CAN INTERNAL STATES SIGNAL TROUBLE AND GUIDE RECOVERY IN VLAS?

Anonymous authors
Paper under double-blind review

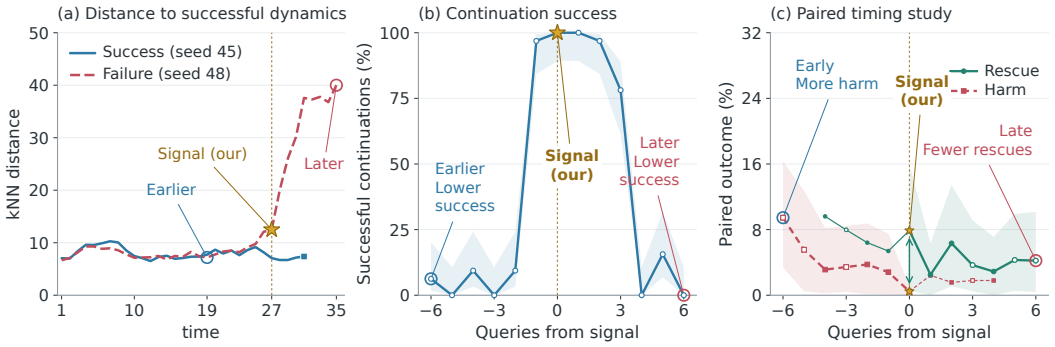


Figure 1: When to recover. **(a)** k-NN distance to successful executions for a success and a failure from one initial state (chosen for its rise); past the first signal (star) it is the recorded continuation, not a recovery. **(b)** Recovery $\mathcal{R}(0.30, 3)$ on this failure around the signal, each time continued by 32 noise seeds to the end: on this example success is highest near the signal. **(c)** Rescue and harm over all $1{,}280$ paired branches per time (Figure 4 cohort): earlier interventions harm more, later ones rescue fewer; the aggregate does not establish an optimal time. Shaded: 95% intervals.

ABSTRACT

Frozen vision-language-action (VLA) policies can keep acting after progress has stalled. We introduce NAR, a test-time harness around a frozen mixture-of-experts VLA, driven by the expert routing already computed during action generation. On LIBERO-10, HiMoE-VLA failures exhibit stronger routing recurrence despite similar numbers of experts used per inference. NAR measures recurrence relative to the episode's opening and detects deviations from successful source executions using nearest-neighbour scoring, temporal smoothing, and episode-level conformal calibration. Alarms trigger a fixed release-and-lift primitive before control returns to the same policy, requiring no parameter optimization or additional policy forward passes. Under a shared monitoring pipeline, routing achieves higher recall than the evaluated state and action readouts at their calibrated operating points. In a retrospective cross-suite evaluation, a profile fitted on native and perturbed LIBERO-Goal successes detects 97.2% of failures at a 1.32% episode-level false-alarm rate across $4{,}320$ episodes from the other three suites, without target-task reference trajectories. With a separate pooled profile, paired replay of $1{,}608$ completed alarms yields 175 net additional successes: 10.88 percentage points per alarm and 0.40 points when normalized by the recorded corpus. A separate timing study finds the lowest observed harm at the alarm, while later interventions lose opportunities to act. Transfer depends on source-success coverage, with specificity degrading on CALVIN. These results show how a policy's existing routing trace can support a test-time harness that connects execution monitoring with the measured benefits and harms of intervention.

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.

1 INTRODUCTION

At test time, a vision-language-action (VLA) system must manage the execution of its policy as well as generate actions. A frozen policy may keep attempting an unsuccessful grasp or pushing against an obstruction after progress has stalled. We view the surrounding runtime process as a test-time harness: it observes execution, triggers a temporary intervention upon an alarm, and returns control to the same policy. Building such a harness requires an execution signal that is useful for intervention. An alarm can identify a failing episode, yet the response it triggers can also disrupt an episode that would otherwise succeed.

Runtime monitoring has drawn on action consistency and progress (Agia et al., 2024), references from successful executions (Xu et al., 2025), and internal VLA features (Gu et al., 2025; Seligmann et al., 2026). Internal anomaly signals have also been used to initiate recovery of frozen policies (Zhang et al., 2026d). For mixture-of-experts (MoE) VLAs, the architecture provides a particularly structured observation interface: expert routing (Du et al., 2025). Each inference selects which experts participate in action generation, and expert identities remain consistent across successive inferences. The policy's action-generation mechanism therefore produces a trace of its own computation that a harness can read without additional sensors or policy forward passes. A simulator exposes the full world state, but a real robot must reconstruct the world-frame poses of its arm and the objects with costly external instruments; the routing trace comes from the policy's own computation.

We find a useful signal in the temporal recurrence of routing. In HiMoE-VLA rollouts on LIBERO-10, successful executions and those ending in timeout use nearly the same number of experts within an inference, yet failures exhibit stronger recurrence across inferences. This observation motivates describing how closely recent routing patterns resemble one another relative to the episode's opening. The resulting descriptor summarizes relations within an execution and can be compared across tasks without assigning shared semantic labels to individual experts. A bank of these descriptors provides a reference for successful execution dynamics without requiring a successful trajectory matched to the current task.

We introduce **NAR, a test-time harness around a frozen MoE VLA, driven by its internal expert routing** (Figure 2). NAR connects observation, alarm, and recovery through a fixed routing descriptor. Its k-nearest-neighbour distance to successful source descriptors is smoothed over time and compared with an episode-level split-conformal threshold. The reference bank and threshold are fitted from successes without parameter optimization; task identity and elapsed time are not predictive inputs. At an alarm, a fixed release-and-lift primitive temporarily replaces the policy's actions. Control then returns to the same frozen policy, whose subsequent routing remains available for reassessment. This interface allows us to measure both the monitoring value of the internal trace and the consequences of intervening on its alarms.

Experiments with frozen HiMoE-VLA policies support the utility of this interface within LIBERO. Under a shared monitoring pipeline, routing achieves higher failure recall than the evaluated state and action readouts at their calibrated operating points. Its alarms follow physical stalls but often leave time for recovery. A profile fitted on native and perturbed LIBERO-Goal successes detects 97.2% of failures at a 1.32% episode-level false-alarm rate across $4{,}320$ episodes from the other three suites, without target-task reference trajectories. These episodes were excluded from fitting but inspected before profile selection. The conformal guarantee requires exchangeability with calibration successes; the observed transfer rates are empirical and depend on source coverage, with specificity degrading on CALVIN (Section 5).

The intervention results show why the harness must be evaluated beyond detection accuracy. Using a separately frozen pooled profile, paired replay of $1{,}608$ completed alarms yields 175 net additional successes for a fixed one-chunk release-and-lift primitive: 10.88 percentage points per alarm and 0.40 points when normalized by the recorded corpus. A separate Goal-profile timing study with a three-chunk variant finds greater harm before the alarm and fewer opportunities to intervene afterward, with the lowest observed harm at the alarm. Per-alarm benefit depends on alarm precision and the response's rescue and harm rates. These results quantify the value and cost of acting on an internal signal without establishing an optimal intervention time (Sections 4 and 5).

Our contributions are threefold: (i) we characterize expert-routing recurrence as an observable trace of policy computation and evaluate its usefulness for monitoring frozen MoE VLA executions;

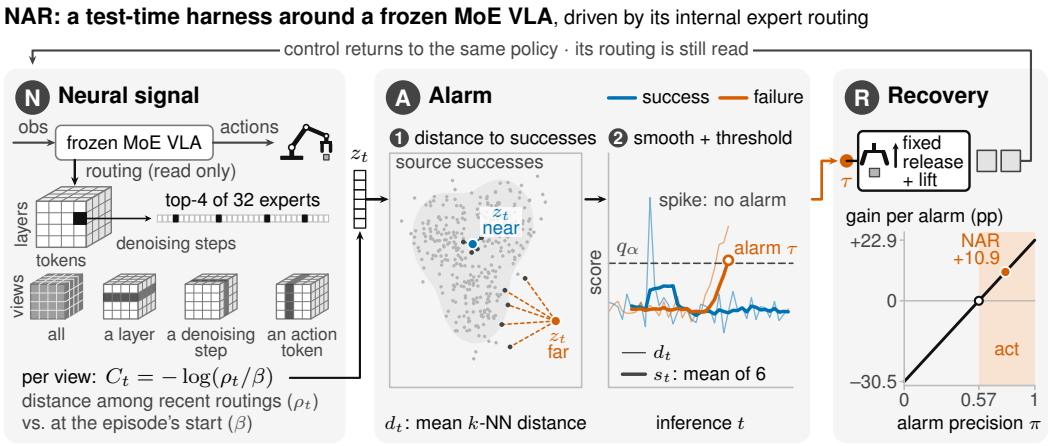


Figure 2: NAR (schematic; the gain plot is measured). **(N)** Per view (a cube slice), C_t compares routing distances at lags ≥ 3 among recent inferences (ρ_t) with those at the episode's start (β); larger C_t means more recurrent routing. With related terms (Appendix A) it forms z_t. **(A)** s_t, the six-inference mean of the k-NN distance d_t from z_t to successful executions, first crosses a split-conformal threshold q_α fitted on successes at τ. **(R)** A fixed release-and-lift primitive replaces c chunks, then the same policy resumes. With the replayed rescue $r = 22.9\%$ and harm $h = 30.5\%$ (Section 5.4), acting pays above precision $h/(r+h) = 0.571$, and the alarm's 0.775 gains $+10.9$ pp.

(ii) we instantiate NAR as a test-time harness driven by internal routing, using successful source executions to support monitoring across tasks and suites; (iii) we quantify the benefits and harms of alarm-triggered intervention through state-matched replay and analyze how intervention timing changes their balance.

2 Related Work

Runtime failure monitoring for robot policies. Agia et al. (2024) separate temporally inconsistent actions from consistent actions that make no progress, and Xu et al. (2025) calibrate detectors on successful executions alone. SAFE (Gu et al., 2025) reads VLA hidden states and trains a predictor on successes and failures; later monitors vary the supervision, from episode-level contrastive labels to world-model latents with conformal prediction (Park et al., 2026; Ma et al., 2026b; Huang et al., 2026; Zhang et al., 2026b). Others delegate the verdict to a vision-language model (Duan et al., 2025; Qi et al., 2026; Pacaud et al., 2025; Grislain et al., 2025) or study monitors under input shift (Rajaprakash et al., 2026; Zhu et al., 2026). Closest to us, VLA-FAIL (Seligmann et al., 2026) fits a last-layer Mahalanobis distance on successful data, and PFD (Lee & Har, 2026) perturbs weights at the cost of extra forward passes. NAR focuses on a fixed statistic of expert-routing recurrence across policy inferences. It combines established nearest-neighbour scoring (Sun et al., 2022), episode-level split-conformal calibration (Vovk et al., 2005; Lei et al., 2018; Angelopoulos & Bates, 2021) and temporal averaging, a relative of change detection (Page, 1954). We evaluate this routing signal through comparisons with external monitors, cross-suite transfer and interventions at its alarms.

Recovery and test-time intervention. Many recovery methods freeze the backbone but train another component: a 3D object probe (Zhang et al., 2026a), a residual recovery policy (Hu et al., 2026) or a language feedback policy (Jeong et al., 2026); others add external models or planning (Song et al., 2026; Ma et al., 2026a; Kwok et al., 2026) or learn from recovery data (Lin et al., 2025; Hu et al., 2025). CoRe (Zhang et al., 2026d) detects success-calibrated internal anomalies and realigns the scene with reconstruction and planning; B2FF (Shin et al., 2026) selects pre-imagined milestones, FaRe (Li et al., 2026) repairs causal history, and RTC (Black et al., 2025) addresses continuity across chunks. We evaluate whether routing-based alarms provide useful trigger times for a fixed recovery primitive, separating detection quality from paired intervention benefits and harms; exploratory response-selection results are in Appendix E.3.

Internal signals for steering and selection. Mixture-of-experts layers (Shazeer et al., 2017; Fedus et al., 2022) give VLAs an explicit routing interface (Du et al., 2025); skill-level routing (Zhang et al., 2026c) is coarser than per-inference dynamics. VLA features can be read and written, by activation intervention (Häon et al., 2025), linear observers (Buurmeijer et al., 2026), sparse autoencoders (Swann et al., 2026), conceptor subspaces (Miao et al., 2026) or closed-loop steering (Babu et al., 2026); we only read. In language models, internal agreement guides choices among candidates for one query (Chen et al., 2026a;b;c); a policy's executed actions cannot be undone.

3 Routing recurrence as an internal sensor

The sensor stream. We monitor frozen HiMoE-VLA policies (Du et al., 2025), one checkpoint per LIBERO suite (Liu et al., 2023). An inference runs ten flow-matching denoising steps (Lipman et al., 2023), emits a chunk of ten actions, and exposes eight hierarchical-MoE routed layers with 32 experts under top-4 routing. At each routing site $\sigma = (\ell, s, a)$ of inference t, a layer, a denoising step and an action token, the four selected experts $E_t(\sigma)$ form the site's *support*; over the hundred sites of a layer, the counts u_t^ℓ of how often each expert is selected form its *participation*; with the routing probabilities p_t these are three views of one computation. The monitor reads nothing else, no image, hidden activation, robot state, task or elapsed time; the information is in how these views *change* between inferences (Section 5.2).

Recurrence and contraction. Let $R_{ij}^d = d(R_i, R_j)$ compare two inferences under one routing view d (metrics in Appendix A) and define

$$C_t^d = -\log \frac{\max(\rho_t^d, \epsilon)}{\max(\beta^d, \epsilon)}, \qquad \beta^d = \operatorname*{med}_{0 \le i < j < P} R_{ij}^d, \qquad \rho_t^d = \operatorname*{med}_{\substack{t-W < i < j \le t \\ j - i \ge L}} R_{ij}^d, \tag{1}$$

with prefix $P = 5$, window $W = 6$, minimum lag $L = 3$ and $\epsilon = 10^{-7}$: $C_t > 0$ means that recent inferences resemble each other more than the episode's own opening did. The lag of three, fixed before our analysis, matches the decorrelation scale later measured on successful routing (Section 5.2); this operator carries most of the discrimination (Section 5.3).

