When Does an Observable Lose Failure Evidence? Exact Ranking Losses for Runtime Detection in Vision–Language–Action Policies

Anonymous authors
Paper under double-blind review

Abstract

No runtime failure detector for a vision–language–action (VLA) policy can beat the Bayes-optimal ranking of the observable it reads. This ceiling is standard; what a practitioner needs to know is what separates it from the score actually deployed. We decompose that gap exactly into a representation loss from compressing the observable (for example, into sparse-autoencoder atoms), a projection bias from fitting a misspecified likelihood family, and an estimation term, and we characterize each. A compression is lossless if and only if it preserves the success posterior; otherwise its AUROC cost has an exact form and is bounded by the missing conditional information with a sharp constant $\sqrt{2/3}$. Ignoring the direction of time costs at most $\sqrt{(\Sigma_1 + \Sigma_0)/12}$ AUROC, where Σ_y is the path-reversal divergence of outcome class y; the constant is sharp, yet irreversibility does not identify this value, and fully reversible paths can separate the outcomes perfectly. Low-order Markov likelihood projections can erase all evidence, and score consistency yields AUROC consistency exactly when the limit ties only states with equal likelihood ratio, a condition that a consistent maximum-likelihood fit can violate. For evaluation, early termination adds an exact bonus to prefix AUROC, and without structural assumptions no finite-sample procedure can certify a low ceiling. On seven LIBERO simulation interfaces spanning OpenVLA, OpenVLA-OFT, and $\pi_{0.5}$ (7,000 report rollouts; no physical robot), episode length alone reaches AUROC 0.9996–1.0000 as the termination identity predicts, other atom-level families exceed the pre-specified Grammar-LLR witness by up to 0.42 on the same rollouts, and the ordering of atom-level and hidden-state witnesses changes across interfaces. Project page: `https://anonymous.4open.science/r/experience-94DE`

1 Introduction

Vision–language–action (VLA) policies map camera images and a language instruction to robot actions (Brohan et al., 2022; Zitkovich et al., 2023; Driess et al., 2023; Team et al., 2024; Kim et al., 2024; Black et al., 2024). Because they still fail on some episodes, deployment needs a runtime failure detector: a score, computed while the episode runs, that decides when to stop the robot or hand control to a human. Such detectors use uncertainty estimates, embedding novelty, or the policy's internal features (Xu et al., 2025; Römer et al., 2025; Yu et al., 2025; Gu et al., 2025), and they are compared by AUROC, the probability of ordering a random success and a random failure correctly (Hendrycks & Gimpel, 2017; Lee et al., 2018; Fawcett, 2006).

Yet a detector's AUROC depends on more than the detector. Every detector reads an observable of the episode, such as the policy's hidden states or sparse-autoencoder (SAE) atoms computed from them, and ranks outcomes only as well as that observable allows. With OpenVLA on LIBERO-Spatial, a likelihood-ratio score on SAE atoms beats the SAFE hidden-state detector (Gu et al., 2025) (AUROC 0.983 vs. 0.967). With the same policy on LIBERO-Goal, the order reverses (0.849 vs. 0.997). This flip raises our question: is one detector worse, or do the two read different evidence?

The standard answer settles only part of this question. Once the policy, horizon, atomizer, task family, and evaluation law are fixed, no detector can beat the Bayes likelihood-ratio ranking of the

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.

observable's two outcome-conditioned laws (Neyman & Pearson, 1933; Fawcett, 2006; Menon & Williamson, 2016). This ceiling bounds every detector by one number, so it cannot say whether a low AUROC comes from weak evidence or from a detector that uses the evidence badly. The open questions lie between the ceiling and the deployed score: which compressions lose evidence, how much evidence lies in the direction of time, why a consistently estimated score can still rank badly, and what prefix AUROC measures when successful episodes stop early.

This paper characterizes what separates a deployed detector from the ceiling of the observable it reads. More precisely, we split the gap exactly into a representation loss from compressing the observable, a projection bias from fitting a misspecified likelihood family, and an estimation term from finite data (Figure 1). Each term calls for a different remedy: a new observable, a richer family, or more data. It turns out that each has a sharp answer. A compression loses nothing if and only if it keeps the success posterior; otherwise its loss has an exact form and is bounded by the information it discards (Theorem 2). Ignoring time direction costs at most $\sqrt{(\Sigma_1 + \Sigma_0)/12}$ AUROC, where Σ_y is the path-reversal divergence of class y; the constant is sharp, yet irreversibility alone does not fix the value (Theorem 4). A low-order Markov family can erase evidence that separates the outcomes perfectly (Proposition 2). A consistent score yields a consistent AUROC exactly when its limit ties only states with equal likelihood ratio, which maximum likelihood can violate (Theorem 5). For evaluation, early stopping adds an exact bonus to prefix AUROC (Proposition 3), and without structural assumptions no finite-sample procedure can certify a low ceiling (Proposition 4).

The proofs rest on one classical identity: the ceiling is one half plus the mean absolute difference of the success posterior (Lemma 1). From it, Pinsker's inequality bounds compression; splitting each path into its reversal orbit and its direction reduces time direction to the same bound; a perturbation of tied states gives rank stability; and the two-point argument of Antos et al. (1999) gives the impossibility result. We read these results on seven LIBERO simulation interfaces spanning OpenVLA, OpenVLA-OFT, and $\pi_{0.5}$ (Kim et al., 2024; Liu et al., 2023; Kim et al., 2025; Physical Intelligence et al., 2025), three primary and four extension, each under a locked protocol (7,000 report rollouts; no physical robot). Episode length alone reaches AUROC 0.9996–1.0000 at the full horizon, as the termination identity predicts. Two likelihood families on the same atoms differ by up to 0.42 AUROC, entirely through the projection and estimation terms. We report each fitted score as a *witness*, a lower bound on its observable's ceiling, with paired intervals; by Proposition 4, we estimate the ceiling only at the first step.

