# ROLLOUT-LEVEL CALIBRATION FOR FAILURE MONITORING IN GENERATIVE ROBOT POLICIES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Failure monitors for generative robot policies must raise timely alarms while limiting false positives on executions that would otherwise succeed. We introduce RoLCA (Rollout-Level Conformal Alerting), a modular layer that aggregates progress-matched score evidence over time and across channels, then split-conformally calibrates its maximum over eligible monitoring times using successful rollouts only. Under exchangeability of successful rollouts, RoLCA provides finite-sample rollout-level false-positive control for causal scoring, aggregation, fusion, and monitoring rules fixed independently of threshold calibration. The guarantee is marginal over the calibration draw and the next successful rollout, without requiring within-rollout independence. Across 30,502 rollouts in six task–backbone cells, calibrated monitors operate near the nominal false-positive rate in two sealed confirmations, but detection gains depend on the monitoring convention. On the second confirmation, a post-hoc comparison yields a 7.6-percentage-point paired recall gain over matched running-average OR under fixed-horizon monitoring; this advantage is not established after stop-at-success recalibration. Fusion improves early recall over a rollout-trained single channel but does not consistently outperform execution consistency alone. A clock-only baseline detects all timeout failures in this population under stop-at-success evaluation, motivating assessment of early warning beyond eventual coverage. These results separate false-alarm validity from detection power and underscore the need for explicit monitoring intervals, matched comparators, and timing baselines.

## 1 INTRODUCTION

An action-chunking robot policy can produce plausible motions without progressing toward success: a missed grasp may be followed by smooth but ineffective actions. Diffusion policies and flow-based vision–language–action models execute a predicted action prefix before replanning (Chi et al., 2025; Black et al., 2025a; Physical Intelligence et al., 2025), balancing temporal coherence and responsiveness (Black et al., 2025b; Nie et al., 2026). A runtime monitor should detect unsuccessful executions while limiting alarms on rollouts that would otherwise succeed.

We study *success-only anytime monitoring*: threshold calibration uses successful rollouts, and any eligible alarm within a horizon $T_{\max}$ on such a rollout is a false positive. Dependent observations and overlapping action chunks mean that marginal per-cycle calibration alone does not control the rollout-level false-positive rate (FPR). The monitoring interval also matters: scoring through a fixed horizon (E) and stopping at first success (S) define different false-alarm events. Validity and detection power must therefore be assessed under an explicit monitoring convention.

We introduce RoLCA (**Ro**llout-**L**evel **C**onformal **A**lerting), a modular layer that split-conformally calibrates a monitoring statistic's maximum over eligible times. Our instantiation transforms progress-matched score ranks into evidence, aggregates over time, and fuses channels. Under exchangeability of successful rollouts, RoLCA provides finite-sample FPR control for any causal statistic and monitoring convention fixed independently of threshold calibration. The guarantee is marginal over the calibration draw and the next successful rollout, requiring neither within-rollout independence nor valid per-cycle $p$- or $e$-values.

Rollout-maximum calibration is established in FAIL-Detect and related monitors (Xu et al., 2025; Rao et al., 2026); RoLCA uses the same split-conformal principle. Our contribution separates statis-



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.

tic construction, monitoring eligibility, and threshold calibration, and examines their implications for detection. We evaluate 30,502 rollouts from discovery, two sealed confirmations, and a separate published-monitor population across three tasks and two feature backbones within a rectified-flow actor family. Confirmation protocols were fixed before seed opening; subsequent analyses are labeled post hoc. Evaluations use uninterrupted trajectories to observe outcomes without alarm-triggered intervention.

**False-alarm control.** Prespecified RoLCA monitors operate near nominal FPR across both confirmations. At $\alpha = 0.10$, multiplying per-cycle evidence and applying a Ville threshold yields FPR approximately $0.26$, illustrating the consequence of using that threshold without the required sequential validity assumptions (Section 4.2).

**Channels and decision rules.** The discovery gain from flow-inversion energy does not replicate. On the second confirmation, joint mean improves early recall over RND-OE alone under both conventions, but its full-horizon increment is inconclusive. Against matched running-average OR, its post-hoc paired gain under (E) is $+0.076$; the original comparison also changes temporal aggregation (Section 4.4).

**Monitoring convention and early warning.** Post-hoc recalibration under (S) changes detector rankings: joint mean has no established full-horizon advantage over execution alone or matched OR, and execution leads by step 100. All failures in this population are timeouts, and an S-calibrated clock achieves full eventual recall at FPR $0.072$. Learned monitors can alarm earlier, with benefits varying across cells, motivating explicit stopping rules and timing baselines (Section 4.3).

## 2 Related Work

**Runtime failure monitoring.** Robot monitors use reconstruction, density estimation, ensemble disagreement, and task-aware OOD scores (Richter & Roy, 2017; Sinha et al., 2022; Lakshminarayanan et al., 2017; Farid et al., 2022). For generative policies, Sentinel combines temporal plan inconsistency with VLM reasoning (Agia et al., 2025), while FIPER separately thresholds observation-embedding RND and action-chunk entropy (Römer et al., 2025). Other approaches use robot-conditioned densities, visual-semantic reasoning, world-model latents, or supervised VLM critics (Zhou et al., 2026; Rolland et al., 2026; Zhang et al., 2026; Qi et al., 2026). We distinguish score information from aggregation and calibration, documenting training access and implementation differences in Appendix E.

**Rollout calibration and multichannel decisions.** Split conformal prediction provides marginal finite-sample guarantees under exchangeability (Vovk et al., 2005; Lei et al., 2018; Angelopoulos & Bates, 2023), with applications to outlier detection and robot decision-making (Bates et al., 2023; Lindemann et al., 2023; Ren et al., 2023). FAIL-Detect calibrates rollout maxima of success-only scores (Xu et al., 2025); Rao et al. calibrate episode maxima of CUSUM statistics (Rao et al., 2026). Tri-Info averages supervised GRU outputs before functional conformal thresholding (Yang et al., 2026). RoLCA builds on this calibration principle, separating statistic construction from monitoring eligibility and threshold selection. We compare fused statistics with channel-wise budget splitting under explicit monitoring intervals, without claiming a general power advantage for either rule.

