OBSERVED ACCURACY IS NOT VISUAL CAPABILITY: AUDITING DECISION-LEVEL CONFOUNDS IN VIDEO ANOMALY BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

A rule reading only the video header, meaning resolution, duration and frame count, predicts a video anomaly benchmark's labels at 0.970. We had adopted that benchmark because it looked clean.

How much of such a shortcut survives a strong model? We formalize the question as *residual side-channel capacity*, the drop in optimal predictive risk from adding an unintended channel to a frozen model score, which under log loss is $I(Y; Z \mid S)$; our plug-in estimate is a lower bound on it. Above the strongest model we measure, video headers close 42% of the label uncertainty its score leaves open, lifting AUC from 0.947 to 0.982. On a rendered benchmark the same audit returns capacity indistinguishable from zero, so this is a property of how footage was collected, not of the estimator.

A benchmark score is an intermediate observable, and reading capability off it means first showing which layer moved. We therefore separate what one score conflates, acquisition shortcuts, visual ranking and the decision operating point, probing each by its own intervention and releasing the audit as a harness with cached scores that reproduces every headline number on CPU in seconds. Task structure makes part of any score unidentifiable, which gives an exact account of when repairing an operating point helps, hurts or does nothing. Because ranking transfers across domains and absolute scale does not, a certified rank-based selective rule retains 1.2–$1.7\times$ the coverage of an equally certified score-threshold rule at matched risk, with paired intervals excluding zero in three of four settings. We also report four revisions to our own conclusions and 18 pre-registered negative verdicts.

1 INTRODUCTION

A benchmark score is a claim about capability. In video anomaly understanding, we find much of it traces to decision thresholds, question order in a file, and video encoding. Figure 1 summarises the audit that follows.

We began with a reproduction: a state-of-the-art checkpoint matched its six published metrics to within 1.1 points. Stratifying the same predictions by anomaly *semantics* revealed a gap the aggregate cannot express (§4). Pursuing it exposed a metric that moves with the decision threshold whenever an intervention does (§5).

The paper is one instrument, a *capability decomposition audit* (Appendix B). Separate interventions probe three diagnostic axes, not an additive decomposition: shortcut susceptibility, controlled visual ranking, decision operating-point error. Separating them changes what a number means: a rule-based recall that rises from 0.292 to 0.667 with more visual compute is flat, 0.2083 against 0.2083, once the false-positive rate is matched, and a $+37$-point "improvement" proves empty. Our contributions are:

- **A released audit protocol.** Each of the three applies wherever its own input exists: metadata, covariate overlap, a structural constraint, calibration labels. Here it finds a benchmark whose labels a video-header rule predicts at 0.970.

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

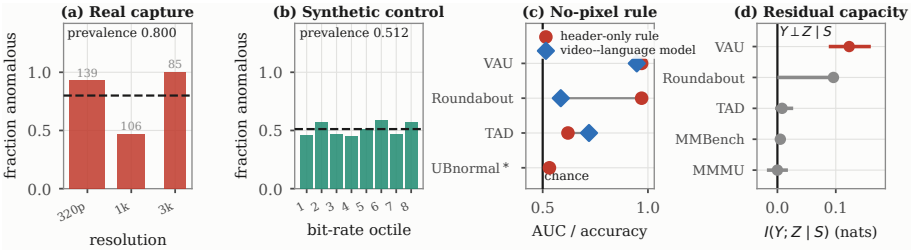


Figure 1: **Acquisition metadata carries label information on benchmarks collected from real deployments, and none on a synthetically rendered one.** (a) Roundabout clips by header resolution; bar height is the fraction anomalous, the dashed line the prevalence a label-blind rule returns, the number above each bar n. (b) The same audit on UBnormal, a rendered benchmark, over bit-rate octiles (n=68 each): every bucket sits on that line. (c) Header-only rule and a video–language model on one axis against chance; $*$ marks the synthetic benchmark, where no model score is compared. (d) What the channel carries *above* a frozen model score, $I(Y; Z \mid S)$, clustered bootstrap intervals, solid line conditional independence. The last two entries are image benchmarks, included only as controls; their intervals contain zero.

- **Residual side-channel capacity, and an active audit of it.** We define the label information an unintended channel keeps *above* a frozen model score ($= I(Y; Z \mid S)$ under log loss) and estimate it by cross-fitting. It separates three regimes: 0.123 nats above a 0.947 model, indistinguishable from zero on a second benchmark, 0.278 on a third. Controlled evaluation removes a weak channel, not a near-deterministic one.
- **A controlled decision diagnosis.** BISI is a ranking-preserving intervention on the decision layer. Its guarantees say exactly when correcting the aggregate decision cardinality helps, hurts, or does nothing. We audit our own earlier claims under it, revising four.
- **A certified practical consequence.** Ranking stays stable when the absolute score scale shifts. Under independent certification, RankGuard's rank-based selective decisions retain 1.2–$1.7\times$ the coverage of score-based ones at matched certified risk. It correctly abstains where the ranking fails.

What we release. The audit is the artifact: zero-pixel floors, the exploitability attack, propensity matching, the operating-point identities, the certification code. Each is cheap in its inputs, carries a pre-registered criterion, and ships with the negative controls that made us trust it, including the grid-searched certificate that passes nominally while violating its risk bound. Appendix B tabulates what each needs and what its verdict licenses.

A note on what we do not claim. Constrained inference is not new, and the complementary-pair instantiation is known in the associated competition community. Our contribution is the route to it: task constraints identify which score directions cannot be read as capability. It yields an exact formula for operating-point displacement, a degradation law under approximate constraints, and an identity fixing when the repair helps or hurts.

2 Related Work

Measurement validity for benchmarks. A parallel line asks, normatively, what a score licenses. Jo & Wilson (2026) would ground evaluation in a theory of capability and treat benchmarks "not as fixed datasets, but as experimental protocols whose degrees of freedom must be explicitly modeled". Alaa et al. (2025) call for "principled tools to evaluate the construct validity". Bean et al. (2025) survey 445 benchmarks and find statistical testing rare. We read that work as a complement, not a competitor: it says which questions to ask. We measure one quantity, $I(Y; Z \mid S)$ above a frozen incumbent score, with an estimator, intervals, a null arm and a negative control. Frameworks say what to look for; a capacity estimate, how much is there.

Video anomaly reasoning and adaptive visual compute. Recent systems couple video–language models with chain-of-thought supervision. Huang et al. (2025) introduce a perception-to-cognition chain-of-thought with reinforcement, and Zhu et al. (2025) release VAU-Bench, the benchmark we audit. The former's released checkpoint is our main object of study: we reproduce all six reported numbers to within 1.1 points, then ask what they measure. A large literature also allocates visual budget adaptively, scoring accuracy at each budget. For that line, §5 is a methodological caution, not a competing method. Raising the budget also shifts the decision point, so accuracy-versus-budget curves conflate evidence gain with calibration drift in either direction.

Benchmarks solvable through unintended channels. Our audit joins an established genre founded by Torralba & Efros (2011) on dataset bias and Geirhos et al. (2020) on shortcut learning. NLI is partly solvable from the hypothesis alone (Poliak et al., 2018; Gururangan et al., 2018) or from syntactic heuristics (McCoy et al., 2019). VQA leaned on language priors until rebalancing (Goyal et al., 2017), and single-frame evidence largely serves video-language understanding (Buch et al., 2022). We do not audit the video-anomaly benchmarks that define the task (Sultani et al., 2018; Wu et al., 2020; Liu et al., 2018); our conclusions cover the five corpora we measure. Concurrent video-VLM auditing uses controlled channel conflict to separate visual sequence evidence from positional encoding in temporal scores (Rahman, 2026). We inherit it and add two things. Our channel is *acquisition metadata*: invisible to text-only floors, untouched by rebalancing, outside the input pipeline. And detection is not the endpoint: we test whether the channel can manufacture apparent performance and whether controlled evaluation removes it, the distinction between benchmark exploitability and model reliance that a blind baseline cannot draw. This is also orthogonal to contamination, which asks whether a model has seen the test items.