Ultimately, a detector's AUROC describes an interface as much as a detector. The opening flip is then no paradox: the two scores read different observables whose ceilings need not be ordered, and either may sit far below its own. Comparing detectors therefore means reporting the interface and asking which term a gap comes from.

2 Related work

We position the results against three lines of work.

Runtime failure detection for VLA policies. FAIL-Detect calibrates uncertainty signals using successful rollouts only (Xu et al., 2025); FIPER combines embedding novelty with action uncertainty (Römer et al., 2025); ARMADA connects online detection to shared control (Yu et al., 2025); and SAFE learns task-agnostic internal-feature scores (Gu et al., 2025). SAFE also observes that LIBERO failures always reach the maximum length and truncates evaluation to remove this shortcut; Proposition 3 quantifies exactly what that shortcut adds. These works improve the detector; we characterize what separates a detector from the ceiling of the observable it reads.

Ranking theory and comparison of experiments. Bayes-optimal bipartite ranking and its risk representations are developed in Menon & Williamson (2016), surrogate regret bounds for AUROC in Agarwal (2014), and the comparison of experiments, under which a garbled observation is never more informative, goes back to Blackwell (1953). Lower bounds for estimating Bayes error are given by Antos et al. (1999). We build on these tools; our results concern the specific losses that arise when VLA activations are compressed into atoms, reversed in time, and scored by low-order likelihood projections.

Information-theoretic views of policies and agents. Information-theoretic tools have been used for representation learning, inverse-RL lower bounds, embodiment, human–robot communication, and deployment diagnostics (Tishby & Zaslavsky, 2015; Komanduru & Honorio, 2021; Pitti et al., 2025; Obayashi et al., 2025; Reid et al., 2025). Our object is narrower: the outcome-conditioned laws of a fixed observable.

3 PRELIMINARIES

This section fixes the observable, its outcome-conditioned laws, and the analysis distribution that all later results share. We fix a finite task set $\mathcal{T}$, a VLA policy π, and a horizon T. One rollout produces residual-stream activations $h_{1:T}$ at a fixed layer and an outcome $Y \in \{0, 1\}$, where $Y = 1$ denotes success. The evaluation interface $\mathcal{I} = (\pi, T, \sigma, \mathcal{T}, \mu)$ collects the policy, horizon, atomization map, task family, and the law μ over tasks, initial states, and environment randomness.

Atom trajectories and termination. Let $E : \mathbb{R}^d \to \mathbb{R}^K$ be a fixed sparse-autoencoder encoder (Olshausen & Field, 1996; Gao et al., 2024; Bricken et al., 2023; Cunningham et al., 2023) and $\mathcal{A} = [K]$. The atomization map and trajectory are defined as:

$$\sigma(h) := \arg\max_{i \in [K]} E(h)_i, \qquad A_t := \sigma(h_t) \in \mathcal{A}, \qquad A_{1:t} := (A_1, \ldots, A_t), \tag{1}$$

with ties broken lexicographically. Concretely, each step keeps its most active dictionary atom. An episode may stop before T; after it stops we set $A_s := \perp$, an absorbing symbol, so that every prefix $A_{1:t}$ is well defined and reveals whether the episode is still running. The map σ is part of the interface; Section 5 records which atomizer each audited interface uses.

Outcome laws, scores, and the balanced analysis law. Fix a prefix length t and write X for the observable (for example $X = A_{1:t}$), with finite range $\mathcal{X}$ and outcome-conditioned laws:

$$P_y := \mathcal{L}(X \mid Y = y), \qquad p_y(x) := P_y(x), \qquad y \in \{0, 1\}. \tag{2}$$

A detector is a score $s : \mathcal{X} \to \mathbb{R}$, and $\mathrm{AUROC}(s)$ is tie-aware and oriented so that larger scores rank successes above failures; it depends only on P_1 and P_0 (Fawcett, 2006). AUROC ignores the class prior, so we may analyze it under the balanced law and the success posterior:

$$M := \tfrac{1}{2}(P_1 + P_0), \qquad \eta(x) := \frac{p_1(x)}{p_1(x) + p_0(x)}, \qquad D_{xz} := p_1(x)p_0(z) - p_1(z)p_0(x). \tag{3}$$

Informally, $\eta(x)$ is how strongly prefix x points to success when both outcomes are equally likely, and D_{xz} is the net evidence for ranking x above z. Mutual informations written I_{bal} are computed under M in nats. Nothing here assumes that the real success rate is $1/2$.

4 MAIN RESULTS

We now state what separates a deployed score from the ceiling of the observable it reads. Figure 1 summarizes the chain: the observable is compressed, a likelihood family is projected onto, and the projection is estimated.

4.1 BACKGROUND: THE CEILING IN POSTERIOR FORM

We first record the standard ceiling in the form the new results use. For $x \in \mathcal{X}$, the likelihood ratio is $\ell(x) := p_1(x)/p_0(x) \in [0, \infty]$, with $\ell(x) := 0$ when $p_1(x) = p_0(x) = 0$.