**Sequential validity and accumulation.** Conformal martingales, test supermartingales, and $e$-processes support anytime-valid inference under sequential assumptions (Vovk et al., 2003; Fedorova et al., 2012; Ville, 1939; Howard et al., 2021; Grünwald et al., 2024; Ramdas et al., 2023). RoLCA obtains validity by calibrating a scalar maximum over a bounded monitoring interval. Its evidence construction draws on $p$-to-$e$ calibration (Vovk & Wang, 2021), without requiring valid per-cycle $e$-values or an $e$-process. CUSUM and $e$-detectors provide related accumulation mechanisms (Page, 1954; Lorden, 1971; Shin et al., 2024); Appendices B.4 and F.2 detail the evaluated rules and fusion comparisons.

## 3 RoLCA: Construction and rollout-level validity

RoLCA calibrates a complete monitoring statistic at the rollout level. Construction data and settings $\mathcal{D}$ fix the score models, transforms, temporal aggregation, fusion, and monitoring convention

without using the threshold-calibration rollouts. A disjoint set of successful CAL rollouts then determines the alarm threshold. We first give this general construction and its guarantee, then describe the statistics used in our experiments (Figure 1). Outer calibration supplies validity; the statistic determines detection power.

### 3.1 MONITORING INTERVALS AND SCORE AVAILABILITY

At policy cycle $j$, the actor observes context $C_j$, samples $a_{j,1:H} \in \mathbb{R}^{H \times d_a}$, and executes its first $H' \leq H$ actions, where $d_a$ is the action dimension. The cycle starts at action $t_j = 1 + (j-1)H'$; $o_t$ denotes the observation before action $t$. For a rollout with $J_R$ statistic records, we distinguish record index $j$ from its alarm-opportunity time $\tau_j$, measured in executed environment steps. A statistic $Z_j(R; \mathcal{D})$ may use only information available by $\tau_j$; channels requiring longer executed windows become available later. The opportunity schedule and all missing-score rules are part of $\mathcal{D}$.

Let $T_{\mathrm{succ}}(R)$ be the first success step, or $+\infty$ if success is not reached within $T_{\max}$. We consider fixed-horizon monitoring (E) and stop-at-success monitoring (S), with endpoints

$$L_{\mathrm{E}}(R) = T_{\max}, \qquad L_{\mathrm{S}}(R) = \min\{T_{\mathrm{succ}}(R), T_{\max}\}.$$

For a convention $\kappa \in \{\mathrm{E}, \mathrm{S}\}$ fixed before calibration, the eligible records are

$$\mathcal{I}_\kappa(R) = \{j \in \{1, \ldots, J_R\} : \tau_j(R) \leq L_\kappa(R)\}. \tag{1}$$

Under (S), an alarm tied with success is eligible. Calibration and evaluation use the same rule. A false positive is an eligible alarm on a rollout that would succeed without intervention. Our evaluation uses uninterrupted trajectories; intervening at an alarm censors that outcome. Anytime validity here covers the specified interval within a bounded rollout.

### 3.2 ROLLOUT CALIBRATION AND FALSE-ALARM CONTROL

Fix $\kappa$ as part of $\mathcal{D}$. For any fixed causal statistic $Z_j$, define

$$V(R; \mathcal{D}) = \max_{j \in \mathcal{I}_\kappa(R)} Z_j(R; \mathcal{D}), \tag{2}$$

with $\max \varnothing = -\infty$. The default statistic below is $Z_j = \log M_j$, but calibration does not require that form. Let $\mathcal{K} = (R_1, \ldots, R_m)$ contain $m \geq 1$ successful calibration rollouts, and write $V_r = V(R_r; \mathcal{D})$. For $\alpha \in (0, 1)$, set

$$k = \lceil (m+1)(1-\alpha) \rceil, \qquad h = \begin{cases} V_{(k)}, & k \leq m, \\ +\infty, & k = m+1, \end{cases} \tag{3}$$

where $V_{(k)}$ is the $k$th smallest calibration statistic. RoLCA alarms at the earliest eligible opportunity with $Z_j > h$. Strict exceedance handles ties conservatively. Changing the monitoring convention changes the calibration statistic and generally requires recalibrating $h$.

**Proposition 1** (Rollout-level FPR control)**.** *Suppose that, conditionally on $\mathcal{D}$, the test rollout $R$ and calibration rollouts $R_1, \ldots, R_m$ are exchangeable successful rollouts. Then*

$$\mathbb{P}(\text{alarm during } R \mid \mathcal{D}) = \mathbb{P}(V(R; \mathcal{D}) > h \mid \mathcal{D}) \leq \frac{m+1-k}{m+1} \leq \alpha. \tag{4}$$

*Probability is joint over $\mathcal{K}$ and the next successful rollout. No within-rollout independence or per-cycle score-validity assumption is required. Averaging over $\mathcal{D}$ gives the unconditional bound.*

This is the standard split-conformal order-statistic argument: conditional on $\mathcal{D}$, the $m+1$ rollout statistics are exchangeable, and at most $m+1-k$ can strictly exceed the $k$th smallest of the remaining $m$ values. Ties only reduce this number; $k = m+1$ gives no alarm (Appendix A). The guarantee holds for either fixed monitoring convention, including trajectory-dependent success stopping, but is marginal over the calibration draw. It does not ensure FPR at most $\alpha$ for every realized calibration set or under deployment shift. Appendix A.2 gives a stronger calibration-conditional alternative under conditional i.i.d. sampling; all primary results use Eq. (3).