Selection bias, calibration, and constrained decoding. Answer-position and answer-token biases are well documented, with permutation- and prior-based remedies (Zheng et al., 2024). We add no debiasing estimator, finding instead that our dominant bias is semantic, a preference for NO that permutation cannot address. Label-shift corrections that estimate a target prior from unlabeled data (Lipton et al., 2018) are BISI's closest alternative. They *estimate* the nuisance, which requires it to transfer; BISI eliminates it by construction, a contrast Appendix N tests. Cardinality constraints at inference are long-standing (Brukhim & Globerson, 2018) and our complementary-pair decoding is known in that community; we claim neither. We add a reading of that machinery as an *identification* statement: the constraint row space is exactly the set of score directions that cannot influence a structured decision, hence the set to quotient out before reading a score as capability.

3 SETUP

Models and data. We study a released state-of-the-art video anomaly reasoning checkpoint (7B, chain-of-thought) on its own benchmark, VAU-Bench (439 held-out videos: 204 anomalous, 235 normal). On AIC Track 3 TAR, a large transportation reasoning benchmark, we build a leakage-clean subset by deduplicating video IDs against the model's training corpus. Source-level exclusion is insufficient: we drop 804 of $3{,}669$ videos, including 88% of the only source with rule-violation labels, leaving $n{=}24$ rule-based items. Sampling ≤ 200 videos per source yields 847 videos and $2{,}541$ exact-match questions, split video-disjointly.

Continuous decision scores. Threshold-free analysis needs a scalar score; free-text generation exposes none. We re-feed the prompt with the model's *own* chain-of-thought truncated at the decision token and read the next-token distribution. One forward pass, no generation, reproducing the hard decision on $1756/1756$ items. We vary frames $\times$ resolution in visual tokens: $1\times$ (16f@$128 \cdot 28^2$), $2\times$ (32f), $4\times$ (16f@$512 \cdot 28^2$), $8\times$ (32f@$512 \cdot 28^2$). The $2\times \to 8\times$ step is strictly nested: same frame indices at higher resolution, no silent clamping ($1{,}872 \to 7{,}776$ tokens).

4 LAYER 1: AGGREGATE ACCURACY HIDES SUBGROUP COLLAPSE

Anomaly semantics split the reproduced predictions in two: visually dramatic anomalies (fire, explosion, riot) are recalled at 0.952, rule-based traffic violations (wrong-way driving, red-light running,

illegal parking) at 0.292. The rule-based subgroup is small (n=24; §3), so we read its *level* as indicative and rest the argument on the gap's persistence. The gap reproduces on every model tested (§8.2) and survives every remedy below. Both figures share one operating point, so it is no thresholding artifact.

Neither rarity nor the standard remedies explain it. Wrong-way driving has 26 training examples, more than several classes detected perfectly, yet is recalled at 0.14. Scale does not help: two backbone generations plateau at exactly 0.333. Fine-tuning does not: the published pipeline adds 9.8 aggregate points while *widening* the semantic gap from 0.522 to 0.660. Prompting does not: verbalizing the missing evidence repairs 1 of 31 misses.

The model is confidently wrong. In the decision-token distribution, 59% of items saturate at $|s| > 15$. One true illegal-parking case is called normal with $P(\text{Normal}) = 0.999991$. Only 1 of 17 missed rule-based anomalies hedges at all in its chain-of-thought (Appendix A). That removes the gradient a confidence-based adaptive-compute policy needs, and foreshadows Appendix N: the budget effect's sign proves unpredictable from the model's scores.

5 Layer 2: Apparent Compute Gains Are Operating-Point Drift

§4 invites more visual compute, and on the surface it works: rule-based recall rises from 0.292 to 0.667 as the budget goes $1\times \rightarrow 8\times$. Most of this is the decision threshold moving, not the model seeing more.

The budget moves the operating point. True- and false-positive rates rise *together*: $0.848 \rightarrow 0.951$ and $0.115 \rightarrow 0.183$. Net label flips relative to $1\times$ are $+6$, $+13$, $+37$, every one NORMAL→ABNORMAL. All breakage sits in the normal class, 21 of 208 at $8\times$, with *zero* true anomalies broken. More visual detail does not degrade the evidence; it makes the model readier to call anomaly.

At a matched false-positive rate the gain is exactly zero. Fixing $\text{FPR} = 0.10$ at every budget, rule-based recall is 0.2083 at $1\times$ and 0.2083 at $8\times$: identical at eight times the cost. Overall TPR there is *worse* at $8\times$ than at $1\times$ (0.667 vs. 0.740), a different quantity that coincidentally shares its value.

Genuine discriminability saturates early. Threshold-free, AUC rises $0.9016 \rightarrow 0.9291$ from $1\times$ to $2\times$, then stops at 0.9282 for $8\times$. These are compute-sweep operating points, below the 0.947 reported elsewhere here: the checkpoint's released full-resolution configuration. The sweep fixes every other decoding setting so only visual tokens vary; comparisons within a table share a protocol. We report an equivalence bound, not a failure to reject: four times more compute beyond $2\times$ buys at most $+0.014$ AUC (95% one-sided upper bound). This rules gains out; it does not assert their absence.

The distortion runs in both directions. On the transportation benchmark the same mechanism reverses sign. The raw budget effect is $+0.7$ points against a bias-invariant $+3.2$, so the operating point *hides* a real gain. This confounding is not a one-directional inflation to discount: recovering any compute benefit's sign and magnitude needs a bias-invariant scale. We report this direction as an *illustration rather than a certified result*: a single paired comparison without an interval, not reproduced by the harness, carrying no weight beyond the certified results. We include it because the sign reversal is the point.

6 Layer 3: The Benchmark Admits Non-Visual Solutions

The first two layers concern how outputs become scores; the third, the benchmark. We found it by accident: a constraint-corruption control (§8) failed to degrade performance, which should have been impossible. That thread produced three independent channels: the transportation benchmark is solvable without watching a video.

Order, wording, options. Binary questions come two per video as a complementary pair, stored positive-first. "YES to the first, NO to the second" scores 0.9667 over all $3{,}669$ videos; combining it with model evidence gives 0.9622, *below* the prior alone. Independently, TF–IDF on one question's text reaches 0.9008 leave-one-source-out when both questions of a pair are scored. In the multiple-choice subtask, position-only (0.262) and stem-only (0.253) priors sit at chance, but scoring each *candidate option* independently reaches 0.7792 video-disjoint, matching the 0.776–0.799 of video models.

Scope and repair. Randomizing pair order repairs order leakage (0.5014). Wording leakage has none, being a property of how the items were generated, so we report a text-only floor alongside every number here. Scope, the redacted-split caveat and the full protocol are in Appendix G. We have notified the organizers.

A fourth channel, on a benchmark we called clean. The three above are textual and one shared prompt closes them, but not a visual one. On Roundabout, scene identity alone reaches $\mathrm{AUC} = 0.8381$: two intersections supply 129 uniformly anomalous clips, the rest are uniformly normal. A rule reading only the video header, resolution $\neq 1080\times720$, scores 0.970 with no pixels at all. This is how the footage was collected, not an artifact of our processing. Closing a text channel does not certify a benchmark.

Residual side-channel capacity. Detection is not yet harm, so we measure two quantities (Table 1). *Benchmark exploitability*: can a shortcut-seeker manufacture a gain? *Model reliance*: does its ranking survive evaluation matched on the measured covariates? The first is a dataset property, the second a model property; conflating them is the error this paper is about. To quantify the first, write $\mathcal{R}_\ell(X) = \inf_f \mathbb{E}[\ell(Y, f(X))]$ for a proper loss ℓ. The *residual side-channel capacity* of an unintended channel Z above a frozen model score S is

$$\mathrm{RSCC}_\ell(Z \to Y \mid S) \;=\; \mathcal{R}_\ell(S) - \mathcal{R}_\ell(S, Z) \qquad \text{so that} \qquad \mathrm{RSCC}_{\log} = I(Y; Z \mid S), \tag{1}$$

the identity following from $\mathcal{R}_{\log}(X) = H(Y \mid X)$. It is zero exactly when $Y \perp Z \mid S$, and under Brier loss equals $\mathbb{E}[(\eta_{SZ} - \eta_S)^2]$. The identity is standard and the restricted-class estimate below is a conditional usable-information quantity (Xu et al., 2020; Hewitt et al., 2021). New here is reading it as a benchmark property: how much label information the channel still carries *after* the incumbent's score is spent. Cross-fitting $\hat{h}(S)$ and $\hat{g}(S, Z)$ on the audited labels and differencing held-out losses estimates it in a fixed hypothesis class: capacity, not a participant holding no labels. A frozen-split variant closes even that gap. Fitted on a video-disjoint 80% split and read *once* on the untouched remainder, the attacker still gains $+0.057$ AUC (95% one-sided LCB $+0.021$; TAD null arm $+0.004$).