Proposition 1 (Bayes-AUROC ceiling; standard)**.** *For any fixed observable:*

$$A^\star(X) := \sup_s \mathrm{AUROC}(s) = \mathbb{E}\big[\mathbf{1}\{\ell(X^+) > \ell(X^-)\} + \tfrac{1}{2}\mathbf{1}\{\ell(X^+) = \ell(X^-)\}\big], \tag{4}$$

where $X^+ \sim P_1$ and $X^- \sim P_0$ independently; any score that preserves the strict order of ℓ attains it.

The best detector thus sorts observations by how much more likely they are under success than under failure; the new results use the following posterior form of that optimum.

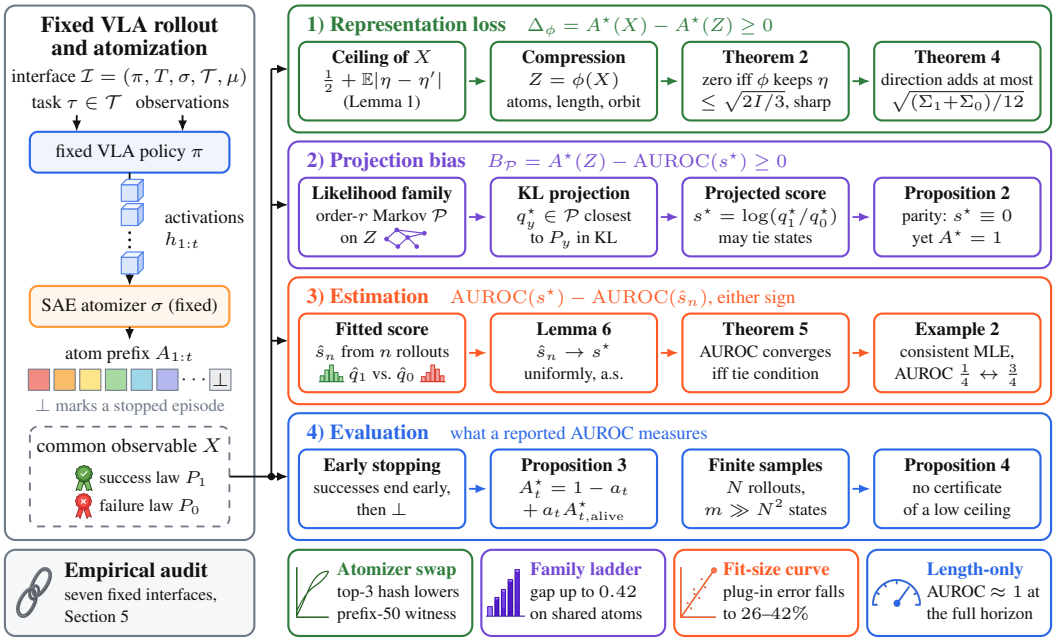


Figure 1: **From the ceiling to a deployed score.** A fixed interface produces the common observable X, with success law P_1 and failure law P_0 (left). The gap between its ceiling and a fitted score splits into the three terms of (6), one per layer and written beside its title: compression can only lose evidence (green), a likelihood family can hide all of it (purple), and estimation may or may not settle the ranking (orange). The blue layer concerns what evaluation itself measures. The bottom row names the audit in Section 5 that reads each layer.

Lemma 1 (Posterior form; standard)**.** *With X, X' independent draws from M:*

$$A^\star(X) = \tfrac{1}{2} + \mathbb{E}\,|\eta(X) - \eta(X')| = \tfrac{1}{2} + \tfrac{1}{4}\sum_{x,z} |D_{xz}|. \tag{5}$$

In words, the ceiling exceeds chance by the average disagreement in success posterior between two random prefixes; a detector can only rank well where prefixes differ in how strongly they point to success. For a representation $Z = \phi(X)$, a KL-projected score $s^\star$ on Z, and a fitted score $\hat{s}_n$, the gap to the ceiling splits exactly as:

$$\begin{aligned} A^\star(X) - \mathrm{AUROC}(\hat{s}_n) = {} & \underbrace{A^\star(X) - A^\star(Z)}_{\Delta_\phi\text{: representation}} + \underbrace{A^\star(Z) - \mathrm{AUROC}(s^\star)}_{B_\mathcal{P}\text{: projection}} \\ & + \underbrace{\mathrm{AUROC}(s^\star) - \mathrm{AUROC}(\hat{s}_n)}_{\text{estimation}}. \end{aligned} \tag{6}$$

The first two terms are nonnegative; the third can have either sign. Each term responds to a different remedy: a new observable, a richer family, or more data. The next four subsections characterize them in turn.

4.2 What compression costs

The first question is when a representation, such as top-1 atoms or episode length, loses evidence. Write $C_a := \phi^{-1}(a)$ for the fibers of ϕ and $\eta_Z(z) := \mathbb{E}_M[\eta(X) \mid Z = z]$ for the posterior after compression.