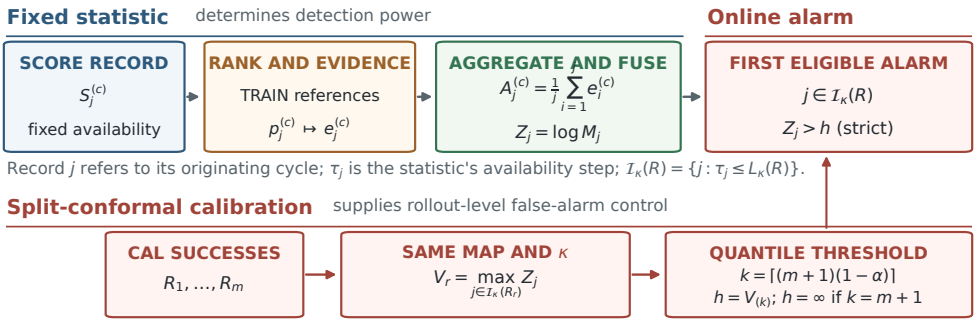



Figure 1: RoLCA calibrates a complete rollout statistic. Construction data fix the causal score-to-fusion map and monitoring convention $\kappa$ before threshold calibration. Record $j$ is indexed by its originating cycle, while $\tau_j$ denotes when its statistic is available. Each successful CAL rollout contributes the maximum $Z_j$ over eligible records $\mathcal{I}_\kappa(R)$; a conformal order statistic of these maxima sets threshold $h$. Online monitoring uses the same map and eligibility rule, alarming at the first eligible $Z_j > h$. Running-average evidence with mean fusion illustrates one choice of statistic.

### 3.3 MONITORING SCORES

Scores increase with departure from successful behavior. We index each score by its originating policy cycle and use outcome-independent progress

$$u_j = \frac{t_j - 1}{T_{\max} - 1} \in [0, 1]. \tag{5}$$

Progress remains anchored to that cycle even if the score becomes available later. Our actors use $H = 8$, $H' = 4$, and $T_{\max} = 400$.

**Execution consistency.** A fixed feature extractor $\phi_o$ maps the observation after executing the prefix to $f_j^+ = \phi_o(o_{t_j+H'}) \in \mathbb{R}^{d_f}$. A diagonal-Gaussian predictor $q_{\text{exec}}$, fitted on demonstrations, gives per-dimension predictive surprisal

$$S_{\text{exec},j} = -\frac{1}{d_f} \log q_{\text{exec}}\left(f_j^+ \mid C_j, a_{j,1:H'}\right). \tag{6}$$

This score is available after $jH'$ executed steps.

**Flow-inversion energy.** For $H = 2H'$, concatenate the executed prefixes from cycles $j$ and $j+1$, and approximately invert this realized continuation under the flow model conditioned on $C_j$. If $x_{0,j} \in \mathbb{R}^{Hd_a}$ is the resulting base latent, the standard-Gaussian prior gives

$$S_{\text{inv},j} = \frac{\|x_{0,j}\|_2^2}{2Hd_a}. \tag{7}$$

The score becomes available after $(j+1)H'$ steps. It is an energy proxy, not an exact action likelihood: numerical inversion omits the change-of-variables term, and replanning means the executed window is not a single sample from the earlier conditional policy (Appendix B.2). The second confirmation adds a rollout-trained RND-OE score (Section 4.1); the calibration rule is unchanged.

### 3.4 EVIDENCE AGGREGATION AND CHANNEL FUSION

For each channel, five progress bins fixed from successful TRAIN rollouts normalize changing score distributions. Let $b(j)$ contain $u_j$ and let $\mathcal{R}_{b(j)}$ be the multiset of available training scores in that bin. Suppressing channel indices, define the upper-tail rank and transformed evidence

$$p_j = \frac{1 + \sum_{s \in \mathcal{R}_{b(j)}} \mathbf{1}\{s \ge S_j\}}{|\mathcal{R}_{b(j)}| + 1}, \qquad e_j = \frac{1}{2\sqrt{p_j}}. \tag{8}$$

The rank resembles a conformal $p$-value (Vovk et al., 2005; Bates et al., 2023), but pooled, dependent cycle scores do not establish cycle-level exchangeability. The $p$-to-$e$ transform (Vovk & Wang, 2021) emphasizes small ranks; we require neither super-uniform $p_j$ nor valid $e$-values. Structurally missing terminal scores receive the fixed value $e_j = 0.5$ (Appendix B.1).

For a nonempty channel set $\mathcal{C}$, the running averages and default mean fusion are

$$A_j^{(c)} = \frac{1}{j}\sum_{i=1}^{j} e_i^{(c)} \quad (c \in \mathcal{C}), \qquad M_j = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} A_j^{(c)}. \tag{9}$$

Fusion pairs the same originating record across channels and waits until its required inputs are available; record indices are not common completion times. Running averages favor persistent evidence, but need not form an $e$-process. We calibrate $Z_j = \log M_j$ once at level $\alpha$; the logarithm is a monotone rescaling. Other fixed aggregation and fusion rules, including joint maxima, receive the same guarantee after their own rollout calibration. The matched OR instead calibrates each $\log A_j^{(c)}$ at $\alpha/|\mathcal{C}|$ and alarms at the earliest component crossing. It changes the multichannel decision rule; neither construction guarantees higher recall (Section 4.4).

# 4 Experiments

We examine false-alarm control, monitoring horizons, and the contributions of score channels and multichannel decision rules.

## 4.1 Evaluation design

**Platform and populations.** We evaluate *lift*, *can*, and *square* in robomimic (Mandlekar et al., 2022; Zhu et al., 2020), with BC-RNN and Diffusion Policy visual-feature backbones (Mandlekar et al., 2022; Chi et al., 2023). All six actors are rectified-flow chunk policies (Liu et al., 2023), with $H = 8$, $H' = 4$, and $T_{\max} = 400$; cells vary task and feature backbone within this actor family. All actors were retrained before collecting these populations. The study comprises 30,502 rollouts: 11,752 for discovery (1,081 failures), and three separate populations of 6,250 each for flow-inversion confirmation (654 failures), published-monitor comparison, and rollout-trained-channel confirmation (676 failures). The last population supplies all E/S comparisons below.

