randomized procedure uses global seed 23. Seeding fixes every sampled quantity but not the floating-point reduction order inside SDPA attention, so for Molmo, LLaVA, and Gemma the attack's 8-bit endpoint is not bit-reproducible across runs; the released code documents which endpoints reproduce exactly and which reproduce as a different valid endpoint with matching answer preservation and budget. All hidden-state probes read the final transformer layer, fixed in advance and not selected on validation: SAPLMA at the last answer token, P(IK) at the last prompt token, and CCPS and CCPS-D on final-layer answer tokens; only the hidden width differs across models (2048, 2560, 3584, 3840). Probes are trained on a 20,000-item GQA training split at its natural class ratio (between 15,832/4,168 and 16,961/3,039 correct/incorrect across models) with class-balanced binary cross-entropy, positive weight n_-/n_+, and early-stopped on a disjoint 5,000-item validation split. Aggregation and table generation reproduce from the released artifacts on CPU.

The attack is projected gradient descent through the frozen model, run per item and direction with a fixed number of steps and restarts, returning only answer-preserving iterates. The four probe objectives use 30 steps and 2 restarts and the answer-coupled objective 60 steps and 3, a budget assigned before launch by a label-free rule that grants the larger search only where it materially widens the reachable range; the first restart starts from the clean image and later ones from random starts, and the best feasible iterate over all restarts is kept (Table 8). For an item whose answer is to be held fixed, it takes signed-gradient steps on the image against the target estimator, with the harmful direction set by item correctness: the score is pushed down when the answer is correct and up when it is wrong. At every step the attack operates on the raw RGB image in $[0, 1]$ and projects onto $[x - \epsilon, x + \epsilon] \cap [0, 1]$, so the iterate stays within both the budget and the valid pixel range; endpoints are stored as 8-bit images (Appendix A.3). At each iterate every emitted answer token, including the terminator, must remain the argmax at its position ($\tau_{\text{safe}} = 0$). Among feasible iterates the attack keeps the one with the largest harmful score movement. Optimization may pass through infeasible points, but only a feasible iterate is returned, and the final answer is regenerated by fresh greedy decoding and checked for byte identity before the item enters any analysis. An item that

never finds a feasible harmful iterate keeps its clean score and zero perturbation, which weakens the measured collapse. All parameters, including the ones swept, are given in Table 8. The full procedure is Algorithm 1.

Algorithm 1 Answer-preserving readout attack (one readout g, one direction $\sigma \in \{+1, -1\}$)

Require: clean image $x_0 \in [0,1]^d$; budget ϵ; step α; steps T; restarts R; margin τ_{safe}; committed answer tokens $a^\star_{1:n}$

$\Pi(x) \triangleq \text{clip}\big(\text{clip}(x,\, x_0 - \epsilon,\, x_0 + \epsilon),\, 0,\, 1\big)$; $\quad \mathcal{C} \leftarrow \emptyset$
for $r = 0$ to $R - 1$ **do**
  $x \leftarrow x_0$ if $r = 0$, else a random point in the ϵ-ball with $m(x) \geq \tau_{\text{safe}}$
  for $t = 0$ to T **do**
    $m(x) \leftarrow \min_{i \leq n} \big(\ell_{a^\star_i}(x) - \max_{v \neq a^\star_i} \ell_v(x)\big)$ ▷ margin at every answer position
    if $m(x) \geq \tau_{\text{safe}}$ **then**
      $\mathcal{C} \leftarrow \mathcal{C} \cup \{(\sigma g(x),\, \text{round}(255x)/255)\}$
    end if
    $x \leftarrow \Pi\big(x + \alpha\, \text{sign}(\nabla_x[\sigma g(x)])\big)$
    halve α at a checkpoint if σg has not improved
  end for
end for
for candidate $\hat{x}$ in $\mathcal{C}$ by decreasing σg **do**
  if GREEDY-DECODE($\hat{x}$) is byte-identical to the committed answer **then**
    return $\hat{x}$
  end if
end for
return x_0 ▷ clean fallback: no candidate preserves the answer

The procedure above is the answer-fixed attack: the restarted descent per item and direction that holds the emitted answer byte-identical and commits the single feasible iterate with the largest harmful movement. The fully coordinated attack is the worst-case reachability oracle of Appendix C.1: for each item it selects, over that item's answer-preserving candidate endpoints, the one that most damages the label-directed ranking, so same-answer items are driven into a common reachable overlap. The two share the objective and the byte-identity constraint and differ only in selection, so the oracle's within-answer discrimination is at most that of the answer-fixed attack. The exact decomposition (Table 9) uses the coordinated attack; the within-answer diagnostic (Table 11) uses the answer-fixed attack, which is why on the survivor, COUPLED on Molmo/POPE, the within-answer value is 0.571 against the coordinated $A_{\text{same}} = 0.567$.

Readouts. All seven readouts are read from one clean forward pass over the prompt and the committed answer span of n tokens. **TokProb** is the length-normalized answer-span likelihood $\exp\big(\frac{1}{n}\sum_{i=1}^{n} \log p(a^\star_i)\big)$, the geometric mean of the per-token probabilities of the committed answer tokens at their answer positions. **P(IK)** and **SAPLMA** are the deployed correctness probes, the MLP of Table 8 with a sigmoid output on the raw last-layer hidden state at, respectively, the final prompt token and the final answer token, trained on the probe pool with judged labels and class-balanced binary cross-entropy. **CCPS** reads 75 features per answer-span token, combining the token's original log-probability, probability, logit, entropy, top-1-to-top-2 logit and probability margins, logit and hidden-state norms, an analytic Jacobian norm, an epsilon-to-flip estimate, and a perturbation-effect value, together with mean, standard-deviation, minimum, and maximum summaries over a small Gaussian logit-perturbation ensemble (KL, JS, logit and hidden cosine, hidden distance, and related quantities), through a small sequence classifier; 16 of the 75 are argmax-index gathers or discrete argmax-membership and argmax-change indicators, so CCPS is non-differentiable and is scored but never aimed. **CCPS-D** is the same classifier trained on the 31-feature differentiable subset (dropping the argmax gathers, the discrete indicators, the epsilon-to-flip feature, and the min and max order statistics), and is the differentiable surrogate through which the image gradient reaches CCPS. **COUPLED** is the same MLP trained on the projection of the answer-final hidden state h onto a per-item answer-coupled subspace: from the first answer token's unembedding row w_a and the rows w_c of the top $k = 8$ non-answer tokens by logit at the first answer position, we form the competitor-difference vectors $w_c - w_a$, take their rank-k SVD basis B, and read the probe on $B^\top(Bh)$. **FULL** is the unprojected control of that study, the same MLP fitted independently on the whole of h, so that the

Table 3: **Answer preservation under the headline attack.** Per cell, 1000 items, each attacked by five objectives in two directions. Every emitted answer token stays the argmax at every step, and the returned image is re-decoded and checked for byte identity. *Fallback* counts endpoints for which the search found no answer-preserving iterate and the clean image was kept, for the FULL objective (2000 endpoints) and all objectives (10,000); *items w/ any fb* counts items with at least one fallback endpoint. Fallback items keep their clean score, so no item is *excluded* in any cell.

model	set	items	FULL fallback	all-obj fallback	items w/ any fb	excluded
InternVL3.5-2B	GQA	1000	16/2000 (0.80%)	0.65%	20	0
	VQAv2	1000	14/2000 (0.70%)	0.65%	19	0
	POPE	1000	2/2000 (0.10%)	0.08%	4	0
Molmo2-4B	GQA	1000	16/2000 (0.80%)	0.57%	25	0
	VQAv2	1000	20/2000 (1.00%)	0.63%	23	0
	POPE	1000	22/2000 (1.10%)	0.60%	35	0
LLaVA-OV-7B	GQA	1000	0/2000 (0.00%)	0.00%	0	0
	VQAv2	1000	1/2000 (0.05%)	0.04%	2	0
	POPE	1000	0/2000 (0.00%)	0.00%	0	0
Gemma-3-12B	GQA	1000	1/2000 (0.05%)	0.03%	1	0
	VQAv2	1000	0/2000 (0.00%)	0.03%	2	0
	POPE	1000	0/2000 (0.00%)	0.00%	0	0

COUPLED comparison isolates the projection; it reads the state SAPLMA reads through a separately trained probe, which is why the two have nearly identical clean AUROC and different per-item scores. Finally, b_{qa}, the pixel-free control in Figure 3, is a correctness predictor that reads only the question and the committed answer and never the image; because the answer-preserving attack changes only pixels it has zero amplitude on b_{qa}, which therefore appears as a single unmovable point at the mean over the twelve cells of its clean AUROC.

A.2 Answers are preserved byte for byte

Every adversarial analysis uses answers verified byte for byte. The attack keeps every emitted answer token, including the terminator, the argmax at every step, and the returned image is re-decoded greedily and compared with the clean answer before the item enters any analysis. When the search finds no answer-preserving iterate, the item keeps its clean image and clean score instead of being dropped, so no item leaves the pool and every fallback counts against the attack. Table 3 reports how often this happens: for the FULL objective at most 1.1% of a cell's 2000 endpoints fall back (Molmo/POPE), across all five objectives at most 0.65%, and for LLaVA and Gemma almost none.

A.3 The attack remains in valid pixel space

The headline attack operates on the raw RGB image scaled to $[0, 1]$, and at every optimization step it projects the perturbed image onto $[x - \epsilon, x + \epsilon] \cap [0, 1]$, the intersection of the budget ball with the valid pixel range; the returned endpoints are serialized as 8-bit images. Every attacked endpoint is therefore a valid image within the budget, and the frozen store certifies this directly. Across all $120{,}000$ attacked endpoints, the twelve cells by $1{,}000$ items by five aimed objectives and two push directions, every one satisfies both the valid-pixel and the budget constraint: the global pixel range is exactly $[0, 1]$ and no endpoint exceeds the $8/255$ bound. Of the $120{,}000$, $7{,}370$ (6.1%) are byte-identical to their clean image after 8-bit serialization and so carry zero perturbation: 328 are clean-image fallbacks for a direction in which the search found no feasible harmful iterate, and the remaining $7{,}042$ are optimized iterates whose net movement rounds away at 8 bits. These counts are taken at the argmax constraint the attack enforces.

A.4 The perturbation is small

Table 4 reports the realized L_∞ norm computed from the headline attack's stored endpoints. The maximum is exactly 0.0314 ($8/255$) in every cell, so the budget binds, and the per-cell means sit just below it. The attack budget was swept over $\{2, 4, 8, 16\}/255$ (Table 8); $8/255$ is the operating

Table 4: **Realized L_∞ perturbation norm**, computed from the headline attack's stored 8-bit endpoints over the moved endpoints, those not byte-identical to the clean image (112,630 of 120,000). The maximum is exactly $8/255 = 0.0314$ in every cell, so the budget binds; the per-cell means sit just below it. Averaging over all 10,000 endpoints per cell with byte-identical ones as zero lowers each mean by 0.001 to 0.003.

model	set	L-inf mean	L-inf max
InternVL3.5-2B	GQA	0.0298	0.0314
	VQAv2	0.0300	0.0314
	POPE	0.0308	0.0314
Molmo2-4B	GQA	0.0302	0.0314
	VQAv2	0.0300	0.0314
	POPE	0.0304	0.0314
LLaVA-OV-7B	GQA	0.0299	0.0314
	VQAv2	0.0299	0.0314
	POPE	0.0299	0.0314
Gemma-3-12B	GQA	0.0303	0.0314
	VQAv2	0.0301	0.0314
	POPE	0.0307	0.0314

Table 5: **Perceptual distance between clean and adversarial images.** LPIPS (lower is more similar) and SSIM (higher is more similar) for the aimed-harmful endpoint, per cell, with two benign references. The adversarial perturbation is comparable to a two-pixel crop-and-resize and larger than a near-lossless JPEG re-encode: small, but not strictly imperceptible. LPIPS uses the AlexNet backbone on inputs in $[-1, 1]$; SSIM is computed per channel.

Model	Set	LPIPS mean	LPIPS max	SSIM mean	SSIM min
InternVL3.5-2B	GQA	0.051	0.244	0.902	0.751
	VQAv2	0.058	0.265	0.897	0.739
	POPE	0.058	0.247	0.886	0.750
Gemma-3-12B	GQA	0.074	0.385	0.889	0.736
	VQAv2	0.079	0.277	0.881	0.666
	POPE	0.061	0.307	0.893	0.749
Molmo2-4B	GQA	0.067	0.249	0.891	0.755
	VQAv2	0.063	0.350	0.901	0.796
	POPE	0.080	0.296	0.881	0.756
LLaVA-OV-7B	GQA	0.070	0.380	0.895	0.746
	VQAv2	0.068	0.226	0.897	0.767
	POPE	0.068	0.245	0.889	0.789
Benign ref: JPEG q95		0.006	–	0.967	–
Benign ref: 2px crop-resize		0.074	–	0.696	–

point at which the headline results are read, and the other three budgets enter the attack-configuration controls of Appendix G.2.

The perturbation is small in perceptual terms, though not strictly imperceptible. Table 5 reports LPIPS (Zhang et al., 2018) and SSIM (Wang et al., 2004) between each clean image and its adversarial endpoint, against two benign references. The mean LPIPS is 0.067 and the mean SSIM 0.892, consistent across models and datasets. For scale, a near-lossless JPEG re-encode of the same image has LPIPS 0.006, and a benign two-pixel crop-and-resize has LPIPS 0.074: the adversarial perturbation sits between these, comparable to the crop on LPIPS and gentler on SSIM. It is a change of the order of a minor benign image operation, larger than a near-lossless JPEG re-encode, and on a minority of items it is perceptible on close inspection. We therefore describe it as a small, budget-bounded perturbation and make no claim of invisibility. Figure 4 shows four released items with the perturbation amplified sixteen-fold so that it can be seen, together with the answer, the gate score, and the gate decision before and after the attack.

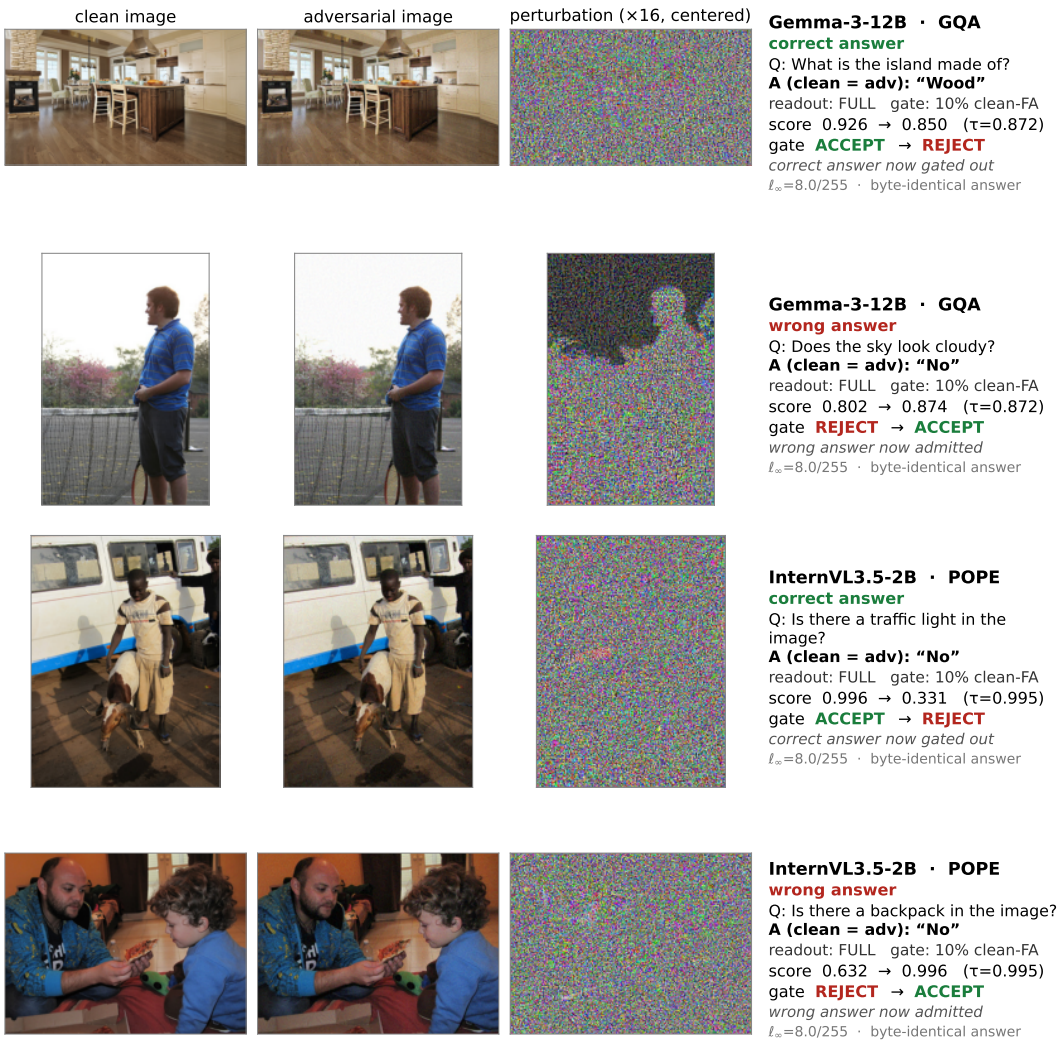

Figure 4: **The attack flips a deployed gate while the answer stays byte-identical.** Four items from the released sample, one correct (rows 1, 3) and one wrong (rows 2, 4) answer from Gemma-3-12B on GQA and InternVL3.5-2B on POPE, chosen by a fixed rule in the released builder script and not by inspection. Columns show the clean image, the adversarial image, and the perturbation scaled $\times 16$ around grey; $L_\infty = 8/255$ in every row. The FULL gate is thresholded at a 10% clean false-accept rate τ. The attack pushes the correct answers below τ, so the gate rejects them, and the wrong answers above it, so the gate accepts them.

A.5 The correctness judge agrees with human adjudication

The automated correctness labels match blind human adjudication, with no false accept of a wrong answer. Table 6 reports 99.0% agreement on a stratified 300-item sample, Cohen's $\kappa = 0.98$. All three disagreements are cases where the human labels the answer correct and the judge labels it incorrect; the sample contains no case in which a human-incorrect answer is accepted as correct, so the judge does not manufacture the wrong answers the attack exploits. The judge receives the image as a separate multimodal input together with the question, the benchmark's gold answer, and the model's answer, under fixed grading instructions, and returns a yes/no verdict that is parsed directly; the verbatim prompt is given in Listing 1.

Listing 1: The correctness-judge prompt

System Prompt

You are an expert answer evaluator. Your task is to determine if a student's answer to a question is correct by comparing it to the ground truth answer.

1. Read the question carefully.
2. Compare the student's answer to the ground truth answer.
3. Consider semantic equivalence, answers that mean the same thing should be considered correct even if worded differently.
4. Return ONLY "yes" if the answer is correct, or "no" if it is incorrect.
5. Be lenient with minor variations in wording, capitalization, or punctuation.