Descriptor, score and alarm. The frozen descriptor $x_t = [a_t, Q(C_t^{\text{supp}}), Q(\rho_t^{\text{supp}}), C_t^{\text{part}}] \in \mathbb{R}^{10+17+17+10}$ stacks ten local statistics a_t of the probabilities with relative contraction, absolute recurrence and participation contraction (pooling Q in Appendix A); every coordinate is causal. After robust standardization against a bank $\mathcal{M}$ of 8,192 descriptors from successful reference episodes, the score d_t is the mean distance to the $k = 20$ nearest bank points and the evidence s_t the mean of the last $W_s = 6$ scores, a finite-memory rule rather than a sequential test (Section 5.3). The alarm is the first crossing $\tau = \inf\{t \ge 12 : s_t > q_\alpha\}$, where q_α is the split-conformal quantile ($\alpha = 0.015$) of the whole-episode maximum of s_t over n held-apart successful calibration episodes: under exchangeability $\Pr(\tau < \infty) \le \alpha$ on a successful episode (Appendix B), a bound on episode-level false alarms only; the alarm fires once per episode (re-arming: Section 5.4) and never reads an outcome. Our default *Goal* profile is fitted on Goal-suite successes, native and LIBERO-Plus (Fei et al., 2026) ($n = 942$ calibration successes), and tested on suites it never saw; the *pooled* profile, fitted on all four suites, was the frozen alarm of the control study (Appendix C). No model is trained: the assets are a bank, 54 medians and scales, one threshold.

4 Acting on the alarm

A fixed primitive. At τ a release-and-lift primitive $\mathcal{R}(\lambda, c)$ replaces the policy's next c chunks: each opens the gripper and lifts along $+z$ with normalized amplitude λ for five actions, then holds for five. Control then returns to the frozen policy, whose routing the monitor keeps reading (Appendix E.3); the primitive takes no argument from the signal. We report $\mathcal{R}(0.60, 1)$ in the full-corpus control study and $\mathcal{R}(\lambda, 3)$, $\lambda \in [0.15, 1.0]$, in later studies.

Replay-branch evaluation. Two online rollouts of one configuration diverge long before any alarm, so we replay the parent episode's recorded actions to the first-alarm inference, verify the simulator state across arms, and branch into a *floor* arm (the policy continues) and a *recovery* arm that share the noise seed (Appendix D.5), native cap and success predicate (terms: Appendix D.1).

When is acting on an alarm worth it? With $A = \{\tau < \infty\}$, paired success outcomes Y_0, Y_1 of floor and recovery, precision $\pi = \Pr(Y_0 = 0 \mid A)$, rescue rate $r = \Pr(Y_1 = 1 \mid Y_0 = 0, A)$ and harm rate $h = \Pr(Y_1 = 0 \mid Y_0 = 1, A)$, and since both arms coincide without an alarm,

$$\Delta = \Pr(A)\big[\pi r - (1-\pi)h\big], \qquad \pi_\star = \frac{h}{r+h}. \tag{2}$$

Conditional on an alarm the paired gain is $(r+h)(\pi - \pi_\star)$: for fixed rescue and harm rates with $r + h > 0$, intervening pays when $\pi > \pi_\star$, and the break-even depends on the response's rescue and harm rates.

5 Experiments

5.1 Setup, baselines and statistical protocol

Policy and data. We monitor HiMoE-VLA, one checkpoint per LIBERO suite (Section 3), on every recorded native episode with complete routing (Appendix D.1; the cohorts each question uses are listed in Table 5): the policy succeeds in 96.83%, and all 1,373 failures are time-outs at the caps of 52, 30, 28 and 22 inferences (LIBERO-10, Goal, Object, Spatial). The seeds 42–49 grid splits configurations into reference, calibration, development and validation roles, two tasks per suite held out (hold-outs refer to monitor fitting; each suite runs its own checkpoint). The block of seeds 50–57 reuses the same 2,000 configurations with new noise seeds; no frozen profile is refitted on it, and Table 1a cross-fits every monitor by configuration within it. LIBERO-Plus enters only where marked (Appendix F); CALVIN D→D (Mees et al., 2022) is only a cross-benchmark check (Appendix D.4). Routing is read by forward hooks on the MoE gates, leaving the forward pass unchanged; hardware and software are listed in Appendix D.1.

Monitor and metrics. Profiles are fitted on successes only (Goal profile: 2,094 reference and 942 calibration successes; Appendix C). We report episode-level recall, precision, F1 and false-alarm rate on successes (FPR) and, as 40.4% of successes end before the thirteenth inference, the rate among those reaching it (FPR_{13}), and alarm timing against other detectors and a hindsight end-effector stall.

Baselines. For RQ1, the complete simulator state, the 8-dimensional proprioceptive state, the 70-dimensional predicted action chunk and end-effector motion become recurrence descriptors of NAR's form (or snapshot variants) and pass through NAR's pipeline, and a causal stall detector's displacement score shares its mean and threshold; all are refitted by five-fold cross-fitting over configurations that keeps the evaluated fold out of every fit. A task-aware monitor, the best of 81 rules on the full simulator state and goal predicates, chosen on held-out folds, faces the frozen NAR; an earlier detector and elapsed-time sentinels are further references (Appendices D.2, D.3). For RQ2 each row of Table 3 changes one part of the alarm; for RQ3 the reference is the unmodified twin.

Statistics. Episodes sharing a configuration are dependent, so intervals resample configurations with all their seeds: 4,000 paired resamples for RQ1 (2,000 for Table 1b), 2,000 for Table 3, 5,000 suite-stratified ones for the timing curve and 20,000 with Bonferroni correction for its contrasts (95% percentile intervals). Comparisons at a detector's own conformal threshold are deployable; those at a matched false-alarm rate set the threshold on the evaluation successes and are descriptive.

5.2 RQ1: Does routing reveal failure as well as privileged state, and earlier?

RQ1.1: Failing executions exhibit stronger routing recurrence. On native LIBERO-10, successes and failures use almost the same number of distinct experts per layer and inference (27.59 and 27.32 of 32) but not in the same way over time (Figure 3): at 75% of each episode's own length, the turnover of expert usage between consecutive inferences over the 100 action sites is 0.41 in successes and 0.19 in failures, and over common windows of running episodes turnover alone separates the two with AUROC 0.75–0.82. Across lags (all 5,768 LIBERO-10 episodes, native and LIBERO-Plus), successful routing decorrelates within three inferences, reaching 99.4% of its lag-nine distance, whereas failing routing is closer at every lag (0.80 vs 0.87 at lag one) and still drifting at lag nine. An alarm marks such a computation: 79% of the successes whose end effector stalled (ad hoc displacement rule) raise an alarm, against 0.7% of the others, and 27 of 41 false alarms are stalls that later recovered.

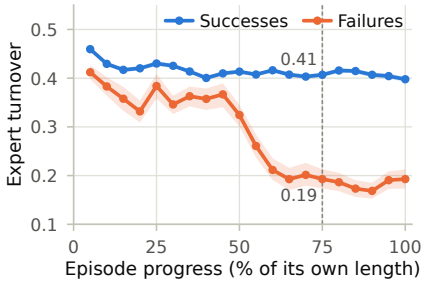


Figure 3: RQ1.1: expert-usage turnover $1 - \cos(u_t, u_{t-1})$ between consecutive inferences (100 action sites) against each episode's own progress, native LIBERO-10 of seeds 42–49 (3,725 successes, 275 failures, NAR's fitting episodes included). Every episode enters every point (no survival selection); 95% bands from 2,000 configuration resamples.

Table 1: RQ1.2: routing achieves higher recall than the state- and action-based monitors, with no resolved false-alarm difference in (a) and 0.6 points more false alarms than the task-aware rules in (b). Seeds 50–57, episodes with at least 13 inferences (509 failures, 9,169 successes); each threshold set without the evaluated episodes; false-alarm rate among those successes; bold: best per panel; 95% paired intervals. In (a) every monitor, routing included, is refitted by five-fold cross-fitting; (b) uses the deployed NAR. †Privileged state; ‡knows the goal; §scores displacement directly. More: Table 6.

	Recall	Precision	False-alarm	F1	Routing − monitor (pp)	
Monitor	(%) ↑	(%) ↑	rate (%) ↓	(%) ↑	Recall ↑	False-alarm rate ↓
(a) Outside inputs through NAR's pipeline						
Routing (ours, cross-fitted)	**98.0**	**67.2**	2.66	**79.7**	ref.	ref.
Complete simulator state†	88.2	62.4	2.94	73.1	+9.8 [5.7, 14.4]	−0.28 [−1.01, 0.45]
Proprioception (8-d)	88.6	61.5	3.08	72.6	+9.4 [6.4, 12.9]	−0.41 [−0.95, 0.10]
Predicted action chunk	94.5	67.1	**2.57**	78.5	+3.5 [1.2, 6.2]	+0.09 [−0.34, 0.53]
End-effector motion	91.0	61.4	3.17	73.3	+7.1 [3.9, 10.5]	−0.51 [−1.21, 0.17]
Causal stall detector§	89.0	63.6	2.82	74.2	+9.0 [5.0, 13.4]	−0.16 [−0.72, 0.41]
(b) Task-aware rules on the full simulator state and goal predicates						
Routing, deployed NAR (ours)	**97.1**	72.9	2.01	**83.2**	ref.	ref.
Task-aware physical rules†‡	84.7	**77.1**	**1.40**	80.7	+12.4 [8.1, 17.3]	+0.61 [0.15, 1.11]

RQ1.2: Routing achieves higher recall than the state- and action-based monitors at their calibrated thresholds. Through one pipeline (Table 1a), routing catches 3.5 to 9.8 points more failures than every outside input while no false-alarm difference is resolved; matched false-alarm rates keep the gaps except, at 1.55%, those to the action chunk and full state, and most of the advantage sits in LIBERO-10 (Appendix D.3). The task-aware baseline reads the full simulator state and the goal predicates: rules that track goal progress, action response, recurrence and grasp loss (Table 1b). Against them the deployed NAR trades 0.61 points more false alarms for 12.4 points more recall; precision and F1 are not resolved, and on LIBERO-10 the rules reach the higher F1 (0.82 against 0.76). The detections overlap imperfectly: no comparator in Table 1 catches more than 12 failures missed by routing. Such privileged state is free in simulation but hard to obtain on a real robot, where the world-frame poses of the arm and the objects must be reconstructed with costly, calibrated external instruments; reading the model's internal mechanism needs no added sensor, only the forward passes the policy already runs (3.14 ms per update on CPU; Appendix C). On jointly detected failures the outside monitors fire a median of 0–1 inference after routing and the task-aware rules two after the deployed NAR; routing follows the hindsight physical stall by a median of four inferences and leaves a median of 14 before termination, and a per-suite timer tuned on evaluation successes can match or exceed NAR, which needs neither task identity nor duration calibration (Appendix D.3).

5.3 RQ2: Does the alarm generalize across suites, and what carries it?

RQ2.1: The alarm transfers across LIBERO states, tasks and suites; a CALVIN stress test exposes a limit of cross-benchmark specificity. In Table 2, F1 stays at 0.83–0.88 from unseen initial states to unseen tasks and to a suite left out of the fit, and fitting on Goal successes alone still gives 0.82 on the three other suites (97.2% recall at 1.32% false alarms), weakest on LIBERO-10. We made the Goal profile our default after seeing this set and the 12,000 non-Goal episodes of seeds 50–57 (Appendix C), so it is not a blind test. Perturbed LIBERO-Plus episodes are absorbed (F1

Table 2: RQ2.1: each panel is one step further from the fit (native LIBERO unless marked; conformal threshold, $\alpha = 1.5\%$; FPR over all successes). (a) Final validation on the role grid, *two tasks per suite held out of the fit; (b) seen before the Goal profile became our default. More: Tables 8, 7.

Evaluated on	Fitted on	Episodes / fail.	Recall (%)	FPR (%)	F1 (%)
(a) Within LIBERO: new initial states, tasks, suite (pooled profile; last row leave-suite-out)					
Seen tasks, new initial states	all four suites*	2,560 / 99	86.9	0.93	82.7
Two held-out tasks per suite	all four suites*	3,200 / 85	95.3	0.61	87.6
A suite left out of the fit	the other three	5,760 / 184	88.6	0.72	84.2
(b) One suite in the fit, the other three at test (Goal profile)					
LIBERO-10 (weakest suite)	Goal only	1,440 / 91	96.7	3.34	78.6
All three	Goal only	4,320 / 141	97.2	1.32	82.3
(c) Beyond native LIBERO (LIBERO-Plus: pooled profile; CALVIN: Goal profile)					
LIBERO-Plus perturbations	all four suites	25,902 / 5,410	97.4	3.67	92.2
CALVIN D→D (subtasks)	Goal only	548 / 82	100.0	9.87	78.1

Table 3: RQ2.3: representation, scoring and temporal evidence on the 4,320 held-apart episodes (141 failures) of Table 2b; all variants recalibrated on Goal successes (post hoc). (a) Reduced descriptors, no bank; (b) distance functions; (c) temporal evidence rules. ΔF1: to the frozen row, 95% intervals except for (a). Rec.$_{55}$: recall at the frozen row's 55 false alarms, threshold set on the evaluation successes (descriptive). More: Table 9.

Variant	Rec.	FPR (%)	F1	ΔF1	Rec.$_{55}$
Frozen ($k = 20$, 8,192 bank points, mean of six)	0.972	1.32	0.823	–	0.979
(a) 17 lag-three contraction coordinates, no bank	0.872	1.08	0.796	−0.027	0.943
(a) Lag-one routing change, no bank	0.667	2.68	0.542	−0.281	0.582
(b) Distance to the bank median	0.858	1.01	0.796	−0.027 [−0.064, +0.015]	0.936
(b) Single Gaussian (Mahalanobis)	1.000	3.59	0.653	−0.170 [−0.250, −0.102]	0.922
(c) Raw score, no smoothing	0.950	1.51	0.793	−0.030 [−0.054, −0.006]	0.915
(c) Mean of three	0.965	1.36	0.814	−0.008 [−0.024, +0.007]	0.950
(c) Exponential average, newest weight 0.2	0.993	1.53	0.812	−0.011 [−0.034, +0.009]	0.979
(c) e-CUSUM of conformal p, $\kappa = 0.2$	1.000	2.68	0.716	−0.107 [−0.173, −0.052]	0.872

0.92; the pooled fit includes that family), a new benchmark is not: on CALVIN the unchanged Goal profile catches all 82 failures but also flags 46 of the 49 successes that reach the earliest alarm time (9.87% of all successes; Appendix D.4).