**Splits and analysis status.** Within each analysis pool, 50 shared rollout-level splits partition successes into 50% TRAIN, 25% CAL, and 25% evaluation, after removing any reserved score-fit rollouts. Failures are used only for evaluation; they neither fit score models nor construct rank references or calibration thresholds. Progress bins are TRAIN quintiles, and the nominal FPR budget is $\alpha = 0.10$. Each confirmation's candidate monitors and decision criteria were frozen before opening its seed. The matched running-average OR, clock baseline, stopping-convention comparison, and early-recall analyses are post hoc; they reuse frozen models, scores, and splits and leave the original gate outcomes unchanged (Appendices C and D.2).

**Monitors.** We form running-average evidence from execution consistency ($S_{\text{exec}}$) and RND-OE. The joint rules calibrate the mean or maximum of execution, inversion, and RND-OE evidence. RND-OE measures a predictor–random-target discrepancy on policy context features; it is trained on successful discovery rollouts and frozen before confirmation (Appendix B.3). Two OR comparators allocate $\alpha/3$ per channel: the original comparator uses instantaneous clipped, progress-standardized scores, whereas the post-hoc matched OR uses the same running-average evidence channels as the joint rules. Both alarm at the first component crossing.

**Monitoring conventions.** The original evaluator, **(E)**, monitors every rollout through step 400, including cycles after success; any alarm on a successful rollout is a false positive. The post-hoc **(S)** convention uses simulator first-success annotations to stop monitoring at that step, including an alarm tied with success. Failures remain active through step 400. Under (S), each monitor's threshold is recalibrated using only eligible times on CAL successes; the same eligibility rule is applied during evaluation. Score trajectories and TRAIN references remain fixed: (S) changes the operational false-alarm event and its calibration. Alarm times reflect when the statistic is computable, including the final-cycle availability convention (Appendix B.1).

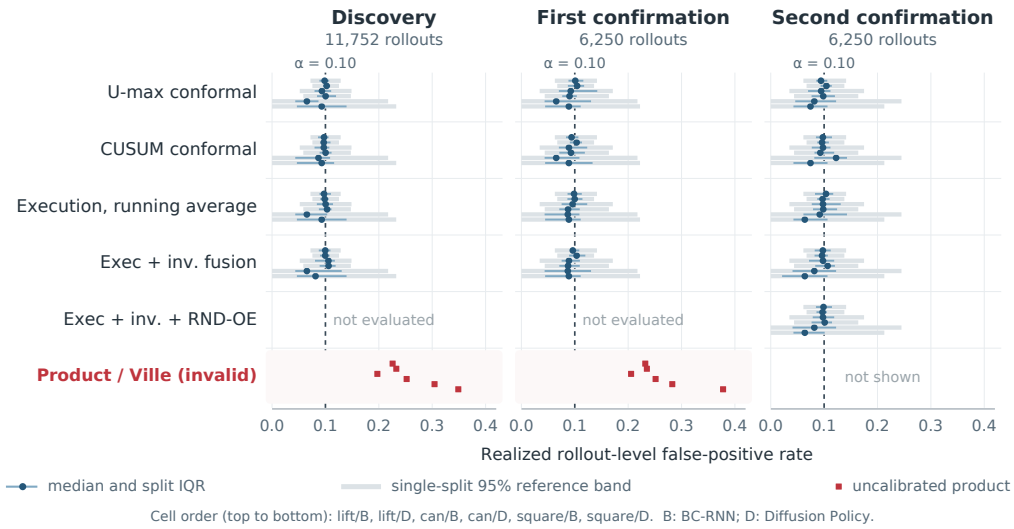



Figure 2: Rollout-level false-positive rates under fixed-horizon monitoring (E), at nominal $\alpha = 0.10$. Each evaluated row shows six task–backbone cells. Blue points and thin bars give cell medians and interquartile ranges over 50 overlapping splits; pale bars are single-split 95% Beta-binomial reference bands, not confidence intervals for the medians. The dashed line marks the nominal level. Red squares show the uncalibrated execution-channel product/Ville diagnostic. Appendix C.5 defines the reference bands and the separate acceptance gates.

**Endpoints and uncertainty.** FPR is the fraction of successful rollouts with an eligible alarm; recall is the corresponding fraction of failures. We report full-horizon recall, recall by step 100, and recall strictly before the first alarm of an S-calibrated clock. This clock is the reference under *both* conventions; its cutoff is determined separately in each cell and split from CAL successes. Missed failures remain in every denominator. Cell summaries are medians over splits; macro summaries average the six cell medians. Paired gains instead average the cell medians of within-split recall differences and need not equal differences of macro summaries. Second-confirmation contrasts use 2,000 paired failure-bootstrap resamples with independent streams across cells and shared resamples across monitors and splits within a cell. These pointwise 95% intervals condition on the observed success pool and fixed splits; they do not adjust for post-hoc comparisons. Overlapping splits are not independent replications (Appendix C.4).

### 4.2 ROLLOUT-LEVEL FALSE-ALARM CONTROL

Rollout calibration supports different temporal statistics: instantaneous maxima and CUSUM use progress-standardized execution scores, while running-average monitors use progress-matched evidence. Under fixed-horizon monitoring (E), these monitors and their fusions operate near nominal FPR (Figure 2). On the flow-inversion confirmation, macro FPR ranges from $0.089$ to $0.093$. All 30 prespecified cell–monitor checks on that population and all 18 on the second confirmation fall within their frozen binomial acceptance bands. These bands use a single-split reference model; they provide neither individual nor simultaneous 95% coverage guarantees for the reported medians over overlapping splits (Appendix C.5).