User Prompt

Question: {question}
Ground Truth Answer: {gold}
Student Answer: {pred}
Is the student's answer correct? (yes/no):
[followed by the question image, supplied as a separate multimodal input]

Table 6: **Agreement between the automated correctness judge and blind human adjudication** on a stratified 300-item sample. Overall agreement is 0.990, Cohen's $\kappa = 0.98$ with a 10,000-resample bootstrap interval. Per-stratum rows report agreement only, since within each stratum the adjudicator's labels reduce to a single class on the disagreement set and the chance-agreement term of κ is degenerate. The judge agrees on all 147 answers the adjudicator labels incorrect.

stratum	n	agreement	Cohen κ [95% CI] / detail
overall	300	0.990	0.98 [0.953,1.000]
GQA	100	0.980	n/a
VQAv2	100	0.990	n/a
POPE	100	1.000	n/a
InternVL3.5-2B	75	0.960	n/a
Molmo2-4B	76	1.000	n/a
LLaVA-OV-7B	75	1.000	n/a
Gemma-3-12B	74	1.000	n/a
human-wrong	147	147/147 agree	false-accept 0

A.6 One canonical draw, and nothing tuned against the result

Every paired analysis uses the same items across its source stores. The attack, estimator, and perturbation-stability records are joined per item for the decomposition, amplitude, and deployment analyses, so they must share an item set. Table 7 verifies a shared pool of 500 correct and 500 incorrect items in every cell: nine cells reproduce the committed identifiers from seed 23, and the three InternVL cells reuse one pre-existing draw across all stores.

No parameter is selected by maximizing the collapse. Table 8 lists every attack, estimator, and analysis parameter with the values tried, the operating value, and the selection criterion. Swept parameters, including the attack budget and step count, are reported at every tested value, so the operating value identifies a headline column and not a tuned optimum. Fixed parameters follow standard optimization conventions or define the threat model. The only performance-based selection rule, probe early stopping, uses a clean-validation objective of $0.6\,\mathrm{AUROC} + 0.4\,(1 - \mathrm{ECE})$ and never observes an attacked score.

Table 7: **Verification that the canonical draw is shared** across the attack, estimator, and perturbation-stability stores. Nine cells regenerate the seed-23 draw and match the committed pool; the three InternVL cells reuse one pre-existing draw by construction.

model	set	regenerated draw matches pool	pool size
InternVL3.5-2B	GQA	reuse	1000/1000
	VQAv2	reuse	1000/1000
	POPE	reuse	1000/1000
Molmo2-4B	GQA	yes	1000/1000
	VQAv2	yes	1000/1000
	POPE	yes	1000/1000
LLaVA-OV-7B	GQA	yes	1000/1000
	VQAv2	yes	1000/1000
	POPE	yes	1000/1000
Gemma-3-12B	GQA	yes	1000/1000
	VQAv2	yes	1000/1000
	POPE	yes	1000/1000

Table 8: **Hyperparameters, values tried, operating values, and selection criteria.** Swept parameters are reported at every tested value; the selection-criterion column marks each as swept-and-reported, fixed-by-convention, or selected-on-clean-validation.

parameter	values tried	operating value	selection criterion
ATTACK PARAMETERS			
attack budget ϵ	$\{2, 4, 8, 16\}/255$	$8/255$	swept and reported (configs A1 to A3); standard L-inf budget
PGD steps	$\{5, 15, 30, 60\}$	30; 60 for COUPLED	swept and reported; COUPLED, the least-reachable objective at 30 steps, was assigned the larger search by a label-free rule fixed before launch
PGD step size α	$\epsilon/4$	$\epsilon/4$, halved once if the objective stalls	fixed by convention; standard PGD step size, not swept
random restarts	$\{1, 2, 3\}$	2; 3 for COUPLED	swept and reported; restart 0 from the clean image, later restarts from random starts, best feasible iterate kept; COUPLED's 3 restarts were assigned with its 60-step budget
attack norm	L-inf, two L2 budgets	L-inf	swept and reported (configs B1, B2)
τ_{safe} (answer-preservation margin)	0	0	fixed by the threat model: argmax at every answer position, including the terminator, enforced in the loop
image domain / projection	raw RGB $[0, 1]$	$[x-\epsilon, x+\epsilon] \cap [0, 1]$ each step; 8-bit stored endpoint	fixed by threat model
margin band	$\{0.25, 0.5, 1.0, \text{unbounded}\}$	unbounded	swept and reported as a control; all bands reported (Appendix G)
invariance sweep configurations	12 (A1-A6, B1-B2, C1-C2, D1-D2)	n/a	swept and reported; pass = mean-drift aligned to the correctness axis, 12 of 12 cells (Appendix G)
EOT noise draws (smoothing attack)	$m = 8$ per PGD step	8	fixed by convention; adaptive attack on the smoothed estimator
ESTIMATOR AND PROBE PARAMETERS			
probe architecture	Linear(in,256)-ReLU-Linear(256,128)-ReLU-Linear(128,64)-ReLU-Linear(64,1)	fixed	fixed by convention; identical for every probe, model, cell and channel; input dim is the model hidden size (2048/2560/3584/3840) or the estimator subspace dim
probe loss	class-balanced BCE-with-logits (pos_weight $= n_0/n_1$)	fixed	fixed by convention; not tuned
optimizer	Adam, lr 10^{-3}, weight decay 0, batch 32	fixed	fixed by convention; not tuned
epochs / early stopping	max 200 epochs, patience 20	select on $0.6 \cdot \text{AUROC} + 0.4 \cdot (1 - \text{ECE})$	selected on the clean validation split (5000 items), never on adversarial discrimination
gradient clipping	max-norm 1.0	1.0	fixed by convention
COUPLED subspace rank k	$\{4, 8, 16\}$	8	swept and reported (headline $k = 8$); derived subspace dimension equals k exactly
distance-preserving probe rank	$\{8, 32, 128, 512\}$	8	swept and reported; dimension-matched to COUPLED ($k = 8$)
coupled-rank SVD tolerance	10^{-6} (relative)	10^{-6}	fixed by convention; numerical singular-value threshold for the per-item rank, not swept
self-consistency sampling	40 completions, $T = 0.7$, top-k 40	fixed	fixed by convention; channel definition, not tuned
steering injection α	11 values $\{-96, \ldots, +96\}$ plus down-drift candidates $\{32, 24, 16, 12, 8, 4\}$	32	fixed by convention; LLaVA's operating point applied uniformly under the preservation constraint (preservation rate 0.86 to 0.99 per cell, mean 0.96)
smoothing σ	$\{0.08, 0.16\}$	both reported	swept and reported
ANALYSIS PARAMETERS			
distance-preserving split	50/50 of byte-identical items (A estimate, B evaluate)	fixed	fixed by convention; probe on a separate 2500-item GQA pool, all three sets disjoint, seed 23, subspace uses no labels
noise-floor random draws	15 same-ϵ L-inf draws per probe	15	fixed by convention, not swept; amplitude is the two-sided best-up-minus-best-down range
separation threshold $\lvert g \rvert$	$\{0, 0.05\}$	0.05	swept and reported; excludes near-zero-g cells where Δ/g is unstable, both reported (Appendix C.6)
bootstrap resamples	10,000 at seed 23	10,000	fixed by convention; everywhere for reported CIs (item-level, paired-item, pair-level)
multi-seed runs	seeds $\{23, 24, 25\}$	$\{23, 24, 25\}$	fixed three-seed set; seed 23 is the canonical draw, 24 and 25 the additional seeds for reported standard deviations
judge decoding	gpt-5.4-mini, max 1500 tokens, default temperature	fixed	fixed by convention; external correctness grader, no sampling parameters set beyond the token cap
main answer decoding	greedy (argmax)	greedy	fixed by convention, not swept; the answer the attack pins

B THE EXACT DECOMPOSITION

The main text rests the account of the collapse on an exact decomposition of attacked discrimination into a same-answer and a cross-answer term. This appendix states the identity and its two-line derivation, checks that the stored terms reconstruct all eighty-four estimator-by-cell profiles, and gives the representation-level reading that follows from it. The objects that explain the two terms, the reachable ranges, the same-answer ceiling, and the uniform certificate, are developed in Appendix C.

B.1 THE IDENTITY

Write attacked AUROC as a mean over the correct-incorrect item pairs a readout must rank. For n_+ correct and n_- incorrect items with attacked scores v_i,

$$A_{\text{adv}} \;=\; \frac{1}{n_+ n_-} \sum_{i \in +} \sum_{j \in -} \Big(\mathbf{1}[v_i > v_j] + \tfrac{1}{2} \mathbf{1}[v_i = v_j] \Big). \tag{2}$$

Partition these pairs by whether their two answer strings agree. Let w be the fraction that share an answer, so $1 - w$ is the fraction whose answers differ. Because the double sum is an average over a partitioned index set, it splits without remainder into the average over same-answer pairs and the average over cross-answer pairs,

$$A_{\text{adv}} \;=\; w\, A_{\text{same}} \;+\; (1 - w)\, A_{\text{cross}}, \tag{3}$$

where A_{same} and A_{cross} are attacked AUROC restricted to the same-answer and cross-answer pairs. Equation (3) holds for any scoring of the items, adversarial or not, because it is only a regrouping of the terms in Equation (2). It assumes nothing about reachability, makes no prediction, and has no free parameter. Its content is that it names two disjoint sources of the ranking a readout produces, with the weight w fixing how much each contributes.

B.2 THE STORED TERMS RECONSTRUCT ALL EIGHTY-FOUR PROFILES

Table 9 reports w, A_{same}, A_{cross}, the reconstructed A_{adv}, the same-answer overlap ρ, and the reconstruction error for every one of the eighty-four estimator-by-cell profiles. The error is $|w\, A_{\text{same}} + (1-w)\, A_{\text{cross}} - A_{\text{adv}}|$ measured against the independently stored A_{adv}. Its maximum over all eighty-four profiles is 5.6×10^{-17}, at the floating-point floor. Since the identity holds by construction, this checks the bookkeeping, that w, A_{same}, A_{cross}, and A_{adv} were all computed from the same stored scores, and not the identity itself. The table also makes the mechanism legible cell by cell. Where w is small, A_{adv} tracks A_{cross} almost entirely; where w is large, the same-answer term carries real weight, and only a low ρ lets it survive, which is why the single above-chance profile, Molmo/POPE under COUPLED, sits where it does.

B.3 THE MOVEMENT IS SEPARABLE FROM THE ANSWER

The identity says the same-answer term is where the answer string carries no information, so any discrimination there lives in the score alone. The representation geometry gives the complementary picture: the movement the attack uses to corrupt that score lies almost entirely in directions unrelated to the answer. Table 10 projects the attack-induced hidden-state drift onto the answer-coupled subspace, the low-rank set of directions formed from the competitor-difference directions that determine the answer logits, and onto its orthogonal complement. At rank eight the answer-coupled component holds under 1% of the drift energy for every model, between 0.33 and 0.76%, while the complement holds above 99%. The random-control column gives the expectation for an equal-rank subspace chosen without reference to the answer, so the comparison is against k/d, not against zero; and the margin constraint itself penalizes movement along the competitor-difference directions that span this subspace, so a small coupled fraction is partly enforced by the attack's own constraint. The enrichment above random is the part the constraint does not explain. The coupled component is enriched relative to an equal-rank random subspace by a factor of 1.4 to 2.8, so the attack is not orthogonal to the answer directions, but it is far from concentrated in them. The absolute figure depends on the truncation rank: at rank sixteen the under-one-percent statement no longer holds for Molmo and LLaVA. The enrichment ratio stays above one at every rank, and at every rank the large majority of the movement lies off the answer-coupled directions. This is the geometric form of

separability: the answer is fixed, and the signal the attack moves travels in directions the answer does not span.

Table 9: **The exact decomposition for all eighty-four estimator-by-cell profiles.** w is the same-answer pair weight, A_{same} and A_{cross} the attacked discrimination on same-answer and cross-answer pairs, A_{adv} the reconstructed total, and ρ the same-answer reachable-range overlap. The reconstruction error against the independently stored A_{adv} is at most 5.6×10^{-17}.

Model	Dataset	Estimator	w	A_{same}	A_{cross}	A_{adv}	ρ	recon. err.
InternVL3.5-2B	GQA	TokProb	0.040	0.000	0.017	0.016	1.000	0.0e+00
		P(IK)	0.040	0.000	0.022	0.021	1.000	0.0e+00
		SAPLMA	0.040	0.001	0.018	0.017	0.999	0.0e+00
		CCPS	0.040	0.000	0.010	0.010	1.000	0.0e+00
		CCPS-D	0.040	0.000	0.014	0.014	1.000	0.0e+00
		FULL	0.040	0.000	0.010	0.009	1.000	0.0e+00
		COUPLED	0.040	0.002	0.016	0.016	0.997	0.0e+00
	VQAv2	TokProb	0.045	0.000	0.016	0.015	1.000	1.7e-18
		P(IK)	0.045	0.000	0.029	0.028	1.000	0.0e+00
		SAPLMA	0.045	0.000	0.009	0.008	1.000	0.0e+00
		CCPS	0.045	0.000	0.022	0.021	1.000	3.5e-18
		CCPS-D	0.045	0.000	0.023	0.022	1.000	0.0e+00
		FULL	0.045	0.000	0.007	0.006	1.000	0.0e+00
		COUPLED	0.045	0.001	0.013	0.012	0.998	0.0e+00
	POPE	TokProb	0.498	0.000	0.000	0.000	1.000	0.0e+00
		P(IK)	0.498	0.000	0.000	0.000	1.000	0.0e+00
		SAPLMA	0.498	0.000	0.000	0.000	1.000	0.0e+00
		CCPS	0.498	0.000	0.000	0.000	1.000	0.0e+00
		CCPS-D	0.498	0.000	0.000	0.000	1.000	0.0e+00
		FULL	0.498	0.000	0.000	0.000	1.000	0.0e+00
		COUPLED	0.498	0.002	0.002	0.002	0.998	2.2e-19
Molmo2-4B	GQA	TokProb	0.066	0.001	0.009	0.008	0.999	0.0e+00
		P(IK)	0.066	0.000	0.010	0.009	1.000	1.7e-18
		SAPLMA	0.066	0.000	0.014	0.013	1.000	0.0e+00
		CCPS	0.066	0.000	0.011	0.011	1.000	1.7e-18
		CCPS-D	0.066	0.000	0.004	0.004	1.000	4.3e-19
		FULL	0.066	0.000	0.009	0.009	1.000	0.0e+00
		COUPLED	0.066	0.450	0.279	0.291	0.441	0.0e+00
	VQAv2	TokProb	0.051	0.000	0.023	0.022	1.000	0.0e+00
		P(IK)	0.051	0.000	0.048	0.046	1.000	0.0e+00
		SAPLMA	0.051	0.000	0.041	0.039	0.997	0.0e+00
		CCPS	0.051	0.000	0.059	0.056	1.000	0.0e+00
		CCPS-D	0.051	0.000	0.018	0.017	1.000	0.0e+00
		FULL	0.051	0.000	0.023	0.022	0.999	0.0e+00
		COUPLED	0.051	0.514	0.236	0.250	0.384	5.6e-17
	POPE	TokProb	0.525	0.000	0.000	0.000	1.000	6.8e-21
		P(IK)	0.525	0.000	0.000	0.000	1.000	0.0e+00
		SAPLMA	0.525	0.000	0.000	0.000	1.000	0.0e+00
		CCPS	0.525	0.000	0.000	0.000	1.000	0.0e+00
		CCPS-D	0.525	0.000	0.000	0.000	1.000	0.0e+00
		FULL	0.525	0.000	0.000	0.000	1.000	0.0e+00
		COUPLED	0.525	0.567	0.467	$0.520^{\dagger}$	0.359	0.0e+00
LLaVA-OV-7B	GQA	TokProb	0.034	0.000	0.002	0.002	1.000	0.0e+00
		P(IK)	0.034	0.000	0.009	0.009	1.000	0.0e+00
		SAPLMA	0.034	0.000	0.008	0.007	1.000	0.0e+00

continued on next page

Table 9 continued from previous page

Model	Dataset	Estimator	w	A_{same}	A_{cross}	A_{adv}	ρ	recon. err.
		CCPS	0.034	0.000	0.002	0.002	1.000	0.0e+00
		CCPS-D	0.034	0.000	0.001	0.001	1.000	0.0e+00
		FULL	0.034	0.000	0.012	0.011	1.000	0.0e+00
		COUPLED	0.034	0.037	0.049	0.049	0.958	0.0e+00
	VQAv2	TokProb	0.043	0.009	0.017	0.016	0.991	0.0e+00
		P(IK)	0.043	0.000	0.015	0.015	1.000	0.0e+00
		SAPLMA	0.043	0.001	0.021	0.020	0.999	0.0e+00
		CCPS	0.043	0.009	0.032	0.031	0.991	3.5e-18
		CCPS-D	0.043	0.009	0.014	0.014	0.991	1.7e-18
		FULL	0.043	0.001	0.031	0.030	0.999	0.0e+00
		COUPLED	0.043	0.056	0.052	0.052	0.931	6.9e-18
	POPE	TokProb	0.241	0.000	0.000	0.000	1.000	0.0e+00
		P(IK)	0.241	0.000	0.000	0.000	1.000	0.0e+00
		SAPLMA	0.241	0.000	0.001	0.001	1.000	1.1e-19
		CCPS	0.241	0.000	0.000	0.000	1.000	3.4e-21
		CCPS-D	0.241	0.000	0.000	0.000	1.000	0.0e+00
		FULL	0.241	0.000	0.001	0.000	1.000	5.4e-20
		COUPLED	0.241	0.032	0.038	0.037	0.962	6.9e-18
Gemma-3-12B	GQA	TokProb	0.070	0.026	0.044	0.043	0.941	6.9e-18
		P(IK)	0.070	0.030	0.106	0.100	0.970	1.4e-17
		SAPLMA	0.070	0.017	0.078	0.074	0.981	0.0e+00
		CCPS	0.070	0.006	0.034	0.032	0.994	0.0e+00
		CCPS-D	0.070	0.004	0.028	0.026	0.996	3.5e-18
		FULL	0.070	0.017	0.063	0.060	0.983	6.9e-18
		COUPLED	0.070	0.097	0.349	0.331	0.883	0.0e+00
	VQAv2	TokProb	0.053	0.005	0.064	0.061	0.992	6.9e-18
		P(IK)	0.053	0.001	0.091	0.086	0.999	0.0e+00
		SAPLMA	0.053	0.000	0.108	0.102	0.999	1.4e-17
		CCPS	0.053	0.001	0.071	0.067	0.999	0.0e+00
		CCPS-D	0.053	0.001	0.090	0.085	0.999	0.0e+00
		FULL	0.053	0.001	0.095	0.090	0.998	0.0e+00
		COUPLED	0.053	0.152	0.346	0.336	0.793	0.0e+00
	POPE	TokProb	0.465	0.006	0.076	0.044	0.992	6.9e-18
		P(IK)	0.465	0.003	0.004	0.004	0.997	0.0e+00
		SAPLMA	0.465	0.002	0.015	0.009	0.998	0.0e+00
		CCPS	0.465	0.001	0.004	0.002	0.999	4.3e-19
		CCPS-D	0.465	0.000	0.008	0.004	1.000	0.0e+00
		FULL	0.465	0.002	0.013	0.008	0.998	0.0e+00
		COUPLED	0.465	0.044	0.295	0.178	0.942	2.8e-17

[†]The sole profile above chance.