A synthetic negative control. If header capacity came from the estimator rather than collection it would appear everywhere; it does not. UBnormal (Acsintoae et al., 2022) is a rendered benchmark of 543 clips over 29 scenes, and the identical audit there returns AUC 0.532 ± 0.052 against a chance level of 0.5, and accuracy 0.520 against a 0.512 baseline. Real deployments encode anomalous and normal footage through different capture paths; one rendering pipeline does not. Above the strongest model we measure, six header features close 0.123 nats, 42% of the 0.293 nats its score leaves open, lifting AUC from 0.947 to 0.982. On Roundabout capacity is 0.278 nats (56%) and a chance-level model becomes an apparent 0.956 detector; on TAD-bench it is indistinguishable from zero. We report these only as attacker gains, never as scores we claim. Propensity matching leaves the ranking essentially unchanged on VAU and TAD-bench, so the shortcut exists but no reliance is detectable under matched evaluation. On Roundabout neither control neutralizes it, the resolution rule being *global*: a near-deterministic channel can only be disclosed. The limiting case is so-tad (duration alone 0.9996), excluded from every controlled comparison here.

What this costs us. This disqualifies our own headline demonstration: we measured a $+18.4$ point gain from structural decoding on the binary subtask and called it recovered visual evidence. But a text-only system reaches 0.9008 there, so the gain mixes visual evidence with benchmark-specific textual priors inseparably. Our clean demonstration moves to §8, on a benchmark whose format admits no such channel.

Table 1: The shortcut layer, both directions. "Attack" fuses the VLM score with six video-header features out-of-fold; refitting the score alone on the same folds changes nothing (-0.002 to -0.010), so the gains come from the metadata. "Matched" evaluates on propensity-matched subsets. † marks the one $\mathrm{RSCC}_{\log}$ interval containing zero. Appendix H gives the fold construction.

	metadata	unconstrained eval		matched eval		$\mathrm{RSCC}_{\log}$
Benchmark	only	VLM	+metadata attack	VLM	attack	(nats)
VAU	0.865	0.947	0.982	0.971	0.967	**0.123**
TAD-bench	0.810	0.986	0.988	0.978	0.972	0.012†
Roundabout	0.961	0.586	**0.956**	0.437	0.651	**0.278**

7 Method: Bias-Invariant Structured Inference

7.1 A. Identification: which score directions are nuisance

Let valid outputs be $\mathcal{Y} = \{y : Ay = b\}$ and let s be the model's score vector. For any λ and any $y \in \mathcal{Y}$,

$$(s + A^\top \lambda)^\top y \;=\; s^\top y + \lambda^\top A y \;=\; s^\top y + \lambda^\top b, \tag{2}$$

i.e. the perturbation adds the *same constant* to every valid output. The component of s in the row space of A is therefore unidentifiable from structured decisions: we cannot estimate it from them, and it should not influence them.

7.2 B. Repair: the inference rule this implies

Write $[s] = s + \mathrm{row}(A)$ for scores differing only by an unidentifiable component.

Theorem 1 (Constraint-respecting decisions depend only on $[s]$) *Let $\hat{y}(s) = \arg\max_{Ay=b} s^\top y$. Then $\hat{y}(s + A^\top \lambda) = \hat{y}(s)$ for all λ.*

Theorem 1 implies the repair: stop reading s coordinate-wise and decide jointly within the feasible set:

$$\hat{y} \;=\; \arg\max_{y:\, Ay=b} \; s^\top y \tag{3}$$

Equation (3) is ordinary constrained argmax. New is the identification statement that licenses it: a repair for an identification problem about measurement, not a decoding convenience. By Theorem 1 the canonical representative $\tilde{s} = P_A s$, with $P_A = I - A^\top (AA^\top)^\dagger A$, cannot change the decision, a verified no-op in both instantiations. The rest of the section makes the route pay: §C1 the distortion, §C2 degradation under an approximate constraint, §C3 when it helps or hurts.

7.3 C1. Guarantee: how far an operating point displaces accuracy

Take the pairwise-complement case. Let Z_+ and Z_- be the scores of a pair's truly-positive and truly-negative question under an unbiased operating point, with CDFs $F_\pm$. The model applies a shared additive bias β.

Theorem 2 (Exact distortion) *Independent thresholding at zero attains per-question accuracy*

$$A_{\mathrm{raw}}(\beta) = \tfrac{1}{2}\left[1 - F_+(-\beta) + F_-(-\beta)\right], \qquad A'_{\mathrm{raw}}(\beta) = \tfrac{1}{2}\left[f_+(-\beta) - f_-(-\beta)\right], \tag{4}$$

whereas constrained decoding attains $A_{\mathrm{pair}} = \Pr[Z_+ > Z_-]$, which does not depend on β.

Two consequences matter. A'_{raw} takes the sign of $f_+(-\beta) - f_-(-\beta)$. An operating-point shift can move measured accuracy in *either* direction, fabricating an apparent gain or suppressing a real one: the bidirectional behaviour of §5. And the recoverable gap $A_{\mathrm{pair}} - A_{\mathrm{raw}}(\beta)$ is largest where β sits far from the density-balancing value, so recovery tracks bias severity (§8.2). The formula is exact to five decimals at every β tested (Appendix C): $A_{\mathrm{pair}} = 0.8383$ against a measured 0.8394, the difference being tied pairs.

7.4 C2. GUARANTEE: DEGRADATION UNDER A MISSPECIFIED CONSTRAINT

If a fraction ε of pairs violate the exactly-one-positive constraint, any decoder forced to emit one positive scores $1/2$ there, so $A(\varepsilon) = r - \varepsilon(r - \frac{1}{2})$. With $r = 0.8383$ this predicts $0.8383 - 0.3383\varepsilon$, and measured values at $\varepsilon \in \{0, .05, .10, .20, .50\}$ track it to within 0.0095 (Appendix L). Degradation is linear and predictable, not catastrophic. So the corruption experiment is evidence that the gain comes from a *true* constraint.

7.5 C3. GUARANTEE: EXACTLY WHEN STRUCTURED DECODING WINS OR LOSES

Theorem 3 (Cardinality regime) *Let independent thresholding predict m positives while the constraint requires K, and let $q = |m - K|$ be the number of boundary items the constraint forces to flip, of which C correct an error and $B = q - C$ break a correct prediction. Then*

$$\mathrm{Acc}_K - \mathrm{Acc}_{\mathrm{raw}} = \frac{2C - q}{n}, \qquad |\mathrm{Acc}_K - \mathrm{Acc}_{\mathrm{raw}}| \le \frac{|m - K|}{n}. \tag{5}$$

Constrained decoding therefore improves accuracy if and only if $C > q/2$. The constraint dictates *how many* boundary items must move, not that moving them helps. Two predictions follow. Where the operating point nearly matches prevalence, q is small and the method provably cannot change much: VAU at $2\times$ has m=206, K=204, q=2, bound ± 0.0046, measured 0.0000. Where it is far off, the forced flips can go badly: at $8\times$, q=33 with C=10, B=23, yielding exactly the -0.0296 measured. The identity holds to machine precision in all four cells (Appendix C).

Corollary 4 (Regret of a misestimated cardinality) *For top-k decoding on a fixed score vector, $\left|\mathrm{Acc}(\hat{K}) - \mathrm{Acc}(K)\right| \le |\hat{K} - K|/n$.*

The two decodings differ on exactly $|\hat{K} - K|$ items, each moving accuracy by at most $1/n$. The bound is tight and, unlike Theorem 3, *actionable in the wrong direction*: it converts a tolerance on accuracy into a precision requirement on $\hat{K}$, which §7.7 uses to decide when to stop acquiring labels.