RQ2.2: A smaller bank preserves performance; transfer remains sensitive to source-success coverage. A bank of 2,048 points keeps recall 0.957 and F1 0.821, one of 1,024 points 0.915 and 0.801 (Table 9d), and either half of the reference episodes gives F1 0.818 and 0.828. The threshold is the conformal quantile of the calibration successes' maximum smoothed distance (Figure 5), so it moves with the bank (9.63–12.77) while the held-apart false-alarm rate stays at 1.24–1.44% (Figure 7). The bound covers successes exchangeable with the calibration pool: among successes long enough to be flagged, the realized rate is 1.79% on the 4,320 held-apart episodes and 3.83% on LIBERO-10 (Table 7; Appendix E.1). The pool needs perturbed successes: 12 of the 13 calibration maxima above the threshold are LIBERO-Plus, and a native-only refit gives threshold 6.56 and FPR 59.9%; as the refit changes bank, scaling and pool together, it shows sensitivity to source-success coverage without isolating a component.

RQ2.3: Recurrence carries most of the discrimination. The recurrence carries the score: without a bank, at the frozen 55 false alarms, each layer's change between consecutive inferences reaches recall 0.582 and the 17 lag-three contraction coordinates 0.943 (Table 3a); the lag is where successful routing decorrelates (RQ1.1), and the descriptor reads which expert is which (consistent renaming keeps 96.81% recall, renaming anew at every inference $\leq 0.17\%$; Appendix E.1). Scoring all 54 coordinates against the bank lifts recall to 0.979, though descriptor and scorer change together; with the descriptor fixed, the bank median and the k-NN distance are not resolved, while a Gaussian or a single neighbour costs 0.17–0.26 F1, mostly through the threshold (Table 3b). The mean trades time

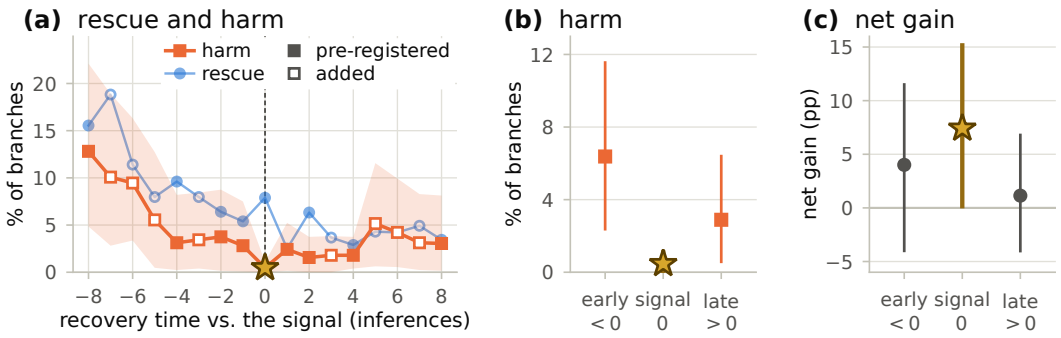


Figure 4: RQ3.2: recovery $\mathcal{R}(0.30, 3)$ at 17 times around the first Goal-profile signal (dashed line, star), each paired with the unmodified continuation from the same state (1,280 branches per time). **(a)** Harm (unmodified succeeds, recovery fails) and rescue (unmodified fails, recovery succeeds) per time; filled: pre-registered times, open: added after the first results were seen (Appendix D.8). **(b)** Harm and **(c)** net gain (rescue − harm) averaged over the early (-8 to -1) and late ($+1$ to $+8$) times, a grouping chosen after the curves were seen. Shading and bars: 95% intervals from 5,000 suite-stratified configuration bootstraps. Per-time values: Table 10.

for false alarms: raw scores raise more false alarms and catch fewer failures, yet fire earlier on most shared detections; a mean of three, an exponential average or a Fisher combination stays within noise, and e-CUSUM or Shiryaev–Roberts costs 0.11–0.12 F1 and fires later (Table 3c; Appendix E.1).

5.4 RQ3: Is the alarm time the right moment to act, and what decides the gain?

RQ3.1: Acting at the signal yields a positive paired effect in the full-corpus replay. We executed the replay-branch protocol at all 1,644 first alarms that the pooled profile, the frozen alarm at the time, raises in the corpus, with primitive $\mathcal{R}(0.60, 1)$ and the recorded noise stream (one seed; Appendix D.5); 1,608 cases completed, and the 36 incomplete ones, all in LIBERO-10, have no paired outcome and are excluded. Recovery rescues 22.9% [17.0, 28.7] of alarmed failures and harms 30.5% [21.6, 40.0] of alarmed successes, both against the twin. Because the alarm's precision against the twin (0.775) exceeds the break-even $\pi_\star = 0.571$ of Eq. 2, the mean paired gain is +10.88 pp [4.67, 16.83] per completed alarm: 175 additional successful episodes. Normalized by the 43,345 recorded episodes, this net count from completed pairs corresponds to +0.40 pp; it equals 12.7% of the corpus's 1,373 failures; 160 of them are in LIBERO-10 (22.6% of its failures; Appendix E.2).

RQ3.2: Recovery timing shifts the balance of rescue, harm and delivery. We applied $\mathcal{R}(0.30, 3)$ from 8 inferences before to 8 after the first Goal-profile signal on 40 alarm configurations, 8 of them false alarms, each with 32 noise streams shared across times (Figure 4; Appendix D.8). At the signal, recovery rescues 7.9% of branches and breaks 0.5% (net +7.4 pp [0.2, 15.2], the only pointwise interval excluding zero; not after Bonferroni correction). Earlier, the execution has not yet settled: 28% of unmodified continuations still succeed 8 inferences before the signal, against 19.5% at it, and recovery breaks more of them (harm 2.8–12.8%; the mean over earlier times exceeds the signal by +5.9 pp [+0.9, +13.2], Bonferroni-corrected over five contrasts fixed after the curves were seen). Later, rescue falls to 2.5–6.3% of branches and net gain to -0.9 to +4.8 pp, partly because fewer interventions fit before the cap (53% at +8; undelivered branches keep their native outcome): among delivered ones rescue is 2.6–8.6%, around the signal's, while harm rises to 1.6–7.4%. Earlier times rescue more (net +8.8 pp at -7): the signal has the lowest observed harm in this cohort, and the study does not establish an optimal time.

RQ3.3: The rescue–harm balance sets the gain; budget and depth go with rescue. In the twins the sign of the gain is set by whether precision exceeds the break-even: every suite's point precision exceeds its own break-even, but only LIBERO-10's gain excludes zero ($r = 40.9\%$, precision 0.662 against 0.415). Rescue is associated with the budget left (the bands are confounded with suite): alarms with fewer than ten inferences left rescue 2.3%, against 36–48% with more (Table 12). How far the evidence has crossed the threshold matters too: among the true alarms of 1,326 twins with

a recorded depth, rescue rises with it (Spearman $+0.195$ $[+0.094, +0.303]$), though depth does not separate true from false alarms (AUROC 0.517). Another trigger changes r and h as well as π, so these twins do not establish the control value of a timer or of a random trigger (precision 3.2%, the base failure rate).

Repeated recovery depends on the trigger rule. A second pulse six inferences later reduces success in the mixed cohort (-13.6 pp $[-20.9, -7.0]$). In an exploratory pilot (40 timing-study configurations, two new noise seeds each), recovering at every threshold exceedance cuts successes from 21 to 6 at the native cap B. An edge rule re-arms at $s_t \leq q_\alpha$ and intervenes at the next strict upward crossing; relative to one recovery it causes no additional failures in the 80 paired branches, with outcomes unchanged at B and two more successes at $4B$ ($+2.5$ pp $[0.0, 7.5]$; Table 4). The edge rule was selected after inspecting $2B$ results, and the extra successes occur beyond the native cap (Appendix E.3).

Table 4: RQ3.3: repeating the recovery (exploratory): successes of 80 paired branches at the step cap B, $2B$, $4B$.

After the first alarm	B	$2B$	$4B$
No recovery	18	21	23
Recover once	21	27	30
At every $s_t > q_\alpha$	6	6	–
Re-armed crossing	21	28	32

6 Limitations and scope

All results concern one mixture-of-experts policy family on LIBERO, one checkpoint per suite; dense VLAs expose no discrete routing, on CALVIN, in a capture below the benchmark score, the frozen profile keeps recall but not specificity (Appendix D.4), and no image- or language-model observer was run. The sensor reports that progress has stopped, after the physical stall; its advantage over the outside monitors rests on one seed block and on our adapters for the outside inputs, and elapsed time remains the strongest competitor wherever the task's success lengths are known (Appendix D.3). The monitor is training-free but not data-free: transfer needs perturbed Goal successes in its pools, its threshold rests on the few calibration episodes long enough to be scored (Appendix E.1), and the descriptor and default profile were chosen after results had been seen (Appendix C). The control study has one seed per arm and a primitive that harms a third of the healthy executions it touches, and the timing curve rests on 40 configurations. Whether routing can also select the response is unresolved: the selectors in Table 14 show no resolved improvement over their comparators, and with 66 alarmed failures, mostly LIBERO-Plus, 80% power is reached only around AUROC 0.70 (Appendix E.3).

A re-armable test-time harness. N, A and R close a loop around the frozen policy: control returns to the same policy, whose routing is read again. In the exploratory 80-branch pilot, re-arming preserves one-recovery outcomes at the native cap and adds successes only beyond it (Section 5.4). Re-reading routing after handback flags episodes that will still fail (precision 0.980, recall 0.136; Appendix E.3).

7 Conclusion

Expert-routing recurrence is a useful runtime signal for monitoring frozen HiMoE-VLA policies. Under one protocol it reaches higher failure recall than the outside readouts we tested; NAR transfers a reference fitted on source successes across LIBERO suites, with false-alarm behaviour set by source coverage and calibration; and paired replay shows that acting on an alarm can pay for a fixed recovery, while moving the intervention trades broken recoverable executions against lost chances to act.

Three questions follow, none requiring a change to the frozen policy: whether a dense VLA, which exposes no discrete routing, carries a comparable recurrence signal in a continuous internal trace; whether the perturbed source successes that the threshold needs can be collected on a real robot, where they cannot be simulated; and whether a re-armed harness, which repeated the primitive without added failures in a pilot, can pay within the native step cap. Two properties make these steps tractable: the monitor reads the policy's own forward pass at 3.14 ms per update on CPU (Appendix C), and it needs neither the full world state that a simulator exposes nor a label from any failed run, so moving it to hardware means recording successful runs and refitting the bank and threshold. In the policies we studied, internal states do signal trouble; whether they can also guide the response beyond a fixed primitive is the question that remains.

AI USE STATEMENT

We used generative AI tools, in the form of a large-language-model coding assistant, for writing and refactoring analysis and experiment-orchestration code, for read-only audits that re-derived reported numbers from their source files, for generating figures from recorded data, for retrieving bibliographic metadata of cited works, and for drafting and editing the manuscript text. Research questions, experimental decisions and the choice of what to report were made by the authors. We did not use generative AI to produce or alter any experimental data. AI-assisted code was tested against recorded reference outputs, every reported number was re-derived from its source file by a separate audit, and bibliographic entries were checked against the arXiv and venue pages; the authors reviewed all AI-assisted code, analyses and text and take responsibility for the final content, including claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All experiments are in simulation, on LIBERO, LIBERO-Plus and a CALVIN transfer check; no human subjects or personal data are involved. The alarm bounds false alarms only under exchangeability with its calibration episodes and provides no safety guarantee under distribution shift. The recovery primitive opens the gripper, which on a physical robot can drop a held object; any real-world use would need hardware-appropriate safety checks that we have not studied.

REPRODUCIBILITY STATEMENT

Section 3 and Appendices A–B define the monitor completely; Appendix C specifies the frozen assets (bank, scales and thresholds), and the packaged implementation reproduces the recorded descriptors exactly on the included raw-routing samples. Appendix D gives every control protocol, including amendments made after interim inspection. Per-episode trajectories and routing captures of the control studies were kept in RAM-backed storage by design; several were later mirrored to an archive directory, and code, protocols and aggregate results are persisted.

REFERENCES

Christopher Agia, Rohan Sinha, Jingyun Yang, Zi-ang Cao, Rika Antonova, Marco Pavone, and Jeannette Bohg. Unpacking Failure Modes of Generative Policies: Runtime Monitoring of Consistency and Progress. In *Conference on Robot Learning (CoRL)*, 2024. arXiv:2410.04640.

Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Abhijith Babu, Ramneet Kaur, Nathaniel D. Bastian, Olivera Kotevska, Susmit Jha, Yanzhao Wu, Sumit Kumar Jha, and Anirban Roy. Closed-Loop Neural Activation Control in Vision-Language-Action Models. *arXiv preprint arXiv:2606.00269*, 2026.

Kevin Black, Manuel Y. Galliker, and Sergey Levine. Real-Time Execution of Action Chunking Flow Policies. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.07339.

Hugo Buurmeijer, Carmen Amo Alonso, Aiden Swann, and Marco Pavone. Observing and Controlling Features in Vision-Language-Action Models. *arXiv preprint arXiv:2603.05487*, 2026.

Kang Chen, Yaoning Wang, Kai Xiong, Zhuoka Feng, Wenhe Sun, Haotian Chen, and Yixin Cao. Do LLMs Signal When They're Right? Evidence from Neuron Agreement. In *International Conference on Machine Learning (ICML)*, 2026a. arXiv:2510.26277.

Kang Chen, Fan Yu, Junjie Nian, Shihan Zhao, Zhuoka Feng, Zijun Yao, Heng Wang, Minshen Yu, and Yixin Cao. Thinking Traps in Long Chain-of-Thought: A Measurable Study and Trap-Aware Adaptive Restart. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026b. arXiv:2601.11940.

Kang Chen, Minshen Yu, Junjie Nian, Yaoning Wang, Yixin Cao, and Yugang Jiang. Does the Same Token Mean the Same State? MoE Routing as Signal for Reasoning Control. *arXiv preprint arXiv:2606.22798*, 2026c.

Zhiying Du, Bei Liu, Yaobo Liang, Yichao Shen, Haidong Cao, Xiangyu Zheng, Zhiyuan Feng, Zuxuan Wu, Jiaolong Yang, and Yu-Gang Jiang. HiMoE-VLA: Hierarchical Mixture-of-Experts for Generalist Vision-Language-Action Policies. *arXiv preprint arXiv:2512.05693*, 2025.

Jiafei Duan, Wilbert Pumacay, Nishanth Kumar, Yi Ru Wang, Shulin Tian, Wentao Yuan, Ranjay Krishna, Dieter Fox, Ajay Mandlekar, and Yijie Guo. AHA: A Vision-Language-Model for Detecting and Reasoning Over Failures in Robotic Manipulation. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.00371.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: A Progressive Robustness Benchmark for Visual-Language-Action Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2510.13626.