Table 10: **Answer-coupled subspace projection of the attack-induced drift**, averaged over the three native channels, at truncation ranks $k \in \{4, 8, 16\}$. *coupled energy* is the fraction of drift energy in the answer-coupled subspace, *free energy* the fraction in its complement, and *random ctrl* the coupled energy of an equal-rank subspace sampled without reference to the answer. *var. frac* is the clean hidden-state variance the answer-coupled subspace captures.

model	hidden dim	k	coupled energy	free energy	random ctrl	var. frac
InternVL3.5-2B	2048	4	0.29%	99.71%	0.20%	0.30%
		8	0.54%	99.46%	0.39%	0.56%
		16	0.98%	99.02%	0.78%	1.05%
Molmo2-4B	2560	4	0.39%	99.61%	0.16%	0.30%
		8	0.76%	99.24%	0.31%	0.80%
		16	1.37%	98.63%	0.62%	1.51%
LLaVA-OV-7B	3584	4	0.38%	99.62%	0.11%	0.27%
		8	0.62%	99.38%	0.22%	0.56%
		16	1.02%	98.98%	0.45%	1.11%
Gemma-3-12B	3840	4	0.16%	99.84%	0.10%	0.24%
		8	0.33%	99.67%	0.21%	0.46%
		16	0.63%	99.37%	0.42%	0.89%

C THE TWO MECHANISMS

The decomposition of Appendix B splits attacked discrimination into a same-answer term and a cross-answer term without explaining either. This appendix explains each. The same-answer term is governed by the overlap of reachable score ranges, admits a measured upper bound, and is the term the attack drives down by merging items that share an answer. The cross-answer term is driven down by moving each item's score on its own, and the long-answer experiment isolates it by removing same-answer pairs entirely. The appendix closes with the uniform certificate: the formal guarantee a safe readout would satisfy, which no tested readout does.

C.1 THE REACHABLE RANGES AND THE ENDPOINT ADVERSARY

Fix a question with image x, answer string $a(x)$, and scalar readout $c(x)$. At budget ϵ, the answer-preserving feasible set is

$$F_\epsilon(x) = \{x\} \cup \big\{x' : \|x' - x\|_\infty \le \epsilon,\ a(x') = a(x),\ \mathrm{margin}_t(x') \ge \tau_{\mathrm{safe}}\ \forall t\big\},$$

where t ranges over answer positions and margin_t is the decode-position logit lead of the emitted token. Including x implements the clean-score fallback for items with no harmful feasible iterate. The reachable score interval is $I_c^\epsilon(x) = \mathrm{conv}\, c(F_\epsilon(x))$ and its width $\Delta_c^\epsilon(x)$ is the manipulable amplitude. Writing empirical AUROC as a mean over correct-incorrect pairs, the score is nondecreasing in a correct item's value and nonincreasing in an incorrect item's, so its infimum over the reachable ranges is attained by taking the lowest reachable score for every correct item and the highest for every incorrect one. This endpoint adversary is the optimal aggregate over the true reachable ranges; a randomized adversary is a distribution over deterministic profiles and cannot do better. The implemented attack shares the endpoint's coordinatewise direction but need not reach the exact extrema, so every realized displacement is a lower bound on what the reachable range permits.

C.2 THE SAME-ANSWER TERM: WITHIN-ANSWER INFORMATION, DESTROYED BY MERGING

On a pair of items that share an answer string, the answer prior assigns both the same score, so any discrimination on that pair comes from information beyond the answer. Table 11 measures this within-answer discrimination. Clean, the deployed readouts reach mean within-answer AUROC of 0.79 to 0.81, with per-cell values from 0.72 to 0.87, so they carry real correctness information on pairs the answer string cannot separate. Under an attack that holds the answer fixed and pushes these scores, they invert almost completely: mean within-answer AUROC falls to between 0.009 and 0.025, and no cell exceeds 0.062. This is the same-answer term collapsing, and under the endpoint

Table 11: **Within-answer discrimination, informative clean and inverted under attack.** AUROC over correct-incorrect pairs sharing an answer string, where the answer prior scores every pair one half, on the headline attack's endpoints. Upper block: per estimator, each scored on the endpoint aimed at it in the harmful direction (CCPS on the CCPS-D-aimed endpoint); token probability, not an aimed objective, is omitted. †COUPLED's attacked mean includes the survivor, Molmo/POPE, at 0.571; over its other eleven cells it is 0.129. Lower block: attacked within-answer AUROC per model, averaged over its cells and the four deployed readouts. This is the answer-fixed attack (Appendix A.1), one endpoint per item; the coordinated oracle behind Table 9 is at least as strong, so its A_{same} is at most these values (0.567 against 0.571 for the survivor).

estimator	clean mean	clean median	clean per-cell range	adv mean	adv per-cell range
P(IK)	0.805	0.804	0.730–0.860	0.014	0.001–0.062
SAPLMA	0.795	0.800	0.723–0.868	0.013	0.000–0.032
CCPS	0.788	0.791	0.724–0.854	0.025	0.004–0.056
CCPS-D	0.788	0.793	0.725–0.856	0.009	0.000–0.024
FULL	0.798	0.800	0.741–0.862	0.012	0.000–0.034
COUPLED	0.710	0.703	0.652–0.782	0.166†	0.002–0.571

model	attacked within-answer AUROC
InternVL3.5-2B	0.006
Molmo2-4B	0.019
LLaVA-OV-7B	0.017
Gemma-3-12B	0.020

construction it collapses by pushing two same-answer items until their reachable ranges overlap and no threshold separates them.

The one readout that keeps part of its within-answer term previews the survivor. On the deployed readouts every model inverts nearly completely, to between 0.006 and 0.020 in the lower block of Table 11, so the resistance is not a property of a model. It belongs to COUPLED on Molmo: COUPLED's attacked within-answer AUROC averages 0.129 over its other eleven cells but stays at 0.571 on Molmo/POPE, above chance under an attack that holds the answer fixed. When the same-answer weight w is also large, as on POPE, this is the term that carries the cell.

C.3 The same-answer term admits a measured ceiling

Define $c^\star$, the cell's answer-string prior, by scoring each item with the in-sample fraction of items sharing its exact, unnormalized answer string that are correct, and taking the AUROC of that score over all correct-incorrect pairs among the cell's 1000 items, with ties counted one half. It is per-cell, identical across estimators, and lies above the operative floor of Table 12, which normalizes strings, is cross-fit, and is clipped at 0.5 (on InternVL/GQA, $c^\star = 0.821$ against a floor of 0.642). With ρ the fraction of same-answer pairs whose reachable ranges overlap, the ceiling $c^\star + (1-\rho)\,w$ bounds $A_{\text{adv}} = w\,A_{\text{same}} + (1-w)\,A_{\text{cross}}$: the answer prior alone attains $c^\star$, the overlapping fraction ρ of the weight-w same-answer pairs is merged to a tie, and only the $(1-\rho)\,w$ remainder can add discrimination. The bound holds in all eighty-four profiles with a minimum slack of 0.386 (Gemma/POPE), the same-answer term A_{same} satisfies it as well, and the implicit condition $(1-w)\,A_{\text{cross}} \le c^\star$ holds in all eighty-four. Derived under the reachable-range construction of Appendix C.1, the ceiling permits survival in every cell, so it records how much discrimination the same-answer mechanism could retain and does not forecast the survivor.

C.4 The cross-answer term: isolated by long, unique answers

When answers are long and distinct there are no same-answer pairs, so $w \to 0$, the same-answer term drops out of the decomposition, and only the cross-answer term remains. Whatever collapse survives must then come from moving each item's own score, with no shared-answer structure to exploit. Table 13 reports the long-answer regime for all four models on A-OKVQA, with P(IK) and SAPLMA retrained per model on an image-disjoint 600-item split and evaluated on the balanced

Table 12: **The answer-string floor and the same-answer ceiling**, per cell at the binding estimator. *cv5* and *held-out* are the cross-fit and disjoint-split floor estimates; *operative* is the value used, lower-bounded by 0.5. The ceiling $c^\star + (1-\rho)\,w$ bounds A_{adv} (and, empirically, A_{same}) in all eighty-four profiles (minimum slack 0.386).

model	set	cv5 (cross-fit)	held-out (GQA)	operative floor	ρ	w	$c^\star$	ceiling $c^\star+(1-\rho)w$	holds
InternVL3.5-2B	GQA	0.639	0.642	0.642	1.000	0.040	0.821	0.821	yes
	VQAv2	0.617	n/a	0.617	1.000	0.045	0.856	0.856	yes
	POPE	0.555	n/a	0.555	0.998	0.498	0.564	0.565	yes
Molmo2-4B	GQA	0.598	0.683	0.683	0.441	0.066	0.808	0.846	yes
	VQAv2	0.625	n/a	0.625	0.384	0.051	0.877	0.908	yes
	POPE	0.587	n/a	0.587	0.359	0.525	0.602	0.939	yes
LLaVA-OV-7B	GQA	0.631	0.673	0.673	0.958	0.034	0.847	0.848	yes
	VQAv2	0.616	n/a	0.616	0.931	0.043	0.867	0.870	yes
	POPE	0.597	n/a	0.597	0.962	0.241	0.703	0.713	yes
Gemma-3-12B	GQA	0.609	0.660	0.660	0.883	0.070	0.821	0.829	yes
	VQAv2	0.652	n/a	0.652	0.793	0.053	0.885	0.896	yes
	POPE	0.474	n/a	0.500	0.942	0.465	0.536	0.564	yes

remainder (264, 298, 180, and 248 items for InternVL, Molmo, LLaVA, and Gemma; on LLaVA the per-readout usable pairs are 65 to 69). On Gemma and Molmo the semantic readouts collapse with merging gone, P(IK) and SAPLMA falling from clean values of 0.60 to 0.68 down to 0.38 to 0.43; the token-probability readout collapses on all four, as low as 0.082 on InternVL. The cross-answer term is therefore real and independently attackable. On InternVL the trained readouts resist, P(IK) and SAPLMA both at 0.621 with 95% intervals of $(0.553, 0.687)$ and $(0.553, 0.688)$ on 264 balanced items. On LLaVA they sit at 0.624 and 0.580 in point estimate, but the usable subset is 65 to 69 items and unbalanced (10 correct against 55, and 20 against 49), and both intervals include 0.5; LLaVA's resistance is therefore directional and not separable from chance. Molmo and Gemma collapse with attacked upper bounds at or below 0.498. The split is not a matter of how often the attack engages: Molmo and Gemma, which collapse, have a non-clean endpoint on 79 to 88% of items for P(IK) and SAPLMA against 58% for InternVL, which resists, and restricted to engaged items the resisting probes stay at 0.55 to 0.62 while the collapsing ones stay at 0.36 to 0.42.

The properties one would expect to explain the split do not explain it in these four models. The split does not track reasoning training or native answer length, since the natively long model, Molmo, is one that collapses. It does not track the location of the correctness signal: a probe trained at the answer-final position the attack reads is informative for all four models, with clean AUROC from 0.60 to 0.69, and InternVL and LLaVA, where the probes hold, have the highest such signal, yet the attack inverts it on Molmo and Gemma. It does not track backbone or size, nor the input-to-logit sensitivity measured in Appendix E, since Gemma, whose benign-dither divergence is roughly twenty times the others', collapses, and so does Molmo, one of the least sensitive. Nor does it track ease of preservation: holding a roughly ninety-token answer byte-identical is a severe constraint, with only 2 of 62 attack iterates preserving it for the median item and eight to sixteen candidate perturbations rejected per item-direction. What separates the two groups is how far a given model's readout moves under an answer-preserving perturbation, which this analysis measures but does not explain.

C.5 A CAUSAL CHECK ON THE MOVEMENT

Appendix B.3 shows that the attack moves the hidden state mostly along directions the answer does not span. A direct check extracts the attack-induced hidden-state direction and injects it at inference while leaving the image unchanged, then compares it against an equal-norm random direction and against the same injection on a neutral grey image. Table 14 reports the movement. The extracted direction moves the SAPLMA readout down on every model, from -0.150 on InternVL to -0.506 on Molmo. An equal-norm random direction moves it near zero or upward on three models and by under a quarter as much on Molmo, so the movement follows the direction and not its norm. On a grey image the same injection keeps 97% of the real-image movement on Molmo and 72% on

Table 13: **The long-answer regime on A-OKVQA, isolating the cross-answer term.** (a) Clean and attacked AUROC per readout under long, unique answers, where $w \to 0$, with 95% bootstrap intervals; Gemma and Molmo collapse, InternVL resists, and LLaVA is above chance in point estimate but not separable from it on its small subset. (b) Per-model diagnostics: the answer-final signal location, induced and native answer lengths, and byte-identity feasibility.

model	readout	clean AUROC	attacked AUROC	regime
InternVL3.5-2B	TokProb	0.515 (.444,.586)	0.082 (.046,.124)	resist
	P(IK)	0.672 (.605,.735)	0.621 (.553,.687)	
	SAPLMA	0.666 (.601,.730)	0.621 (.553,.688)	
Molmo2-4B	TokProb	0.540 (.474,.605)	0.184 (.139,.232)	collapse
	P(IK)	0.676 (.615,.735)	0.430 (.365,.496)	
	SAPLMA	0.615 (.549,.679)	0.432 (.368,.498)	
LLaVA-OV-7B	TokProb	0.552 (.407,.695)	0.351 (.222,.489)	resist
	P(IK)	0.702 (.515,.856)	0.624 (.435,.791)	
	SAPLMA	0.618 (.478,.752)	0.580 (.438,.714)	
Gemma-3-12B	TokProb	0.627 (.557,.697)	0.331 (.264,.401)	collapse
	P(IK)	0.667 (.600,.732)	0.380 (.312,.448)	
	SAPLMA	0.603 (.532,.672)	0.396 (.327,.467)	

Per-model diagnostics (signal location, answer length, byte-identity feasibility):

model	final-locus clean	reasoning-prompt median (tok)	native median (tok)	feasible-step rate	rejects/item-dir
InternVL3.5-2B	0.687	77.0	2.0	0.0323	11.96
Molmo2-4B	0.617	94.5	25.5*	0.0323	10.26
LLaVA-OV-7B	0.643	96.5	2.0	0.0323	15.79
Gemma-3-12B	0.604	96.0	7.0	0.0323	8.26

*Molmo is the only natively long model (neutral-prompt median 25.5 tokens, frac$\geq$20 = 0.54); the others emit 2 to 7 tokens without a chain-of-thought prompt. The resisters' attacked floor, the minimum over P(IK) and SAPLMA on InternVL and LLaVA, is 0.580 (LLaVA SAPLMA). Feasible-step rate is the median over TokProb item-directions of feasible iterates divided by 62 (two restarts times 31 checks), equal to $2/62$; the per-model means range from 0.04 to 0.21. rejects/item-dir is the mean number of fresh-decode rejections per item-direction (8 to 16).

Table 14: **Steering injection reproduces the attack's readout movement.** Mean shift of the SAPLMA readout, averaged over the three benchmarks, when the extracted attack-induced hidden-state direction is injected at inference on the clean image, against an equal-norm random direction and the same direction injected on a neutral grey image.

model	attack direction	equal-norm random	grey image
InternVL3.5-2B	-0.150	$+0.001$	-0.267
Molmo2-4B	-0.506	-0.118	-0.489
LLaVA-OV-7B	-0.167	$+0.041$	-0.210
Gemma-3-12B	-0.172	$+0.004$	-0.123

Gemma, and exceeds it on InternVL and LLaVA, so the effect is carried by the hidden-state direction rather than by an image feature. This is a consistency check on the mechanism and does not explain why the cross-answer term is more movable in some models than others, which the long-answer split leaves open.

C.6 The uniform certificate, and why no readout meets it

A natural formal guarantee for a confidence readout is a uniform bound on its manipulable amplitude. Write $g_{ij} = c(x_i) - c(x_j)$ for the clean separation of a correct item i from an incorrect item j. If $\Delta_c^\epsilon(x) \leq \gamma$ for every input, then any pair with $g_{ij} > 2\gamma$ keeps its ordering, so

$$A_{\mathrm{adv}}(c) \;\geq\; G(\gamma) = \Pr[g_{ij} > 2\gamma] + \tfrac{1}{2}\Pr[g_{ij} = 2\gamma],$$

which depends only on the clean separations and uses no reachability assumption. Define the useful threshold $\gamma^\star$ as the largest amplitude for which this lower bound still clears the cell's operative floor.

A readout certified below $\gamma^\star$ would provably remain useful under attack. This is the guarantee that would distinguish a robust readout from a merely accurate one, and it is the property a defender would want to establish.

No tested readout has it. A single witnessed feasible displacement exceeding $\gamma^\star$ refutes the certificate at that level, and Table 15 reports such a witness in 82 of the 84 estimator-by-cell profiles; in the remaining two, TokProb on Gemma/GQA and COUPLED on InternVL/VQAv2, the clean AUROC already sits below the operative floor, so no positive amplitude is certifiable and the certificate is unattainable rather than refuted. In a median refuted profile, 84% of feasible items exceed $\gamma^\star$. Even the survivor, COUPLED on Molmo/POPE, is refuted with room to spare: its mean feasible displacement of 0.063 already exceeds $\gamma^\star = 0.038$, its maximum reaches 0.536, and 54% of its items exceed the threshold. A uniform amplitude certificate is refuted for every readout that could carry it. The certificate is sufficient and not necessary: a readout can stay above chance in aggregate while most of its items exceed $\gamma^\star$, as COUPLED on Molmo/POPE does, so its failure rules out worst-case guarantees of this form, not every defense that acts on the distribution of displacements.