Theorem 2 (Exact cost of compression)**.** *For any deterministic $Z = \phi(X)$ on a finite space:*

(i) the loss has the exact forms:

$$\begin{aligned} \Delta_\phi &= \mathbb{E}|\eta(X) - \eta(X')| - \mathbb{E}|\eta_Z(Z) - \eta_Z(Z')| \\ &= \tfrac{1}{4}\sum_{a,b}\Big[\sum_{x\in C_a, z\in C_b} |D_{xz}| - \Big|\sum_{x\in C_a, z\in C_b} D_{xz}\Big|\Big]; \end{aligned} \tag{7}$$

(ii) $\Delta_\phi = 0$ *if and only if* η *is constant on every fiber (*M*-a.s.), equivalently* $Y \perp X \mid Z$ *under* M*;*

(iii) with I_{bal} *in nats:*

$$0 \le \Delta_\phi \le \min\Big\{\tfrac{1}{2},\ \sqrt{\tfrac{2}{3}\, I_{\mathrm{bal}}(Y; X \mid Z)}\Big\}, \tag{8}$$

and the constant $\sqrt{2/3}$ *cannot be improved uniformly over finite spaces.*

The three parts answer three different questions. (i) Where the loss comes from: a fiber that merges prefixes with opposite ranking evidence lets that evidence cancel, and each bracket is exactly the cancelled amount. (ii) When nothing is lost: a representation is lossless precisely when it keeps the success posterior; the number of states, reconstruction quality, and preserved task identity are all beside the point. (iii) How much is lost: at most the square root of the outcome information that X still has given Z. The proof idea is short: compressing sums D_{xz} over fibers in (5), the triangle inequality on each block gives (i), the diagonal blocks give (ii), and (iii) chains a Gini-mean-difference bound with Pinsker's inequality (Appendix B). Monotonicity follows at once.

Corollary 3 (Post-processing monotonicity; standard)**.** *For any deterministic* ϕ*,* $A^\star(\phi(X)) \le A^\star(X)$*: a coarser atomizer, a prefix map, or a feature extraction can hide outcome evidence but cannot create it.*

Recall the classical fact behind this corollary (Blackwell, 1953): for two outcomes, one observable is a (possibly randomized) garbling of another exactly when its Bayes ROC curve lies below the other's everywhere. Observables that are not nested, such as atoms from one layer and hidden states from another, need not be ordered at all: when neither garbles the other, their Bayes ROC curves cross, and which one "wins" depends on the operating point.

4.3 What time direction is worth

Naturally, one could summarize a path by how irreversible it is and skip the full prefix laws. The next result says exactly how much ranking power the direction of time carries. Let R be an involution of $\mathcal{X}$ (time reversal), let $U := \{X, \mathsf{R}X\}$ be the unordered reversal orbit, and let $\Sigma_y := D_{\mathrm{KL}}(P_y \,\|\, P_y \circ \mathsf{R})$ in nats.

Theorem 4 (Detection value of time direction)**.** *For any finite* $\mathcal{X}$ *and involution* R*, fixed points allowed:*

$$0 \le A^\star(X) - A^\star(U) \le \min\Big\{\tfrac{1}{2},\ \sqrt{\tfrac{\Sigma_1+\Sigma_0}{12}}\Big\}, \tag{9}$$

and $1/\sqrt{12}$ *is the smallest constant valid for all finite spaces.*

The theorem controls one thing and not another: (i) it bounds the *extra* ranking power gained by knowing the direction of each path, which vanishes with the irreversibility of the two outcome classes; (ii) it says nothing about the ranking power of the orbit itself, which entropy production does not constrain. Behind the bound, reversal changes only an orientation bit inside each orbit, so the information lost by forgetting direction is a per-orbit Bernoulli divergence; a sharp binary inequality bounds it by $(\Sigma_1 + \Sigma_0)/8$, and Theorem 2(iii) converts that into AUROC. The surprising part is how little irreversibility identifies.

Example 1 (Irreversibility does not identify direction value)**.** *On two states swapped by* R*, let* $r \ne \frac{1}{2}$*. If* $P_1 = P_0 = \mathrm{Bern}(r)$*, both classes have* $\Sigma_y = D_{\mathrm{KL}}(\mathrm{Bern}(r) \| \mathrm{Bern}(1-r))$ *and direction is worth nothing. If* $P_1 = \mathrm{Bern}(r)$ *and* $P_0 = \mathrm{Bern}(1-r)$*, the* Σ_y *are the same, yet direction is worth* $|r - \frac{1}{2}|$ *AUROC. Conversely, if the classes occupy disjoint orbits symmetrically, then* $\Sigma_1 = \Sigma_0 = 0$ *and* $A^\star(X) = A^\star(U) = 1$*.*

Put differently, irreversibility caps what the arrow of time can add, but a strongly irreversible policy may gain nothing from it, and a perfectly reversible one may still be perfectly monitorable.

4.4 When more data cannot help

The projection term $B_{\mathcal{P}}$ can absorb the ceiling's entire excess over chance. Our audit fits order-r Markov likelihoods, and there is a simple way to hide all evidence from them.

Proposition 2 (Parity blind spot of low-order projections). *Let $X = (X_1, \dots, X_T) \in \{0,1\}^T$, let P_1 be uniform on even-parity paths and P_0 uniform on odd-parity paths. For every $r \le T-2$, the KL projections of P_1 and P_0 onto order-r Markov laws (time-homogeneous or not) coincide and equal $\mathrm{Bern}(\frac{1}{2})^{\otimes T}$. Hence the projected log-likelihood ratio is identically zero, with AUROC $\frac{1}{2}$, while $A^\star(X) = 1$. Both classes are reversal invariant, so $\Sigma_1 = \Sigma_0 = 0$ and $A^\star(U) = 1$.*

Intuitively, every block that an order-r model can see has the same law under success and failure (any $T-1$ coordinates of a parity-constrained uniform path are uniform), so all the evidence sits in dependence beyond the model's order. More data cannot move a population projection; only a larger family can. Together with Theorem 4, the proposition shows that perfectly reversible, perfectly separable evidence can still be invisible to a likelihood family.