Clemence Grislain, Hamed Rahimi, Olivier Sigaud, and Mohamed Chetouani. I-FailSense: Towards General Robotic Failure Detection with Vision-Language Models. *arXiv preprint arXiv:2509.16072*, 2025.

Qiao Gu, Yuanliang Ju, Shengxiang Sun, Igor Gilitschenski, Haruki Nishimura, Masha Itkina, and Florian Shkurti. SAFE: Multitask Failure Detection for Vision-Language-Action Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.09937.

Bear Häon, Kaylene Stocking, Ian Chuang, and Claire Tomlin. Mechanistic Interpretability for Steering Vision-Language-Action Models. In *Conference on Robot Learning (CoRL)*, 2025. arXiv:2509.00328.

Haodi Hu, Chung-Ta Huang, Jing Liu, Ye Wang, Kei Suzuki, Matthew Brand, and Toshiaki Koike-Akino. ReCoVLA: VLM-Guided Reward Compilation for Failure Recovery in Vision-Language-Action Policies. *arXiv preprint arXiv:2606.09630*, 2026.

Zheyuan Hu, Robyn Wu, Naveen Enock, Jasmine Li, Riya Kadakia, Zackory Erickson, and Aviral Kumar. RaC: Robot Learning for Long-Horizon Tasks by Scaling Recovery and Correction. *arXiv preprint arXiv:2509.07953*, 2025.

Dehao Huang, Aoxiang Gu, Chengjie Zhang, Bolin Zou, Wenlong Dong, Zilang Cen, Yue Wang, and Hong Zhang. VLAConf: Calibrated Task-Success Confidence for Vision-Language-Action Models. *arXiv preprint arXiv:2605.29605*, 2026.

Hyun Joe Jeong, Gokul Swamy, and Andrea Bajcsy. Learning What to Say to Your VLA: Mostly Harmless Vision Language Action Model Steering. In *Conference on Robot Learning (CoRL)*, 2026. arXiv:2606.12299.

Jacky Kwok, Xilun Zhang, Mengdi Xu, Yuejiang Liu, Azalia Mirhoseini, Chelsea Finn, and Marco Pavone. Scaling Verification Can Be More Effective than Scaling Policy Learning for Vision-Language-Action Alignment. *arXiv preprint arXiv:2602.12281*, 2026.

Yousung Lee and Dongsoo Har. Perturbation-Based Epistemic Uncertainty for Failure Detection in Vision-Language-Action Models. *arXiv preprint arXiv:2606.20754*, 2026.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Lin Li, Long Chen, Kwunhang Wong, Jiaming Lei, Song Jin, Shucheng Du, Chuhan Zhang, Songchen Ma, Weihao Zhang, Jun Xiao, and Kwang-Ting Cheng. Causal-History Test-Time Scaling for Failure Recovery in Autoregressive World-Action Models. *arXiv preprint arXiv:2609.18016*, 2026.

Zijun Lin, Jiafei Duan, Haoquan Fang, Dieter Fox, Ranjay Krishna, Cheston Tan, and Bihan Wen. FailSafe: Reasoning and Recovery from Failures in Vision-Language-Action Models. *arXiv preprint arXiv:2510.01642*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023. arXiv:2306.03310.

Lin Liu, Zhicheng Bao, Lu Zhang, Ziying Song, Wu Yang, Yuzheng Zhuang, Shuai Tao, Wulong Liu, Caiyan Jia, and Huchuan Lu. LIBERO-RECOVER: Beyond Task Success Towards Failure Recovery in Robotic Manipulation Models. *arXiv preprint arXiv:2609.05178*, 2026.

Chenyang Ma, Kai Lu, Guangyu Yang, Jiuming Liu, Shitong Xu, Bill Byrne, Ioannis Havoutis, Niki Trigoni, and Andrew Markham. CycleVLA: Proactive Self-Correcting Vision-Language-Action Models via Subtask Backtracking and Minimum Bayes Risk Decoding. *arXiv preprint arXiv:2601.02295*, 2026a.

Jie Ma, Zongxi Liu, and Yi Zhu. FailureSpot: Label-Efficient Timestamp-Level Failure Detection for Vision-Language-Action Models. *arXiv preprint arXiv:2609.04277*, 2026b.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. CALVIN: A Benchmark for Language-Conditioned Policy Learning for Long-Horizon Robot Manipulation Tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

Miranda Muqing Miao, Subin Kim, Brandon Yang, and Lyle Ungar. Contrastive Conceptor Activation Steering (COAST): Unlocking Vision-Language-Action Models through Hidden States. *arXiv preprint arXiv:2605.17144*, 2026.

Paul Pacaud, Ricardo Garcia, Shizhe Chen, and Cordelia Schmid. Guardian: Detecting Robotic Planning and Execution Errors with Vision-Language Models. *arXiv preprint arXiv:2512.01946*, 2025.

E. S. Page. Continuous Inspection Schemes. *Biometrika*, 41(1/2):100–115, 1954.

Seongheon Park, Wendi Li, Changdae Oh, Samuel Yeh, Zsolt Kira, Michael Hagenow, and Sharon Li. Hide-and-Seek in Trajectories: Discovering Failure Signals for VLA Runtime Monitoring. *arXiv preprint arXiv:2605.30834*, 2026.

Carl Qi, Xiaojie Wang, Silong Yong, Stephen Sheng, Huitan Mao, Sriram Srinivasan, Manikantan Nambi, Amy Zhang, and Yesh Dattatreya. Self-Refining Vision Language Model for Robotic Failure Detection and Reasoning. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2602.12405.

Harshitha Rajaprakash, Aditeya Prajapati, Rong Xue, Abrar Anwar, and Jesse Thomason. SAFE-CAST: Robust Failure Detection for VLA Policies with Contrast-Set Training and Calibration. *arXiv preprint arXiv:2608.04246*, 2026.

Florian Seligmann, Emiliyan Gospodinov, Enes Ulas Dincer, and Gerhard Neumann. VLA-FAIL: Efficient Task Failure Detection for Finetuned Vision-Language-Action Models. *arXiv preprint arXiv:2606.21386*, 2026.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Suyeon Shin, Juwon Kim, Hyeonbin Park, Hyunseo Kim, Hyundo Lee, Hyung-Sin Kim, and Byoung-Tak Zhang. Back to the Familiar Future: Failure Recovery for VLA Policies via Pre-Imagined Milestone Selection. *arXiv preprint arXiv:2606.09258*, 2026.

Chang Song, Bin Qian, Yan Feng, and Zhijie Song. VLA-Corrector: Stage-Aware Observable State Understanding for Prompt-Based Closed-Loop Recovery of Vision-Language-Action Policies. *arXiv preprint arXiv:2609.06508*, 2026.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-Distribution Detection with Deep Nearest Neighbors. In *International Conference on Machine Learning (ICML)*, 2022.

Aiden Swann, Lachlain McGranahan, Hugo Buurmeijer, Monroe Kennedy III, and Mac Schwager. Sparse Autoencoders Reveal Interpretable and Steerable Features in VLA Models. *arXiv preprint arXiv:2603.19183*, 2026.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

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. In *Robotics: Science and Systems (RSS)*, 2025. arXiv:2503.08558.

Fan Zhang, Seongbin Park, Baharan Mirzasoleiman, Shariar Talebi, and Nader Sehatbakhsh. Probe-Act: Probe-Guided Training-Free Failure Recovery in Vision-Language-Action Models. *arXiv preprint arXiv:2606.09740*, 2026a.

Haoran Zhang, Yifu Lu, Boyang Wang, Xuhui Kang, Yen-Ling Kuo, Zezhou Cheng, Mengdi Wang, and Odest Chadwicke Jenkins. Foresight: Failure Detection for Long-Horizon Robotic Manipulation with Action-Conditioned World Model Latents. *arXiv preprint arXiv:2606.23085*, 2026b.

Likui Zhang, Tao Tang, Zhihao Zhan, Xiuwei Chen, Zisheng Chen, Jianhua Han, Jiangtong Zhu, Pei Xu, Hang Xu, Hefeng Wu, Liang Lin, and Xiaodan Liang. AtomicVLA: Unlocking the Potential of Atomic Skill Learning in Robots. *arXiv preprint arXiv:2603.07648*, 2026c.

Yanyan Zhang, Disheng Liu, Kai Ye, Chaoda Song, Xinpeng Li, Mohsen Hariri, Vikash Singh, Yu Yin, and Vipin Chaudhary. Imagining Recovery: Inference-Time Counterfactual Realignment for Vision-Language-Action Models. *arXiv preprint arXiv:2608.14822*, 2026d.

Xueyang Zhou, Yangming Xu, Guiyao Tie, Yongchao Chen, Guowen Zhang, Duanfeng Chu, Pan Zhou, and Lichao Sun. LIBERO-PRO: Towards Robust and Fair Evaluation of Vision-Language-Action Models Beyond Memorization. *arXiv preprint arXiv:2510.03827*, 2025.

Kaiwen Zhu, Dongfang Liu, and Liangkai Liu. VLA-Scope: Shift-Aware Failure Prediction for Vision-Language-Action Models. *arXiv preprint arXiv:2609.21246*, 2026.

A THE DESCRIPTOR IN FULL

Three routing views. Drop the state token and normalize each recorded probability vector to sum to one. Write $R_t = (p_t, B_t)$ for eight routed layers, ten denoising steps and ten action tokens. For a subset V of sites and a subset G of layers, define

$$d_V(R_i, R_j)^2 = \frac{1}{|V|} \sum_{\sigma \in V} \frac{|E_i(\sigma) \triangle E_j(\sigma)|}{8}, \tag{3}$$

$$d_G(R_i, R_j)^2 = \frac{1}{32|G|} \sum_{\ell \in G} \left\| \frac{u_i^\ell - u_j^\ell}{100} \right\|_2^2, \tag{4}$$

$$m_t^\ell = \frac{1}{10} \sum_a H(p_t^{\ell,10,a}, p_{t-1}^{\ell,10,a}), \qquad H^2(p,q) = 1 - \sum_e \sqrt{p_e q_e}. \tag{5}$$

The support metric is the root mean squared Hellinger distance between uniform top-4 measures, since $H^2(U_S, U_{S'}) = |S \triangle S'|/8$ for two four-element supports. The participation metric compares expert participation counts $u^\ell = \mathrm{diag}((B^\ell)^\top B^\ell)$. Because every site selects exactly four experts,

the weighted degree of expert e in the off-diagonal co-selection graph equals $3u_e^\ell$, so participation count and graph degree differ only by a constant; the frozen variant uses this diagonal and not the off-diagonal residual. The probability view supplies the local-motion block. These are complementary projections, not sufficient statistics of the policy computation.

The 44 recurrence coordinates. Inference ranges include both endpoints. There are fifty support views: all sites, the back four layers, eight single layers, ten action-token views over all layers and ten over the back layers, and the analogous twenty denoising-step views. For each view, β is the median over all ten pairs in inferences $0{:}4$, and ρ_t is the median over the six pairs with lag at least three in inferences $t-5{:}t$. The floor in the log ratio is 10^{-7}, and distances are rounded to seven decimal places before the medians are taken. For either C or ρ, the pooling Q keeps the all-site and back-layer values and takes the 10%, 50% and 90% quantiles of each of the five remaining families, giving $2 + 5 \times 3 = 17$ coordinates per block. The ten participation contractions use the all-layer, back-layer and eight single-layer views. Hence $17 + 17 + 10 = 44$ coordinates become available at $t = 5$. Quantile pooling removes ordering within a family; it does not make the descriptor invariant to arbitrary layer permutations.

The ten local-dynamics coordinates. Let $\mathcal{T}_w$ denote a trailing mean of w finite values, unavailable until the whole window exists. The first eight coordinates are

$$a_t^\ell = \mathcal{T}_6\left[-\log \frac{\max(m_t^\ell, 10^{-6})}{\max(\frac{1}{4}\sum_{j=1}^{4} m_j^\ell, 10^{-6})}\right]. \tag{6}$$

The ninth uses denoising curvature $\gamma_t = \mathrm{mean}_{\ell\in\mathrm{back},s,a} \|\Delta_s^2 \sqrt{p_t^{\ell,s,a}}\|_2/\sqrt{2}$ and equals

$$\mathcal{T}_3\big[\log\big(\max(\gamma_t, 10^{-6})/\max(\mathrm{med}\,\gamma_{2:7}, 10^{-6})\big)\big].$$

For the tenth, let $J_{t,t-k}$ be the mean weighted Jaccard similarity between back-layer final-step action-token probability vectors and set $\eta_t = \max_{k\in\{2,3,4\}} J_{t,t-k} - J_{t,t-1}$ over available lags; the coordinate is $\mathcal{T}_6[-(\eta_t - \mathrm{med}\,\eta_{2:6})/\sigma_0]$ with the inherited scale $\sigma_0 = 0.020707666873931885$. Baseline-relative values are never emitted before their baseline exists, and the local block is masked until $t = 7$.

Causality, invariance and startup. For every t, x_t is a deterministic function of $R_{0:t}$; later observations cannot change it. A renaming of expert identities that is consistent over the whole episode preserves every metric. In exact arithmetic, a positive rescaling of a metric cancels in the contraction ratio while the floors are inactive, but not in the absolute block ρ_t. The descriptor first exists at $t = 7$ and the trailing mean of six scores at $t = 12$, so the earliest alarm uses thirteen observations.

The lag-three rule has an empirical motivation: on LIBERO-10 successes (native and LIBERO-Plus) the routing variogram reaches 94.9%, 98.2% and 99.4% of its plateau at lags one, two and three. This supports a short decorrelation scale; it does not derive a unique lag. The prefix uses lags one to four and the current window lags of at least three, so their ratio is a reference-relative recurrence statistic rather than an unbiased comparison of identical lag distributions.

B THE CONFORMAL ARGUMENT

Condition on the frozen feature map and reference bank. Let $M^{(1)}, \ldots, M^{(n)}$ be the whole-episode maxima of the evidence statistic on the n calibration successes, and $M^\star$ the maximum on a new successful episode. If the $n + 1$ maxima are exchangeable, rank symmetry gives, for $r = \lceil (n + 1)(1 - \alpha) \rceil \le n$,

$$\Pr\big(M^\star > M_{(r)}\big) \le \frac{n + 1 - r}{n + 1} \le \alpha. \tag{7}$$

Ties with a strict crossing are conservative. The threshold is $+\infty$ when $r > n$, and an episode with no scorable inference has maximum $-\infty$ and stays in the pool. The event $M^\star > q_\alpha$ is exactly the event that the causal monitor emits at least once, so the monitor need not observe the future maximum to implement the test. The guarantee concerns false alarms on successful episodes drawn like the

calibration episodes. It gives no protection under distribution shift, no bound on failure recall, and no statement about the value of intervening. Because unscorable short successes stay in the pool with maximum $-\infty$, the nominal level applies to the whole success distribution; on the scorable subset the effective level is higher (Appendix C).