By contrast, multiplying per-cycle evidence and applying a Ville threshold without outer calibration yields macro FPR $0.260$ on discovery and $0.264$ on the first confirmation, exceeding the acceptance limit in every confirmation cell. Marginal validity of individual factors would not suffice to make their product a test supermartingale. This diagnostic illustrates the consequence of using a Ville threshold without the required sequential assumptions. RoLCA instead obtains rollout-level control by calibrating the complete statistic's maximum under exchangeability (Proposition 1).

| | (E) fixed horizon | | | (S) stop at success | | |
|---|---|---|---|---|---|---|
| Monitor | FPR | Recall | $\leq 100$ | FPR | Recall | $\leq 100$ |
| Execution alone | 0.092 | 0.403 | 0.339 | 0.095 | 0.552 | 0.479 |
| RND-OE alone | 0.091 | 0.507 | 0.239 | 0.087 | 0.535 | 0.292 |
| Joint mean | 0.090 | 0.523 | 0.391 | 0.092 | 0.559 | 0.439 |
| Joint max | 0.091 | 0.485 | 0.355 | 0.090 | 0.503 | 0.393 |
| OR, matched averages | 0.071 | 0.443 | 0.312 | 0.072 | 0.560 | 0.424 |
| OR, instantaneous | 0.075 | 0.406 | 0.218 | 0.069 | 0.666 | 0.471 |
| Clock only | 0 | 0 | 0 | 0.072 | 1.000 | 0.333 |

Table 1: Second-confirmation results at nominal $\alpha = 0.10$ under two monitoring conventions. Entries average cellwise medians over 50 shared splits; $\leq 100$ denotes recall by step 100. All (S) thresholds are recalibrated on active opportunities. Matched running-average OR, the clock, early endpoints, and the E/S comparison are post hoc. The (E) clock never alarms. A common nominal budget does not imply equal realized FPR.

### 4.3 Monitoring conventions change detector rankings

Table 1 compares fixed-horizon monitoring (E) with stop-at-success monitoring (S) on the same 6,250 rollouts, holding score trajectories and splits fixed. Recalibrating thresholds on active opportunities raises recall unevenly: execution alone increases from $0.403$ to $0.552$, joint mean from $0.523$ to $0.559$, and instantaneous OR from $0.406$ to $0.666$. Thus, including post-success observations in calibration affects both thresholds and detector rankings.

All 676 failures exhaust the 400-step budget without satisfying the success condition. A clock-only monitor calibrates elapsed time using the last eligible opportunity on each CAL success. Under (S), it achieves recall $1.000$ at FPR $0.072$; under (E), all calibration maxima equal 400, so it never crosses its threshold. Full-horizon recall therefore measures detection by timeout and must be assessed alongside warning time.

Under (S), joint mean detects $0.413$ of failures strictly before the split-specific clock alarm, compared with $0.438$ for execution alone, $0.416$ for matched running-average OR, and $0.495$ for instantaneous OR. These results establish warning beyond elapsed time for some failures, but not a general advantage for fusion. Figure 3 shows substantial variation across cells. Earlier alarms precede timeout; they do not establish detection before independently validated failure onset.

### 4.4 Channel and decision-rule comparisons

**Additional channels.** Under (E), joint mean improves over execution alone by a paired $+0.132$, exceeding the prespecified $0.05$ macro-gain criterion, but fails the per-cell criterion. RND-OE alone reaches recall $0.507$, close to joint mean's $0.523$; their paired full-horizon difference is inconclusive (Table 2). Joint mean improves early recall over RND-OE under both conventions. Its advantage over execution alone is convention-dependent: under (E), all three paired endpoint gains have intervals excluding zero; under (S), the full-horizon and before-clock differences are inconclusive, and execution alone leads by step 100. These early-recall comparisons concern the fraction of failures detected, not whether the same failures are detected sooner. Gains also vary across cells: square/D favors execution alone at every audited endpoint under both conventions, while the full-horizon lift/B contrast reverses from $+0.029$ under (E) to $-0.343$ under (S) (Appendix D.3).

**Joint decisions versus channel-wise OR.** The original $+0.118$ comparison with instantaneous OR ($95\%$ interval $[+0.087, +0.148]$) changes both the temporal statistic and the multichannel decision rule. With running-average channels matched, joint mean gains $+0.076$ $[+0.057, +0.093]$ under (E). This compares a fused threshold with three $\alpha/3$ tests at channel-specific opportunities; it does not isolate calibration from decision geometry or score availability. Under (S), no reliable difference from matched OR is established at any endpoint (Table 2); these intervals do not establish equivalence. The instantaneous OR has higher full-horizon recall than joint mean under (S), further limiting claims of general superiority from joint calibration.

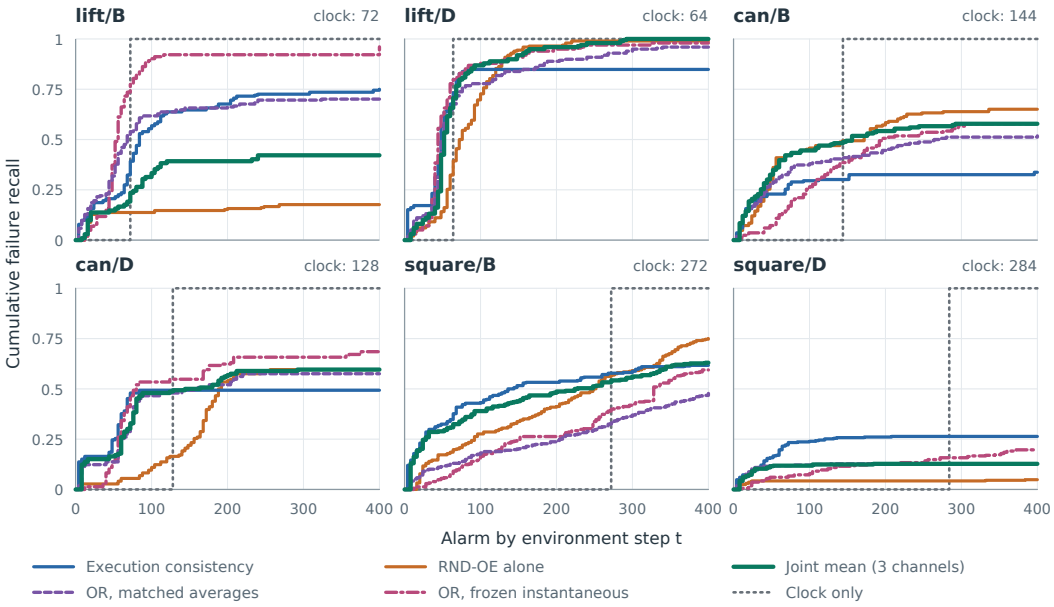



Figure 3: Post-hoc cumulative failure recall under stop-at-success monitoring (S), with thresholds recalibrated on eligible CAL-success opportunities. Curves show cellwise medians over 50 shared splits; missed failures remain in the denominator. Dotted curves show the clock-only monitor, and panel annotations give its median first-alarm step. All failures time out at step 400: earlier alarms indicate warning before timeout, not detection before validated failure onset. B/D denote the BC-RNN/Diffusion Policy feature backbones.