4.5 When score consistency gives ranking consistency

The estimation term concerns a fitted score converging to its projection $s^\star$. Fix any order $\prec$ on the finite space $\mathcal{X}$. The AUROC of a score can be rewritten as:

$$\mathrm{AUROC}(s) = \tfrac{1}{2} + \tfrac{1}{2} \sum_{x \prec z} D_{xz} \operatorname{sgn}\big(s(x) - s(z)\big), \tag{10}$$

so only the signs of score differences matter. Let $\kappa(s^\star) := \frac{1}{2} \sum_{x \prec z:\, s^\star(x) = s^\star(z)} |D_{xz}|$ collect the evidence carried by pairs that the limit ties.

Theorem 5 (Rank stability). *For a fixed limit score $s^\star$ on finite $\mathcal{X}$, the following are equivalent: (a) $\mathrm{AUROC}(s_n) \to \mathrm{AUROC}(s^\star)$ for every sequence with $\|s_n - s^\star\|_\infty \to 0$; (b) the tie condition holds: $s^\star(x) = s^\star(z) \Rightarrow D_{xz} = 0$. Moreover, if $\delta^\star$ is the smallest nonzero gap of $s^\star$, then for $0 < \epsilon < \delta^\star/2$:*

$$\sup_{\|s - s^\star\|_\infty \le \epsilon} \big|\mathrm{AUROC}(s) - \mathrm{AUROC}(s^\star)\big| = \kappa(s^\star). \tag{11}$$

The tie condition essentially requires that the limit may merge states only if they carry the same likelihood ratio. The true log-likelihood ratio satisfies it; a KL-projected one need not, because projection can tie states that the data distinguish. When the condition fails, arbitrarily small score errors can swing AUROC by the full amount $\kappa(s^\star)$. The reason is visible in (10): a perturbation smaller than half the smallest nonzero gap cannot change the sign of any untied pair, so only tied pairs move, each worth $\frac{1}{2}|D_{xz}|$.

Example 2 (A consistent MLE with inconsistent AUROC). *Let $\mathcal{X} = \{0,1,4\}$, $p_1 = (\frac{3}{4}, 0, \frac{1}{4})$, $p_0 = (0,1,0)$, and fit the exponential family $q_\theta(x) \propto e^{\theta x}$, $\theta \in [-1,1]$, to each class by maximum likelihood. Both classes have mean 1, so both projections equal $\theta^\star = -\frac{1}{4}\ln 3$ and the projected log-likelihood ratio is 0. The fits are consistent and the fitted log-likelihood ratio tends to zero uniformly, yet its sign follows the class-1 sample mean. AUROC therefore alternates between $\frac{1}{4}$ and $\frac{3}{4}$, each with probability tending to $\frac{1}{2}$, and does not converge, although the limit score has AUROC $\frac{1}{2}$.*

The culprit is a harmful tie: the limit score merges $x = 0$ and $x = 1$, which carry opposite evidence. Appendix B adds two related facts. The α-smoothed Markov grammar of our audit converges almost surely to its KL projection whenever expected transition counts are positive (Lemma 6); under the tie condition the decomposition (6) is then eventually exact with a zero estimation term. And a KL projection whose mean log-likelihood-ratio gap is positive can still rank backwards (Example 3).

4.6 What evaluation can and cannot certify

The last two results concern the evaluation protocol rather than the detector. In LIBERO a successful episode stops when the task is completed, while a failure runs to the horizon. Let a_t be the probability that a successful episode is still running at step t.

Proposition 3 (Termination identity). *Suppose only successful episodes stop before T and stopped prefixes end in $\perp$. Then:*

$$A_t^\star = (1 - a_t) + a_t\, A_{t,\mathrm{alive}}^\star, \tag{12}$$

where $A_{t,\mathrm{alive}}^\star$ is the ceiling for running successes against failures. The same identity holds for the AUROC of any score that ranks stopped prefixes above running ones.

Table 1: **Result-to-audit map.** Status separates new results, standard facts, and empirical claims.

Result	Status	What the audit reads	Where
Prop. 1, Lem. 1	standard	the ceiling; estimated only at $t = 1$	Sec. 5.4
Prop. 3	new	length-only AUROC at the horizon; a_t from per-rollout ledgers	Sec. 5.1
Thm. 2, Cor. 3	new / std.	representation: atomizer swap; length as a view of the prefix	Secs. 5.1, 5.2
Prop. 2, Thm. 5	new	projection and estimation: family ladder on a fixed observable	Sec. 5.4
Thm. 11	proved	plug-in error decreases with fit size	Sec. 5.4
Prop. 4	new	why no ceiling estimate for $t \geq 2$	Sec. 5.4
Thm. 4	new	not read on outcome labels; task-label analogue only	App. K
Interface-specific ordering	empirical	paired atom-versus-SAFE witnesses on seven interfaces	Tab. 3

That is, every success that has already stopped is a free correct pair. Even the full sequence of prefix ceilings does not settle when to raise an alarm (Example 4). At the horizon $a_T = 0$ and the ceiling is 1 whatever the runtime evidence; truncating each task at its shortest successful episode, as SAFE does (Gu et al., 2025), sets $a_t = 1$ and reports $A^{\star}_{t,\mathrm{alive}}$. Between these two ends a prefix AUROC mixes both, so a_t should be reported with it.