Under $\Pr[y \mid x] \propto \exp(s(x)^\top y)\, \mathbf{1}[Ay = b]$ constrained argmax is Bayes-optimal for exact-match loss; we note this only to delimit the claim. What is not standard is reading the constraint as an identification statement, which turns a decoder into a measurement instrument.

7.6 TWO INSTANTIATIONS

Pairwise complement (N=2, K=1): the benchmark guarantees exactly one YES per question pair, so $A = [1, 1]$ and $b = 1$. Decoding reduces to comparing $d_i = \log p_i(\mathrm{YES}) - \log p_i(\mathrm{NO})$ within the pair. Any shared answer bias cancels: $(d_1 + \beta) - (d_2 + \beta) = d_1 - d_2$.

Global cardinality (N=439, K=204): the benchmark documents its test composition, so $A = \mathbf{1}^\top$ and decoding is top-K. The operative change replaces absolute thresholding with structured argmax; Theorem 3 says precisely how much that can help or hurt. K derives from the evaluation set's composition, so this instantiation is a *prevalence-informed transductive measurement instrument*, not a deployment algorithm.

7.7 D. ACQUISITION: HOW MANY TARGET LABELS THE CONSTRAINT NEEDS

The transductive caveat is the method's real limitation. Corollary 4 bounds the cost of a wrong constraint by $|\hat{K} - K|/n$, turning "estimate the prevalence well" into "estimate it to within εn". How many target labels does that take? Our own experiments falsify the zero-label requirement (Appendix N), so we drop it and acquire labels by rank stratum: sort by d_i, cut into L strata, draw 2–4 per stratum, estimate $\hat{K} = n \sum_\ell w_\ell \bar{y}_\ell$. Stratification makes this cheap, and, consistent with everything else here, cheap exactly when the ranking is good: at $\mathrm{AUC} = 0.998$ it cuts the $\hat{K}$ error $5\times$ over random sampling, at 0.618 nothing. The same labels give an applicability test for free. Let δ be the top-minus-bottom stratum positive rate, from queried labels only. A small δ means a flat

ranking: the items forced across the boundary are no better than coin flips, the empirical form of $C > q/2$ in Theorem 3. The regret bound adds this to active prior estimation and stratified sampling: the cost of being wrong becomes measurable in advance. Section 8.3 measures what this is worth; the answer is more modest than we first believed.

8 EXPERIMENTS

8.1 A CLEAN DEMONSTRATION, WITH THE BASELINES THAT MATTER

The transportation benchmark admits non-visual solutions (§6), so we use Roundabout-TAU: 334 videos, one shared question, no wording channel. The model has never seen it, and its operating point has collapsed: 3 of 334 called anomalous at prevalence 0.802. That gives $q = |m - K| = 265$ forced flips and, by Theorem 3, room for ± 0.79.

Table 2: Roundabout-TAU at 2× (n=334, prevalence 0.802), with the two controls a single accuracy number hides: a constant predictor, and a decoder that gets K but ignores the model.

Decoder	Acc	Bal. Acc	TPR	TNR	F1
Constant "anomalous" (majority)	**0.8024**	0.5000	1.000	0.000	**0.890**
Model's native threshold	0.2066	0.5056	0.011	1.000	0.022
Random top-K (knows K, ignores model)	0.6832	0.5006	—	—	—
BISI top-K (knows K, uses ranking)	0.7305	**0.5751**	0.832	0.318	0.832

What the number is, and what it is not. Structural decoding beats the native threshold by 52.4 points, exactly as Theorem 3 predicts via $(2C - q)/n$, C=220, B=45, q=265. Reporting it alone would repeat the error this paper is about, as two controls show. A constant predictor, blind to the video, scores 0.8024 against our 0.7305 and wins F1. A decoder knowing only K and picking at random reaches 0.6832: most of the 52.4 points come from K, not the model. The ranking adds only +0.047 accuracy and +0.075 balanced accuracy. Balanced accuracy is its one lead, 0.5751 vs. 0.5000, and even there we claim only the best decoder tested, not a state of the art. The ceiling is the ranking. Two protocols read it: AUC 0.618 after the model's own chain of thought, as here, and 0.586 from a single forward pass, as in Tables 1 and 3, on the same model and videos. We first recorded this as a +52.4 point gain without the majority baseline: the same error, made by authors looking for it.

On the same 847 pairs, structural decoding scores 0.8394 against 0.6700–0.7462 for the strongest permutation debiasing and two prior-subtraction estimators, at no extra forward pass, and only 4.4% of answers flip when options swap: the dominant nuisance is semantic, not positional. The cardinality family behaves the same (Appendices J and K).

8.2 GENERALIZATION ACROSS MODELS, ON THE CLEAN BENCHMARK

We repeated the Layer 1 and 2 audits on our system's base model and a different-generation model, same 439 videos, prompt and scoring protocol. All three rank far better than they decide, AUC .87–.93 against accuracy near .78, and all collapse on rule-based anomalies. Fine-tuning raises rule recall to 0.375 but leaves the semantic gap near +0.58, so structured decoding changes nothing here (q=2): a calibration that does not transfer. Table 11 and Appendices R and F give the per-model readings.

8.3 HOW THE ACQUIRED LABELS SHOULD BE SPENT

Given B target labels, why not fit a threshold? We ran it with the B labels as calibration, the rest untouched test, 200 splits; the answer surprised us.

Acquisition must be stratified, or the method fails outright. Uniform draws lose to threshold fitting everywhere (0.8653 vs. 0.8709 on VAU at B=64); stratified draws halve $|\hat{K} - K|/n$ and reverse it (0.8784 vs. 0.8706), flipped by acquisition alone.

Table 3: Stratified target labels spent three ways, untouched test, 200 seeds; budget dependence is in the text and Appendix P. Cardinality estimation helps only where the native operating point is misplaced. The B acquired labels are held out of the test set here, which is why the Roundabout row differs slightly from Table 2, where all 334 videos are scored (0.203 vs. 0.2066 at the native threshold; 0.706 vs. 0.7305 given the oracle cardinality).

		full prediction (B=64)				selective (B=8)	
Benchmark	AUC	native thr.	fitted thr.	top-$\hat{K}$	oracle K	rank rule	score rule
VAU	0.947	0.775	0.871	**0.878**	0.885	**0.920**	0.885
TAD-bench	0.986	0.916	0.928	**0.933**	0.938	**0.974**	0.942
Roundabout	0.586	0.203	**0.784**	0.709	0.706	0.749	**0.726**

The gain is small, and doing nothing is a strong baseline. With stratified acquisition, cardinality estimation beats threshold fitting by $+0.8$ to $+3.2$ points where the native operating point is badly placed, not where it is near-calibrated. Averaged over benchmarks and budgets the two are indistinguishable, 0.869 vs. 0.865. Our earlier $+9.3$ point claim came from decoding the full set including the labeled items against an unregularized threshold baseline, and survives neither correction.

What survives is a statement about rank versus scale. The defensible claim is selective: at matched coverage, a rank rule abstaining on the uncertain band beats a score rule where the ranking is good. It gains $+3.5$ points on VAU and $+3.2$ on TAD-bench at B=8, narrowing to $+0.7$ and $+0.8$ at B=64. On Roundabout, where the ranking carries no signal, the score rule overtakes as labels accumulate. Ranks transfer across domains and absolute scales do not, but that helps only when the ranking is worth transferring. The advantage survives certification: with a design/certification split and one Clopper–Pearson test, the rank rule keeps 1.2–1.7$\times$ the score rule's coverage at equal certified selective risk, with paired bootstrap CIs excluding zero in three of four cells. The regime is narrow: absolute certified coverage is 0.09–0.34, $\varepsilon = 0.05$ is uncertifiable at these budgets, and at B=32 the certificate cannot separate the arms. Appendix P gives the protocol and a cautionary negative: a grid-searched variant passes nominally yet breaks its risk bound on up to 100% of certified runs.

BISI makes five predictions that can fail, all tested and passing in Appendix D.

9 LIMITATIONS

The decoder does not beat the baselines that matter, and we do not claim it does. §8 reports the baselines it loses to: a constant predictor and a fitted threshold both beat it on Roundabout, and on VAU the effect is non-positive in all four budget cells. BISI is a prevalence-informed transductive *instrument* separating decision layer from ranking layer. It is not a detector, nor a repair that improves accuracy: on these benchmarks the operating point is not where the loss is.