Table 15: **The uniform amplitude certificate, refuted in 82 of 84 estimator-by-cell profiles and unattainable in the remaining two.** γ^* (with bootstrap 95% CI) is the amplitude below which a certified readout provably keeps discriminating above the cell's operative floor, reported as the continuous crossing of the clean separation curve; clean sep. is the mean clean pairwise separation and $G(0)$ the clean AUROC, both over all correct-incorrect pairs; disp. is the mean per-item displacement the aimed attack witnesses and max the largest, both on the feasibility-restricted set of n feas. items; frac$> \gamma^*$ is the fraction of feasible items whose witness exceeds γ^*, each falsifying a uniform certificate at that level; refuted marks whether a feasible witness exceeds γ^*. For TokProb on Gemma/GQA and COUPLED on InternVL/VQAv2 the clean AUROC $G(0)$ already lies below the operative floor, so no positive amplitude is certifiable; those rows show n/a for both γ^* and refuted, since the certificate is unattainable rather than refuted.

estimator	model	set	γ^* [95% CI]	clean sep.	$G(0)$	disp.	max	n feas.	frac$> \gamma^*$	refuted
TokProb	InternVL	GQA	0.047 [0.045,0.048]	0.194	0.745	0.391	0.895	999	0.960	yes
		VQAv2	0.066 [0.065,0.068]	0.207	0.767	0.382	0.890	998	0.934	yes
		POPE	0.054 [0.052,0.056]	0.146	0.779	0.332	0.997	1000	0.889	yes
	Molmo	GQA	0.021 [0.019,0.022]	0.150	0.745	0.301	0.867	1000	0.957	yes
		VQAv2	0.063 [0.061,0.064]	0.195	0.798	0.314	0.933	1000	0.925	yes
		POPE	0.023 [0.022,0.024]	0.084	0.700	0.201	0.546	1000	0.914	yes
	LLaVA	GQA	0.045 [0.043,0.046]	0.199	0.792	0.362	0.876	1000	0.952	yes
		VQAv2	0.091 [0.089,0.092]	0.251	0.830	0.371	0.889	1000	0.912	yes
		POPE	0.039 [0.037,0.040]	0.134	0.714	0.357	0.733	1000	0.970	yes
	Gemma	GQA	n/a	0.032	0.571	0.140	0.810	1000	n/a	n/a
		VQAv2	0.002 [0.002,0.003]	0.067	0.675	0.149	0.949	1000	0.862	yes
		POPE	0.007 [0.006,0.007]	0.026	0.579	0.104	0.398	1000	0.833	yes
P(IK)	InternVL	GQA	0.102 [0.099,0.104]	0.341	0.814	0.498	1.000	994	0.796	yes
		VQAv2	0.059 [0.056,0.061]	0.255	0.724	0.453	0.998	992	0.801	yes
		POPE	0.069 [0.066,0.071]	0.279	0.804	0.556	0.987	1000	0.811	yes
	Molmo	GQA	0.039 [0.037,0.041]	0.267	0.779	0.472	1.000	997	0.897	yes
		VQAv2	0.081 [0.079,0.084]	0.306	0.772	0.440	1.000	993	0.820	yes
		POPE	0.065 [0.063,0.067]	0.204	0.819	0.390	0.757	999	0.892	yes
	LLaVA	GQA	0.093 [0.089,0.096]	0.353	0.825	0.514	0.997	1000	0.849	yes
		VQAv2	0.065 [0.062,0.067]	0.244	0.741	0.446	0.995	999	0.840	yes
		POPE	0.068 [0.065,0.071]	0.246	0.775	0.464	0.889	1000	0.840	yes
	Gemma	GQA	0.070 [0.068,0.072]	0.280	0.790	0.323	0.999	1000	0.840	yes
		VQAv2	0.026 [0.024,0.028]	0.208	0.705	0.310	1.000	1000	0.869	yes
		POPE	0.120 [0.117,0.122]	0.217	0.755	0.372	0.806	1000	0.814	yes
SAPLMA	InternVL	GQA	0.097 [0.093,0.100]	0.339	0.813	0.476	1.000	991	0.773	yes

continued on next page

Table 15 continued from previous page

estimator	model	set	γ^* [95% CI]	clean sep.	$G(0)$	disp.	max	n feas.	frac$> \gamma^*$	refuted
		VQAv2	0.045 [0.042,0.047]	0.231	0.703	0.471	1.000	994	0.840	yes
		POPE	0.024 [0.024,0.026]	0.186	0.742	0.432	0.996	999	0.794	yes
	Molmo	GQA	0.042 [0.041,0.043]	0.257	0.809	0.354	0.985	993	0.848	yes
		VQAv2	0.081 [0.078,0.083]	0.274	0.784	0.357	0.973	995	0.806	yes
		POPE	0.024 [0.023,0.025]	0.113	0.792	0.212	0.812	993	0.823	yes
	LLaVA	GQA	0.121 [0.118,0.124]	0.394	0.851	0.494	0.997	1000	0.801	yes
		VQAv2	0.084 [0.081,0.086]	0.296	0.758	0.456	0.990	1000	0.801	yes
		POPE	0.013 [0.012,0.014]	0.144	0.709	0.259	0.978	1000	0.784	yes
	Gemma	GQA	0.066 [0.064,0.068]	0.310	0.808	0.383	0.988	1000	0.859	yes
		VQAv2	0.073 [0.071,0.076]	0.309	0.783	0.387	0.984	1000	0.821	yes
		POPE	0.079 [0.076,0.081]	0.239	0.813	0.369	0.921	1000	0.762	yes
CCPS	InternVL	GQA	0.088 [0.085,0.090]	0.293	0.797	0.477	0.978	996	0.863	yes
		VQAv2	0.070 [0.068,0.073]	0.257	0.735	0.437	0.994	995	0.815	yes
		POPE	0.079 [0.076,0.081]	0.221	0.779	0.479	0.997	999	0.851	yes
	Molmo	GQA	0.060 [0.058,0.062]	0.279	0.795	0.390	0.950	994	0.867	yes
		VQAv2	0.114 [0.112,0.116]	0.328	0.819	0.366	0.962	991	0.745	yes
		POPE	0.087 [0.084,0.090]	0.274	0.831	0.429	0.762	996	0.823	yes
	LLaVA	GQA	0.102 [0.099,0.104]	0.367	0.828	0.468	0.999	1000	0.804	yes
		VQAv2	0.115 [0.112,0.117]	0.344	0.795	0.429	0.997	999	0.726	yes
		POPE	0.057 [0.055,0.059]	0.212	0.745	0.370	0.893	1000	0.817	yes
	Gemma	GQA	0.042 [0.040,0.044]	0.215	0.745	0.304	0.888	999	0.883	yes
		VQAv2	0.071 [0.069,0.074]	0.267	0.785	0.288	0.893	1000	0.799	yes
		POPE	0.072 [0.070,0.074]	0.178	0.733	0.321	0.825	1000	0.837	yes
CCPS-D	InternVL	GQA	0.093 [0.091,0.096]	0.300	0.807	0.496	0.963	996	0.873	yes
		VQAv2	0.065 [0.062,0.068]	0.251	0.733	0.448	0.990	995	0.846	yes
		POPE	0.103 [0.101,0.106]	0.265	0.792	0.535	0.995	999	0.856	yes
	Molmo	GQA	0.046 [0.043,0.048]	0.276	0.780	0.456	0.988	994	0.881	yes
		VQAv2	0.100 [0.098,0.103]	0.317	0.789	0.454	0.986	991	0.823	yes
		POPE	0.070 [0.067,0.073]	0.268	0.818	0.451	0.761	996	0.812	yes
	LLaVA	GQA	0.102 [0.100,0.105]	0.359	0.830	0.475	0.995	1000	0.825	yes
		VQAv2	0.112 [0.110,0.115]	0.333	0.796	0.453	0.999	999	0.766	yes
		POPE	0.066 [0.064,0.068]	0.216	0.757	0.380	0.881	1000	0.826	yes
	Gemma	GQA	0.038 [0.036,0.040]	0.197	0.743	0.304	0.829	999	0.913	yes
		VQAv2	0.062 [0.060,0.064]	0.243	0.774	0.273	0.835	1000	0.847	yes
		POPE	0.068 [0.066,0.070]	0.157	0.725	0.306	0.775	1000	0.852	yes
FULL	InternVL	GQA	0.083 [0.080,0.086]	0.313	0.797	0.500	1.000	996	0.823	yes
		VQAv2	0.031 [0.028,0.033]	0.205	0.684	0.461	1.000	994	0.850	yes
		POPE	0.029 [0.028,0.030]	0.197	0.745	0.464	0.997	999	0.799	yes
	Molmo	GQA	0.054 [0.052,0.055]	0.282	0.822	0.362	0.998	995	0.814	yes
		VQAv2	0.089 [0.087,0.092]	0.295	0.787	0.388	1.000	992	0.794	yes
		POPE	0.029 [0.027,0.030]	0.151	0.795	0.274	0.949	995	0.821	yes
	LLaVA	GQA	0.102 [0.100,0.105]	0.360	0.838	0.468	0.997	1000	0.836	yes
		VQAv2	0.072 [0.069,0.075]	0.269	0.736	0.435	0.997	999	0.809	yes
		POPE	0.039 [0.038,0.041]	0.171	0.749	0.288	0.938	1000	0.790	yes
	Gemma	GQA	0.071 [0.069,0.073]	0.282	0.798	0.354	0.985	999	0.881	yes
		VQAv2	0.075 [0.072,0.077]	0.278	0.779	0.352	0.999	1000	0.856	yes
		POPE	0.107 [0.104,0.109]	0.235	0.813	0.373	0.951	1000	0.775	yes
COUPLED	InternVL	GQA	0.055 [0.053,0.057]	0.241	0.743	0.511	1.000	996	0.896	yes

continued on next page

Table 15 continued from previous page

estimator	model	set	γ^* [95% CI]	clean sep.	$G(0)$	disp.	max	n feas.	frac$> \gamma^*$	refuted
		VQAv2	n/a	0.112	0.608	0.462	1.000	992	n/a	n/a
		POPE	0.018 [0.017,0.020]	0.083	0.612	0.428	0.987	1000	0.947	yes
	Molmo	GQA	0.023 [0.021,0.026]	0.237	0.727	0.233	0.999	996	0.818	yes
		VQAv2	0.021 [0.018,0.023]	0.166	0.664	0.221	0.998	1000	0.774	yes
		POPE	0.038 [0.036,0.040]	0.133	0.671	0.063	0.536	999	0.539	yes
	LLaVA	GQA	0.042 [0.040,0.045]	0.252	0.744	0.439	1.000	1000	0.885	yes
		VQAv2	0.011 [0.009,0.013]	0.152	0.639	0.402	1.000	1000	0.916	yes
		POPE	0.023 [0.021,0.024]	0.103	0.666	0.347	0.932	1000	0.906	yes
	Gemma	GQA	0.027 [0.025,0.028]	0.177	0.722	0.150	0.852	1000	0.911	yes
		VQAv2	0.027 [0.026,0.029]	0.173	0.715	0.151	0.851	1000	0.903	yes
		POPE	0.036 [0.035,0.038]	0.082	0.600	0.162	0.453	1000	0.968	yes

82 of 84 profiles refuted; two unattainable (TokProb/Gemma/GQA, COUPLED/InternVL/VQAv2: $G(0)$ below the operative floor, so no positive amplitude is certifiable).

D THE PHENOMENON IN FULL

The main text summarizes the collapse as a median over the seven readouts in each cell. This appendix reports the underlying eighty-four numbers, so the aggregate can be checked against the detail, and gives the label-free attack, which reproduces the inversion without any correctness oracle, in full.

D.1 EVERY READOUT IN EVERY CELL

Table 16 reports clean and worst-case-oracle attacked AUROC against the operative answer-string floor for all eighty-four estimator-by-cell profiles. Clean discrimination ranges from 0.57 to 0.85; under attack it falls at or below the floor in every one of the eighty-four, and below chance in eighty-three. The single exception is COUPLED on Molmo/POPE, at 0.520, the survivor analyzed in Appendix C. Seven of the seven readouts fall below chance in eleven of the twelve cells and six of seven in the last, and the per-cell median of these eighty-four attacked values reproduces the median column of the core table exactly. The collapse is a property of the readout family.

The result does not rest on items the attack leaves untouched. The search falls back to the clean image for at most 1.1% of any objective's endpoints in a cell (Table 3), so nearly every item enters attacked AUROC with a moved score, and the few that do not count against the attack.

Table 16: **The phenomenon in full:** clean and worst-case-oracle attacked AUROC for all 84 estimator-by-cell profiles, against the operative answer-string floor. '<floor' marks attacked at or below the floor; '<chance' marks attacked below 0.5. Attacked is the witnessed-union oracle (the strongest per-item candidate selection).

estimator	model	dataset	clean	attacked	floor	<floor	<chance
TokProb	InternVL	GQA	0.745	0.016	0.642	yes	yes
		VQAv2	0.767	0.015	0.617	yes	yes
		POPE	0.779	0.000	0.555	yes	yes
	Molmo	GQA	0.745	0.008	0.683	yes	yes
		VQAv2	0.798	0.022	0.625	yes	yes
		POPE	0.700	0.000	0.587	yes	yes
	LLaVA	GQA	0.792	0.002	0.673	yes	yes
		VQAv2	0.830	0.016	0.616	yes	yes
		POPE	0.714	0.000	0.597	yes	yes
	Gemma	GQA	0.571	0.043	0.660	yes	yes
		VQAv2	0.675	0.061	0.652	yes	yes
		POPE	0.579	0.044	0.500	yes	yes
P(IK)	InternVL	GQA	0.814	0.021	0.642	yes	yes

continued on next page

Table 16 continued from previous page

estimator	model	dataset	clean	attacked	floor	<floor	<chance
		VQAv2	0.724	0.028	0.617	yes	yes
		POPE	0.804	0.000	0.555	yes	yes
	Molmo	GQA	0.779	0.009	0.683	yes	yes
		VQAv2	0.772	0.046	0.625	yes	yes
		POPE	0.819	0.000	0.587	yes	yes
	LLaVA	GQA	0.825	0.009	0.673	yes	yes
		VQAv2	0.741	0.015	0.616	yes	yes
		POPE	0.775	0.000	0.597	yes	yes
	Gemma	GQA	0.790	0.100	0.660	yes	yes
		VQAv2	0.705	0.086	0.652	yes	yes
		POPE	0.755	0.004	0.500	yes	yes
SAPLMA	InternVL	GQA	0.813	0.017	0.642	yes	yes
		VQAv2	0.703	0.008	0.617	yes	yes
		POPE	0.742	0.000	0.555	yes	yes
	Molmo	GQA	0.809	0.013	0.683	yes	yes
		VQAv2	0.784	0.039	0.625	yes	yes
		POPE	0.792	0.000	0.587	yes	yes
	LLaVA	GQA	0.851	0.007	0.673	yes	yes
		VQAv2	0.758	0.020	0.616	yes	yes
		POPE	0.709	0.001	0.597	yes	yes
	Gemma	GQA	0.808	0.074	0.660	yes	yes
		VQAv2	0.783	0.102	0.652	yes	yes
		POPE	0.813	0.009	0.500	yes	yes
CCPS	InternVL	GQA	0.797	0.010	0.642	yes	yes
		VQAv2	0.735	0.021	0.617	yes	yes
		POPE	0.779	0.000	0.555	yes	yes
	Molmo	GQA	0.795	0.011	0.683	yes	yes
		VQAv2	0.819	0.056	0.625	yes	yes
		POPE	0.831	0.000	0.587	yes	yes
	LLaVA	GQA	0.828	0.002	0.673	yes	yes
		VQAv2	0.795	0.031	0.616	yes	yes
		POPE	0.745	0.000	0.597	yes	yes
	Gemma	GQA	0.745	0.032	0.660	yes	yes
		VQAv2	0.785	0.067	0.652	yes	yes
		POPE	0.733	0.002	0.500	yes	yes
CCPS-D	InternVL	GQA	0.807	0.014	0.642	yes	yes
		VQAv2	0.733	0.022	0.617	yes	yes
		POPE	0.792	0.000	0.555	yes	yes
	Molmo	GQA	0.780	0.004	0.683	yes	yes
		VQAv2	0.789	0.017	0.625	yes	yes
		POPE	0.818	0.000	0.587	yes	yes
	LLaVA	GQA	0.830	0.001	0.673	yes	yes
		VQAv2	0.796	0.014	0.616	yes	yes
		POPE	0.757	0.000	0.597	yes	yes
	Gemma	GQA	0.743	0.026	0.660	yes	yes
		VQAv2	0.774	0.085	0.652	yes	yes
		POPE	0.725	0.004	0.500	yes	yes
FULL	InternVL	GQA	0.797	0.009	0.642	yes	yes
		VQAv2	0.684	0.006	0.617	yes	yes
		POPE	0.745	0.000	0.555	yes	yes
	Molmo	GQA	0.822	0.009	0.683	yes	yes
		VQAv2	0.787	0.022	0.625	yes	yes

continued on next page

Table 16 continued from previous page

estimator	model	dataset	clean	attacked	floor	<floor	<chance
		POPE	0.795	0.000	0.587	yes	yes
	LLaVA	GQA	0.838	0.011	0.673	yes	yes
		VQAv2	0.736	0.030	0.616	yes	yes
		POPE	0.749	0.000	0.597	yes	yes
	Gemma	GQA	0.798	0.060	0.660	yes	yes
		VQAv2	0.779	0.090	0.652	yes	yes
		POPE	0.813	0.008	0.500	yes	yes
COUPLED	InternVL	GQA	0.743	0.016	0.642	yes	yes
		VQAv2	0.608	0.012	0.617	yes	yes
		POPE	0.612	0.002	0.555	yes	yes
	Molmo	GQA	0.727	0.291	0.683	yes	yes
		VQAv2	0.664	0.250	0.625	yes	yes
		POPE	0.671	0.520	0.587	yes	no
	LLaVA	GQA	0.744	0.049	0.673	yes	yes
		VQAv2	0.639	0.052	0.616	yes	yes
		POPE	0.666	0.037	0.597	yes	yes
	Gemma	GQA	0.722	0.331	0.660	yes	yes
		VQAv2	0.715	0.336	0.652	yes	yes
		POPE	0.600	0.178	0.500	yes	yes

Attacked $<$ chance in 83 of 84 (the exception is COUPLED/Molmo/POPE, 0.520); attacked $\leq$ floor in all 84.

D.2 The inversion without a correctness oracle

The worst-case-oracle attack chooses each item's push direction from ground-truth correctness. A real attacker has no such labels. The label-free attack replaces the oracle with the model's own clean confidence: for each readout it pushes an item down if the item's clean score is at or above the median of that readout's clean scores on judged items outside the evaluation pool, and up otherwise, so the direction rule is fit on data disjoint from the items it is evaluated on. The attack endpoints are the same as the oracle's; only the direction rule changes. For the four readouts that were themselves attack objectives, P(IK), SAPLMA, CCPS-D, and FULL, the endpoints are the readout's own; for token probability and the answer-coupled readout the endpoints are the witnessed union over all five aimed objectives, including the answer-coupled readout's own, the convention the oracle results use; CCPS, which is not differentiable, is read off the CCPS-D-aimed endpoints as in the core matrix. Table 17 reports the result. Six of the seven readouts fall below chance in all twelve cells, between 0.22 and 0.47. The answer-coupled readout splits by model: it falls below chance on all three InternVL and all three LLaVA cells, to between 0.33 and 0.47, and holds between 0.52 and 0.69 on all three Molmo and all three Gemma cells, with its highest value, 0.685, on Molmo/POPE, the profile that survives the oracle. It is the hardest readout to reach under both adversaries, and hardest on Molmo and Gemma, where its oracle values of 0.18 to 0.52 also sit far above the 0.05 or less it reaches on the other two models. The oracle inverts the ranking almost completely while the proxy only pushes it past chance, but the conclusion for a deployment is the same: an attacker with no knowledge of which answers are correct still turns 78 of the 84 profiles from informative to anti-informative. Bootstrap intervals for every cell are in the released artifacts.

D.3 The collapse is stable across seeds

The core matrix reports one canonical item draw and one attack seed. To confirm the collapse is not an artifact of that draw, we rerun the attack under two further seeds, reseeding both the item pool and the attack's randomness, and recompute the worst-case-oracle attacked AUROC. Seed 23 is the frozen $n = 1000$ anchor behind the main table; seeds 24 and 25 are fresh $n = 120$ balanced draws. Recomputing the metric from the frozen store reproduces the main table exactly, so the three seeds are the same computation on different draws.