C FROZEN ARTIFACTS AND HOW THEY WERE CHOSEN

Identity. Both profiles share the descriptor of Appendix A (`dyn_degree54`), robust scaling $z = D^{-1}(x - \mu)$ with coordinate medians μ and $D_{jj} = \max(\mathrm{IQR}_j/1.349, 10^{-3})$ taken from the bank, $k = 20$ mean nearest-neighbour distance, a trailing mean of six scores with a strict crossing, and $\alpha = 0.015$ with the per-episode-maximum split-conformal threshold. A non-finite score clears the evidence window. The *Goal* profile (`graph_degree/goal`) has threshold 9.901234276656902 and the *pooled* profile (`graph_degree/pooled`) threshold 9.8784454575703. The packaged implementation reproduces the recorded descriptors of 16 raw-routing samples with maximum absolute difference 0, and refitting the Goal assets from the packaged data reproduces bank, scales and threshold exactly.

Fitting pools. Roles come from an outcome-blind partition of a 27,168-episode table (native 16,000 and LIBERO-Plus 11,168, seeds 42–49): tasks 4 and 9 of each suite form task validation, and remaining configurations are ordered outcome-blind within family, suite and task and assigned in blocks of ten to reference (4), calibration (2), development (2) and validation (2). All eight seeds of a configuration share a role. The Goal profile uses the Goal suite's 2,094 reference successes (1,256 native, 838 LIBERO-Plus) and 942 calibration successes (625 native, 317 LIBERO-Plus) over tasks 0–3 and 5–8; its sampler balances groups, so 4,096 of the 8,192 bank points are LIBERO-Plus. The pooled profile uses 8,428 reference and 3,863 calibration successes of all suites and both families.

The effective conformal level. Calibration episodes too short to be scored enter with maximum $-\infty$, so the nominal level covers the whole success distribution. Only 289 of the 942 Goal calibration episodes and 2,191 of the 3,863 pooled ones are scorable; on those, the in-sample exceedance is 4.5% and 2.6%. Twelve of the thirteen Goal exceedances are LIBERO-Plus episodes.

How the variant was selected. The frozen comparison contained 40 fitted candidates and three historical anchors. Its pre-registered primary candidate was a different statistic; the degree variant entered as a matched control and was designated the default after results were opened. The Goal profile became our default after its result on the seeds 50–57 block and on the cross-suite validation set had been computed. Candidate constants (prefix, window, pooling quantiles) were chosen, not derived. We therefore present the frozen variant as what it is, and report corpus-wide numbers alongside the held-out ones.

Which parts of the descriptor matter. On the 16,000-episode native block of seeds 50–57 with pooled calibration, dropping the participation block (a 44-coordinate descriptor) changes F1 by $+0.0007$ $[-0.0041, +0.0056]$, while keeping only the 17 relative-contraction coordinates costs -0.0175 $[-0.0251, -0.0100]$ and a single contraction scalar -0.0337 $[-0.0459, -0.0219]$, all through extra false alarms. Under Goal calibration the 44-coordinate variant falls to F1 0.754 against 0.832 at $\alpha = 0.015$, a threshold-placement effect (swept over α the two peak at 0.847 and 0.846); the local-motion block alone never exceeds 0.318. Under pooled calibration on native LIBERO-10 of the same block, $k = 1$ gives F1 0.772 against 0.745 for $k = 20$ (under the Goal profile it costs 0.257, Table 9), and replacing the absolute distance by a local density ratio drops F1 to 0.584. For the contraction block alone, a minimum lag of two classifies the native LIBERO-10 validation episodes exactly as lag three does, while lag one misses two more of the 91 failures (F1 0.764 against 0.772). Adding elapsed-time coordinates helps the pooled profile slightly ($+0.0058$ $[+0.0021, +0.0098]$) but collapses the Goal profile (F1 0.368, FPR 10.9%), which is why the descriptor reads no time.

Cost. The monitor runs on CPU. One update of the packaged pooled monitor takes 3.14 ms on average (p_{99} 6.37 ms), and the online sidecar including inter-process transfer took 3.6 ms per inference, against roughly 2 s per inference for the simulated rollout with policy service.

D PROTOCOLS

D.1 POLICY, CORPUS, COHORTS AND OFFLINE DETECTION REPLAY

Table 5: Cohorts of Section 5 (native LIBERO episodes unless marked). Table 1a uses configuration-level cross-fitting within seeds 50–57, with each evaluated fold excluded from every fitting stage; Table 1b uses the frozen Goal profile. The transfer and control cohorts and their uses are listed below.

Cohort	Episodes / failures	Used in
All recorded episodes with routing (corpus)	43,345 / 1,373	RQ2.1 (Table 7)
Validation roles of LIBERO-10, Object and Spatial	4,320 / 141	RQ2 (Tables 7, 3)
Seeds 42–49, LIBERO-10	4,000 / 275	RQ1.1 (Figure 3)
Seeds 50–57 block, 2,000 configurations	16,000 / 509	RQ1.2–1.3, RQ2 checks
reaching the 13th inference	9,678 / 509	RQ1.2–1.3 (Table 1)
LIBERO-10, native and LIBERO-Plus	5,768 / 680	RQ1.1, RQ2.2
External LIBERO-Plus episodes	25,902 / 5,410	RQ2.1
CALVIN D→D subtasks, 150 sequences	548 / 82	RQ2.1 (Appendix D.4)
First pooled-profile alarms, replay twins	1,608 alarms	RQ3
Goal-profile alarms, recovery at 17 times	40 configs × 32 streams	RQ3 (Figure 4)

The policy is HiMoE-VLA, one fine-tuned checkpoint per suite, served with bf16 autocast and float32 weights. Each inference receives two 224×224 images and an 8-dimensional state, runs ten flow-matching steps and returns a chunk of ten 7-dimensional actions (six end-effector deltas and the gripper, -1 meaning open), all of which are executed before the next inference. Routing is recorded for the eight hierarchical-MoE layers (indices 2–5 and 12–15) as probabilities of shape $8 \times 10 \times 11 \times 32$ and the selected top-4 identities; the monitor drops the state token. Episodes start with ten settling steps. Native caps are 520, 300, 280 and 220 actions (LIBERO-10, Goal, Object, Spatial). The corpus holds 43,477 native episodes: the 16,000-episode role grid (2,000 configurations, 10 tasks $\times$ 50 initial states per suite; flow-noise seeds 42–49), a 16,000-episode block with seeds 50–57 that reuses the same 2,000 configurations, and 11,477 further episodes from earlier and additional runs; 43,345 of them have complete routing, and the other 132 (two of them failures) enter no denominator. Native episodes ran through the LIBERO-PRO (Zhou et al., 2025) code base with its perturbation off. All used the same simulator seed and wrist-camera layout. Because the same configuration recurs across seeds, episodes are not independent at the configuration level; the seeds 50–57 block is a subset of the corpus, not an independent reproduction. Detection replay gives each recorded episode a fresh monitor, feeds recorded routing one inference at a time, stops at the first signal or at the end, and only then reads the recorded outcome; there is no refitting, gating or duration filter, and a last-inference alarm counts.

Terminology. A *configuration* is a (suite, task, initial state) triple; a *seed block* runs each configuration once per noise seed; the episode branched at an alarm is the *parent*; an alarm whose floor arm fails is a *hit*; *handback* returns control to the policy after the primitive; a *screen* draws alarm states in a fixed outcome-blind order; *native* means unperturbed LIBERO.

Hardware and software. All rollouts ran on one node with eight NVIDIA H20-3e GPUs (141 GB each, driver 580.105.08), two Intel Xeon Platinum 8575C CPUs (192 threads) and 2 TB of memory, under Ubuntu 24.04. The policy is the official HiMoE-VLA PyTorch implementation (`github.com/ZhiyingDu/HiMoE-VLA`, commit `27a2c46`, with our serving patches, which every server records) in Python 3.11 with PyTorch 2.6.0 (CUDA 12.4) and transformers 4.48.1. It is served over WebSocket by batched servers that run up to 32 requests per forward pass under bf16 autocast; in the main collection runs sixteen servers, two per GPU, served the four suites. Batched bf16 inference is not bit-identical to single-request inference (action differences up to about 10^{-2}); every corpus episode was recorded through these batched servers. Routing is read by PyTorch forward hooks on the gate submodules of the eight HB-MoE layers: each hook recomputes the gate's 32-way softmax from the gate input, checks on the first four inferences that its top-4 agrees with the gate's own selection, keeps the top-4 indices and the full probabilities on the device, and copies them to the host once per inference into a Zarr store; the forward pass is otherwise unchanged. The simulator runs on CPU in Python 3.8 with MuJoCo 3.2.3 and robosuite 1.4.0, through the LIBERO-PRO code base

(commit `eafdb80`) with its perturbation off over the original LIBERO assets (commit `8f1084e`), rendering with EGL; LIBERO-Plus episodes use the LIBERO-plus code base (commit `4976dc3`). The monitor and all analyses run on CPU.

D.2 The earlier detector used for the privileged-input comparison

An earlier exploratory comparison used a detector from an earlier stage of this project, not NAR, so that every input was scored by one fixed pipeline; it is distinct from the NAR-based comparison of Section 5.2 and Table 1. For each group (suite, base task) of the seeds 50–57 block it builds an atlas from the inferences of successful episodes in other configuration blocks than the scored episode; each inference is mapped to its nearest-neighbour distance d and to the spread s of the relative episode times of its ten nearest atlas points, both z-standardized on the group's successes. Eight phase bands of these coordinates feed a CUSUM, whose threshold is the $(1-\alpha)$ quantile of the CUSUM peaks on the success bank with $\alpha = 0.01$; an alarm is the first crossing from the fifth inference on. Unlike NAR it conditions on task identity through the atlas. Inputs were the $12{,}800$-dimensional recorded top-4 routing points, the complete simulator state before the action (robot and object positions and velocities, clock removed), object pose alone, the 8-dimensional proprioceptive state and the flattened 70-dimensional predicted chunk. Its length-only sentinel replaces every inference's input by the mean of successful episodes at the same inference index and runs the identical pipeline. Intervals are configuration-cluster bootstraps. On the $9{,}678$ episodes of seeds 50–57 with at least thirteen inferences (509 failures; realized false-alarm rates 1.9–2.2%), routing reaches recall 0.916 $[0.872, 0.955]$, the complete simulator state 0.782 $[0.724, 0.838]$, proprioception 0.758, the predicted action chunk 0.786 and object pose 0.550 (routing minus state $+0.133$ $[+0.089, +0.182]$); the union of state and object pose reaches 0.806, 69 failures are caught by routing alone and 13 only by a privileged input. The length-only sentinel reaches 0.507 from routing (precision 0.42) and 0.367 from full state, and over its own sentinel each input gains about the same ($+0.409$ and $+0.416$). Where both detect a failure, routing fires a median two inferences before the state-fed detector.

D.3 Inside versus outside under NAR's pipeline

The comparison of Section 5.2 uses the $16{,}000$ native episodes of seeds 50–57 (509 failures), whose traces store the simulator state before each inference, the 8-dimensional proprioceptive state and the predicted chunk (no images). The $2{,}000$ configurations are split into five folds by a fixed outcome-blind assignment; for each evaluation fold, successes of the other four are split by configuration about 2:1 into bank and calibration episodes, and the evaluation fold never enters bank, scaling, sentinel means, calibration or the choice of variant. Routing's realized false-alarm rate on all successes of the block is 1.58% $[1.24, 1.94]$ for $\alpha = 1.5\%$. Every arm then uses NAR's recipe: robust scaling, mean distance to the $k = 20$ nearest of $8{,}192$ bank points, the mean of six scores, and the episode-maximum split-conformal threshold at $\alpha = 0.015$, with the first alarm possible at the thirteenth inference. Routing uses the frozen 54-dimensional descriptor. Each outside input becomes a descriptor of the same form: per view (the whole vector and fixed semantic blocks, such as robot and scene positions and velocities for the simulator state), the contraction and the log of the median distance between inferences at least three apart in the last six, against the median over the first five inferences; a second variant appends the scaled current input, and the variant is chosen per evaluation fold by an inner cross-fit on the other four folds. The causal stall detector scores the log ratio of the trailing five-inference end-effector displacement to the episode's early displacement. The sentinel of an arm replaces every input by the mean input of bank successes of the same suite and task at the same inference index. The hindsight stall time follows the ad hoc rule of Section 5.2 (future five-inference displacement below a quarter of the early typical value, sustained). Intervals resample configurations with all eight seeds together. All checks pass except the conformal-validity check of one unselected variant (action chunk, snapshot; realized false-alarm rate 1.95% against a bound of 1.91%). Per suite, with each input's false alarms matched to routing's within the suite, routing minus the input in recall is $+0.043$ to $+0.122$ on LIBERO-10 (255 failures; every interval excludes zero); on Goal (95), Object (39) and Spatial (120) it is 0.000, $+0.051$ and -0.042 for the action chunk, $+0.295$, 0.000 and -0.033 for proprioception and $+0.126$, $+0.205$ and $+0.158$ for the complete simulator state.

Table 6: Details of Table 1. Only R / M: failures caught by routing but not by the monitor, and the reverse (of 509). ΔRec.: routing's recall minus the input's, 95% paired intervals, at matched false-alarm rates: thresholds set on the evaluation successes so that each input raises as many false alarms as routing (2.66%) or as the frozen pooled-profile NAR (1.55%); descriptive. Later: median inferences by which the monitor fires after routing on failures both catch. Stall: median alarm index minus the hindsight stall index.