| Joint mean minus comparator | Conv. | Full horizon | By step 100 | Before clock |
|---|---|---|---|---|
| Execution alone | E | $+0.132$ $[+0.100, +0.164]$ | $+0.061$ $[+0.031, +0.092]$ | $+0.053$ $[+0.024, +0.081]$ |
| | S | $+0.015$ $[-0.016, +0.047]$ | $-0.035$ $[-0.066, -0.002]$ | $-0.010$ $[-0.042, +0.024]$ |
| RND-OE alone | E | $+0.012$ $[-0.016, +0.039]$ | $+0.152$ $[+0.120, +0.185]$ | $+0.130$ $[+0.097, +0.161]$ |
| | S | $+0.021$ $[-0.006, +0.043]$ | $+0.145$ $[+0.114, +0.178]$ | $+0.121$ $[+0.091, +0.154]$ |
| Matched OR | E | $+0.076$ $[+0.057, +0.093]$ | $+0.076$ $[+0.056, +0.093]$ | $+0.059$ $[+0.043, +0.077]$ |
| | S | $-0.001$ $[-0.021, +0.021]$ | $+0.011$ $[-0.012, +0.035]$ | $+0.002$ $[-0.023, +0.023]$ |

Table 2: Paired recall differences on the second confirmation, with pointwise 95% failure-bootstrap intervals. Positive values favor joint mean. Intervals use independent cell streams and condition on the success pool and fixed splits. The E full-horizon comparison with execution alone was prespecified; the remaining contrasts and early endpoints are post hoc. All intervals shown use the audited bootstrap calculation; original gate outcomes are unchanged.

### 4.5 Replication and comparisons with published monitors

Adding flow-inversion energy to execution consistency gives a paired discovery gain of $+0.023$, but this becomes $-0.003$ on the first confirmation and both prespecified G-C1 criteria fail (Tables 7 and 8 in the appendix). The discovery advantage therefore does not replicate. Other ablations also show that additional information or complexity need not help: adding action-chunk entropy

reduces two-channel recall by $0.135$ on the separate published-monitor population, and an apparent advantage on weaker actors largely disappears after actor retraining (Appendices E and F.1).

On the published-monitor population under (E), demonstration-trained FAIL-Detect logpZO has a macro-recall difference of $-0.002$ $[-0.040, +0.037]$ from execution consistency, without establishing superiority of either score. Training logpZO and RND on successful deployed-policy rollouts improves their reported recall on that population, but the comparison changes training-data access and does not isolate score-family superiority (Appendix E). Together, these findings distinguish the validity supplied by rollout calibration from detection power, which depends on the available signals, their aggregation, and the operational monitoring interval.

## 5 Limitations and scope

**Empirical scope.** The main experiments cover three robomimic tasks with rectified-flow actors and two visual-feature backbones; generalization to other policy families or deployment environments remains unestablished. Exploratory evaluations of reference BC-RNN and Diffusion Policy implementations on two additional tasks appear in Appendix G. Published monitors are evaluated on a separate population with documented implementation deviations and training-data access differences (Appendix E); these comparisons do not establish general superiority over published scores.

**Guarantee.** Proposition 1 requires conditional exchangeability of successful calibration and test rollouts, with the statistic and monitoring convention fixed independently of threshold calibration. The guarantee is marginal over the calibration draw and the next successful rollout; it neither bounds the conditional FPR of every realized calibration set nor protects against deployment shift. Under conditional i.i.d. sampling, the alternative in Appendix A.2 provides a high-probability bound on calibration-conditional FPR. Both guarantees concern the specified monitoring interval. Outcomes censored by alarm-triggered intervention cannot automatically be reused for calibration.

**Detection power and monitoring convention.** False-alarm control does not ensure informative scores or beneficial fusion. The flow-inversion gain does not replicate, and three-channel fusion fails its prespecified per-cell criterion. On the second confirmation, the joint mean's advantages over execution alone and matched running-average OR depend on fixed-horizon evaluation (E). After stop-at-success recalibration (S), neither full-horizon advantage is established, and execution alone has higher recall by step 100. Higher early recall than RND-OE alone therefore does not demonstrate a general advantage of fusion over simpler monitors. These findings also do not isolate a causal mechanism for the observed gains.

**Timing and statistical interpretation.** All 676 failures in the second confirmation exhaust the time budget without achieving success. Early recall and before-clock recall therefore quantify warnings before timeout, without establishing detection before failure onset. Discovery onset annotations were not independently validated, so onset-based timing metrics remain descriptive (Appendix C.6). The stopping-convention, matched-OR, and early-recall analyses are post hoc. Their pointwise intervals condition on the observed success pool and fixed splits; they do not account for exploratory analysis selection, and the overlapping splits are not independent replications.

## 6 Conclusion

RoLCA separates monitoring-statistic construction from rollout-level false-alarm calibration. Applying standard split-conformal calibration to a fixed statistic's maximum over eligible monitoring times controls FPR under exchangeability, without within-rollout independence or valid per-cycle $p$- or $e$-values. Across 30,502 rollouts, calibrated monitors operate near the nominal FPR, while detection gains depend on score quality and the monitoring convention. Joint mean outperforms matched running-average OR under fixed-horizon evaluation, but this advantage is not established after stop-at-success recalibration. Additional channels improve early recall over RND-OE alone without consistently outperforming execution consistency. These results support calibrating the complete operational alarm event and evaluating detection power with explicit stopping rules, matched comparators, and timing baselines.