Proposition 4 (No universal certificate of a low ceiling)**.** *Draw n_1 successes and n_0 failures, $N = n_0 + n_1$, on m states (m even), and let $\varepsilon_m := \min\{1, \max[\frac{N(N-1)}{2m}, \frac{n_0(n_0-1)+n_1(n_1-1)}{m}]\}$. If a data-dependent $\Upsilon \in [\frac{1}{2}, 1]$ satisfies $\Pr(\Upsilon \geq A^{\star}) \geq 1 - \delta$ for every pair of class laws, then under $P_1 = P_0 = \mathrm{Unif}$ on the m states (where $A^{\star} = \frac{1}{2}$):*

$$\Pr(\Upsilon = 1) \geq 1 - \delta - \varepsilon_m. \tag{13}$$

The consequence is sobering: without structural assumptions an honest upper bound on the ceiling must report 1 even when both classes are uniform and the truth is chance, as soon as $m \gg N^2$. With $K = 1024$ atoms and the 250-rollout report sets of our extension OFT cells, $t = 2$ already gives $m = 1{,}048{,}576$ and $\varepsilon_m \leq 0.05$. Lower witnesses, possibly combined with structural assumptions, are therefore the only general route to certified statements about the ceiling, which is why Section 5 estimates the ceiling only at $t = 1$.

5 Finite-rollout audit on seven simulated interfaces

We now read the results on real rollouts, including where the data leave a question open. All experiments are LIBERO simulation rollouts; there is no physical-robot or RoboTwin experiment. Every interface follows one locked-split discipline: models are fit only on a fit split, orientations and thresholds are chosen only on a calibration split, and each number is computed once on a report split.

Interfaces. The three *primary* interfaces come from the original locked study: OpenVLA on LIBERO-Spatial (3,000 report rollouts) and LIBERO-Object (1,000), atomized by one frozen top-1 SAE ($K = 1024$) trained on Spatial and transferred to Object, and an OpenVLA-OFT row on LIBERO-Spatial (1,000). Archived records show that the OFT row uses a LIBERO-10 checkpoint, a 220-step horizon, and its own random-projection atomizer, so it changes policy, checkpoint, horizon, and atomizer together. The primary study's raw pools and weights are no longer accessible, so the four *extension* interfaces are new instances, not reproductions: OpenVLA-OFT (joint four-suite checkpoint) on LIBERO-Spatial and LIBERO-Object (250 report rollouts each, one shared new SAE), $\pi_{0.5}$ on LIBERO-Spatial (500; own SAE), and OpenVLA on LIBERO-Goal (1,000). Appendix H gives checkpoints, atomizers, horizons, and splits.

Protocol. Grammar-LLR is first-order Markov with additive smoothing $\alpha = 0.5$; n-gram, KT, unigram+bigram (L2), and GRU scores are further atom-level witnesses. SAFE (Gu et al., 2025) reads VLA hidden states, a different observable with a different ceiling. Task-residualized AUROC subtracts each score's within-task mean. Paired comparisons use DeLong tests (DeLong et al., 1988) and 10,000 class-stratified paired-bootstrap replicates, with Holm correction over four pre-specified atom-versus-SAFE contrasts per interface.

5.1 Termination: an exact bonus at the horizon

We start with Proposition 3, because it explains the most striking number in the audit. On the three primary interfaces a length-only score reaches full-horizon AUROC 0.9996, 1.0000, and 1.0000

Table 2: **Evidence localization on the primary interfaces.** Locked-report AUROC point estimates.

Interface	Len. full	Gram. full	Gram. p50	2-gram p50	SAFE full	SAFE p50	Gram. resid.	SAFE resid.
OpenVLA–Spatial	0.9996	0.9833	0.7122	**0.7855**	0.9673	0.6686	0.9801	0.9541
OpenVLA–Object	1.0000	0.9706	0.6317	**0.6749**	0.9592	0.6323	0.9784	0.9612
OpenVLA-OFT–Spatial	1.0000	0.8904	0.5824	0.6394	0.9573	**0.7015**	0.9103	0.9385

(Table 2). Per-rollout ledgers were not retained for the primary interfaces, but those of two extension interfaces show that the premise of the proposition holds exactly there: on LIBERO-Goal all 223 failures run to the 300-step horizon and none of the 777 successes do, and on $\pi_{0.5}$ all 8 failures reach the 220-step horizon and no success does. Hence $a_T = 0$, and the identity gives a full-horizon ceiling of 1 for any observable that reveals termination, whatever the runtime evidence; the length score attains it, and the near-perfect primary rows are what the identity predicts when $a_T \approx 0$. At the median success length (step 99 on Goal, 103 on $\pi_{0.5}$) we have $a_t \approx \frac{1}{2}$, so even an observable with no runtime evidence has a prefix ceiling near 0.75 there. SAFE removes this bonus by truncation; the identity quantifies it. Conversely, every successful episode on Goal and $\pi_{0.5}$ lasts at least 58 and 69 steps, so $a_{50} = 1$: their prefix-50 rows carry no bonus at all, and there the atom witness leads SAFE on Goal (0.735 vs. 0.650), while on $\pi_{0.5}$, with only 8 failures, both fall below $\frac{1}{2}$. Length is a deterministic view of the full prefix, so by Corollary 3 it adds no evidence beyond the prefix; we read prefix rows as runtime evidence and full-horizon rows as statements about complete episodes (Appendix M).