Monitor	Only R	Only M	ΔRec. at 2.66%	ΔRec. at 1.55%	Later	Stall
Routing, cross-fitted	–	–	–	–	–	+4
Complete simulator state	56	6	+0.100 [0.057, 0.154]	+0.102 [−0.005, 0.165]	+1	+6
Proprioception (8-d)	50	2	+0.114 [0.069, 0.194]	+0.194 [0.092, 0.272]	+1	+5
Predicted action chunk	23	5	+0.028 [0.006, 0.069]	+0.033 [−0.021, 0.078]	0	+4
End-effector motion	40	4	+0.106 [0.051, 0.157]	+0.122 [0.029, 0.177]	+1	+7
Causal stall detector	52	6	+0.098 [0.051, 0.165]	+0.175 [0.069, 0.257]	+1	+7
Routing, deployed NAR	–	–	–	–	–	–
Task-aware physical rules	75	12	–	–	+2	–

Alarm timing. On failures both detect, the outside inputs fire a median of 0 to 1 inference after routing (every median's interval within $[0, 2]$), later on 43–67% of these failures and earlier on 20–36%; the task-aware rules fire a median of two after the frozen NAR (later on 62%). Routing leaves a median of 14 inferences, at least ten on 61% of its detections. None of the inputs anticipates the physical event: all fire after the stall of the end effector (median 4 $[4, 5]$ inferences for routing, 4–7 for the others; a hindsight reference). NAR's raw score needs about three inferences to rise and the six-score mean adds two (pooled profile; 228 alarmed LIBERO-10 failures). In the retrospective timing cohort of Section 5.4, intervention at the alarm has the lowest observed harm.

Task-aware physical rules. To give the outside view the strongest reading we could build, a monitor reads the full simulator state at every recorded inference, restored from the recording without stepping the simulator, together with the task's own goal predicates. For each unfinished goal it tracks progress (object-to-target distance, or the goal joint's distance to its threshold), gripper contact and the end effector's response to the executed chunk. Four rule families were fixed before any result: a stage stall (no progress on the same goal and phase over a window), a weak action response (commanded motion without observed motion or progress), a combined rule (either, or recurrence without progress, or grasp loss under a closing command) and a collapse of motion relative to the episode's start; windows of 4, 8 or 12 inferences, two progress tolerances and one to three consecutive conditions give 81 rules. A rule is chosen by F1 on four of five task-stratified configuration folds (eight seeds together) and applied to the fifth; every fold chose the combined rule with a window of 12. Single families reach recall 0.64–0.66. The frozen NAR (Goal profile) is not refitted; its fit includes other seeds of some Goal configurations of this block, but its recall advantage comes from Object and Spatial (1.000 and 0.917 against 0.718 and 0.600), while on LIBERO-10 the rules reach the higher F1 (0.819 against 0.755) with 85 false alarms to NAR's 157. The recordings had been inspected before, so this is retrospective cross-validation, not a blind test. Nothing was re-simulated: the comparison re-reads recorded first signals. This reading exists only in simulation. A real robot does not observe its world-frame coordinates or the poses, contacts and goal predicates of the objects it handles; they must be reconstructed with expensive, calibrated external instruments (motion capture, multi-camera tracking, fiducial markers) that are complex to maintain. The routing trace is part of the policy's own computation and needs none of them.

Elapsed-time competitors. Every failure runs to a known cap. Under the same pipeline, the sentinel of an input observes nothing but the task and the inference index; it reaches recall 0.71–1.00 at FPR_{13} 1.9–2.7% depending on the input it replaces, and the proprioceptive one exceeds routing (0.998; routing minus it -0.018 $[-0.031, -0.006]$). A per-suite timer tuned on the evaluation successes matches or beats NAR on the native held-apart suites and their perturbed counterparts. These competitors need the task and its success lengths; NAR reads neither: one threshold serves caps from 22 to 52 inferences, a single timer calibrated on Goal successes has a 14.2% false-alarm rate on the other suites, and elapsed-time coordinates in the descriptor break the Goal profile (F1 0.368; Appendix C).

D.4 CALVIN: CROSS-BENCHMARK CHECK

CALVIN D→D (Mees et al., 2022) chains five language-specified subtasks in one environment, each capped at 360 steps (36 inferences). We use the released HiMoE-VLA CALVIN checkpoint, which has the same MoE layout as the LIBERO checkpoints (HB layers 2–5 and 12–15, 32 experts, top-4, ten denoising steps) but outputs seven joint angles and the gripper, with the released camera layout and CALVIN commit `fa03f01`. Run through the released evaluation on all 1,000 sequences, without routing hooks, it completes 3.968 subtasks per sequence serially and 4.000 in a four-shard run, against 3.98 reported by its authors. These two runs only check that the checkpoint reproduces its score: they recorded no routing, so the alarm was not evaluated on them.

Detection capture. Routing was recorded on the first 150 sequences through the batched capture server, under settings that differ from the released evaluation: a CRC32 mapping from sequence to initial state, CPU rendering, OpenCV resizing, a newer `calvin_env` commit and no tactile camera. The policy completes 3.11 subtasks per sequence there, so this capture is not a reproduction of the benchmark score. It holds 548 attempted subtasks, 466 successful and 82 failed. Following a protocol registered before any CALVIN number was computed, the unit is the subtask: a fresh monitor starts at each subtask boundary, the frozen pooled profile is applied unchanged (threshold 9.8784, first alarm at the thirteenth inference, no refitting on CALVIN), and replay stops at the first signal. The Goal profile, our default, was replayed later on the same records under a second protocol, frozen before that replay but after the pooled result was known (threshold 9.9012, same start, no refitting).

Result. Under the Goal profile the monitor catches all 82 failures and raises 46 false alarms among the 466 successes (TP 82, FP 46, FN 0, TN 420; recall 1.000, precision 0.641, FPR 9.87%, F1 0.781); the pooled profile raises 43 (TN 423; precision 0.656, FPR 9.23%, F1 0.792). Under both, true and false alarms fire at a median of the thirteenth inference. The headline hides the limitation. Successful subtasks are short (median 8 inferences) and failures run to the cap (median 36), so 417 of the 466 successes end before the monitor can fire, and 46 of the 49 that reach the thirteenth inference are flagged (43 under the pooled profile). Recall transfers across benchmark, embodiment and action space; specificity does not, and a recalibration on CALVIN successes was not evaluated. A pilot with the wrist image in the other input slot (20 sequence pairs) gives the same picture: recall 1.000 in both arms, false alarms on 11.9% and 10.9% of successes under the Goal profile (11.9% and 7.3% under the pooled profile).

D.5 FULL-CORPUS CONTROL AT POOLED ALARMS

All 1,644 first alarms of the pooled profile in the corpus were declared. For each, the recorded actions were replayed to the alarm inference with the policy image, wrist image and state checked at every inference against the recording, using the rendering backend of the original run (1,200 cases on EGL, 408 retried on OSMesa after an image mismatch). Seven arms were run from the resulting state with the recorded noise stream: the unmodified policy (the twin), $\mathcal{R}(0.60, 1)$, three instruction rewrites held for four chunks (a polite paraphrase, an "escape the trap" suffix and a "release and lift" prefix), an instruction rewrite released when the alarm score clears, and a continuous blend of policy and primitive weighted by the alarm score. All seven arms start from the same alarm state in every completed case. 1,608 cases completed; all 36 missing cases are LIBERO-10, alarm earlier than completed ones (median 29 against 21 inferences left) and are more often false alarms (17 of 36). The protocol forbids extrapolating to unexecuted cases. A hit is a case whose twin fails; rescue and harm are defined on the twin. Intervals are percentile bootstraps over physical configurations (4,000 draws); the per-alarm effect is the mean of (arm − twin) success. The twin reuses the recorded noise stream and reproduces the historical outcome in 98.4% of alarms: of the 1,248 alarmed historical failures, 13 (1.0%) succeed in the twin and 288 (23.1%) with the primitive, so rescue is measured against one noise realization (with fresh noise, 35 of the 58 recorded failures in Table 13 succeed when re-executed).

Noise streams across studies. Arms of a replay branch always share the policy noise seed. The full-corpus study above reuses the recorded noise stream; the re-execution and screen studies key noise to the absolute inference index; the mixed and natural cohorts (Section 5.4; Appendix E.3) share a seed but not inference-aligned noise after the primitive.

D.6 Online re-execution of historical failures

The study froze all 58 recorded failures of acquisition seed 50 in one complete 2,000-record native batch (28 LIBERO-10, 12 Goal, 3 Object, 15 Spatial). Each configuration was executed again with its recorded task, initial state, simulator seed and suite checkpoint, with freshly specified policy noise; historical stochastic actions were not assumed to reproduce. Both calibration profiles monitored the same unmodified parent and stopped independently at their first signal. At each first signal, floor and recovery branches replayed the new parent's actions to an identical simulator state (maximum discrepancy exactly 0.0 on every branch). Floor executed the chunk already proposed at the alarm; recovery substituted $\mathcal{R}(0.30, 3)$ immediately. Three continuation seeds were averaged within each parent, and no-alarm parents contribute their natural outcome to both arms. Intervals use 4,000 bootstrap resamples of parents within suites, keeping arms and seeds together. The original plan used amplitude 0.60; after 36 parents had completed and interim outcomes had been inspected, a documented amendment changed the amplitude to 0.30 and all 58 configurations were executed again. The amplitude choice is therefore not independent of this cohort, and the incomplete 0.60 run does not provide a complete comparison. The cohort is selected on historical failure, so it estimates neither population false-alarm rates nor benchmark-wide gains; some task/initial-state configurations also occur in fitting roles.

D.7 The 96-configuration alarm screens

Cohort. From the 1,644 recorded native first alarms of the pooled profile, each suite contributed the first 24 distinct physical configurations in a fixed outcome-blind order, with one recording per configuration chosen the same way. No outcome, difficulty, remaining budget or branch availability entered selection, and all 96 configurations remain in every denominator. Every selected routing prefix was replayed incrementally and reproduced its recorded first signal before any continuation.

Arms. Floor (policy continuation at the alarm state with the matched seed) and $\mathcal{R}(\lambda, 3)$ for $\lambda \in \{0.15, 0.225, 0.30, 0.45, 0.60\}$, later extended by $\lambda \in \{0.75, 0.90, 1.00\}$ after inspection of the first ladder (the controller clips normalized inputs at 1.0). Native caps including the recorded prefix are 300, 220, 280 and 520 actions for Goal, Spatial, Object and LIBERO-10; late and truncated interventions are kept. Three continuation seeds per configuration share noise keyed by configuration, seed and absolute inference index. All arms of a configuration start from identical full simulator-state and observations; new continuations use OSMesa rendering, and historical EGL images are not assumed to be pixel-identical.

Goal-profile screen. The same 96 configurations were replayed under the Goal profile. For 55 the Goal alarm falls on the same inference as the pooled alarm and the existing paired branches were reused after provenance checks; the other 41 alarm earlier and all their branches were executed afresh at the earlier boundary. This screen also evaluated prespecified rules that use a first-five-inference routing risk (cluster failure rates fitted on a disjoint 8,000-episode development split, so the risk is outcome-calibrated) to decide whether to intervene and at which amplitude, fitted on four configuration folds and applied to the fifth.

Uncertainty. Rates weight suites equally (288 seed-pairs per arm). Intervals are 95% percentiles of 2,000 suite-stratified bootstrap resamples of physical configurations with repeats kept together; adaptive rules are refitted inside every replicate with fixed fold membership. The effective sample is 24 configurations per suite, and the many exploratory comparisons are not adjusted for multiplicity.

D.8 Recovery timing around the signal

The timing study of Section 5.4 is conditional on an existing first signal of the Goal profile in the native corpus scan (1,830 alarms). Of the 417 alarms whose recording could be replayed exactly under OSMesa rendering, 40 distinct configurations were drawn, 10 per suite over 36 tasks, by task balance in a fixed outcome-blind order, without reading outcomes, scores or signal times; one recording per configuration, all in every denominator, including 8 whose recorded episode succeeded (false alarms). The corpus scan contains development and reference roles. Each configuration received 32 flow-noise streams keyed by configuration, stream and absolute inference and shared across times.

At each time the primitive $\mathcal{R}(0.30, 3)$ replaced three chunks and the policy then resumed. Negative times replay the recorded prefix to the intervention state and pair recovery with a fresh unmodified continuation from that state; the signal time starts from the exactly replayed first-alarm state; positive times follow a new unmodified continuation from the signal state and intervene on it. Native caps and success termination are kept, so a late intervention that no longer fits is recorded as undelivered with the paired unmodified outcome. The nine times $-8, -4, -2, -1, 0, +1, +2, +4, +8$ were fixed before any outcome; the other eight integer times were added after those results had been inspected, under an analysis frozen before their outcomes, and the emphasis on harm is post hoc. A repeated run at the signal reproduced its outcomes (343 against 345 recovery successes). Intervals resample configurations within suites with all streams and times kept together; the five planned contrasts use 20,000 resamples with Bonferroni correction. All prefixes and branch boundaries passed exact replay checks; batched policy arithmetic is not bit-identical across runs.

D.9 Unreported exploratory runs

Additional exploratory runs were initiated for the 1,830 Goal-profile first alarms with $\mathcal{R}(0.30, 3)$ and three seeds (5,490 declared pairs; 186 alarms lack saved prefixes), a Goal-suite arm at amplitude 0.60, and an amplitude-by-duration factorial. These runs were incomplete at the reporting cutoff, and none of their partial outcomes is used in this paper.

E Additional results

E.1 Detection

Table 7: First-alarm detection on 43,345 recorded native LIBERO episodes (percent), fitting episodes included. *Validation-role episodes of the three suites outside the Goal profile's fit, seen before it became our default. FPR_{13}: among successes long enough to be flagged. The * row by suite: Table 8 (Appendix E.1).

		Goal profile (default)				Pooled profile			
Episodes	n / fail.	Rec.	Prec.	FPR	FPR_{13}	Rec.	Prec.	FPR	FPR_{13}
LIBERO-10	11,120 / 707	97.2	63.3	3.83	3.83	89.5	65.8	3.16	3.16
Goal	10,778 / 257	99.6	92.1	0.21	0.61	99.6	93.8	0.16	0.47
Object	10,826 / 122	100.0	64.9	0.62	0.72	100.0	81.3	0.26	0.30
Spatial	10,621 / 287	94.8	97.8	0.06	0.34	89.2	98.8	0.03	0.17
All suites	43,345 / 1,373	97.4	73.1	1.17	1.97	92.3	77.1	0.90	1.51
Non-Goal suites	32,567 / 1,116	96.9	69.7	1.50	2.20	90.6	73.7	1.14	1.68
Validation roles, non-Goal suites*	4,320 / 141	97.2	71.4	1.32	1.79	87.9	74.7	1.01	1.37

Table 8 gives the Goal profile's per-suite results on the 4,320 validation and task-validation episodes of the three non-Goal suites (the * row of Table 7) and the fit-source comparison on their LIBERO-10 part. Across the corpus the two profiles trade recall for precision at nearly equal F1 (0.835 Goal, 0.840 pooled), every pooled alarm is also a Goal-profile alarm, never an earlier one, and excluding the Goal profile's native fitting episodes leaves 97.4% / 72.7% / 1.22% on 41,425 episodes.