### AI USE STATEMENT

In preparing this work, the authors used OpenAI and Anthropic language models to assist with manuscript editing and reference checking, figure and table preparation, experiment scripts, and supplementary documentation. The tools also provided feedback on experimental design, presentation, and interpretation of results. The authors reviewed and revised the assisted work, checked references and reported results against their source records, and take full responsibility for the methods, data, claims, and final manuscript.

### ETHICS STATEMENT

This work studies failure monitoring in simulated robotic manipulation using existing datasets. Human review is limited to annotating simulated rollouts. The proposed monitors may inform intervention decisions, but their false-positive guarantees require exchangeability and do not ensure failure detection or safe operation under deployment shift. Evaluation uses uninterrupted, shadow-mode rollouts and does not establish the safety or effectiveness of alarm-triggered interventions. We report negative findings, implementation deviations, and post-hoc analyses alongside confirmatory results to clarify the evidence and its limitations.

### REPRODUCIBILITY STATEMENT

The release will include source code, environment specifications, evaluation protocols and amendments, seed and checkpoint manifests, and split-level outputs with scripts to regenerate the tables and figures. Appendix A provides the validity proofs; Appendix B details score construction, availability, and computation; and Appendix C specifies evaluation, annotation, and uncertainty estimation. Appendix D distinguishes confirmation outcomes from post-hoc analyses. Appendices E, F, and G provide published-monitor comparisons, sensitivity analyses, and the exploratory external-actor evaluation, respectively.

## REFERENCES

Christopher Agia, Rohan Sinha, Jingyun Yang, Ziang 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)*, volume 270 of *Proceedings of Machine Learning Research*, pp. 689–723, 2025.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi_0$: A vision-language-action flow model for general robot control. In *Robotics: Science and Systems (RSS)*, 2025a. doi: 10.15607/RSS.2025.XXI.010.

Kevin Black, Manuel Y. Galliker, and Sergey Levine. Real-time execution of action chunking flow policies. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025b.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion Policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10–11):1684–1704, 2025. doi: 10.1177/02783649241273668.

Alec Farid, Sushant Veer, and Anirudha Majumdar. Task-driven out-of-distribution detection with statistical guarantees for robot learning. In *Conference on Robot Learning (CoRL)*, volume 164 of *Proceedings of Machine Learning Research*, pp. 970–980, 2022.

Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, and Vladimir Vovk. Plug-in martingales for testing exchangeability on-line. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 86(5):1091–1128, 2024.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

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.

Lars Lindemann, Matthew Cleaveland, Gihyun Shim, and George J. Pappas. Safe planning in dynamic environments using conformal prediction. *IEEE Robotics and Automation Letters*, 8(8): 5116–5123, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Gary Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, 1971.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning (CoRL)*, volume 164 of *Proceedings of Machine Learning Research*, pp. 1678–1690, 2022.

Junnan Nie, Jiayi Li, Jiachen Zhang, Junyi Lao, Chenghao Liu, Tianle Zhang, Liang Lin, and Songfang Huang. PACE: Phase-aware chunk execution for robot policies with action chunking. *arXiv preprint arXiv:2606.00537*, 2026. doi: 10.48550/arXiv.2606.00537.

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

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. doi: 10.48550/arXiv.2504.16054.

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.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

Ziyang Rao, Yiren Zhao, Weiyu Guo, Ben Fei, Yandong Guo, and Hui Xiong. The geometry of flow-matching uncertainty: A cost-free uncertainty proxy and its application in flow-based VLA failure detection. arXiv preprint arXiv:2607.27933, 2026. URL `https://arxiv.org/abs/2607.27933`.

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *Conference on Robot Learning (CoRL)*, volume 229 of *Proceedings of Machine Learning Research*, pp. 661–682, 2023.

Charles Richter and Nicholas Roy. Safe visual navigation via deep learning and novelty detection. In *Robotics: Science and Systems (RSS)*, 2017.

Quentin Rolland, Fabrice Mayran de Chamisso, and Jean-Baptiste Mouret. Failure identification in imitation learning via statistical and semantic filtering. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Ralf Römer, Adrian Kobras, Luca Worbis, and Angela P. Schoellig. Failure prediction at runtime for generative robot policies. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. E-detectors: A nonparametric framework for sequential change detection. *The New England Journal of Statistics in Data Science*, 2(2): 229–260, 2024.

Rohan Sinha, Apoorva Sharma, Somrita Banerjee, Thomas Lew, Rachel Luo, Spencer M. Richards, Yixiao Sun, Edward Schmerling, and Marco Pavone. A system-level view on out-of-distribution data in robotics. *arXiv preprint arXiv:2212.14020*, 2022.

Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning (ACML)*, volume 25 of *Proceedings of Machine Learning Research*, pp. 475–490, 2012.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. Testing exchangeability on-line. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 768–775, 2003.

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

Jinghan Yang, Yunchao Zhang, Wang Yuan, Haolun Wan, Jiaming Zhang, Zhengyang Hu, and Yanchao Yang. Tri-Info: Generalizable, interpretable failure prediction for VLA models via information theory. arXiv preprint arXiv:2606.19998, 2026. URL `https://arxiv.org/abs/2606.19998`.

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, 2026. URL `https://arxiv.org/abs/2606.23085`.

Shijie Zhou, Bin Zhu, Jiarui Yang, Xiangyu Zhao, Jingjing Chen, and Yu-Gang Jiang. RC-NF: Robot-conditioned normalizing flow for real-time anomaly detection in robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43050–43060, 2026.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Kevin Lin, Abhiram Maddukuri, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.