All 18 candidate methods had a kill criterion and cost-matched control fixed before they ran and none survived (Table 12), several killed by the very diagnostic this paper proposes; Appendix O lists three claims we withdraw. The audit covers three models and five video benchmarks, minus so-tad as unmeasurable, and matching controls only measured covariates, so we claim control, not causal removal. Exploitability is interventional, since we build the attacker; reliance is only observational. We report duration-matched evaluation ($0.947 \rightarrow 0.952$, unchanged) and read reliance as no detectable dependence, not demonstrated independence. The rule-based subgroup rests on n=24, its only source 88% contaminated by the model's training corpus; the cardinality instantiation is transductive; and BISI guarantees invariance, not accuracy. The three layers are those our instruments separate, not all the ways a score can mislead, and the rule-based blind spot is measured rather than repaired.

10 CONCLUSION

Observed accuracy conflates acquisition shortcuts, visual ranking and decision placement, and separating them changes the sign and magnitude of conclusions. Measure the layer the claim is about.

REFERENCES

Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. UBnormal: New benchmark for supervised open-set video anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. Position: Medical large language model benchmarks should prioritize construct validity. In *International Conference on Machine Learning (Position Track)*, 2025. arXiv:2503.10694.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv:2110.01052*, 2021.

Andrew M. Bean et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks)*, 2025. arXiv:2511.04703.

Nataly Brukhim and Amir Globerson. Predict and constrain: Modeling cardinality in deep structured prediction. In *International Conference on Machine Learning (ICML)*, pp. 658–666, 2018.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jinwoo Choi, Chen Gao, Joseph C. E. Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *Advances in Neural Information Processing Systems*, pp. 851–863, 2019.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. In *Science*, volume 349, pp. 636–638, 2015.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *International Conference on Machine Learning (ICML)*, pp. 5988–6008, 2022.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *NeurIPS*, 2017.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT*, pp. 107–112, 2018.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning. Conditional probing: Measuring usable information beyond a baseline. In *Proceedings of EMNLP*, pp. 1626–1639, 2021.

Chao Huang, Benfeng Wang, Jie Wen, Chengliang Liu, Wei Wang, Li Shen, and Xiaochun Cao. Vad-r1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought. *arXiv preprint arXiv:2505.19877*, 2025.

Nathanael Jo and Ashia Wilson. Position: AI evaluations should be grounded on a theory of capability. In *International Conference on Machine Learning*, 2026. arXiv:2509.19590.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning (ICML)*, 2018.

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection—a new baseline. In *CVPR*, 2018.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of ACL*, pp. 3428–3448, 2019.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning (ICML)*, 2020.

Susan A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–366, 2003.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of *SEM*, pp. 180–191, 2018.

Farrukh Rahman. What does a temporal benchmark score measure? decomposing channel use in video vlm evaluation. *arXiv preprint arXiv:2607.12304*, 2026.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, 2018.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *ECCV*, 2020.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations (ICLR)*, 2020.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations (ICLR)*, 2024.

Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning. *arXiv preprint arXiv:2505.23504*, 2025.

USE OF AI ASSISTANCE

An AI assistant (Claude, Anthropic) was used throughout this work under the authors' direction: to implement and debug the experiment harness and the cache construction, to write the scripts that generate every table from the committed run records, to run pre-registered analyses on cached tensors, to check claims in this manuscript against those records, and to draft and edit prose, including this section. All experimental designs, pre-registration rules, verdicts and decisions about what the paper claims were made by the authors, who read and are responsible for every sentence and every number. No experimental result was produced or altered by an AI system without a committed record.

A QUALITATIVE EXAMPLES



Figure 2: Two failures and one correction, on the two benchmarks whose task format admits no textual channel. **(a)** A rule-based anomaly missed with near-certainty, which removes the gradient a confidence-based policy would need. **(b)** On an unseen domain the same model's operating point collapses: it calls 3 of 334 videos anomalous where prevalence is 0.802. **(c)** The identical score, decoded under the cardinality constraint, recovers the label. The italic line states the control that a single accuracy number hides.

B THE AUDIT AS A WORKFLOW

Each layer below has its own input requirement and its own pre-registered criterion; the last column states what the benchmark score can still be read to support once that layer returns, not what a practitioner must do.

Table 4: The audit as a workflow. Each row is a layer with its own input requirement and its own pre-registered criterion; the last column states what the benchmark score can still be read to support, not what a practitioner must do.

Layer	Needs	Observed regime	What the score can support
Shortcut detect / exploit (§6)	item metadata	near-deterministic channel	nothing about visual capability; disclose the floor
Acquisition control (§6)	covariate overlap	no overlap	the comparison is not identifiable here
Controlled ranking (§6)	matched evaluation	AUC near chance	ranking-limited: repairing the threshold cannot help
Decision diagnosis (§7)	a true $Ay = b$	large $q = \lvert m - K \rvert$	decision-limited: Thm. 3 gives the exact effect
Certified action (§8.3)	calibration labels	certificate fails	evidence insufficient for deployment at that risk

C NUMERICAL VERIFICATION OF THE PROPOSITIONS

All three empirically checkable propositions were verified against the measured scores before being written into §7.

Table 5: Theorem 2: the closed form for $A_{\text{raw}}(\beta)$ against measurement, on 847 complementary pairs at $2\times$. The two agree to five decimals at every β. $A_{\text{pair}} = 0.8383$ throughout, independent of β as the theorem requires.

β	-4	-2	-1	0	$+1$	$+2$	$+4$
closed form	.5024	.5360	.5584	.6499	.7456	.7226	.6092
measured	.5024	.5360	.5584	.6499	.7456	.7226	.6092

Table 6: Theorem 5: predicted $A(\varepsilon) = 0.8383 - 0.3383\,\varepsilon$ against measurement under same-label pair corruption.

ε	0	0.05	0.10	0.20	0.50
predicted	.8383	.8214	.8044	.7706	.6692
measured	.8394	.8264	.8111	.7801	.6761

The non-monotonicity of the measured row is itself the content of Theorem 2: accuracy peaks near $\beta = +1$ and falls on both sides, so an operating-point shift can raise or lower measured accuracy with no change in the underlying ranking, whose quality stays fixed at 0.8383.

The $8\times$ row is the informative one: the constraint forces 33 flips, but only 10 of them are corrections, so structured decoding loses by exactly the predicted amount. This is a regime, not a defect—the constraint fixes how many boundary items must move, not that moving them helps.

D FALSIFICATION TESTS IN FULL

BISI makes predictions that can fail. We test five of them here in full.

(i) The accuracy identity. For exactly-one-positive pairs, BISI accuracy must equal the within-pair ranking rate $\text{PairRank} = \Pr[d_{\text{YES}} > d_{\text{NO}}]$. Across 7 sources $\times$ 2 budgets the two quantities agree in 14 of 14 cells, the single deviation being one tied pair, which the theorem excludes.

(ii) Translation invariance. Adding a constant c to every score leaves BISI output unchanged by Theorem 1. Empirically, thresholding swings between 0.500 and 0.723 for $c \in [-10, 10]$ while BISI is constant at 0.8394; the same holds for the cardinality family (0.8861 throughout). Note that the best achievable global recalibration (0.7226) still trails BISI by 11.7 points, so what is removed is not only a global offset.

(iii) Constraint corruption. If the benefit came from label leakage rather than from the constraint being true, corrupting the constraint would not hurt. Pairing same-label questions for a fraction f of pairs degrades accuracy as $0.839 \to 0.826 \to 0.811 \to 0.780 \to 0.676$ for $f = 0, 0.05, 0.10, 0.20, 0.50$, matching the predicted $0.839 - 0.34f$ pointwise.

(iv) Structural inventory. We enumerate which constraints are real design invariants rather than learned correlations. The complementary-pair structure holds in 3,666 of 3,669 videos (three annotation errors) and is preserved in the official evaluation split; the multiple-choice exactly-one-correct

Table 7: Theorem 3: the cardinality identity on VAU-Bench ($n = 439$, $K = 204$). m is the number of positives predicted by independent thresholding, $q = |m - K|$ the forced flips, C and B the corrections and regressions among them. The identity $(2C - q)/n$ matches the measured change exactly in all four cells.