Other detection results. On the seeds 50–57 block (16,000 native episodes, 509 failures), the pooled profile reaches recall 0.916, precision 0.766, FPR 0.92% (F1 0.834; FPR_{13} 1.55% [1.11, 2.02]) and the Goal profile 0.971, 0.729, 1.19% (F1 0.832). On a mixed task-held-out set of 5,440 episodes (3,200 native, 2,240 LIBERO-Plus) the pooled profile reaches F1 0.943, and 0.876 on its native part. A threshold chosen without labels, as the $1 - \alpha$ quantile of a log-normal fitted to the bulk of unlabeled calibration peaks (9.694), gives the same native F1 (0.798) as the conformal threshold on the development role.

Evidence width on the seeds 50–57 block. Under the pooled profile on the 16,000 native episodes of seeds 50–57 (509 failures; FPR over all successes), with each width recalibrated by the same conformal rule and alarming once its window fills, means of two, three and six scores reach recall

Table 8: Goal profile on the 4,320 validation and task-validation episodes of the three non-Goal suites, and the fit-source comparison on the LIBERO-10 part of that set (1,440 episodes, 91 failures). Median alarm inference is over detected failures (0-based).

Target	n	Fail.	Recall	FPR (%)	Precision	F1	Median alarm
LIBERO-10	1,440	91	0.967	3.34	0.662	0.786	32
Object	1,440	14	1.000	0.35	0.737	0.848	17
Spatial	1,440	36	0.972	0.36	0.875	0.921	19
All	4,320	141	0.972	1.32	0.714	0.823	30
LIBERO-10 targets, bank and threshold fitted on:							
Goal suite only	1,440	91	0.967	3.34	–	0.786	
Goal, Object, Spatial	1,440	91	0.868	2.67	–	0.767	
All four suites (pooled)	1,440	91	0.857	2.67	–	0.761	

Table 9: Further rows of Table 3 and (d) the bank size of RQ2.2 (Goal profile; the same 4,320 episodes and 141 failures; post hoc; Earl./later: of the failures both flag at the frozen row's 55 false alarms, those flagged before/after the frozen rule). Sequential tests turn each score into a conformal p-value against all finite scores of the 942 Goal calibration successes and into an e-value $\kappa p^{\kappa-1}$; alarms start at the thirteenth inference, as for the frozen rule, and the Shiryaev–Roberts recursion starts from $R = 1$. Conformal: episode-maximum threshold at $\alpha = 1.5\%$ (deployable); nominal: $\ln(1/\alpha)$ on the log statistic. ΔF1 without an interval is a point difference. Every failure runs to its step cap, so recall 1.000 alone does not separate a test from a timer.

Variant	Rec.	FPR (%)	F1	ΔF1	Rec.$_{55}$	Earl./later
Frozen ($k = 20$, mean of six)	0.972	1.32	0.823	–	0.979	–
(b) Nearest neighbour only, $k = 1$	1.000	5.17	0.566	-0.257 $[-0.342, -0.182]$	0.922	0/74
(b) Wide neighbourhood, $k = 320$	0.915	1.24	0.801	-0.022 $[-0.047, +0.005]$	0.943	34/15
(c) Fisher, six conformal p	0.972	1.32	0.823	$+0.000$ $[-0.008, +0.007]$	0.979	2/26
(c) Shiryaev–Roberts, $\kappa = 0.2$	1.000	2.85	0.703	-0.120 $[-0.189, -0.062]$	0.894	26/87
(c) e-CUSUM, $\kappa = 0.5$	1.000	4.40	0.605	-0.218	0.844	29/78
(c) Shiryaev–Roberts, $\kappa = 0.5$	1.000	3.78	0.641	-0.182	0.851	29/77
(c) e-CUSUM, $\kappa = 0.2$, nominal	1.000	20.32	0.249	-0.573 $[-0.643, -0.504]$		
(c) Shiryaev–Roberts, $\kappa = 0.2$, nominal	1.000	30.44	0.181	-0.641 $[-0.705, -0.577]$		
(c) e-CUSUM, $\kappa = 0.5$, nominal	1.000	27.73	0.196	-0.627		
(c) Shiryaev–Roberts, $\kappa = 0.5$, nominal	1.000	46.95	0.126	-0.697		
(d) Bank of 2,048 points	0.957	1.27	0.821	-0.002		
(d) Bank of 1,024 points	0.915	1.24	0.801	-0.022		

0.917, 0.923 and 0.916 with 160, 153 and 142 false alarms (F1 0.822, 0.830, 0.834; point estimates; this block was seen before the Goal profile became our default). The six-score mean trims false alarms in total but not in every suite: LIBERO-10 and Object favour six (129 against 142 and 7 against 13 false alarms with two), Goal and Spatial two or three (4 against 6 false alarms; 109 against 104 detections).

Alarm timing of the evidence rules (Table 3). At matched false alarms the raw score fires earlier than the six-score mean on 79 of 129 shared detections (it raises 63 false alarms where the mean raises 55 and catches three fewer failures); e-CUSUM and Shiryaev–Roberts ($\kappa = 0.2$) fire later on 87–88 of 122–124 shared detections, and at the nominal $1/\alpha$ threshold they give 20–30% false alarms; $k = 320$ stays within noise of $k = 20$ (Table 9).

Sensitivity to removing perturbed source successes. Refitting the Goal profile on its native episodes only (1,256 reference and 625 calibration successes) gives threshold 6.56, FPR 59.9% and F1 0.101 on the 4,320 target episodes. Length is not the reason: native and LIBERO-Plus Goal successes are equally short (about 30% of either reach the thirteenth inference). The perturbed successes reach higher score maxima in the calibration pool (95th percentile of scorable maxima 5.97 native against 11.06 LIBERO-Plus), and 12 of the 13 maxima above the default threshold

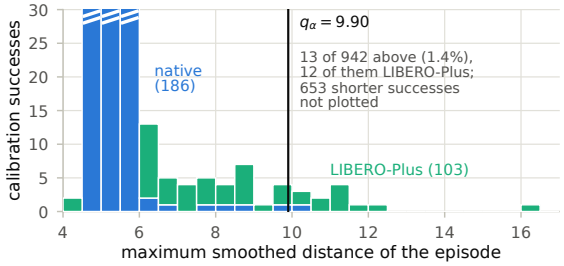


Figure 5: Where the threshold comes from (Goal profile; RQ2.2): maximum smoothed distance of each Goal calibration success long enough to be scored (289 of 942), stacked by family; the three tallest bars are cut at 30 (58, 121, 53). The threshold is the conformal quantile at $\alpha = 1.5\%$ over all 942, so 13 (1.4%) exceed it, 12 of them LIBERO-Plus.

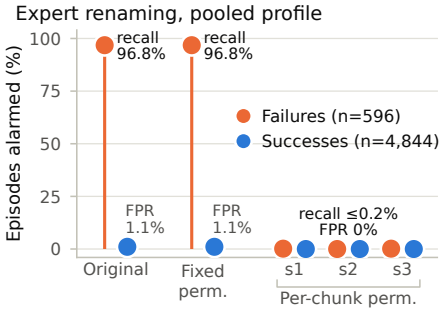


Figure 6: Alarms of the frozen pooled profile on 5,440 task-held-out recordings (3,200 native, 2,240 LIBERO-Plus; 596 failures, 511 of them LIBERO-Plus) when expert identities are renamed once per episode or independently at every inference (three seeds), without recalibration (Section 5.3).

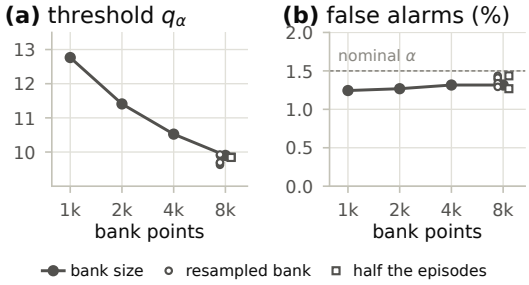


Figure 7: Across the evaluated banks the recalibrated threshold varies **(a)** while the false-alarm rate on the 4,320 held-apart episodes stays at 1.24–1.44% **(b)** (Goal profile; RQ2.2).

are LIBERO-Plus episodes (Figure 5), so without them the conformal rank lands far lower. The native-only refit changes the reference bank, robust scaling and calibration pool together. It therefore shows sensitivity to source-success coverage without isolating a single component. Resampling the mixed-source bank moves F1 only within 0.814–0.828, but this is a different perturbation from removing perturbed source successes.

The threshold is the fragile part. The conformal level bounds false alarms only for successes exchangeable with the calibration episodes, and only 289 of the Goal profile's 942 calibration successes are long enough to be scored (Appendix C). Across all 32,567 non-Goal episodes the realized rate is 2.20% among flaggable successes and 3.83% on LIBERO-10, which carries most false alarms (precision 0.63–0.66). Resampling the calibration pool moves F1 on one seed block to $[0.68, 0.85]$ for the Goal profile, against $[0.832, 0.840]$ for pooled; on 25,902 external LIBERO-Plus episodes the pooled profile keeps 97.4% recall at 3.67% FPR.

E.2 Control

Table 11 gives the full-corpus control study by suite and Table 12 by remaining budget.

Other responses at the same alarm states. Against the twin, the four instruction rewrites change success by at most ±0.12 pp per alarm: they rescue only 1.4–3.0% of alarmed failures and harm 4.7–10.2% of alarmed successes, against 22.9% and 30.5% for the primitive. Blending the primitive continuously by the alarm score loses 3.23 pp $[1.07, 5.24]$ per alarm, with opposite signs by suite (LIBERO-10 negative, Object positive). A gate on remaining budget, fitted on half of the configu-

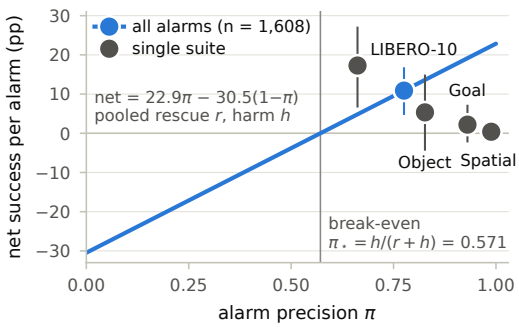


Figure 8: Control value against alarm precision on the 1,608 native pooled-profile alarms of Table 11: Eq. 2 per alarm with the overall rescue and harm rates (line), and measured suites with 95% intervals; along the line acting pays right of $\pi_\star = 0.571$, and each suite has its own break-even.

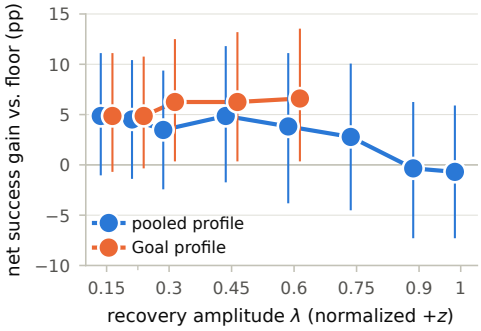


Figure 9: $\mathcal{R}(\lambda, 3)$ on 96 outcome-blind pooled-alarm configurations (24 per suite, three paired seeds): success change against the no-recovery floor, 95% configuration-bootstrap intervals, for the pooled alarm (0.75–1.0 added after inspecting ≤ 0.60) and the Goal-profile alarm on the same configurations (41 re-run at an earlier alarm). Amplitude 0.30 minus 0.60 under the Goal alarm is -0.35 pp $[-4.51, 4.17]$; under the pooled alarm, lifts of 0.90 or more harm LIBERO-10 (-13.9 pp $[-29.2, -1.4]$ at 0.90; exploratory).

Table 10: Recovery timing around the first Goal-profile signal (Section 5.4, Figure 4; protocol in Appendix D.8). Per time: percent of the 1,280 branches in which the unmodified continuation succeeds (Unmod.), in which recovery rescues or harms, and in which recovery succeeds (Recov.); net gain in pp with pointwise 95% intervals (5,000 suite-stratified configuration bootstraps); Delivered: branches in which recovery could still be delivered before the cap (undelivered branches keep their native outcome), with rescue and harm as a percent of those. *The signal. †Times added after the pre-registered nine had been inspected.

	% of all 1,280 branches				Net gain (pp)	Delivered	% of delivered	
Time	Unmod.	Rescue	Harm	Recov.	[95% interval]	(branches)	Rescue	Harm
−8	28.0	15.5	12.8	30.7	+2.7 [−11.6, +17.1]	1280	15.5	12.8
−7†	23.2	18.8	10.1	32.0	+8.8 [−5.9, +23.0]	1280	18.8	10.1
−6†	20.4	11.4	9.5	22.3	+2.0 [−9.1, +13.0]	1280	11.4	9.5
−5†	20.9	8.0	5.5	23.4	+2.4 [−7.0, +11.6]	1280	8.0	5.5
−4	20.8	9.6	3.1	27.3	+6.5 [−2.7, +15.9]	1280	9.6	3.1
−3†	19.6	8.0	3.4	24.1	+4.5 [−3.4, +12.9]	1280	8.0	3.4
−2	20.1	6.4	3.8	22.7	+2.7 [−4.9, +10.7]	1280	6.4	3.8
−1	19.5	5.4	2.8	22.0	+2.6 [−4.6, +10.2]	1280	5.4	2.8
0*	19.5	7.9	0.5	27.0	+7.4 [+0.2, +15.2]	1280	7.9	0.5
+1	19.5	2.5	2.4	19.6	+0.1 [−4.1, +4.6]	1248	2.6	2.5
+2	19.5	6.3	1.6	24.3	+4.8 [−0.9, +12.0]	1216	6.7	1.6
+3†	19.5	3.7	1.8	21.4	+1.9 [−2.4, +7.7]	1074	4.4	2.1
+4	19.5	2.9	1.8	20.6	+1.1 [−2.6, +5.7]	963	3.8	2.4
+5†	19.5	4.3	5.2	18.7	−0.9 [−8.5, +6.5]	896	6.1	7.4
+6†	19.5	4.2	4.2	19.5	+0.0 [−7.2, +7.3]	830	6.5	6.5
+7†	19.5	4.9	3.1	21.3	+1.8 [−5.5, +9.0]	729	8.6	5.5
+8	19.5	3.4	3.0	19.9	+0.4 [−6.2, +6.3]	678	6.5	5.8

rations and evaluated on the other half (400 splits), gains -0.24 pp $[-2.16, +0.00]$ over acting on every alarm; its chosen threshold is typically zero, i.e. no gate.