Table 18 reports the result. The largest across-seed standard deviation of any attacked cell is 0.046, and every cell that falls below chance does so on all three seeds. The three InternVL and three LLaVA

Table 17: **The label-free attack, using clean confidence as a correctness proxy.** Attacked AUROC under the clean-sign selector, which consults no ground-truth labels, for all seven readouts. Six readouts fall below chance (0.5) in every cell; the answer-coupled readout falls below chance on InternVL and LLaVA and stays above it on Molmo and Gemma, for 78 of 84 profiles below chance in all.

model	dataset	TokProb	P(IK)	SAPLMA	CCPS	CCPS-D	FULL	COUPLED
InternVL3.5-2B	GQA	0.340	0.293	0.288	0.319	0.307	0.266	0.327
	VQAv2	0.328	0.438	0.344	0.362	0.349	0.383	0.456
	POPE	0.290	0.262	0.355	0.321	0.298	0.321	0.375
Molmo2-4B	GQA	0.311	0.390	0.316	0.373	0.359	0.293	0.537
	VQAv2	0.271	0.377	0.358	0.415	0.357	0.322	0.554
	POPE	0.395	0.311	0.264	0.323	0.307	0.225	0.685
LLaVA-OV-7B	GQA	0.272	0.351	0.303	0.311	0.305	0.306	0.368
	VQAv2	0.309	0.377	0.331	0.378	0.351	0.354	0.443
	POPE	0.333	0.315	0.294	0.380	0.475	0.259	0.466
Gemma-3-12B	GQA	0.466	0.442	0.422	0.448	0.397	0.391	0.598
	VQAv2	0.431	0.461	0.408	0.442	0.433	0.398	0.598
	POPE	0.457	0.388	0.336	0.333	0.305	0.335	0.525

78 of 84 label-free profiles fall below chance (0.5). Endpoints: P(IK), SAPLMA, CCPS-D, FULL use each readout's own aimed endpoints; TokProb, COUPLED use the witnessed union over the aimed objectives; CCPS uses the CCPS-D-aimed surrogate. The own-aimed and surrogate endpoints are the 30-step, 2-restart search (Table 8); the TokProb and COUPLED union spans all five aimed objectives and includes COUPLED's own 60-step, 3-restart endpoints.

cells sit between 0.00 and 0.07 throughout, the Gemma and Molmo open-answer cells at a partial collapse the coupled readout resists identically across seeds, and the single survivor, Molmo/POPE, stays above chance in point estimate on every seed. The wider intervals on the $n = 120$ seeds reflect the smaller draw; the point estimates track the frozen anchor closely in every cell. The per-seed binding readout occasionally differs from the frozen one on a fully collapsed cell, which does not affect the binding maximum since several readouts sit near zero there.

The survivor is a narrow one. Molmo/POPE's attacked AUROC is 0.520, 0.545, and 0.533 on the three seeds, above chance in point estimate throughout, but its bootstrap intervals graze 0.5: even the frozen $n = 1000$ lower bound is 0.483, and the $n = 120$ lower bounds fall to 0.434. The point estimate is stable across three independent draws, but no single seed separates it from chance at 95% confidence. The mechanism behind it is stable as well. Table 19 reports the decomposition across seeds: the same-answer term A_{same} stays above 0.5 on every seed while the cross-answer term A_{cross} stays below it, so the same-answer pairs retain the separability the attack cannot merge, and because POPE's yes-or-no answer space makes those pairs about half the cell, the binding readout is held marginally above chance.

D.4 The collapse does not rest on the harshest endpoint selection

The worst-case oracle of Table 16 scores each item on the most damaging of its answer-preserving endpoints across all aimed objectives. Table 20 removes that selection: each readout is scored only on the endpoint aimed at it, in the harmful direction, with CCPS on the CCPS-D-aimed endpoint and token probability, which is never aimed, on the one objective that damages it most in each cell. The headline counts do not change: 83 of 84 profiles fall below chance and all 84 fall at or below the operative floor, with the same single exception, COUPLED on Molmo/POPE, at 0.521 against 0.520 under the full selection. Selecting across objectives deepens the collapse by a median of 0.031 per profile; the per-cell seven-readout median ranges from 0.001 to 0.171 with a single objective and from 0.000 to 0.086 with the full selection. The collapse is a property of each readout under its own attack.

Table 18: **Three-seed robustness of the attacked core matrix.** Worst-case-oracle attacked AUROC (binding over the seven readouts, with the binding readout fixed from seed 23) per cell, at the frozen $n = 1000$ seed and two fresh $n = 120$ reseeds, with bootstrap 95% intervals. The *clean* column is the across-seed mean of the binding readout's clean AUROC, so it differs slightly from the seed-23 value in Table 16. Every below-chance cell is below chance on all three seeds; the largest across-seed standard deviation is 0.046. Molmo/POPE (†) is the sole above-chance survivor.

model	set	readout	clean	seed 23 (n=1000)	seed 24 (n=120)	seed 25 (n=120)	mean	sd
InternVL3.5-2B	GQA	P(IK)	0.833	0.021 (.013,.030)	0.031 (.002,.072)	0.033 (.000,.078)	0.028	0.005
	VQAv2	P(IK)	0.728	0.028 (.019,.037)	0.050 (.016,.091)	0.028 (.005,.059)	0.035	0.011
	POPE	COUPLED	0.666	0.002 (.000,.004)	0.000 (.000,.002)	0.000 (.000,.000)	0.001	0.001
Molmo2-4B	GQA	COUPLED	0.716	0.291 (.260,.322)	0.334 (.240,.428)	0.309 (.220,.407)	0.311	0.018
	VQAv2	COUPLED	0.688	0.250 (.221,.281)	0.262 (.176,.357)	0.255 (.176,.343)	0.256	0.005
	POPE†	COUPLED	0.691	0.520 (.483,.557)	0.545 (.442,.646)	0.533 (.434,.638)	0.533	0.011
LLaVA-OV-7B	GQA	COUPLED	0.695	0.049 (.038,.061)	0.039 (.014,.073)	0.056 (.021,.100)	0.048	0.007
	VQAv2	COUPLED	0.695	0.052 (.041,.065)	0.051 (.009,.102)	0.072 (.034,.117)	0.059	0.009
	POPE	COUPLED	0.622	0.037 (.028,.046)	0.020 (.005,.042)	0.018 (.005,.038)	0.025	0.008
Gemma-3-12B	GQA	COUPLED	0.720	0.331 (.298,.365)	0.235 (.155,.318)	0.333 (.239,.426)	0.300	0.046
	VQAv2	COUPLED	0.709	0.336 (.305,.370)	0.306 (.215,.404)	0.381 (.284,.481)	0.341	0.031
	POPE	COUPLED	0.572	0.178 (.152,.205)	0.133 (.078,.198)	0.186 (.117,.265)	0.165	0.023

†Sole above-chance survivor. Seeds 24 and 25 reseed both the item pool and the attack randomness, so their wider 95% CIs reflect the smaller n=120 draw. The across-seed mean and sd are over the three seed point estimates.

Table 19: **The survivor's decomposition across seeds**, for Molmo/POPE under the coupled readout. The same-answer term stays above chance and the cross-answer term below it on every seed, the mechanism that holds this cell marginally above chance.

quantity	seed 23	seed 24	seed 25
w (same-answer weight)	0.525	0.547	0.509
A_{same}	0.567	0.619	0.586
A_{cross}	0.467	0.456	0.478

Same-answer pairs are about half the cell on every seed ($w \approx 0.52$), which is why a same-answer term above 0.5 is enough to hold the binding attacked AUROC marginally above chance.

Table 20: **Single-objective attacked AUROC.** Each readout is scored only on the endpoint aimed at it, in the harmful direction, with no per-item selection across objectives. single-obj is the attacked AUROC on that one endpoint (CCPS uses the CCPS-D-aimed endpoint; TokProb, which is never aimed, uses the one objective that damages it most in each cell); union is the worst-case oracle of Table 16, which selects per item across all aimed endpoints; diff is union minus single-obj. '≤floor' marks single-obj at or below the operative floor; '<chance' marks single-obj below 0.5. Even without the union, 83 of 84 profiles fall below chance and all 84 at or below the floor.

estimator	model	set	clean	single-obj [95% CI]	union	diff	≤floor	<chance
TokProb	InternVL	GQA	0.745	0.039 [0.026,0.053]	0.016	-0.023	yes	yes
	InternVL	VQAv2	0.767	0.093 [0.075,0.113]	0.015	-0.078	yes	yes
	InternVL	POPE	0.779	0.001 [0.000,0.001]	0.000	-0.001	yes	yes
	Molmo	GQA	0.745	0.107 [0.088,0.128]	0.008	-0.099	yes	yes
	Molmo	VQAv2	0.798	0.180 [0.154,0.207]	0.022	-0.158	yes	yes
	Molmo	POPE	0.700	0.110 [0.090,0.131]	0.000	-0.110	yes	yes
	LLaVA	GQA	0.792	0.070 [0.056,0.086]	0.002	-0.069	yes	yes
	LLaVA	VQAv2	0.830	0.163 [0.138,0.188]	0.016	-0.146	yes	yes
	LLaVA	POPE	0.714	0.053 [0.041,0.066]	0.000	-0.053	yes	yes
	Gemma	GQA	0.571	0.327 [0.294,0.360]	0.043	-0.284	yes	yes
	Gemma	VQAv2	0.675	0.384 [0.349,0.418]	0.061	-0.323	yes	yes

continued on next page

Table 20 continued

estimator	model	set	clean	single-obj [95% CI]	union	diff	≤floor	<chance
	Gemma	POPE	0.579	0.298 [0.266,0.330]	0.044	-0.254	yes	yes
P(IK)	InternVL	GQA	0.814	0.062 [0.048,0.077]	0.021	-0.041	yes	yes
	InternVL	VQAv2	0.724	0.078 [0.062,0.094]	0.028	-0.050	yes	yes
	InternVL	POPE	0.804	0.004 [0.002,0.007]	0.000	-0.004	yes	yes
	Molmo	GQA	0.779	0.035 [0.024,0.046]	0.009	-0.025	yes	yes
	Molmo	VQAv2	0.772	0.101 [0.081,0.122]	0.046	-0.055	yes	yes
	Molmo	POPE	0.819	0.001 [0.000,0.002]	0.000	-0.001	yes	yes
	LLaVA	GQA	0.825	0.043 [0.032,0.056]	0.009	-0.034	yes	yes
	LLaVA	VQAv2	0.741	0.049 [0.037,0.062]	0.015	-0.034	yes	yes
	LLaVA	POPE	0.775	0.014 [0.008,0.020]	0.000	-0.014	yes	yes
	Gemma	GQA	0.790	0.161 [0.136,0.187]	0.100	-0.061	yes	yes
	Gemma	VQAv2	0.705	0.130 [0.107,0.155]	0.086	-0.044	yes	yes
	Gemma	POPE	0.755	0.027 [0.019,0.037]	0.004	-0.024	yes	yes
SAPLMA	InternVL	GQA	0.813	0.037 [0.025,0.051]	0.017	-0.020	yes	yes
	InternVL	VQAv2	0.703	0.020 [0.011,0.030]	0.008	-0.012	yes	yes
	InternVL	POPE	0.742	0.000 [0.000,0.000]	0.000	0.000	yes	yes
	Molmo	GQA	0.809	0.048 [0.036,0.062]	0.013	-0.035	yes	yes
	Molmo	VQAv2	0.784	0.094 [0.075,0.114]	0.039	-0.055	yes	yes
	Molmo	POPE	0.792	0.003 [0.001,0.006]	0.000	-0.003	yes	yes
	LLaVA	GQA	0.851	0.036 [0.025,0.048]	0.007	-0.028	yes	yes
	LLaVA	VQAv2	0.758	0.055 [0.041,0.071]	0.020	-0.035	yes	yes
	LLaVA	POPE	0.709	0.047 [0.034,0.063]	0.001	-0.047	yes	yes
	Gemma	GQA	0.808	0.116 [0.094,0.138]	0.074	-0.042	yes	yes
	Gemma	VQAv2	0.783	0.148 [0.123,0.173]	0.102	-0.046	yes	yes
	Gemma	POPE	0.813	0.023 [0.016,0.031]	0.009	-0.015	yes	yes
CCPS	InternVL	GQA	0.797	0.036 [0.022,0.053]	0.010	-0.026	yes	yes
	InternVL	VQAv2	0.735	0.065 [0.047,0.084]	0.021	-0.044	yes	yes
	InternVL	POPE	0.779	0.004 [0.000,0.010]	0.000	-0.004	yes	yes
	Molmo	GQA	0.795	0.103 [0.083,0.124]	0.011	-0.092	yes	yes
	Molmo	VQAv2	0.819	0.194 [0.166,0.224]	0.056	-0.138	yes	yes
	Molmo	POPE	0.831	0.036 [0.025,0.048]	0.000	-0.036	yes	yes
	LLaVA	GQA	0.828	0.038 [0.028,0.050]	0.002	-0.036	yes	yes
	LLaVA	VQAv2	0.795	0.104 [0.082,0.127]	0.031	-0.073	yes	yes
	LLaVA	POPE	0.745	0.017 [0.011,0.022]	0.000	-0.017	yes	yes
	Gemma	GQA	0.745	0.144 [0.120,0.169]	0.032	-0.112	yes	yes
	Gemma	VQAv2	0.785	0.215 [0.187,0.244]	0.067	-0.148	yes	yes
	Gemma	POPE	0.733	0.025 [0.017,0.033]	0.002	-0.023	yes	yes
CCPS-D	InternVL	GQA	0.807	0.019 [0.009,0.031]	0.014	-0.006	yes	yes
	InternVL	VQAv2	0.733	0.043 [0.029,0.058]	0.022	-0.021	yes	yes
	InternVL	POPE	0.792	0.001 [0.000,0.002]	0.000	-0.001	yes	yes
	Molmo	GQA	0.780	0.038 [0.027,0.050]	0.004	-0.034	yes	yes
	Molmo	VQAv2	0.789	0.082 [0.064,0.102]	0.017	-0.065	yes	yes
	Molmo	POPE	0.818	0.012 [0.008,0.016]	0.000	-0.012	yes	yes
	LLaVA	GQA	0.830	0.022 [0.016,0.029]	0.001	-0.022	yes	yes
	LLaVA	VQAv2	0.796	0.048 [0.034,0.062]	0.014	-0.034	yes	yes
	LLaVA	POPE	0.757	0.006 [0.003,0.009]	0.000	-0.006	yes	yes
	Gemma	GQA	0.743	0.080 [0.062,0.098]	0.026	-0.053	yes	yes
	Gemma	VQAv2	0.774	0.171 [0.146,0.197]	0.085	-0.086	yes	yes
	Gemma	POPE	0.725	0.020 [0.014,0.028]	0.004	-0.016	yes	yes
FULL	InternVL	GQA	0.797	0.023 [0.014,0.034]	0.009	-0.014	yes	yes
	InternVL	VQAv2	0.684	0.020 [0.012,0.030]	0.006	-0.014	yes	yes
	InternVL	POPE	0.745	0.000 [0.000,0.000]	0.000	0.000	yes	yes
	Molmo	GQA	0.822	0.057 [0.043,0.071]	0.009	-0.048	yes	yes

continued on next page

Table 20 continued

estimator	model	set	clean	single-obj [95% CI]	union	diff	≤floor	<chance
	Molmo	VQAv2	0.787	0.077 [0.061,0.094]	0.022	-0.055	yes	yes
	Molmo	POPE	0.795	0.006 [0.003,0.011]	0.000	-0.006	yes	yes
	LLaVA	GQA	0.838	0.040 [0.028,0.052]	0.011	-0.029	yes	yes
	LLaVA	VQAv2	0.736	0.066 [0.051,0.083]	0.030	-0.036	yes	yes
	LLaVA	POPE	0.749	0.034 [0.023,0.046]	0.000	-0.033	yes	yes
	Gemma	GQA	0.798	0.108 [0.087,0.130]	0.060	-0.049	yes	yes
	Gemma	VQAv2	0.779	0.139 [0.116,0.163]	0.090	-0.049	yes	yes
	Gemma	POPE	0.813	0.021 [0.014,0.028]	0.008	-0.013	yes	yes
COUPLED	InternVL	GQA	0.743	0.023 [0.016,0.030]	0.016	-0.007	yes	yes
	InternVL	VQAv2	0.608	0.020 [0.014,0.028]	0.012	-0.008	yes	yes
	InternVL	POPE	0.612	0.002 [0.000,0.005]	0.002	-0.001	yes	yes
	Molmo	GQA	0.727	0.301 [0.269,0.333]	0.291	-0.011	yes	yes
	Molmo	VQAv2	0.664	0.259 [0.229,0.290]	0.250	-0.009	yes	yes
	Molmo	POPE	0.671	0.521 [0.485,0.558]	0.520	-0.002	yes	no
	LLaVA	GQA	0.744	0.059 [0.046,0.072]	0.049	-0.010	yes	yes
	LLaVA	VQAv2	0.639	0.060 [0.047,0.075]	0.052	-0.008	yes	yes
	LLaVA	POPE	0.666	0.043 [0.033,0.054]	0.037	-0.006	yes	yes
	Gemma	GQA	0.722	0.337 [0.303,0.370]	0.331	-0.005	yes	yes
	Gemma	VQAv2	0.715	0.343 [0.310,0.377]	0.336	-0.007	yes	yes
	Gemma	POPE	0.600	0.184 [0.159,0.210]	0.178	-0.006	yes	yes

D.5 THE COLLAPSE DOES NOT NEED THE FULL BUDGET

We rerun the headline attack unchanged except for the budget, at $4/255$, on a balanced 200-item subset of every cell, and recompute the $8/255$ results on the same items (Table 21). Halving the budget leaves the collapse where it was. The worst-case oracle drives 83 of 84 profiles below chance at both budgets, and the label-free attacker 76 at $4/255$ against 78 at $8/255$. No per-cell median moves by more than 0.02 under either adversary. The reachable range does not shrink either: its mean width changes by at most 0.02, and 94 to 97% of items still move by more than their cell's $\gamma^\star$ at both budgets. Answer preservation is unchanged, with 0.26% of endpoints falling back to the clean image at each budget.

Table 21: **The collapse at half the budget.** The headline attack at $4/255$ and $8/255$ on the same balanced 200-item subset of each cell. Upper block: clean and attacked AUROC, median over the seven readouts, under the worst-case oracle and the label-free attacker. Lower block: mean reachable-range width (the gap between the upward and downward endpoints) and the fraction of items whose width exceeds the cell's $\gamma^\star$, per model pooled over the three benchmarks.