Budget	m	q	C	B	$(2C - q)/n$	measured Δ
$1\times$	200	4	1	3	$-.0046$	$-.0046$
$2\times$	206	2	1	1	.0000	.0000
$4\times$	213	9	4	5	$-.0023$	$-.0023$
$8\times$	237	33	10	23	$-.0296$	$-.0296$

constraint is already used implicitly by argmax; the remaining task types carry no hard structural constraint we could identify.

(v) Cross-family and cross-model generality. §8.2 and the cardinality instantiation above.

E DIAGNOSIS OF THE DEBIASING COMPARISON

Why permutation methods underperform here, and where they do not. Swapping the order in which the two options are named changes the model's answer on only 4.4% of questions: there is little positional bias to remove. The dominant distortion is a semantic preference for NO, which permutation cannot address because it is invariant to option order. Prior-subtraction methods do better precisely because they *estimate* that semantic offset ($\hat{\beta} = -1.31$), and they recover most of the way; structural decoding reaches 0.8394 against their 0.7462 by not needing to estimate it at all.

This is not a general claim that permutation debiasing is ineffective. On the multiple-choice half of the same benchmark, where the option positions carry real degrees of freedom, averaging over all four cyclic permutations improves accuracy from 0.7532 to 0.7769 ($+0.024$, CI $[+0.006, +0.041]$). The two nuisance types call for different instruments, and identifying which one dominates is itself part of the audit.

F WHAT FINE-TUNING CHANGED, AND WHAT IT DID NOT

The base model recalls rule-based anomalies at 0.042—one of the twenty-four—and adaptation raises this to 0.375. Fine-tuning therefore does teach the rule-based category. What it does not do is close the gap, which stays near $+0.58$: the improvement is in absolute level, not in the structure of the deficit. Adaptation also moves the operating point. The untuned models under-report anomaly badly (0.248 and 0.282 against a prevalence of 0.465), while the fine-tuned model sits at 0.469, almost exactly on prevalence. This is worth naming because it explains a null result of ours: structural decoding changes nothing for this model on VAU-Bench precisely because adaptation happened to calibrate it, leaving $q = 2$ (Theorem 3). The calibration is a byproduct of fine-tuning on this distribution, not a property that transfers—on Roundabout-TAU the same model labels 3 of 334 videos anomalous.

G LEAKAGE PROTOCOL, SCOPE, AND REPAIR

Table 8: Three non-visual channels in the transportation benchmark. Each row is a system that never receives a video frame. The rightmost column is the best video model we measured on the same subtask.

Subtask	Channel	Predictor	Accuracy	Video model
Binary	question order	"first is YES"	**0.9667**	0.656 / 0.839
Binary	question wording	TF–IDF, leave-one-source-out	**0.9008**	0.656 / 0.839
Multiple choice	option wording	TF–IDF, video-disjoint CV	**0.7792**	0.776–0.799

We verify all three on the released training split; the official evaluation split preserves the same generating structure but has redacted answers, so we can confirm the structure without confirming the leakage, and we flag that asymmetry rather than assume the worst. Order leakage is repaired by randomizing pair order at evaluation time, which returns that heuristic to chance (0.5014) and leaves model scores untouched, because our protocol presents each question in isolation. Wording leakage has no such repair: it is a property of how the items were generated. We therefore report a text-only floor alongside every number on this benchmark, and we recommend that others do the same. We have notified the organizers.

H RESIDUAL SIDE-CHANNEL CAPACITY: ESTIMATOR AND VALUES

The estimator behind Table 1's last column differences held-out losses of two cross-fitted predictors: $\hat{h}(S)$ on the frozen model score alone and $\hat{g}(S, Z)$ on score plus the six header features, both unweighted logistic models (the attack's class-balanced weighting is kept for AUC but deliberately *not* used here, since balancing distorts the probabilities a proper-loss risk difference needs). Five stratified folds, averaged over 10 fold seeds; intervals are paired bootstrap over items (10,000 resamples) on the per-item loss differences.

Table 9: Residual side-channel capacity of the header channel above each frozen model score. $H(Y)$ is label entropy; $\mathcal{R}(S)$ estimates $H(Y \mid S)$, the uncertainty the model's score leaves open; RSCC is the part of that residue the header closes. The log-loss plug-in is one-sided: a positive value exhibits an attacker, a near-zero value means none was found in the logistic class.

Benchmark	$H(Y)$	$\mathcal{R}(S)$	$\mathrm{RSCC}_{\log}$ [95% CI]	share of $\mathcal{R}(S)$	$\mathrm{RSCC}_{\mathrm{Brier}}$
VAU	0.691	0.293	0.123 $[+0.088, +0.160]$	42%	0.038
TAD-bench	0.623	0.154	0.012 $[-0.004, +0.028]$	8%	0.002
Roundabout	0.497	0.494	0.278 $[+0.220, +0.338]$	56%	0.099

Units are nats per item. The population limit of the log-loss quantity is $I(Y; Z \mid S)$, zero iff $Y \perp Z \mid S$; the Brier form is $\mathbb{E}[(\eta_{SZ} - \eta_S)^2]$, the mean squared movement of the optimal label probability when the channel is added. Both orderings agree. On Roundabout the model's score is nearly uninformative ($\mathcal{R}(S) \approx H(Y)$), so capacity there is close to the channel's unconditional information; on VAU it is strictly residual—the header still closes 42% of what a 0.947-AUC score leaves open—and on TAD-bench the interval contains zero, matching the $+0.002$ AUC null. The anonymous supplement reproduces this cell from cached scores on CPU in seconds.

Relation to usable information. The restricted-class risk difference is the conditional $\mathcal{V}$-information of Hewitt et al. (2021), with $\mathcal{V}$ the logistic class over the concatenated inputs, in the framework of Xu et al. (2020); Ethayarajh et al. (2022) use the unconditional form to measure dataset difficulty. We inherit the quantity and its one-sidedness. What this paper adds is the object it is pointed at and the role it plays: Z is acquisition metadata that no capability claim licenses, S is a frozen incumbent's score rather than a representation being probed, and a positive value is read as benchmark side-channel capacity—an attacker exhibited above the strongest available model—rather than as information content of a representation. The audit's control arm (propensity matching on the same covariates) then asks the separate question of whether the incumbent itself relies on the channel, a distinction the probing literature does not need but a benchmark audit cannot omit.

I INDEPENDENT VISUAL CHANNEL: FULL RESULTS

The preceding result invites an obvious extension: if the language model's head discards evidence, supply evidence from outside the language model. We tested this in the strictest setting we could construct—estimating the target prevalence K with no target labels at all—using a Kinetics-pretrained 3D convolutional encoder, whose inductive bias is motion rather than semantics. Class prototypes fitted on three source datasets and transferred by soft mixture return $\hat{p} = 0.905$ against a true 0.802, which looks like a near miss until one notices that the same prototypes score $\mathrm{AUC} = 0.4474$ on the target: the estimate is expectation maximization collapsing onto one class, not a signal. A source-trained linear probe transferred to the target returns $\hat{p} = 0.284$, and black-box shift estimation (Lipton et al., 2018) corrects it to the same value.

The informative number is the upper bound. Granting the method every target label and training the probe on the target itself yields $\mathrm{AUC} = 0.7580$ under random splits—above the 0.618 the video–language model achieves—but 0.4860 once scenes are disjoint, against 0.8381 for scene identity alone. The encoder never saw the anomaly; it recognized the intersection. Repeating the experiment with an appearance encoder of an orthogonal inductive bias (DINOv2 ViT-S/14) reproduces the pattern more sharply still: 0.8711 under random splits—higher than the scene-identity baseline itself, which is what the shortcut hypothesis predicts—collapsing to 0.5358 under scene-disjoint splits.

Table 10: Published debiasing families against structural decoding, on the same 847 complementary pairs at 2×. Permutation methods remove a bias that is largely absent here; the dominant nuisance is semantic rather than positional.