Table 11: Replay-branch control at all completed native first alarms of the pooled profile, primitive $\mathcal{R}(0.60, 1)$, one seed. π: alarm precision; $\pi_\star = h/(r+h)$: break-even; r, h: rescue and harm rates; Δ: success change per alarm with 95% configuration-bootstrap interval; Net: rescued minus harmed episodes; failures and successes are defined by the unintervened twin. †Three alarmed successes only.

Suite	Alarms	π	$\pi_\star$	r (%)	h (%)	Δ per alarm (pp)	Net
LIBERO-10	926	0.662	0.415	40.9	29.1	$+17.28$ $[+6.56, +27.21]$	$+160$
Goal	273	0.930	0.878	5.1	36.8	$+2.20$ $[-2.34, +7.28]$	$+6$
Object	150	0.827	0.734	15.3	42.3	$+5.33$ $[-4.44, +14.97]$	$+8$
Spatial	259	0.988	0.977	0.8	33.3†	$+0.39$ $[-0.83, +1.75]$	$+1$
All	1,608	0.775	0.571	22.9	30.5	$+10.88$ $[+4.67, +16.83]$	$+175$

Table 12: The control study of Table 11 by remaining budget at the alarm (inferences left before the cap), with the alarms of each band by suite; symbols as in Table 11 (no intervals were computed by band). Bands are confounded with suite: the lowest holds nearly all Spatial alarms, the two highest are LIBERO-10 only. Acting pays where $\pi > \pi_\star$; the highest band loses because its precision, not its rescue rate, is low.

		Alarms by suite								
Remaining	Alarms	LIBERO-10	Goal	Object	Spatial	π	$\pi_\star$	r (%)	h (%)	Net
0–10	548	81	139	70	258	0.962	0.926	2.3	28.6	$+6$
10–20	437	222	134	80	1	0.886	0.522	36.7	40.0	$+122$
20–30	322	322	0	0	0	0.742	0.435	36.0	27.7	$+63$
≥ 30	301	301	0	0	0	0.312	0.381	47.9	29.5	-16

Three-chunk variants. With $\mathcal{R}(\lambda, 3)$ under frozen protocols the primitive gains $+10.9$ to $+13.2$ pp per parent on re-executions of 58 recorded failures, a failure-selected cohort, but on 96 blindly drawn alarm configurations only the Goal-profile alarm at amplitude 0.30 excludes zero ($+6.25$ pp $[0.35, 12.50]$, one of ten unadjusted comparisons), and no amplitude is distinguishable from another (Table 13, Figure 9).

The failure-selected re-executions are all LIBERO-10 ($+27.5$ to $+30.7$ pp per alarmed continuation); only four or five alarmed continuations there could have been harmed, and the amplitude (0.30) was chosen after an interim run at 0.60 had been inspected (Appendix D.6). On the 96 blindly drawn configurations the pooled alarm gains $+3.47$ pp $[-2.43, 9.38]$ at 0.30; on 200 naturally drawn LIBERO-10 executions only 20 alarm and the primitive nets -2 successes (-1.0 pp per execution $[-4.0, 2.0]$).

Three-chunk variants by suite. In the failure-selected online re-execution, rescues occur in eight (pooled) and ten (Goal) distinct LIBERO-10 parents; the LIBERO-10 gains are $+22.62$ pp $[9.52, 36.90]$ and $+27.38$ pp $[13.10, 42.86]$; Goal, Object and Spatial show no paired change, and Object has no alarm at all. On the 46 configurations outside the Goal suite the gains are $+13.77$ and $+16.67$ pp. In the outcome-blind screen with the pooled alarm, $\mathcal{R}(0.30, 3)$ gains $+5.56$ pp $[-6.94, 18.09]$ on LIBERO-10, $+4.17$ $[0, 12.50]$ on Goal (one configuration), $+4.17$ $[-12.50, 22.22]$ on Object and 0 on Spatial; no suite or amplitude excludes zero. Selecting the amplitude per suite,

Table 13: The three-chunk primitive across cohorts (native LIBERO): success change versus the paired floor, pp with 95% interval, over the unit stated (no-alarm units contribute zero; in the natural cohort the floor's noise is not inference-aligned with the primitive).

Cohort	Selection	Alarm	Unit $\times$ seeds	$\mathcal{R}(0.30, 3)$ vs. floor
58 recorded failures, re-run	on failure	pooled	58 parents (23 alarmed) $\times$ 3	$+10.92$ $[+4.60, +18.39]$
		Goal	58 parents (25 alarmed) $\times$ 3	$+13.22$ $[+6.32, +20.69]$
96 pooled-alarm configs (24/suite)	blind draw	pooled	96 alarms $\times$ 3	$+3.47$ $[-2.43, +9.38]$
		Goal	96 alarms $\times$ 3	$+6.25$ $[+0.35, +12.50]$
200 LIBERO-10 executions	natural	pooled	200 runs (20 alarmed) $\times$ 1	-1.0 $[-4.0, +2.0]$

Table 14: Using a readout to choose the response: Δ in pp of success (hit side where noted) against the comparator, 95% interval (cluster bootstrap; row 3 paired McNemar; row 7 percentiles over 400 cross-fits). Mixed: 83 LIBERO-10 alarm parents (74 LIBERO-Plus), pooled alarm, $\mathcal{R}(0.60, 3)$; native 96: Goal-profile alarm, $\mathcal{R}(\lambda, 3)$; native 1,608: pooled alarm, $\mathcal{R}(0.60, 1)$. Rows 2 and 4 use routing of simulated continuations and report the best of 3 horizons (row 2; at $H = 3$: -16.1 $[-23.6, -9.2]$) and of 6 variants (row 4: three horizons $\times$ mean or minimum evidence); row 6's risk is outcome-calibrated.

Decision rule	Cohort	Comparator	Δ
13 primitives, routing kNN at the alarm	mixed, 66 hits	always primitive	$+3.0$ $[-2.7, +9.7]$
8 motion directions, routing on simulated branches	mixed, 83	fixed $+z$	$+0.24$ $[-5.30, +5.78]$
Best of 256 action candidates, routing phase	mixed, 66 hits	first candidate	$+2.5$ $[-1.0, +6.1]$
Continuation seed, post-handback routing	mixed, 83	random seed	$+2.86$ $[-0.60, +6.63]$
Steer generation toward successful routing	mixed, 83	random correction	$+0.00$ $[-3.61, +3.61]$
How hard, early routing risk (never abstained)	native, 96	fixed $\mathcal{R}(0.30, 3)$	-3.47 $[-7.99, +2.78]$
Whether, remaining budget (not routing)	native, 1,608	always act	-0.24 $[-2.16, +0.00]$

or by the historical difficulty of a configuration, on held-out folds does not beat a fixed 0.30 (suite rule $+1.04$ pp $[-2.08, 3.82]$; the difficulty rule changes no held-out assignment). Under the Goal alarm, the prespecified comparison of an early-routing-risk rule that chooses whether and how hard to act gives -3.47 pp $[-7.99, 2.78]$ against fixed 0.30; the rule never abstained. An earlier enriched cohort of 46 native LIBERO-10 pooled-alarm parents (five seeds, 226 complete four-arm branches, its floor and 0.60 arms run a day before its 0.30 and 0.15 arms) favoured $\mathcal{R}(0.30, 3)$ over $\mathcal{R}(0.60, 3)$ by $+23$ net successes $[+3, +43]$ over the 226 branches (net $+20$ and -3 against floor); this was not that study's pre-registered contrast, and neither the natural cohort nor the screens reproduce it.

E.3 The mixed cohort used for response selection

Most rows of Table 14 use 83 LIBERO-10 parents on which the pooled alarm fired (74 LIBERO-Plus, 9 native; 66 failures and 17 false alarms; ten base tasks, two of which contribute 41 parents), with five continuation seeds for the fixed-primitive arm (one to eight in the selection tests) and $\mathcal{R}(0.60, 3)$. On this cohort the primitive raises success on alarmed failures from 4.8% to 23.3% ($+18.5$ pp $[10.0, 27.9]$) and lowers it on false alarms by 36.5 pp $[11.8, 61.2]$. A 2×2 factorial shows both parts matter: success on alarmed failures is 5.0% (closed, still), 7.5% (open, still), 11.5% (closed, lift) and 23.3% (open, lift), with the unintervened floor at 4.7%. Among nine motion directions only $+z$ with an open gripper is net positive. Kicking later than the alarm hurts (-20.4 pp $[-32.1, -10.0]$ when a slower trigger moves the kick later), and a second pulse six inferences after the first costs -13.6 pp $[-20.9, -7.0]$. With 66 alarmed failures, the power calculation reaches 80% only around AUROC 0.70; detecting AUROC 0.60 would need about 330 parents. The routing kNN readout reached AUROC 0.71 for the eventual response outcome when neighbours could share the task (0.58 across tasks), yet its choices did not beat always acting.

Reassessment after handback. On native LIBERO-10, re-reading routing novelty in the four inferences after the one-chunk primitive hands back control flags alarmed failures that will still fail with precision 0.980 $[0.92, 1.00]$ and recall 0.136 (pre-registered; 611 branches). Applied to all 926 LIBERO-10 branches, stopping at that signal would stop 108 attempts, 103 of which fail anyway (95.4% $[89.0, 98.9]$), saving 5.8% of action steps for a 0.54 pp loss in success. This is retrospective accounting; a re-armed trigger, which caused no additional failures relative to one recovery, is described next.

Re-arming the alarm after handback. We restarted from the recorded first Goal-profile alarm of each of the 40 configurations of the timing study (Appendix D.8; 10 per suite) with two new noise seeds, 80 paired branches per arm, and ran each arm to success or the cap. Besides no intervention and one $\mathcal{R}(0.30, 3)$, a *level* rule recovers at every policy inference with $s_t > q_\alpha$, and an *edge* rule (the re-armed crossing of Table 4) re-arms once an observed s_t is at or below q_α and recovers at the next strict upward crossing. Re-arming copies the alarm's causal histories and uses the frozen Goal bank; the primitive adds no routing samples, and nothing reads outcomes or the remaining budget. Caps are

the benchmark cap B (300, 220, 280 and 520 actions for Goal, Spatial, Object and LIBERO-10), $2B$ and $4B$, nested views of one continuation; the level rule was run to $2B$ only (Table 4).

Recovering at every exceedance lets the primitive crowd out the policy: at $2B$ it takes 97.9% of post-alarm actions, 58 of 80 branches never run a policy action again and fail, and against one recovery it rescues none and breaks 15 at B (-18.75 pp $[-27.5, -10.0]$) and 21 at $2B$ (-26.25 pp $[-38.75, -15.0]$). The edge rule repeats in 4, 11 and 14 of 80 branches at B, $2B$ and $4B$ (at most 10 recoveries; 6.2% of post-alarm actions at $4B$). Against one recovery it is identical at B and rescues 1 and 2 branches without harm at $2B$ and $4B$ ($+1.25$ pp $[0.0, 3.75]$; $+2.5$ pp $[0.0, 7.5]$); against no intervention at $4B$ it nets $+11.25$ pp $[0.0, 23.75]$, one recovery $+8.75$ pp $[-1.25, 20.0]$. Time alone also rescues: from $2B$ to $4B$ no intervention gains 2 successes, one recovery 3 and the edge rule 4. These configurations had been examined before, the edge rule was chosen after the $2B$ results, and caps beyond B exceed the benchmark, so this is exploratory, not an independent validation. Every $4B$ branch reproduces its $2B$ prefix exactly. Intervals: $10{,}000$ suite-stratified bootstraps over configurations, both noises kept together.

F Extended related work

What "no failure data" does and does not mean. Hide-and-Seek (Park et al., 2026) learns from episode-level rather than per-step labels, which removes annotation cost but not the need for failure outcomes; FailureSpot (Ma et al., 2026b) addresses the same label noise by active timestamp labelling. FAIL-Detect (Xu et al., 2025) calibrates on successful executions, although several of its candidate scorers are fitted, and VLA-FAIL (Seligmann et al., 2026) needs neither failure data nor a learned predictor, at the cost of an action-chunk consistency term that requires overlapping action intervals. Our setting is not zero-data either: the reference bank and threshold are estimated from successful episodes, and the choice of the frozen variant used development outcomes (Appendix C).

Risk scores are not success probabilities. VLAConf (Huang et al., 2026) fits a step-conditioned predictor with a post-hoc calibrator and reports a probability of success. Our readout is an unnormalized distance to successful dynamics and is never reported as a probability.

Input shift is not execution failure. SAFECAST (Rajaprakash et al., 2026) trains and calibrates risk probes on visual and language contrast sets, and VLA-Scope (Zhu et al., 2026) characterizes input shift before fitting a failure predictor. A policy can be out of distribution and succeed, or in distribution and fail; we measure transfer rather than assume it.

A frozen backbone is not a training-free pipeline. ProbeAct (Zhang et al., 2026a) trains a 3D object probe, ReCoVLA (Hu et al., 2026) trains a residual recovery policy under rewards compiled by a vision-language model, and Jeong et al. (2026) train an external feedback policy with a harm gate. COAST (Miao et al., 2026) makes the parallel point on the representation side: gradient-free is not data-free when conceptor subspaces are estimated from successes and failures.

Compute budget separates otherwise similar recovery methods. CoRe (Zhang et al., 2026d) detects success-calibrated internal anomalies, intervenes and hands control back, reaching its intervention through reconstruction, a counterfactual continuation and a planned realignment. CycleVLA (Ma et al., 2026a), CoVer-VLA (Kwok et al., 2026) and VLA-Corrector (Song et al., 2026) rely on external models or verifiers. Our primitive is a fixed ten-row action pattern with no model call.

Benchmarks are not interchangeable. LIBERO-Plus (Fei et al., 2026) applies controlled perturbation families to separate success from robustness; LIBERO-PRO (Zhou et al., 2025) varies objects, initial states, instructions and environments; LIBERO-RECOVER (Liu et al., 2026) builds recovery scenarios from recorded failures. They share a base suite but not a protocol; we report no LIBERO-PRO or LIBERO-RECOVER results, and CALVIN (Mees et al., 2022) enters only in Appendix D.4. We evaluate on the four original LIBERO suites (Liu et al., 2023); LIBERO-Plus enters only where marked: both profiles' reference and calibration pools, one external detection set, $2{,}240$ of the $5{,}440$ renaming-test episodes, part of the variogram cohort (Section 5.2) and 74 of the 83 response-selection parents (Appendix E.3); all other results are native, and the two families are reported separately.