5.2 Observables over time on the primary interfaces

Table 2 compares observables on the primary interfaces over time. At prefix 50 the atom witnesses lead SAFE on both OpenVLA rows (n-gram-2 0.7855 vs. 0.6686 on Spatial) and trail it on the OFT row (0.6394 vs. 0.7015); the same reversal appears at full horizon and after task residualization. Because the OFT row also changes checkpoint, horizon, and atomizer, we read this as an interface effect, not as a policy effect in isolation. Replacing the top-1 SAE atomizer by a top-3 tuple hash with four times as many states lowers the prefix-50 Grammar witness by 0.09–0.23 (Appendix J). The hash is not a coarsening of the top-1 map, so Corollary 3 does not order the two ceilings; the drop is consistent with Theorem 2(ii), under which what matters is whether a representation keeps the success posterior, not how many states it has. Our witnesses cannot tell whether the ceiling itself dropped. Full baseline ladders, including FAIL-Detect (Xu et al., 2025) and FIPER (Römer et al., 2025), are in Appendix I.

5.3 The atom-versus-hidden ordering across seven interfaces

If the ordering were a property of the detectors, it would persist across interfaces. It does not. Table 3 reports the pre-specified full-horizon contrast on every interface: the atom witness is ahead on OpenVLA–Spatial, statistically indistinguishable on OpenVLA–Object and extension OFT–Spatial, and behind on the other four. Unfortunately, the per-rollout label manifest of the primary OFT row was lost; its labels were reconstructed from the retained score arrays (the exact length separation fixes them up to one integer, and the reconstruction reproduces all eleven originally reported AUROCs), and we mark the row accordingly. Two suite changes (OFT: Spatial→Object; OpenVLA: Spatial→Goal) and one policy change (OpenVLA→ $\pi_{0.5}$ on Spatial) each change the ordering of the two observables, so neither coordinate preserves the ordering in general. Prefix readings move again: on LIBERO-Goal the atom witness leads at prefix 50 (0.735 vs. 0.650) but trails at full horizon (Appendix I). These are orderings of witnesses, not of ceilings. The atoms and SAFE read different layers, so their ceilings need not be ordered (Section 4.2), and by (6) a witness flip can come from any of the three terms.

5.4 Projection, estimation, and what can be certified

Can a weak witness reflect the likelihood family rather than the interface? It can. On extension OFT–Object the pre-specified first-order Grammar witness is 0.475 while unigram n-gram-1 reaches 0.893 $[0.801, 0.962]$ on the same rollouts; on $\pi_{0.5}$ the two give 0.634 and 0.998. Both scores read the same atoms, so in (6) their representation terms coincide and the entire difference, up to 0.42 AUROC, lies in the projection and estimation terms. The first-order grammar has on the order of 10^6 transition parameters on $K = 1024$ atoms, fitted here against 23 and 8 failures; the unigram has about 10^3. Our data cannot separate the two terms, but the theory says which remedy addresses

Table 3: **Paired atom-versus-hidden contrasts at full horizon.** P: primary, E: extension interface. Δ = atom Grammar − SAFE with 95% paired-bootstrap CI; Holm p over four contrasts per interface. Best atom: largest full-horizon AUROC among the atom-level families scored on that interface (selection-optimistic). [a]archived DeLong, raw p, no interval retained. [b]labels reconstructed from retained score arrays (App. H).

	Interface	Report S/F	Atom	SAFE	Δ [95% CI]	Holm p	Best atom
P	OpenVLA–Spatial	2352/648	0.9833	0.9673	$+0.0160^{a}$	$1.8 \times 10^{-5\,a}$	2-gram 0.9870
P	OpenVLA–Object	863/137	0.9706	0.9592	$+0.0115\,[-0.0135, 0.0369]^{b}$	1.000	Grammar 0.9706
P	OpenVLA-OFT–Spatial	779/221	0.8904	0.9573	$-0.0669\,[-0.0936, -0.0410]^{b}$	2.8×10^{-6}	2-gram 0.8982
E	OpenVLA-OFT–Spatial	133/117	0.9954	0.9924	$+0.0030\,[-0.0043, 0.0107]$	1.000	3-gram 0.9967
E	OpenVLA-OFT–Object	227/23	0.4750	0.9941	$-0.5191\,[-0.6771, -0.3654]$	6.8×10^{-10}	1-gram 0.8931
E	$\pi_{0.5}$–Spatial	492/8	0.6341	1.0000	$-0.3659\,[-0.5864, -0.1888]$	3.3×10^{-3}	1-gram 0.9982
E	OpenVLA–Goal	777/223	0.8485	0.9971	$-0.1486\,[-0.1883, -0.1099]$	3.9×10^{-13}	GRU 0.9488

which: by Proposition 2 more data need not close a projection gap, and by Theorem 5 a consistent fit need not stabilize its ranking. Within a family the witness does stabilize as Theorem 11 predicts: plug-in error against the largest-fit reference falls to 26–42% of its initial value on all three primary interfaces (Appendix I).