Decoder	BCQ acc	Labels	Extra forwards
Model, independent threshold	0.6558	no	0
Permutation averaging (strongest form)	0.6700	no	1×
Prior-subtraction debiasing ($\hat{\beta} = -1.31$)	0.7462	no	1×
Prior-matching / label-shift ($\hat{\tau} = -1.37$)	0.7414	no	0
BISI (structural)	**0.8394**	no	**0**

So the failure is not that zero-label transfer is hard, and not that we chose the wrong encoder: this concept has no scene-independent encoding in off-the-shelf visual representations, and no estimator built on them can succeed. This is the sharpest form of the identifiability limit in §9, and the only one where we can locate the obstruction in the representation rather than in the algorithm.

J PUBLISHED DEBIASING BASELINES IN FULL

A natural objection is that label-free debiasing already exists. We implement two families and give them the advantage: rather than estimating a prior from a subset as is usual, we run *every* permutation for *every* item, which is the strongest form of permutation-based debiasing and costs one extra forward pass per item.

Only 4.4% of answers flip when the options are swapped, so there is little positional bias to remove; the dominant distortion is a semantic preference for NO, to which permutation is invariant. Prior-subtraction *estimates* that offset and recovers most of the way; structural decoding reaches 0.8394 against its 0.7462 by not needing to estimate it. On the multiple-choice half, where positions do carry degrees of freedom, permutation helps ($0.7532 \to 0.7769$)—identifying which nuisance dominates is part of the audit (Appendix E).

K SECOND CONSTRAINT FAMILY IN FULL

On VAU-Bench with a global cardinality constraint, BISI is again exactly invariant to score translation (accuracy 0.8861 for all $c \in [-10, 10]$, versus 0.813–0.886 for thresholding), and behaves as the theory predicts: where the native operating point is already aligned (predicted positive rate 0.469 versus prevalence 0.465) it changes nothing ($0.8861 \to 0.8861$). We state as an empirical observation, not a theorem, that the benefit is largest where operating-point bias is severe and vanishes where it is not.

Measured on this bias-invariant scale, the budget curve reads $0.8633\ (1\times) \to 0.8861\ (2\times) \to 0.8497$ ($8\times$): the $8\times$ condition is 3.6 points *worse* than $2\times$, whereas on the raw scale the two are indistinguishable. This reproduces the saturation conclusion of §5 with an entirely different instrument.

L GRACEFUL DEGRADATION: STATEMENT AND MEASUREMENTS

Theorem 5 (Graceful degradation) *If a fraction ε of pairs violate the exactly-one-positive constraint by containing two identical labels, then any decoder forced to emit one positive per pair scores exactly $1/2$ on those pairs, so $A(\varepsilon) = r - \varepsilon\,(r - \frac{1}{2})$, where r is the accuracy on uncorrupted pairs.*

With $r = 0.8383$ this predicts $A(\varepsilon) = 0.8383 - 0.3383\,\varepsilon$; measured values at $\varepsilon \in \{0, .05, .10, .20, .50\}$ are $0.839, 0.826, 0.811, 0.780, 0.676$ against predicted $0.838, 0.821, 0.804, 0.771, 0.669$. The method degrades linearly and predictably rather than catastrophically, and—more importantly for §8—this is why the corruption experiment is evidence that the gain comes from the constraint being true.

M POLICY LEARNING AND REGRESSION-FREE UPDATES

The failure of our routing candidates instantiates a known distinction: optimal individualized treatment rules depend on the *sign* of a conditional effect (Murphy, 2003), and estimating the effect accurately does not imply recovering that sign. Deferring to a distinct expert rather than to more of the same model (Mozannar & Sontag, 2020) remains open here; our monotonicity-training candidate is a budget analogue of positive-congruent training (Yan et al., 2021) and underperforms ordinary fine-tuning under a matched-step control.

N TWO REVISED CONCLUSIONS IN FULL

Representation versus head, and the independent channel, in full. We originally reported that a hidden-state ranker beats structural decoding (0.9599 vs. 0.8442 leave-one-source-out); two controls withdraw it—a plain classifier given the same labels matches the ranker ($\Delta = -0.0035$, CI $[-0.013, +0.006]$), and the like-for-like probe with the head's own target is *negative* ($0.7836/0.8246$ vs. native $0.7882/0.7745$, inflating to $0.925/0.948$ under random splits). Evidence from outside the language model fares no better: zero-label K estimation from a Kinetics 3D encoder fails for every estimator, and the supervised upper bound explains why—0.7580 under random splits, 0.4860 once scenes are disjoint (DINOv2: $0.8711 \rightarrow 0.5358$). The encoders never saw the anomaly; they recognized the intersection (Appendix I). This is scene bias in the sense of Choi et al. (2019), measured on a frozen encoder rather than trained against.

Oracle versus actionable headroom (strengthened). Optimal routing depends only on $\operatorname{sign} \mathbb{E}[Y_{8\times} - Y_{2\times} \mid X]$, weakly predictable on the raw scale (AUC 0.712) yet never deployable. On the corrected scale it is not predictable at all (0.4898 ± 0.03) and the strongest cross-fitted policy turns actively harmful (-0.0085 against $+0.0031$) while oracle headroom remains ($+0.0275$): the residual signal was itself an operating-point artifact.

Training substitutes for inference compute (weakened). Budget-matched continued training at $2\times$ reaches 0.9290 across seeds against 0.9256 for the $8\times$ teacher. We originally read the single-seed difference ($+0.008$) as "parity or better", but the seed-to-seed range is 0.0093, exceeding it. The supportable claim is parity at one quarter of the inference cost, not superiority—the one place where adding seeds changed a conclusion of ours.

O RETRACTED CLAIMS AND REPRODUCIBILITY

Claims we retract. We also withdraw three claims we made earlier in this project. A spatial–temporal super-additivity effect ($I = +0.167$, $n = 24$) collapsed to $I = +0.009$ on $n = 116$. A claim that models can report what they cannot perceive was contradicted by our own mediation test (0 of 21 UNCLEAR$\rightarrow$definite transitions at higher budget, against 8 of 187 in a control group). A mechanistic claim that the missing evidence never survives tokenization was not supported by any measurement we could perform.

Reproducibility note. Every learned variant in this project was evaluated against independently reproduced fixed baselines under pre-registered accept/kill criteria; three results that initially appeared positive were withdrawn when they failed cross-verification against those baselines. Across $14{,}121$ duplicated evaluations, 69 predictions (0.49%) differed across hardware runs, concentrated near score ties under bf16; all aggregate conclusions were unchanged. Deterministic post-processing on fixed scores—the audit, the propositions and BISI—has no seed variance to report. For the one stochastic result in the main text we ran additional seeds, and the seed range exceeded the effect we had claimed (Appendix N).

P CERTIFIED RANK-SELECTIVE PREDICTION

The comparison in §8.3 is empirical; here both selection rules receive an identical finite-sample certificate and are compared on the deployment-relevant axis: *at the same certified selective risk*

ε, which rule answers more? The certificate is a standard distribution-free risk control instance in the learn-then-test family (Angelopoulos et al., 2021), applied to selective prediction (Geifman & El-Yaniv, 2017); we claim the comparison, not the machinery. The adaptivity failure reported below — searching band widths against shared certification labels manufactures nominally valid certificates — is the classical adaptive-data-analysis effect (Dwork et al., 2015), and the fixed-sequence correction is that literature's standard remedy.

Protocol. The B target labels are split. The design half (stratified by rank) fixes the policy: the prevalence estimate $\hat{K}$ and an abstention band for the rank arm, a fitted threshold t and a score band for the score arm, with the band width chosen so the design-estimated selective risk is at most $\varepsilon/2$. The certification half (drawn uniformly, untouched by design) performs one binomial test: the policy deploys only if the Clopper–Pearson upper bound (at $\delta = 0.1$) on the error rate of its answered region is at most ε; otherwise the system refuses. The guarantee is joint: $\Pr[\text{deploy} \wedge \text{risk} > \varepsilon] \leq \delta$, which holds empirically in every cell we ran (worst case 0.01 against a budget of 0.10).