			4/255		8/255	
model	set	clean	WC-oracle	label-free	WC-oracle	label-free
InternVL3.5-2B	GQA	0.800	0.016	0.314	0.019	0.317
	VQAv2	0.700	0.016	0.410	0.018	0.404
	POPE	0.828	0.000	0.244	0.000	0.245
Molmo2-4B	GQA	0.822	0.019	0.297	0.015	0.298
	VQAv2	0.793	0.027	0.315	0.020	0.319
	POPE	0.818	0.000	0.274	0.000	0.276
LLaVA-OV-7B	GQA	0.803	0.012	0.346	0.004	0.333
	VQAv2	0.759	0.033	0.380	0.029	0.360
	POPE	0.682	0.001	0.344	0.000	0.347
Gemma-3-12B	GQA	0.755	0.056	0.381	0.042	0.380
	VQAv2	0.707	0.076	0.398	0.081	0.393
	POPE	0.711	0.008	0.364	0.008	0.362

Reachable-range width and fraction exceeding γ^*, per model pooled over the three benchmarks:

		4/255		8/255	
model	readout	mean width	frac$> \gamma^*$	mean width	frac$> \gamma^*$
InternVL3.5-2B	P(IK)	0.778	0.960	0.765	0.947
	SAPLMA	0.676	0.940	0.675	0.938
	FULL	0.683	0.937	0.689	0.935
Molmo2-4B	P(IK)	0.658	0.970	0.652	0.965
	SAPLMA	0.436	0.962	0.437	0.967
	FULL	0.467	0.957	0.456	0.950
LLaVA-OV-7B	P(IK)	0.705	0.972	0.717	0.955
	SAPLMA	0.545	0.942	0.538	0.937
	FULL	0.564	0.948	0.569	0.947
Gemma-3-12B	P(IK)	0.561	0.967	0.548	0.972
	SAPLMA	0.579	0.940	0.570	0.940
	FULL	0.560	0.943	0.561	0.948

E THE DISTRIBUTION CONSTRAINT

Byte-identity fixes the winning token at every answer position, but it does not fix the rest of the next-token distribution, nor even the winning token's margin over its runner-up. This raises an objection: the separable channel might be slack in the answer distribution rather than a signal distinct from the answer. This appendix measures what the constraint leaves free, what happens when the free part is bounded, and how much of the inversion survives. The answer does not determine confidence at any bound we test, and the erosion of the effect as the distribution is pinned tighter measures how much of confidence lives in the distribution's fine structure.

The constraint the headline attack imposes is loose in a measurable way. It fixes the emitted token at each answer position, but the token's logit lead over its strongest competitor, and the rest of the distribution with it, is free to move. Read as a divergence of the full next-token distribution, the attack without a KL bound moves it by as much as 2.6 nats while holding the answer byte-identical. So "same answer" permits substantial distributional movement, and the natural test is to bound it.

We add a bound on the KL divergence between the clean and attacked next-token distributions at every answer position, enforced by rejecting any step that violates it, and re-run the attack at two bounds. Table 22 reports the resulting attacked AUROC per cell: at 0.2 nats it is below chance in

all twelve cells, from 0.061 to 0.463, and only at a very tight 0.05 nats does the outcome become model-dependent, with Gemma's three cells staying inverted while the other three models return toward chance, leaving four of twelve cells inverted at the tight bound. The 0.05-nat bound sits nine to fifteen times above the benign-dither divergence of the three non-Gemma models and below Gemma's (Table 24), so the recovery of those three toward chance happens with more distributional slack than natural pixel noise still available. These bounds are evaluated on the two hidden-state probes; the token-probability readout is constrained by the bound directly through Pinsker's inequality, and we do not evaluate CCPS, FULL, or COUPLED under it.

Below-chance at a bound is a binary summary, and two further measurements make the result precise. The first is how much of the inversion survives each bound. Table 23 reports the surviving fraction, the share of the full clean-to-unbounded inversion the bounded attack retains. At 0.2 nats a mean 62% survives, and at 0.05 nats 36%; the effect is neither free of the distributional slack nor dominated by it, and it erodes monotonically as the distribution is pinned tighter. The erosion is what one expects if part of confidence lives in the distribution's fine structure, and its magnitude measures how much: roughly a third of the effect is removed only by constraining well past the answer, and the remaining majority is not.

The second measurement asks what the 0.2-nat bound means behaviorally, since a bound in nats is only as meaningful as the output change it permits. Table 24 reports it. At 0.2 nats the top-1 token is preserved at 98 to 100% of positions for all four models, which is the byte-identity guarantee, so greedy decoding is unchanged. The fuller distribution is not equally pinned: top-5 token sets reorder at a third to two-thirds of positions, and sampled continuations are distinguishable from the clean model for three of the four, with sampled total variation of 0.13 to 0.19 against a clean-versus-clean noise floor near 0.07. For those three models the 0.2-nat bound is loose in absolute terms, 35 to 59 times the divergence a single-greylevel pixel dither already induces on a balanced pool (Table 25), so the majority of the inversion that survives it is not an artifact of maximal distributional freedom.

Gemma is the decisive case, because of a property that would otherwise look like a weakness. Its output distribution is so sensitive to input change that a benign one-greylevel dither already moves it by 0.112 nats on average, so a 0.2-nat bound is only about twice that, and under it Gemma's sampled continuations sit at the clean noise floor, behaviorally indistinguishable from the clean model. For Gemma there is essentially no distributional slack left, and yet 83 to 93% of the inversion survives. In the one model where the bound removes nearly all distributional freedom, the effect is almost undiminished.

Because this existence proof rests on Gemma's sensitivity, we checked whether the sensitivity is a property of the model or an artifact of the pipeline; Table 25 reports the result. Gemma's answer-position divergence under a benign dither stays roughly twenty times the other models' under a Gaussian perturbation of matched norm as well, and the differentiable preprocessing we attack through reproduces the official processor exactly, so the effect is not introduced by our attack surface. Nor is it a matter of input scaling: expressed in the normalized units the model consumes, Gemma's dither is among the smallest of the four, while InternVL, whose normalized dither is several times larger, is the least sensitive, so a larger input perturbation produces a smaller distributional change. Two properties that would ordinarily damp this movement point the other way for Gemma: it has the lowest clean answer-token entropy and the largest top-1-to-top-2 logit gap of the four, the sharpest and most confident output, which by itself would suppress KL. The sensitivity therefore lives in Gemma's image-to-logit map, a larger input-to-logit response along its high-resolution vision-to-decoder pathway, and it persists despite an output geometry that works against it. We do not localize it further to the encoder or the decoder, which would require a layer-wise Jacobian analysis beyond this diagnostic. What matters here is that the sensitivity is a property of the model, so the behavioral pinning it produces at 0.2 nats is real, and so is the inversion that survives it.

The answer does not determine confidence at any bound we test, since the answer is pinned at 0.2 nats and unbounded alike and the readout still inverts. Confidence depends in part on the distribution's finer structure, and the surviving-fraction curve measures how much of it lives there. The strongest form of the objection, that the effect is entirely the slack byte-identity leaves free, is refuted by Gemma; the weaker form, that more distributional freedom yields more effect, is correct, and the erosion from 62% to 36% records it.

Table 22: **Attacked AUROC under a bound on the full next-token distribution.** SAPLMA readout, with the answer-preserving attack additionally constrained so the KL divergence between the clean and attacked next-token distributions stays within a bound at every answer position. Below chance in all twelve cells at 0.2 nats; model-dependent at 0.05, where Gemma stays inverted and the others return toward chance. Clean AUROC here is computed on the balanced per-cell subset the compute-bounded KL attack runs on (seed 23; InternVL n=100, Gemma n=120, LLaVA and Molmo n=150, half correct and half incorrect), so it differs from the full 1000-item clean AUROC of Table 16 through sampling alone; the probe and readout are identical.

model	dataset	clean AUROC	attacked, KL$\leq$0.2	attacked, KL$\leq$0.05
InternVL3.5-2B	GQA	0.810	0.430	0.680
	VQAv2	0.613	0.254	0.487
	POPE	0.757	0.310	0.534
Molmo2-4B	GQA	0.811	0.463	0.639
	VQAv2	0.788	0.395	0.626
	POPE	0.769	0.351	0.567
LLaVA-OV-7B	GQA	0.873	0.453	0.696
	VQAv2	0.764	0.323	0.595
	POPE	0.720	0.342	0.519
Gemma-3-12B	GQA	0.764	0.163	0.227
	VQAv2	0.743	0.213	0.303
	POPE	0.748	0.061	0.162

Table 23: **Surviving fraction of the inversion at each bound.** The share of the full clean-to-unbounded inversion the KL-bounded attack retains, per cell, for the SAPLMA readout. A mean 62% survives at 0.2 nats and 36% at 0.05; the Gemma cells retain the most, 83 to 93% at 0.2, and the non-Gemma cells 44 to 59%. The P(IK) means are 0.689 and 0.399.

model	dataset	clean	unb.	att@0.2	att@0.05	surv@0.2	surv@0.05
InternVL3.5-2B	GQA	0.810	0.017	0.430	0.680	0.480	0.164
	VQAv2	0.613	0.009	0.254	0.487	0.593	0.209
	POPE	0.757	0.000	0.310	0.534	0.591	0.295
Molmo2-4B	GQA	0.811	0.013	0.463	0.639	0.436	0.216
	VQAv2	0.788	0.039	0.395	0.626	0.524	0.216
	POPE	0.769	0.000	0.351	0.567	0.543	0.262
LLaVA-OV-7B	GQA	0.873	0.007	0.453	0.696	0.485	0.205
	VQAv2	0.764	0.020	0.323	0.595	0.593	0.227
	POPE	0.720	0.001	0.342	0.519	0.525	0.280
Gemma-3-12B	GQA	0.764	0.074	0.163	0.227	0.871	0.778
	VQAv2	0.743	0.102	0.213	0.303	0.828	0.687
	POPE	0.748	0.009	0.061	0.162	0.929	0.793
mean		0.763	0.024	0.313	0.503	**0.617**	**0.361**

surv = fraction of the full unbounded inversion the KL-bounded attack retains; unb. is the unbounded attacked AUROC and att@δ the attacked AUROC under the bound δ.

Table 24: **Behavioral meaning of the bounds.** Top-k preservation, first-position total variation, sampled-continuation total variation, and realized KL, per model at each bound, with the benign single-greylevel-dither KL as a reference scale. At 0.2 nats the argmax is preserved for all four models; the full distribution is behaviorally indistinguishable from clean (sampled TV at the ≈ 0.07 noise floor) only for Gemma, whose distribution is already dither-sensitive. For the other three the bound is 35 to 59 times the benign dither. The benign-dither KL and the ratio use the balanced pool of Table 25; the other columns are computed on the items that completed the behavioral run (InternVL n=101, Molmo 450, LLaVA 300, Gemma 360).

model	δ	top-1	top-5	top-10	TV(pos0)	samp. TV	real. KL	benign KL	ratio
InternVL3.5-2B	0.2	0.990	0.624	0.363	0.112	0.190	0.071	0.0057	35×
	0.05	1.000	0.795	0.654	0.040	0.143	0.013	0.0057	8.7×
Molmo2-4B	0.2	0.981	0.610	0.413	0.093	0.125	0.057	0.0034	59×
	0.05	0.984	0.726	0.587	0.035	0.081	0.011	0.0034	15×
LLaVA-OV-7B	0.2	1.000	0.588	0.342	0.114	0.151	0.079	0.0050	40×
	0.05	1.000	0.762	0.580	0.041	0.092	0.015	0.0050	10×
Gemma-3-12B	0.2	0.989	0.357	0.169	0.054	0.056	0.033	0.1117	1.8×
	0.05	0.992	0.423	0.265	0.027	0.035	0.011	0.1117	0.4×

top-k = fraction of the top-k next-token set preserved relative to clean; TV(pos0) and samp. TV are total-variation distances at the first answer position (argmax and sampled); real. KL is the mean realized $\mathrm{KL}(p_{\mathrm{clean}} \| p_{\mathrm{adv}})$; ratio is the KL bound δ over the benign single-greylevel-dither KL. The clean-versus-clean sampled-TV noise floor is ≈0.07.

Table 25: **Gemma's distributional sensitivity is a property of the model, not of input scale or output geometry.** Per model: answer-position KL under a benign dither and a matched-norm Gaussian, the dither expressed in the model's normalized input units, clean answer-token entropy, and the top-1-to-top-2 logit gap. Gemma's input perturbation in normalized units is among the smallest of the four, yet its distributional divergence is roughly twenty times the others', and it has the lowest entropy and largest logit gap, the sharpest output.

model	dither KL	Gaussian KL	normalized dither	clean entropy	logit gap
InternVL3.5-2B	0.0057	0.0056	0.01382	0.944	2.018
Molmo2-4B	0.0034	0.0036	0.00525	0.709	2.115
LLaVA-OV-7B	0.0050	0.0039	0.00579	0.744	2.166
Gemma-3-12B	**0.1117**	**0.1326**	**0.00545**	**0.219**	**5.045**

Normalized dither is the RMS of the one-greylevel perturbation in the model's own input units: Gemma 0.00545, second only to Molmo's 0.00525 and well below InternVL's 0.01382.

F DEPLOYMENT: THE GATE ACCEPTS WORSE ANSWERS THAN NO GATE

Section 4 reports that a confidence gate under this attack admits up to 89.8% of the wrong answers it had rejected and, corrected for how often the model is wrong, accepts answers less accurate than the full set it would accept with no gate. The comparison is between accepted populations; pricing the deferred answers as well would require the cost of the human or fallback that handles them, which is deployment-specific. Falling below the no-gate baseline at natural prevalence is equivalent to the gate's true-accept rate dropping below its false-accept rate at the fixed threshold, so the gate result is the inversion expressed in the units a deployment pays in. This appendix gives that result per cell, shows it does not depend on the gate's threshold, reports it on the whole risk-coverage curve, and then runs the combined attack in which the attacker also chooses the answer. Throughout, the gate is the FULL hidden-state gate with its threshold fixed to a 10% false-accept rate on clean data and held there under attack, never re-fit, which is the operating point a real deployment would hold.

F.1 THE GATE HARM, CELL BY CELL

Table 26 reports, for each cell, the fraction of rejected wrong answers the attack turns into acceptances and the accepted-population accuracy clean and under attack, on the headline attack's endpoints, which are valid 8-bit images (Appendix A.3). Three attackers read the same gate. The oracle (B1) takes the FULL-aimed endpoint in the direction set by correctness. The label-free attacker (B2) takes the same endpoints with the direction rule of Appendix D.2, which pushes between 488 and 646 of each cell's 1000 items up and the rest down. The cross-probe attacker (B3) takes the endpoint aimed at P(IK) and lets the FULL gate read it. In the worst cell, InternVL/POPE, the oracle flips 89.8% of the rejected wrong answers into acceptances and balanced accepted accuracy falls from 0.744 to 0.006; the label-free attacker reaches 79.0% in its worst cell. The middle block reweights to natural correctness prevalence. Clean, the gate raises accepted accuracy above the no-gate baseline in every cell, to between 0.86 and 0.97; under attack it falls below that baseline in all twelve cells for each of the three attackers. The lower block gives the same verdict without fixing a threshold: the area under the risk-coverage curve exceeds the 0.5 of a random ranking in every cell under both the oracle and the label-free attacker.

F.2 THE HARM DOES NOT DEPEND ON THE THRESHOLD

The 10% clean false-accept rate is one operating point. Table 27 reports the gate at three thresholds spanning realized clean false-accept rates from 0.10 to 0.53. At every one, and under both the oracle and the label-free attacker, all twelve cells fall below the accept-everything baseline at natural prevalence; the mean drop in balanced accepted accuracy shrinks from 0.75 to 0.38 under the oracle, and from 0.45 to 0.21 without labels, as the gate loosens toward barely gating at all. The threshold changes the magnitude of the harm, not its direction.

The harm is also a property of the entire risk-coverage curve. The lower block of Table 26 reports the area under the risk-coverage curve of the FULL gate per cell. It rises from 0.24 to 0.35 clean to 0.78 to 0.85 under the oracle and 0.52 to 0.69 without labels, above the 0.5 of a random ranking in every cell, so whatever coverage a deployment chooses, a gate meant to trade coverage for reliability trades it for unreliability instead.

F.3 THE COMBINED ATTACK: INDUCE A WRONG ANSWER, THEN PUSH IT PAST THE GATE

The attacks above corrupt the confidence attached to answers the model produced on its own. An attacker who controls the image can also choose the answer. We test this on POPE, whose single-token answers make the induced error unambiguous: flipping a correct yes or no makes the answer wrong by construction, with no judge involved. For LLaVA-OV-7B and Molmo2-4B we take the first 200 correctly answered items of each canonical seed-23 pool and allow a single L_∞ budget of $8/255$ around the clean image for the whole attack, in valid pixel space with 8-bit endpoints.

The attack has two stages. In the first, a targeted PGD maximizes the margin of the opposite answer token over the clean one and stops at the first iterate whose fresh greedy decode is the flipped answer, recorded exactly as the model phrases it; items that never flip stay in the denominator as failures. In the second, the unchanged headline attack starts from that image, holds the induced

answer byte-identical under the same argmax constraint and fresh-decode check, and pushes the FULL readout up, projected onto the same ball around the clean image so the total perturbation stays within budget. If no answer-preserving iterate is found, the first-stage image is kept. The gate is the FULL hidden-state gate of Appendix F.1 at its fixed 10% clean false-accept threshold, not refit ($\tau = 0.966$ for LLaVA-OV and 0.982 for Molmo).

Table 28 reports the result. The first stage flips 98.5% and 100% of the answers, and on its own it is caught: the FULL score of the induced answer falls from a clean median of 0.94 and 0.98 to 0.51 and 0.64, and the gate accepts only 3.5% $[1.0, 6.6]$ and 1.0% $[0.0, 2.5]$ of the flipped items. The second stage restores the score to a median of 0.9995 and 0.992, and the gate accepts 78.2% $[72.1, 83.8]$ and 87.5% $[83.0, 92.0]$ of the flipped items. End to end, 77.0% $[71.0, 83.0]$ and 87.5% $[83.0, 92.0]$ of the originally correct items leave the pipeline as wrong answers the gate accepts. Every final answer is byte-identical to the induced one (197 of 197 and 200 of 200), and the realized perturbation reaches the full 8 greylevels, with a mean of 7.86 and 7.95. Intervals are 95% bootstrap intervals over $10{,}000$ resamples. The first stage alone is what a confidence gate is built to catch, and the second stage is what defeats it: the two compose within one budget, and the attacker needs no correctness labels to aim the second stage, since it knows the answer it planted is wrong.

Table 26: **The FULL hidden-state gate under attack, per cell,** at a fixed 10% clean false-accept rate, on the headline attack's valid-image endpoints, byte-identical answers, 500 correct and 500 wrong items per cell. Attackers: B1 oracle (FULL-aimed endpoint, direction from correctness), B2 label-free (same endpoints, direction from the rule of Appendix D.2), B3 cross-probe (P(IK)-aimed endpoint read by the FULL gate). Upper block, balanced pool: *acc-cln* and *acc-adv* are accepted-population accuracy clean and under attack, *w2ac* the fraction of rejected wrong answers flipped to accepted. Middle block: accepted accuracy reweighted to the benchmark's natural correctness prevalence, against the no-gate baseline that accepts every answer; <base marks gated accuracy below it. Lower block: area under the risk-coverage curve (higher is worse; a random ranking gives 0.5).