Fortunately, the ceiling can still be bounded from below: the lower end of a fixed score's bootstrap interval is an approximate lower confidence bound, because $\mathrm{AUROC}(s) \leq A^{\star}$: 0.984 for the OFT–Object SAFE hidden state, whereas the pre-specified Grammar interval is uninformative for its atoms and the best ladder score's lower end (0.801) would need a multiplicity correction. The ceiling is estimable at $t = 1$: with 87 and 59 of 1,024 atoms observed on the two extension OFT cells, plug-in estimates are 0.932 and 0.913, but cross-fitted values are only 0.700 and 0.597, so the plug-in is optimistic. For $t \geq 2$ Proposition 4 applies with $\varepsilon_m \leq 0.05$: any upper bound with 95% coverage would output 1 with probability at least 0.90 even if both classes were uniform, which is why we report none. The KL envelopes of Appendix F, evaluated with projected divergences, are uninformative in the same regime: they reach 1 to four decimals by $t = 10$–50 (Appendix J).

6 Discussion, Limitations, and Conclusion

We close with what the results imply for practice. That a detector cannot beat the ceiling of what it reads is standard; the results here say what stands between the two. A representation is lossless exactly when it keeps the success posterior, time direction can add at most $\sqrt{(\Sigma_1 + \Sigma_0)/12}$, low-order likelihood witnesses can miss evidence entirely and need not stabilize, prefix AUROC carries an exact termination bonus, and low ceilings cannot be certified without structure. Each result turns into a reporting practice:

1. report the policy and checkpoint, suite, horizon, atomizer, and observable with every detector AUROC, because the ceiling belongs to that interface (Proposition 1) and a representation may lose evidence (Theorem 2);
2. report a_t with every prefix AUROC, because stopped successes inflate it exactly (Proposition 3);
3. report a ladder of witness families rather than one score, because a single family can miss the evidence entirely (Proposition 2) or fail to stabilize (Theorem 5);
4. use paired intervals before claiming an ordering between observables, and treat any upper bound on the ceiling with suspicion unless it rests on stated structure (Proposition 4).

Limitations. There are, of course, limitations. The theory concerns finite observables at a fixed prefix. All evidence is from LIBERO simulation, with no RoboTwin or physical-robot study, and Theorem 4 is not instantiated on outcome labels. The primary OFT paired test rests on reconstructed labels, and that row changes the atomizer together with the policy. Extension OFT cells are small (250 report rollouts), and the $\pi_{0.5}$ and OFT–Object rows have only 8 and 23 failures. Splitting the remaining gap into intrinsic unpredictability and observation loss would require re-simulating from saved states, which we leave to future work. Appendix L discusses scope.

AI USE STATEMENT

In this work, we used generative AI tools to assist with carrying out the experiments and to help complete the initial draft of the manuscript. We have reviewed all AI-assisted work: the authors checked the AI-assisted experiments and text directly by hand, and cross-checked them with independent runs of multiple AI agents. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Each theoretical result states its assumptions where it appears, complete proofs are in Appendices B–G, and the counterexamples are explicit enough to check by hand. Section 5 states the locked-split protocol and the statistical tests and records which assets of the primary study are no longer accessible. Appendix H lists the checkpoint, atomizer, horizon, and splits of every interface, and Appendix I gives the full detector ladders.

REFERENCES

Shivani Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15(49):1653–1674, 2014.

András Antos, Luc Devroye, and László Györfi. Lower bounds for Bayes error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(7):643–645, 1999.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24 (2):265–272, 1953.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. RT-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, pp. 837–845, 1988.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. PaLM-E: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Qiao Gu, Yuanliang Ju, Shengxiang Sun, Igor Gilitschenski, Haruki Nishimura, Masha Itkina, and Florian Shkurti. SAFE: Multitask failure detection for vision-language-action models. *arXiv preprint arXiv:2506.09937*, 2025.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

Abi Komanduru and Jean Honorio. A lower bound for the sample complexity of inverse reinforcement learning. In *International Conference on Machine Learning*, pp. 5676–5685. PMLR, 2021.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Aditya Krishna Menon and Robert C. Williamson. Bipartite ranking: A risk-theoretic perspective. *Journal of Machine Learning Research*, 17(195):1–102, 2016.

Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694–706):289–337, 1933.

Nana Obayashi, Arsen Abdulali, Fumiya Iida, and Josie Hughes. Embodied intelligence paradigm for human-robot communication. *Science Robotics*, 10(105):eads8528, 2025.

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Physical Intelligence, Kevin Black, Noah Brown, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Alexandre Pitti, Max Austin, Kohei Nakajima, and Yasuo Kuniyoshi. Informational embodiment: Computational role of information structure in codes and robots. *Physics of Life Reviews*, 53: 262–276, 2025.

Cameron Reid, Wael Hafez, and Amirhossein Nazeri. Mutual information tracks policy coherence in reinforcement learning. *arXiv preprint arXiv:2509.10423*, 2025.

Ralf Römer, Adrian Kobras, Luca Worbis, and Angela P Schoellig. Failure prediction at runtime for generative robot policies. *arXiv preprint arXiv:2510.09459*, 2025.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5. IEEE, 2015.

Chen Xu, Tony Khuong Nguyen, Emma Dixon, Christopher Rodriguez, Patrick Miller, Robert Lee, Paarth Shah, Rares Ambrus, Haruki Nishimura, and Masha Itkina. Can we detect failures without failure data? Uncertainty-aware runtime failure detection for imitation learning policies. *arXiv preprint arXiv:2503.08558*, 2025.

Wenye Yu, Jun Lv, Zixi Ying, Yang Jin, Chuan Wen, and Cewu Lu. ARMADA: Autonomous online failure detection and human shared control empower scalable real-world deployment and adaptation. *arXiv preprint arXiv:2510.02298*, 2025.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.