Results. At B=64, over 200 splits: on VAU the rank arm answers 9.1% against 5.3% at ε=0.10 and 20.6% against 12.1% at ε=0.15; on TAD-bench, 17.7% against 11.0% and 33.9% against 28.5%. Paired bootstrap CIs exclude zero in three of four cells ($+0.038$ $[+.006, +.070]$; $+0.085$ $[+.039, +.132]$; $+0.067$ $[+.021, +.114]$; $+0.055$ $[-.004, +.113]$). At B=32 the certificate's resolution is insufficient to separate the arms, and ε=0.05 is uncertifiable at any budget here: a zero-error Clopper–Pearson bound on the achievable certification sample already exceeds it. The binding constraint on certified selective prediction at these budgets is the certification sample, identically for both arms. On Roundabout both arms refuse in $\approx 99\%$ of splits—the certified system correctly declines the domain whose ranking is uninformative.

A cautionary variant. Our first implementation searched a 30-point grid of band widths for the smallest passing the same test, reusing the certification labels across the search. It reports nominally valid certificates and appears to enlarge the rank advantage to $+17$ points of coverage—yet on Roundabout its certified policies violate their risk bound on 90–100% of certified runs, and on VAU at ε=0.10 on 22%. Selection over shared calibration data manufactures certificates the same way benchmark shortcuts manufacture capability, and we report the failure because the pre-registered validity check is what caught it. A fixed-sequence correction restores validity but starts testing where the certification sample is smallest and certifies almost nothing; the single-test protocol above is both valid and the more powerful of the two.

Q GROUNDING AVAILABILITY IS NOT EVIDENCE SUFFICIENCY

A natural repair for the rule-based blind spot of §4 is to add an open-vocabulary grounding module. We probed this on the 297 Roundabout clips (79 direction/maneuver violations, 152 collisions, 66 normal) with a frame-level grounder (GroundingDINO) at 64 frames per clip, linking detections into motion tubes by center distance.

Probe layer	Result	Interpretation
Concept grounding ("wrong direction")	0.407	single frames do not encode motion predicates
Motion-tube recovery	94% of clips	moving-object evidence is largely available
Geometry-free rule verification	0.432	grounded motion alone is insufficient

The verification score is violation-vs-collision AUC on same-camera clips, which excludes the metadata channel by construction. The geometry-free rule—a tube whose angular direction opposes the within-video majority—is not merely weak but directionally wrong: it fires on 48% of normal clips against 30% of violations, because entry and exit traffic renders angular direction meaningless without lane structure. The failure occurs *after* object localization and *before* rule verification, so a natural hypothesis is that the missing variable is scene geometry—lane topology, permitted flow—making the verifier $V(h, E_h, G)$ rather than $V(h, E_h)$. We tested that hypothesis with an oracle and it did not survive, in a way worth reporting because the first look supported it. Estimating each scene's center, ring band and canonical rotation from pooled unlabeled footage raises violation-vs-collision AUC only from 0.432 to 0.505, and substituting the benchmark's own lane-semantic

annotations (controlled-vocabulary entry/circulating/exit locations per object) reaches 0.878 (0.781 scene-disjoint)—which reads as confirmation until the feature set is ablated. Object type and participant count alone reach 0.855 scene-disjoint; lane state alone reaches 0.617; and adding oracle lane state on top of machine-derived motion statistics *lowers* AUC from 0.728 to 0.716. The oracle's apparent success was composition, not geometry. What the labels here actually discriminate is who participates—collisions involve two or more agents by definition—and the geometric variable we set out to identify is a dilution term. This does not show that scene geometry is irrelevant to rule violations in general; it shows that on this benchmark no task exists for which geometry is both necessary and cleanly measurable, and that identifying a missing variable requires ablating the oracle that appears to confirm it. Evidence availability is not evidence sufficiency, and an oracle is not an identification.

R CROSS-MODEL AUDIT TABLE

Table 11: The blind spot is a property of the model class: all three models rank far better than they decide, and all three collapse on rule-based anomalies.

Model	Budget	YES rate	AUC	Rule recall	Dramatic recall	Gap
Vad-R1 (fine-tuned)	$1\times$ / $2\times$	.456/.469	.902/.929	.292/.375	.922/.950	$+.63$/$+.58$
Qwen2.5-VL (base)	$1\times$ / $2\times$	.248/.253	.899/.921	.042/.042	.583/.606	$+.54$/$+.56$
Qwen3-VL-8B	$1\times$ / $2\times$	.282/.269	.867/.865	.125/.167	.644/.594	$+.52$/$+.43$

S FULL METHOD VERDICTS

Every candidate below was specified with a kill criterion before its evaluation and compared against a cost-matched control. "Δ" is the difference from that control on the metric named in the kill criterion. We list them in the order they were run, which is also the order in which they narrowed the hypothesis space.

Two observations survive the list. First, the shared failure mode: every gain a sophisticated method sought to capture was already available from a cheaper control—a threshold, a zero-compute answer editor, or ordinary fine-tuning. Second, several candidates were killed by the very diagnostic this paper proposes. The semantic-threshold method appeared to gain $+11.6$ points on the rule-based subgroup, but its test-time false-positive rate had drifted to 0.119 against a budget of 0.10: the gain was purchased with the operating point, exactly the error §5 describes.

Table 12: All 18 method candidates with pre-registered verdicts. Cost is in units of the $1\times$ visual budget where applicable. Bracketed intervals are 95% bootstrap CIs on paired differences.

Candidate	Kill criterion	Result	Verdict
Family 1: route more visual compute to selected inputs			
Probe-Route (probe answers → escalate)	must beat constant policy at matched FPR	identical to constant $1\times$ at FPR$=$ .10 (.7402/.2083 both)	killed
Four-way action router	4-way oracle $>$ 2-way oracle	oracle gap $+.0137$ overall, $+.042$ rule (CI touches 0); rests on 1 video	killed
Counterfactual instability (CIR)	instability predicts benefit sign	Δ alone AUC .395; $\rho(s_1, s_2) =$ .91	killed
Fix-or-Break dual head	beat FIX-only at matched escalation	.542 vs .583 rule recall at $2\times$ escalation	killed
See-or-Think	THINK must add oracle headroom	$+.0000$; $1/31$ repaired; lowers 3-way oracle	killed
Confidence-gated escalation	beat constant $8\times$ at matched cost	$\Delta = +.0000$, CI $[.0000, .0000]$	killed
Learned router (LR/GBM, 7 feats)	beat constant $8\times$	$\leq$ constant at every escalation rate	killed
Family 2: build a better decision score			
Budget-contrastive scoring	CV AUC $\geq .94$	$.9291 \rightarrow .9067$; ΔAUC $-.022$ CI $[-.038, -.011]$	killed
Semantic threshold calibration	Δrule$>$ 0 within FPR budget	$+.116$ but test FPR $.119 > .10$ budget; CI $[-.50, +.67]$	killed
Family 3: abstain instead of escalating			
Budget-perturbation consistency	AURC gain on both datasets	AIC $-.0042$ (better) but VAU $+.0195$ (worse); direction flips	killed
Family 4: acquire new evidence			
Falsify-Then-See @128p	sign-AUC $>$ .75, then deployment value	gate passed (.785); deployment $+.006$ over free editor, n.s.	not deployable
Falsify-Then-See @512p	hi-res must exceed 128p	$.713 < .779$; CI $[-.078, -.058]$	killed
Multi-scale fusion	beat single-scale critic	$+.0072$, CI $[-.014, +.021]$	killed
Permutation debiasing	positional bias must exist	$\chi^2 = 4.0/6.3$ (n.s.); bias is semantic	killed at precheck
Family 5: change training			
Budget-monotone training (v1)	NFR halved, PFR $\geq .8\times$	vs matched control: Δacc $-.0226$ CI $[-.037, -.009]$; NFR worse	killed
Selective monotone training (v2)	ceiling must justify run	control leaves 9 BREAK; max gain $+1.2\%$; no positive effect to target	not run
Signed cross-budget distillation	$\geq$ 1pp over matched CE	$+.0040$, CI $[-.004, +.013]$	killed
Family 6: learn only in the identifiable quotient space			
Quotient residual network	$\geq$ 1pp over parameter-matched ranker with full information	$-.0024$ CI $[-.012, +.007]$; shared component predicts source at .914 and the quotient removes it (.263), but removal does not aid transfer	killed