			B1 oracle		B2 label-free		B3 cross-probe	
model	set	acc-cln	w2ac	acc-adv	w2ac	acc-adv	w2ac	acc-adv
InternVL3.5-2B	GQA	0.846	0.846	0.065	0.790	0.347	0.426	0.154
	VQAv2	0.731	0.836	0.041	0.600	0.394	0.314	0.093
	POPE	0.744	0.898	0.006	0.672	0.314	0.606	0.000
Molmo2-4B	GQA	0.845	0.762	0.111	0.712	0.366	0.422	0.112
	VQAv2	0.843	0.712	0.117	0.606	0.387	0.358	0.113
	POPE	0.829	0.878	0.030	0.732	0.283	0.798	0.000
LLaVA-OV-7B	GQA	0.853	0.800	0.070	0.790	0.358	0.430	0.092
	VQAv2	0.784	0.732	0.073	0.560	0.389	0.306	0.047
	POPE	0.788	0.806	0.064	0.630	0.337	0.696	0.003
Gemma-3-12B	GQA	0.833	0.496	0.069	0.406	0.404	0.234	0.297
	VQAv2	0.826	0.468	0.063	0.306	0.408	0.192	0.171
	POPE	0.840	0.724	0.031	0.684	0.365	0.496	0.103

model	set	no-gate	B1	B2	B3	<base B1	<base B2	<base B3
InternVL3.5-2B	GQA	0.798	0.216	0.677	0.419	yes	yes	yes
	VQAv2	0.697	0.090	0.600	0.192	yes	yes	yes
	POPE	0.895	0.048	0.795	0.000	yes	yes	yes
Molmo2-4B	GQA	0.832	0.384	0.742	0.386	yes	yes	yes
	VQAv2	0.771	0.309	0.680	0.301	yes	yes	yes
	POPE	0.881	0.184	0.745	0.000	yes	yes	yes
LLaVA-OV-7B	GQA	0.859	0.315	0.773	0.383	yes	yes	yes
	VQAv2	0.784	0.224	0.698	0.152	yes	yes	yes
	POPE	0.878	0.329	0.785	0.018	yes	yes	yes
Gemma-3-12B	GQA	0.752	0.183	0.673	0.561	yes	yes	yes
	VQAv2	0.714	0.143	0.632	0.339	yes	yes	yes
	POPE	0.855	0.157	0.773	0.403	yes	yes	yes

model	set	AURC clean	AURC B1	AURC B2
InternVL3.5-2B	GQA	0.266	0.837	0.576
	VQAv2	0.352	0.838	0.519
	POPE	0.325	0.846	0.526
Molmo2-4B	GQA	0.252	0.816	0.636
	VQAv2	0.288	0.809	0.631
	POPE	0.271	0.842	0.676
LLaVA-OV-7B	GQA	0.239	0.828	0.615
	VQAv2	0.309	0.817	0.603
	POPE	0.328	0.832	0.694
Gemma-3-12B	GQA	0.271	0.791	0.573
	VQAv2	0.293	0.781	0.583
	POPE	0.262	0.836	0.589

Table 27: **Operating-point sensitivity.** The FULL gate at three fixed thresholds, averaged over the twelve cells, on the headline attack's valid-image endpoints: realized clean false-accept rate, balanced-pool accepted accuracy clean and under attack, its drop, and the count of cells whose attacked accepted accuracy falls below 0.5, which is equivalent to falling below the no-gate baseline at natural prevalence. B1 is the oracle and B2 the label-free attacker of Table 26.

			B1 oracle			B2 label-free		
operating point	clean FPR	acc-clean	acc-adv	drop	<0.5	acc-adv	drop	<0.5
clean-10%-FPR	0.100	0.814	0.062	0.752	12/12	0.363	0.451	12/12
Youden J	0.247	0.743	0.132	0.611	12/12	0.380	0.363	12/12
τ=0.5	0.525	0.626	0.245	0.380	12/12	0.413	0.212	12/12

Table 28: **The combined attack on POPE.** Within one $8/255$ budget, stage 1 flips a correct answer and stage 2 raises the gate's confidence in the flipped answer while holding it byte-identical. *Flip rate* is over the 200 correct items; *accepted after stage 1* and *after stage 2* are the fractions of flipped items the fixed FULL gate accepts; *end-to-end* is the fraction of all 200 items that end as an accepted wrong answer.

model	flip rate	accepted after stage 1	accepted after stage 2	end-to-end
LLaVA-OV-7B	0.985	0.035	0.782	0.770
Molmo2-4B	1.000	0.010	0.875	0.875

G NO INTERNAL READOUT ESCAPES, IN FULL

Section 5 shows that training a readout to resist the attack does not produce one that is both informative and robust, and that the attack is not an artifact of its own settings. This appendix gives the full per-cell training sweep behind the trade-off, the configuration controls behind the second claim, and the robust-encoder test.

G.1 THE FULL ADVERSARIAL-TRAINING SWEEP

Table 29 reports the adversarial-training sweep in every cell. For each cell we train a family of probes on the same hidden states with a robustness weight β over $\{0.3, 1, 3, 10\}$, and attack each with the same answer-preserving attack; the table gives clean and worst-case-oracle attacked AUROC at each setting, the three-seed attacked value at the strongest setting, and the escape verdict. The escape region, clean above the cell's answer-string prior and attacked above the pre-specified 0.55 bar, is empty in all twelve cells. The trade-off is visible within each row: on Gemma/POPE the lightest penalty gives clean AUROC 0.865 and attacked 0.061, and raising the penalty to its strongest setting lifts attacked to 0.385 only by dragging clean down to 0.629. Robustness is bought with the discrimination the probe was meant to provide.

The training loss is a TRADES-style objective (Zhang et al., 2019) with the attack in the loop: class-balanced binary cross-entropy on clean features plus β times the mean, over items, of the largest absolute change in the probe's output across a pool of answer-preserving adversarial endpoint features for that item, with the pool regenerated every K epochs against the current probe by an inner PGD of the stated number of steps ($\epsilon = 8/255$, step $2/255$). The anchor cell uses the full budget (K=4, 10 inner steps); the other eleven use a reduced training-time inner attack (K=8, 5 steps) with the evaluation attack held at full strength, so their robustness is if anything understated. The gradient-penalty variant replaces the invariance term with 0.1 times the mean input-gradient norm of the probe output. The clean leg of the escape criterion is the answer-string prior on the evaluation split plus a 0.02 margin (0.734 for LLaVA/POPE), the evaluation-split counterpart of the ceiling's in-sample $c^\star$ (0.703) and of the cross-fit operative floor (0.597).

The nearest approach to an escape misses on a criterion fixed in advance. LLaVA/POPE reaches the highest attacked value anywhere in the sweep, 0.536 at its strongest setting, and it clears the clean leg of the criterion, but 0.536 sits below the 0.55 bar, and the three-seed standard deviation of 0.008 places it there reliably.

The failure holds even where the defender was under-resourced. Eleven of the twelve cells were trained against a weaker inner attack than the anchor, with the full-strength evaluation attack held fixed, so their measured robustness is if anything understated; none escaped. The one cell trained at full strength, the anchor LLaVA/VQAv2, also fails, reaching attacked AUROC 0.517 at its strongest setting, and a gradient-penalty variant of the anchor fails as well. The empty escape region is therefore not an artifact of a reduced training budget. Whether a better-resourced defense could escape is open; none we trained, at either budget, did.

G.2 THE COLLAPSE IS NOT AN ARTIFACT OF THE ATTACK SETTINGS

Two controls check whether the collapse is peculiar to the attack's hyperparameters or to its freedom to move the answer margin (Table 30). The upper block reports, per cell, how many of the swept attack configurations produce drift aligned to an independently estimated correctness axis: across ten coordinated configurations spanning budgets $\{2, 4, 8, 16\}/255$, step counts $\{5, 15, 30, 60\}$, two L_2 norms, and two aim targets, alignment holds in all twelve cells, $120/120$. The two label-agnostic configurations pass in 21 of 24 cell-settings; the three misses are all on Gemma and are directionless, a token-probability saturation effect, and they still pass an unsigned-drift criterion. The pattern that drives the collapse is therefore not confined to the operating budget or step count.

The lower block holds the answer margin near its clean value instead of letting the attack move it freely. Under this constraint the readout still collapses: attacked AUROC stays at or below the operative floor in all twelve cells and below chance in ten, with the two exceptions above chance but still below their floors. The inversion does not depend on the attack inflating the answer margin.

Table 29: **The full adversarial-training sweep.** Clean and worst-case-oracle attacked AUROC at each robustness weight β, the strongest setting in bold, with the three-seed attacked value (*peak att*) and the escape verdict per cell. The escape region is empty in all twelve cells. The nearest miss, LLaVA/POPE, reaches attacked 0.536, the highest anywhere, still below the 0.55 bar.

model	dataset	β=0.3	β=1	β=3	β=10	peak att	escape
		clean / attacked (seed 23; peak β bold)					
InternVL3.5-2B	GQA	0.643/0.198	**0.640/0.250**	0.624/0.205	0.607/0.215	0.207	no
	VQAv2	0.646/0.263	0.634/0.189	**0.621/0.307**	0.618/0.278	0.298	no
	POPE	0.845/0.026	0.808/0.000	0.710/0.257	**0.664/0.318**	0.316	no
Molmo2-4B	GQA	0.644/0.214	0.649/0.210	0.619/0.304	**0.608/0.384**	0.359	no
	VQAv2	0.724/0.270	0.681/0.378	**0.651/0.439**	0.610/0.437	0.377	no
	POPE	0.846/0.021	0.835/0.042	0.771/0.244	**0.668/0.412**	0.424	no
LLaVA-OV-7B	GQA	0.721/0.126	0.707/0.155	**0.683/0.266**	0.669/0.263	0.229	no
	VQAv2[§]	0.754/0.209	0.731/0.321	0.710/0.426	**0.674/0.521**	0.517	no
	POPE[‡]	0.818/0.290	0.828/0.194	0.827/0.190	**0.768/0.544**	0.536	no
Gemma-3-12B	GQA	0.693/0.174	0.646/0.255	0.651/0.316	**0.611/0.449**	0.372	no
	VQAv2	0.706/0.238	0.724/0.359	**0.713/0.540**	0.670/0.537	0.511	no
	POPE	0.865/0.061	0.862/0.059	0.829/0.103	**0.629/0.385**	0.385	no

[§]Anchor cell (LLaVA-OV/VQAv2), trained at full budget K=4, inner steps 10; the other eleven cells use the reduced training-time inner attack K=8, inner steps 5, with the evaluation attack unchanged at full strength. A gradient-penalty variant, run for the anchor only, reaches clean 0.763 and attacked 0.201, also no escape.
[‡]Nearest miss: as three-seed means, peak clean 0.755 clears the clean leg (threshold 0.734) but peak att 0.536 (sd 0.008; seeds 0.544, 0.528, 0.536) stays below the 0.55 bar; the row shows the seed-23 values 0.768 and 0.544; it is the highest attacked AUROC anywhere. Escape (0 of 12) requires peak clean above the clean threshold and peak att above 0.55.

G.3 A ROBUST VISION ENCODER

The probes of Appendix G.1 read a frozen representation, so they can gain invariance only by ignoring the directions the attack moves. A robust vision encoder changes the representation itself, and we test it with a paired swap on LLaVA-1.5-7B. The standard tower is OpenAI CLIP ViT-L/14 and the robust tower is FARE-4 (Schlarmann et al., 2024), the same model adversarially fine-tuned at $\epsilon = 4/255$. Both are built from the same open_clip checkpoint format through an identical conversion, so the only difference between them is FARE's fine-tuned weights, and the language model and projector of LLaVA-1.5-7B are unchanged. We run each tower at two resolutions. At 336 pixels, LLaVA-1.5's native input, both towers' 16×16 position grids are bicubically interpolated to 24×24, giving 576 image tokens. At 224 pixels, FARE's training resolution and the configuration in which Schlarmann et al. (2024) evaluate it, no interpolation is used and the model receives 256 image tokens. Our differentiable preprocessing matches the official processor with zero greylevel difference at both resolutions. Clean GQA accuracy is 0.748 and 0.671 for the standard and robust towers at 336, and 0.735 and 0.679 at 224, the expected cost of robust fine-tuning.

For each tower and resolution we retrain P(IK) and SAPLMA with the unchanged recipe on $10{,}000$ GQA training items labeled by the same judge, build balanced 500-item pools (447 for the standard tower on POPE, which yields only 197 wrong answers), and run the headline attack unchanged, with token probability scored on the endpoints aimed at the two probes. The escape criterion is fixed before any attack result: a readout escapes if its clean AUROC exceeds the cell's cross-fit operative floor and its worst-case-oracle attacked AUROC exceeds 0.55, and any escape is rerun on a stratified 150-item subset with 60 steps and 5 restarts and stands only if it survives. The clean leg uses the operative floor rather than the evaluation-split prior of Appendix G.1; every readout clears its floor, so every verdict turns on the robust leg and none depends on the choice.

Table 31 gives the result at FARE's own training budget. Every readout on both towers is informative, and none is robust. The standard tower is inverted almost completely, with oracle attacked AUROC between 0.00 and 0.08; FARE is inverted less completely, between 0.10 and 0.44, and the two resolutions agree. The label-free attacker separates the towers more sharply: it pushes every standard-tower readout below chance, to between 0.28 and 0.46, while FARE's readouts stay between 0.53

Table 30: **Attack-configuration invariance and the margin-band control.** Upper block: per cell, the number of swept attack configurations whose drift aligns to an independent correctness axis ($120/120$ coordinated, $21/24$ label-agnostic). Lower block: attacked AUROC when the answer margin is held near its clean value versus unbounded; the collapse persists at or below the floor in all twelve cells.

model	dataset	coordinated (of 10)	label-agnostic (of 2)
InternVL3.5-2B	GQA	10/10	2/2
	VQAv2	10/10	2/2
	POPE	10/10	2/2
Molmo2-4B	GQA	10/10	2/2
	VQAv2	10/10	2/2
	POPE	10/10	2/2
LLaVA-OV-7B	GQA	10/10	2/2
	VQAv2	10/10	2/2
	POPE	10/10	2/2
Gemma-3-12B	GQA	10/10	1/2†
	VQAv2	10/10	1/2†
	POPE	10/10	1/2†
total		**120/120**	**21/24**

†The three label-agnostic misses are all Gemma-3-12B and directionless (alignment below the null bar), not reversed: C1/Gemma/POPE, C2/Gemma/GQA, C2/Gemma/VQAv2, a token-probability saturation effect; they pass the unsigned-drift criterion. Coordinated configs: A1–A6 (budget $\{2, 4, 8, 16\}/255$, PGD steps $\{5, 15, 30, 60\}$), B1–B2 (two L_2 norms), D1–D2 (aim P(IK)/SAPLMA).

Margin-band control (SAPLMA; clean and attacked AUROC on the attack-engaged subset):

model	dataset	clean AUROC	attacked (unbounded)	attacked (margin)	floor
InternVL3.5-2B	GQA	0.880	0.029	0.039	0.642
	VQAv2	0.792	0.018	0.085	0.617
	POPE	0.832	0.000	0.028	0.555
Molmo2-4B	GQA	0.911	0.091	0.559	0.683
	VQAv2	0.872	0.130	0.513	0.625
	POPE	0.914	0.010	0.107	0.587
LLaVA-OV-7B	GQA	0.922	0.050	0.367	0.673
	VQAv2	0.798	0.047	0.256	0.616
	POPE	0.805	0.032	0.131	0.597
Gemma-3-12B	GQA	0.812	0.232	0.476	0.660
	VQAv2	0.798	0.209	0.365	0.652
	POPE	0.835	0.082	0.405	0.500

Attacked (margin) holds the answer margin in the 0.25 band; it is at or below the operative floor in 12 of 12 cells (below chance in 10 of 12; Molmo/GQA 0.559 and Molmo/VQAv2 0.513 stay above 0.5 but below their floors). The collapse persists when the margin is not allowed to run free.

and 0.64. A robust encoder therefore blunts the weaker adversary at this budget, and the worst-case one still inverts the ranking.

Table 32 follows the budget on GQA and gives the mechanism. The ratio of the mean reachable range to the clean separation, the median score of correct items minus that of wrong items, stays near two for the standard tower at every budget, and its attacked AUROC stays between 0.02 and 0.09, flat in the budget. For FARE the ratio grows with the budget, from 0.61 to 0.84 at $2/255$ to between 1.4 and 2.1 at $8/255$, and the attacked AUROC falls with it, from about chance to between 0.15 and 0.17. The ratio compares a mean against a median gap and is not a per-pair threshold, as FARE's SAPLMA at 336 and $4/255$ shows, inverted at 0.365 with a ratio of 0.95; across budgets it orders the depth of the inversion. At $2/255$ the label-free attacker leaves every FARE readout between 0.65 and 0.70, with bootstrap intervals above 0.55.

Table 31: **A robust vision encoder at its own training budget.** LLaVA-1.5-7B with only the vision tower changed, standard CLIP ViT-L/14 or FARE-4, at $\epsilon = 4/255$. Clean AUROC, worst-case-oracle attacked AUROC, and label-free attacked AUROC per readout. Every clean AUROC clears its operative floor (standard/FARE: GQA at 336, $0.556/0.583$; POPE at 336, $0.561/0.585$; GQA at 224, $0.554/0.634$), and no oracle value clears 0.55, so no cell escapes. POPE was run at 336 only.

			Standard CLIP			FARE-4		
Set	Input	Readout	clean	oracle	label-free	clean	oracle	label-free
GQA	336	P(IK)	0.741	0.041	0.397	0.683	0.325	0.587
		SAPLMA	0.694	0.046	0.430	0.722	0.365	0.613
		TokProb	0.782	0.050	0.349	0.747	0.321	0.578
GQA	224	P(IK)	0.720	0.075	0.460	0.704	0.314	0.575
		SAPLMA	0.715	0.021	0.370	0.718	0.435	0.637
		TokProb	0.747	0.039	0.370	0.744	0.291	0.539
POPE	336	P(IK)	0.774	0.002	0.284	0.740	0.168	0.552
		SAPLMA	0.622	0.003	0.354	0.781	0.209	0.533
		TokProb	0.797	0.000	0.320	0.729	0.105	0.548

Table 32: **The robust encoder across the perturbation budget on GQA.** Worst-case-oracle attacked AUROC and the ratio of the mean reachable range to the clean separation (median correct score minus median wrong score). The standard tower stays inverted with a ratio near two at every budget; FARE's ratio grows with the budget and its attacked AUROC falls with it. †Crossed 0.55 on the full pool and fell to 0.537 under the pre-registered stronger rerun (text). The $8/255$ budget was run at 336 only.

			oracle attacked AUROC			reach / separation		
Tower	Input	ϵ	P(IK)	SAPLMA	TokProb	P(IK)	SAPLMA	TokProb
Standard	336	2/255	0.040	0.085	0.058	1.97	1.76	1.96
		4/255	0.041	0.046	0.050	2.02	1.92	2.00
		8/255	0.040	0.034	0.045	2.02	2.01	2.01
	224	2/255	0.082	0.041	0.050	2.16	2.16	2.30
		4/255	0.075	0.021	0.039	2.25	2.25	2.38
FARE-4	336	2/255	0.475	0.512	0.488	0.83	0.61	0.81
		4/255	0.325	0.365	0.321	1.38	0.95	1.28
		8/255	0.164	0.171	0.152	2.14	1.39	1.85
	224	2/255	0.482	$0.564^{\dagger}$	0.484	0.74	0.79	0.84
		4/255	0.314	0.435	0.291	1.24	1.32	1.35

One cell crossed the bar. FARE's SAPLMA at 224 and $2/255$ reached an oracle attacked AUROC of 0.564 $[0.513, 0.614]$, the only escape in the study, with a reachable range of 0.079 against a clean separation of 0.101. The pre-registered rerun decides it. On the stratified 150-item subset the headline attack gives 0.551 $[0.459, 0.643]$ and the stronger attack 0.537 $[0.445, 0.631]$, finding slightly more reach (0.081 against 0.075), so the escape does not stand. The two values differ by less than their sampling error, and the cell sits at the edge of the bar rather than clearly on either side of it. At half its training budget and its native resolution, FARE brings this readout to the boundary of robustness without crossing it under the attack strength the claim requires.

H The Reachability Ladder in Full

Section 6 reports that an independent verifier relocates the attack, collapsing as the attacker gains access. This appendix gives the full twelve-cell ladder, the ordering of its rungs, the one verifier family whose differentiable path is inexact, and the disclosures behind the reported numbers.

H.1 THE TWELVE-CELL LADDER

Table 33 reports all twelve base-to-verifier pairs across the five access rungs, each the worst-case-oracle attacked AUROC of the verifier's confidence on a common item pool. The pairs span every model as both base and verifier three times. Discrimination falls from a clean mean of 0.655 to 0.480 under score-only access and 0.287 under white-box. The white-box mean over all twelve includes one pair whose verifier gradient is under-powered, discussed in Appendix H.3; over the nine pairs whose verifier admits an exact differentiable path, white-box drives discrimination to a mean of 0.223, between 0.06 and 0.35, the level of an internal readout. Once the attacker can reach its gradient, the verifier collapses to the same regime as the signals inside the model.

Verifier score and rungs. The verifier is a vision-language model prompted, per item, with `Question:` {q} newline `Proposed answer:` {a} newline `Is this answer correct for the image and question? Answer yes or no.` At the final position we read the next-token distribution, sum its probability mass on the yes tokens {yes, Yes, YES} and on the no tokens {no, No, NO}, and define the verifier score as the two-way normalized yes probability $s = p_{\text{yes}}/(p_{\text{yes}} + p_{\text{no}})$ (and $s = 0.5$ when the combined yes-no mass is zero). The decision-only rung thresholds this score at 0.5, the natural boundary at which the yes mass overtakes the no mass, and the hard-label attack flips the accept-or-reject decision against that threshold. The score-only rung queries the scalar s itself with no gradient. Every rung, from the transferred perturbation through the score query and the hard-label decision to the white-box gradient, reads this same yes-no score under the same prompt; the rungs differ only in what the attacker may access, not in how the verifier's confidence is defined.

The white-box search perturbs the image within the $8/255$ budget while enforcing base-model answer preservation exactly as the internal attack does: an argmax gate at every generated position and a fresh greedy byte-identity acceptance test, with the clean image as fallback. The base answer is byte-identical on every white-box endpoint; per pair, the clean-image fallback accounts for 0.9% to 6.5% of the 800 endpoints, and the rest are perturbations that leave the base answer unchanged.

H.2 THE INTERMEDIATE RUNGS ARE NOT SEPARABLY ORDERED

The intended ordering is that each rung, granting more access, drives discrimination lower. It holds at the ends: clean is the highest rung in all twelve pairs, and white-box the lowest in ten. It does not hold cleanly between the two intermediate rungs. The no-access transfer attack (mean 0.571) is on average a slightly stronger attack than decision-only hard-label search (mean 0.601), and in ten of twelve pairs transfer sits at or below decision-only, by 0.004 to 0.095. The transfer rung uses a label-aware oracle over the clean score and the transferred extremal perturbations, which is a strong baseline, while a decision-only boundary search must discover a flip from hard labels alone and does not reliably beat it. Every one of these inversions is within the overlapping bootstrap intervals of the two rungs, so none is individually significant, but the direction is consistent: a good transfer baseline is not weaker than hard-label access. The movement that carries the section is across the rungs that add real optimization power, from clean through score-only to white-box, where the fall is large and consistent.

The saturation account of Section 6 rests on per-item query statistics. Under decision-only access a mean 53% of items flip across the twelve pairs (range 45 to 69%). On the strongest pair of the four-pair ablation, Gemma to LLaVA on GQA, the 31 items that never flip sit at a mean clean-score distance of 0.335 from the 0.5 boundary and the 4 that flip at 0.047, and the median number of queries to the first flip is 132 against 52, 59, and 63 on the three weaker pairs. Both query rungs keep the last image whose fresh greedy decode reproduces the base answer, re-verified every 20 to 25 accepted iterations, with the clean image as fallback; rerunning them without this check moves final AUROC by at most 0.033 and the unflippable residue by at most one item.

H.3 THE MOLMO VERIFIER'S DIFFERENTIABLE PATH

The white-box rung attacks the verifier's gradient, which requires a differentiable path through the verifier's preprocessing that matches the official processor it is scored with. Table 34 reports the agreement per verifier family. For Gemma, LLaVA, and InternVL the differentiable path is bit-exact: it equals the official processor to 0.000 on the clean image, so the white-box gradient sees the same

Table 33: **The spanning-twelve reachability ladder.** Worst-case-oracle attacked verifier AUROC for all twelve base-to-verifier pairs across five access rungs, from clean through no-access transfer, decision-only, and score-only queries, to white-box gradient access. Clean and transfer carry bootstrap 95% intervals; the query and white-box rungs are two-seed means. The mean rows report all twelve pairs, the eleven excluding the under-powered white-box pair, and the nine whose verifier admits an exact differentiable path.

base→verifier	set	clean [95% CI]	transfer [95% CI]	decision HSJ	score SPSA (600 queries)	white-box	notes
LLaVA→Gemma	VQAv2	0.755 (0.68,0.82)	0.578 (0.50,0.66)	0.581	0.485	0.214	
LLaVA→InternVL	VQAv2	0.654 (0.58,0.73)	0.575 (0.50,0.65)	0.625	0.489	0.123	score n=100
Gemma→LLaVA	GQA	0.712 (0.64,0.78)	0.668 (0.59,0.74)	0.703	0.608	0.342	
Molmo→Gemma	POPE	0.660 (0.58,0.73)	0.514 (0.43,0.60)	0.502	0.431	0.195	
InternVL→LLaVA	GQA	0.681 (0.61,0.75)	0.637 (0.56,0.71)	0.664	0.561	0.270	
InternVL→Molmo	POPE	0.495 (0.41,0.58)	0.442 (0.35,0.53)	0.486	0.355	0.378	near-chance clean; parity
Gemma→Molmo	VQAv2	0.754 (0.68,0.82)	0.717 (0.64,0.79)	0.727	0.611	0.744	parity; WB under-powered
Molmo→InternVL	GQA	0.676 (0.60,0.75)	0.592 (0.51,0.67)	0.636	0.477	0.063	score n=100
LLaVA→Molmo	POPE	0.423 (0.34,0.52)	0.390 (0.30,0.48)	0.417	0.323	0.318	near-chance clean; parity
Gemma→InternVL	POPE	0.610 (0.53,0.69)	0.536 (0.46,0.62)	0.582	0.404	0.347	score n=100
InternVL→Gemma	VQAv2	0.752 (0.68,0.82)	0.616 (0.54,0.69)	0.608	0.434	0.289	
Molmo→LLaVA	VQAv2	0.690 (0.61,0.76)	0.586 (0.51,0.66)	0.681	0.578	0.164	
mean (12 pairs)		**0.655**	**0.571**	**0.601**	**0.480**	**0.287**	
mean, excl. under-powered white-box		**0.646**	**0.558**	**0.590**	**0.468**	**0.246**	
mean, bit-exact verifier (9)		**0.688**	**0.589**	**0.620**	**0.496**	**0.223**	

All AUROC are worst-case-oracle attacked verifier confidence. "mean, excl. under-powered white-box" drops the one under-powered Molmo-verifier white-box pair (Gemma→Molmo), giving a white-box mean of 0.246. "bit-exact verifier" drops the three Molmo-verifier pairs, the only non-bit-exact differentiable paths (Table 34): white-box 0.223, range 0.06 to 0.35. In the notes column, score n=100 marks the three InternVL-verifier pairs, "near-chance clean" the two Molmo-verifier pairs on POPE, and "parity" the Molmo-verifier differentiable-path gap.

function the reported score uses. The Molmo verifier's differentiable preprocessing diverges from its official processor by up to 0.144 on the affected cells, a property of that preprocessing and not of the attack. Because every reported verifier score is recomputed with the official processor, this divergence never enters a reported number; it only under-powers the white-box search on the three pairs whose verifier is Molmo. The effect is visible in one of them, the Gemma-to-Molmo pair, whose white-box AUROC (0.744) stays at its clean value (0.754), since the gradient guiding the search does not match the scored function. For this reason the section's white-box figure is taken over the nine pairs whose verifier path is exact, where the search is driven by the true gradient.

Two of the three Molmo-verifier pairs, both on POPE, carry a second, independent limitation: their verifier is at or below chance on the clean image (0.495 and 0.423), so it fails to separate correct from incorrect answers before any attack. Their ladders have little to collapse and sit near chance throughout. These pairs are reported in the table for completeness and excluded from the bit-exact white-box mean.

H.4 Disclosures

Three reporting choices are stated here. First, the score-only rung of the three InternVL-verifier pairs is computed on a balanced 100-item pool instead of the 200 used elsewhere, because the InternVL verifier's dynamic tiling makes it the slowest per-item rung; the score-only floor converges well before this query budget, so the reduction widens the bootstrap interval on those three pairs modestly without changing their collapse, which matches their neighbours. Second, the score-only rung is reported uniformly at a query budget of 600, at which the floor has converged; some pairs were run to a larger budget and are read at the same 600-query checkpoint, so the rung is a single budget across all pairs. Third, the twelve pairs draw on the three benchmarks in a $5/3/4$ split, which affects only the dataset marginal and no per-cell ladder. On the pairs where we computed it, the common 200-item pool tracks the full 1000-item pool to within about 0.07 AUROC.

Table 34: **Differentiable-path agreement per verifier family.** The maximum absolute gap between each verifier's differentiable path and its official processor on the clean image. Three families are bit-exact; only the Molmo verifier's preprocessing diverges, which under-powers the white-box search on the three Molmo-verifier pairs while leaving every reported score, recomputed with the official processor, unchanged.

verifier family	max parity gap	bit-exact	cells affected
Gemma-3-12B	0.000	yes	none
LLaVA-OV-7B	0.000	yes	none
InternVL3.5-2B	0.000	yes	none
Molmo2-4B	0.144	no	3 Molmo-verifier pairs (max 0.144)

The Molmo verifier's per-item mean gap is ≈0.004 against the 0.144 maximum. Every reported verifier score is recomputed with the official processor, so the gap affects only the white-box search, never a reported AUROC.

I Scope and Negative Results

This appendix records what the paper does not establish. It bounds the threat model with two negative results, reports the defenses we evaluated only partially, and notes an earlier construction that reached the same conclusion as the adversarial-training sweep by a different route.

I.1 The attack is per-item and white-box, not universal or transferable

Two negatives fix the threat model. First, the attack does not transfer across models. Table 35 takes a perturbation optimized against one model and evaluates it on another. Only Gemma and LLaVA serve as attack sources here, for a reason specific to this experiment: a transferable perturbation must be optimized in raw pixel space through a differentiable reimplementation of the source model's preprocessing, and at the time of this experiment only the Gemma and LLaVA front-ends had been built and validated. InternVL's dynamic tiling and Molmo's patch flattening make that path higher effort and it was not built, so both appear only as targets; this is a limitation of the run, not of the models, since InternVL's forward pass is deterministic (Section 3) and its differentiable path is bit-exact (Table 34). The same-model white-box attack removes between 0.44 and 0.76 of token-probability discrimination, while the identical image applied to a different target removes at most 0.05, and usually under 0.03. The corruption is a property of the specific model attacked.

Second, the effect requires per-item optimization. Table 36 applies a single fixed perturbation to every item and compares it against an equal-norm random perturbation. The largest gap between the universal perturbation and the random control is 0.024 across all cells, and for six of seven channels the difference is statistically indistinguishable from zero. The per-image aimed attack removes between 0.41 and 0.71 on the same cells, nineteen to seventy-one times the universal-versus-random gap. No single image-independent perturbation carries the effect; it is built anew for each item.

I.2 Defenses we did not fully evaluate

Two defense families are evaluated only partially, and Table 37 gives what we measured. For input purification, five inexpensive purifiers are evaluated nonadaptively, without a purifier-aware attack, so their numbers are a lower bound on what an adaptive attacker achieves; even so, the strongest of them, JPEG at quality 50, leaves between 38 and 76% of the manipulable range while itself corrupting between 3 and 36% of the model's own answers. No cheap purifier removes the range without a discrimination cost, and none is tested against an adaptive attack.

Randomized smoothing (Cohen et al., 2019) is the more instructive case, because it shows why a defense must be attacked adaptively before it can be called robust. Under a transferred attack, smoothing appears to work: the witnessed displacement collapses and the readout recovers above its floor. An expectation-over-transformation attack (Athalye et al., 2018b) that is aware of the smoothing does not collapse: its aimed displacement stays between 0.18 and 0.37, between 2 and 27 times the transferred value, and it drives the smoothed readout's attacked AUROC to between 0.10 and 0.34, below the floor in every cell, on all four models at $\sigma \in \{0.08, 0.16\}$ on GQA and on LLaVA on POPE as well; in each, the aimed displacement exceeds the cell's $\gamma^\star$, so smoothing does

Table 35: **Cross-model transfer is weak.** Token-probability AUROC drop when a perturbation optimized against a source model is evaluated on other targets. *white-box* is the same-model attack; *transfer mean/max* are the drop when the same image is applied to the three other targets. Only Gemma and LLaVA had a raw-pixel differentiable front-end at the time of this run and serve as sources.

source	set	white-box tok drop	transfer mean	transfer max	worst target
Gemma-3-12B	GQA	$+0.444$	$+0.006$	$+0.011$	Molmo
	VQAv2	$+0.481$	$+0.013$	$+0.023$	LLaVA
	POPE	$+0.514$	$+0.024$	$+0.033$	LLaVA
LLaVA-OV-7B	GQA	$+0.760$	$+0.023$	$+0.051$	Molmo
	VQAv2	$+0.710$	$+0.023$	$+0.039$	InternVL
	POPE	$+0.712$	$+0.014$	$+0.044$	InternVL

Metric: token-probability AUROC drop. The same-model (white-box) attack removes 0.444 to 0.760; the same image transferred to a different target removes at most $+0.051$ (LLaVA to Molmo on GQA, the largest of the 18 off-diagonal transfers), and most are below $+0.03$.

Table 36: **A fixed universal perturbation does not carry the effect.** Upper block: the largest universal-minus-random gap per cell against the per-image aimed drop. Lower block: per-channel universal and random drops on the least-movable cell, with bootstrap intervals. The universal perturbation is indistinguishable from an equal-norm random one.

model	set	max|uni−rand|	per-image SAPLMA drop	ratio
Gemma-3-12B	GQA	0.023	0.433	19×
	VQAv2	0.020	0.410	20×
	POPE	0.024	0.681	28×
LLaVA-OV-7B	GQA	0.010	0.708	71×
	VQAv2	0.008	0.548	68×
	POPE	0.009	0.608	68×

Universal vs equal-norm random delta, per channel (Gemma/GQA):

channel	clean	uni drop	rand drop	uni−rand [95% CI]
token-prob	0.571	-0.031	-0.054	$+0.023$ [-0.002,$+0.048$]
P(IK)	0.790	$+0.037$	$+0.028$	$+0.009$ [-0.008,$+0.025$]
SAPLMA	0.808	$+0.013$	$+0.017$	-0.004 [-0.014,$+0.005$]
verbalized	0.554	-0.007	$+0.010$	-0.017 [-0.032,-0.004]
self-consist.	0.552	-0.037	-0.059	$+0.022$ [-0.001,$+0.046$]
CCPS	0.741	$+0.081$	$+0.082$	-0.001 [-0.028,$+0.026$]
CCPS-D	0.743	$+0.036$	$+0.034$	$+0.002$ [-0.022,$+0.025$]

drop = clean minus perturbed AUROC. The largest universal-minus-random difference is 0.024 over all six cells, and six of seven channels have uni−rand intervals containing 0. The per-image aimed attack removes 0.410 to 0.708 on SAPLMA in the same cells, 19 to 71 times the largest universal-versus-random gap.

not restore the amplitude certificate either. Smoothing raises the attacker's cost and does not prevent the collapse once the attacker accounts for it.

I.3 An earlier construction reached the same conclusion

Before the adversarial-training sweep of Appendix G, we tested whether a readout confined to the directions the attack moves least could be both informative and robust. We built a probe on the trailing subspace of an empirical displacement null, the directions in which the attack produces the least movement, and attacked it. The construction found no readout that was both: a probe restricted to the least-moved directions loses the clean discrimination that made it useful, the same trade-off the trained probes exhibit. We record it here because it shows the conclusion does not depend on the training procedure that produced it; a directly constructed low-movement readout fails in the same way.

Table 37: **Defenses not fully evaluated.** Upper block: five input purifiers, evaluated nonadaptively (a lower bound), by the manipulable range they leave and the own-answer corruption they cost. Lower block: randomized smoothing on all four models under an adaptive expectation-over-transformation attack, which the transferred attack (*tr.*) masks but the aimed attack (*aim*) does not; the aimed displacement exceeds the cell's $\gamma^{\star}$ in every cell, so smoothing does not restore the amplitude certificate.

defense	P(IK) survival range	median	corruption range	median
JPEG q75	0.49–0.90	0.680	0.02–0.38	0.164
JPEG q50	0.38–0.76	0.481	0.03–0.36	0.172
blur	0.53–0.97	0.712	0.02–0.17	0.105
4-bit	0.72–1.03	0.956	0.04–0.27	0.131
resize-crop	0.60–1.05	0.781	0.02–0.14	0.075

survival = fraction of the P(IK) manipulable range the purifier leaves (1.0 = removes nothing; values above 1.0 mean the purifier enlarges the range); corruption = own-answer corruption the purifier costs.

Randomized smoothing under the adaptive EOT attack (SAPLMA):

model	set	σ	clean	tr. disp	aim disp	ratio	tr. adv	aim adv	floor	γ^{*}	disp$>\gamma^{*}$
LLaVA	GQA	0 (base)	0.851	–	0.240	–	–	0.143	0.673	0.125	yes
LLaVA	GQA	0.08	0.836	0.079	0.361	4.6×	0.728	0.176	0.673	0.110	yes
LLaVA	GQA	0.16	0.808	0.019	0.361	19.2×	0.784	0.200	0.673	0.085	yes
LLaVA	POPE	0.08	0.693	0.078	0.182	2.3×	0.513	0.104	0.597	0.015	yes
LLaVA	POPE	0.16	0.671	0.022	0.258	11.8×	0.637	0.123	0.597	0.015	yes
InternVL	GQA	0.16	0.779	0.020	0.368	18.8×	0.755	0.175	0.642	0.085	yes
Molmo	GQA	0.08	0.793	0.044	0.183	4.2×	0.722	0.305	0.683	0.045	yes
Molmo	GQA	0.16	0.776	0.012	0.209	17.8×	0.760	0.317	0.683	0.045	yes
Gemma	GQA	0.08	0.788	0.022	0.295	13.3×	0.775	0.255	0.660	0.070	yes
Gemma	GQA	0.16	0.771	0.009	0.245	27.2×	0.766	0.336	0.660	0.070	yes