# A THEORETICAL GUARANTEES

## A.1 ROLLOUT-LEVEL FALSE-POSITIVE CONTROL

**Lemma 1** (Exchangeable rank bound)**.** *Let $W_0, \dots, W_m$ be exchangeable random variables in $\mathbb{R} \cup \{-\infty\}$, and define $N_j = \#\{i \neq j : W_i < W_j\}$. For $k \in \{1, \dots, m\}$,*

$$\mathbb{P}(N_0 \geq k) \leq \frac{m+1-k}{m+1}.$$

*Proof.* For any realization, at most $m+1-k$ variables strictly exceed at least $k$ others. Exchangeability therefore gives

$$\mathbb{P}(N_0 \geq k) = \frac{1}{m+1}\mathbb{E}\left[\sum_{j=0}^{m} \mathbf{1}\{N_j \geq k\}\right] \leq \frac{m+1-k}{m+1}.$$

Ties can only reduce the number of strict exceedances. □

*Proof of Proposition 1.* Condition on the construction data and settings $\mathcal{D}$. The score models, transformations, aggregation, fusion, availability rules, and monitoring convention $\kappa$ define a fixed measurable map

$$V(R; \mathcal{D}) = \max_{j \in \mathcal{I}_\kappa(R)} Z_j(R; \mathcal{D}), \qquad \max \varnothing = -\infty.$$

Applying this same map to conditionally exchangeable successful calibration and test rollouts preserves exchangeability. This includes trajectory-dependent success stopping when it is part of the common map. Let $k = \lceil (m+1)(1-\alpha) \rceil$. If $k = m+1$, the threshold is $+\infty$ and no alarm occurs. Otherwise, $V(R; \mathcal{D}) > V_{(k)}$ means that the test statistic strictly exceeds at least $k$ calibration values. Lemma 1 gives

$$\mathbb{P}(V(R; \mathcal{D}) > h \mid \mathcal{D}) \leq \frac{m+1-k}{m+1} \leq \alpha.$$

This event is exactly an eligible first-crossing alarm. Averaging over $\mathcal{D}$ yields the unconditional bound. □

**Scope.** This standard split-conformal argument is marginal over the calibration set and the next successful rollout. It covers any fixed causal statistic on the specified bounded monitoring interval, without requiring valid per-cycle $p$-values, $e$-values, or within-rollout independence. Strictly increasing transformations preserve the alarm rule when applied to both statistics and thresholds. The result does not imply validity for an indefinitely running process or under deployment shift. Success refers to the outcome without intervention; alarm-interrupted rollouts cannot automatically be reused as successful calibration data.

## A.2 CALIBRATION-CONDITIONAL CONTROL

Suppose, more strongly, that conditional on $\mathcal{D}$, calibration rollouts and the independent test rollout are i.i.d. from the successful population. For $\delta \in (0, 1)$, define

$$k_\delta = \min\left(\left\{\ell \in \{1, \dots, m\} : F_{\mathrm{Beta}(\ell, m+1-\ell)}(1-\alpha) \leq \delta\right\} \cup \{m+1\}\right), \tag{10}$$

where $F_{\mathrm{Beta}(a,b)}$ is the Beta CDF. Set $h_\delta = V_{(k_\delta)}$ for $k_\delta \leq m$, and $+\infty$ otherwise. With $\mathcal{K}$ denoting the calibration set,

$$\mathbb{P}_{\mathcal{K}}[\mathbb{P}_R(V(R; \mathcal{D}) > h_\delta \mid \mathcal{D}, \mathcal{K}) \leq \alpha \mid \mathcal{D}] \geq 1 - \delta. \tag{11}$$

*Proof.* Let $F_{\mathcal{D}}$ be the conditional CDF of the successful-rollout statistic. At threshold $V_{(\ell)}$, strict exceedance gives conditional FPR $\rho_\ell = 1 - F_{\mathcal{D}}(V_{(\ell)})$. Represent $V_i = F_{\mathcal{D}}^{-1}(U_i)$ using independent uniform variables and the generalized inverse CDF. Then

$$F_{\mathcal{D}}(V_{(\ell)}) \geq U_{(\ell)}, \qquad U_{(\ell)} \sim \mathrm{Beta}(\ell, m+1-\ell),$$

so

$$\mathbb{P}_{\mathcal{K}}(\rho_\ell > \alpha \mid \mathcal{D}) \leq F_{\mathrm{Beta}(\ell, m+1-\ell)}(1-\alpha).$$

Equation (10) bounds this probability by $\delta$; the infinite-threshold case has zero conditional FPR. □

This order-statistic tolerance bound permits ties; continuity is needed only for equality with the Beta law (Vovk, 2012). Primary results use the marginal threshold in Eq. (3). The stronger rule is evaluated separately in Appendix F.4.

# B IMPLEMENTATION DETAILS

## B.1 SCORE RECORDS AND ALARM AVAILABILITY

Record $j$ refers to its originating policy cycle. In the main study, $H = 8$, $H' = 4$, and $T_{\max} = 400$. Execution consistency becomes available after the current prefix, at step $4j$; inversion requires two prefixes and becomes available at $4j + 4$. RND-OE uses only context features, but the evaluated implementation schedules its records on the same delayed grid as inversion. Thus execution alone can first alarm at step 4, while the evaluated inversion, RND-OE, and fused statistics first become available at step 8. Progress references use the originating record, not its completion time.

Execution scores are present in records 1–100; inversion and RND-OE scores are present in records 1–99, with a structurally missing terminal score. For inversion, the final window would exceed the horizon; for RND-OE, this is an implementation convention. A missing terminal score receives $p = 1$, hence $e = 0.5$, and remains included in the running average's denominator. This fixed contribution is not neutral evidence and does not erase earlier evidence.

Fusion combines averages through the same record, with opportunity times

$$\tau_j^{\text{exec}} = 4j, \qquad \tau_j^{\text{fusion}} = \begin{cases} 4j + 4, & 1 \le j \le 99, \\ 400, & j = 100. \end{cases} \tag{12}$$

The final execution score and the predetermined missing-channel updates are all available at step 400, so fused records 99 and 100 can be evaluated at the same step. Calibration and evaluation apply the same eligibility rule in Eq. (1). Under (S), ties with success are included; a future window is excluded, rather than padded merely because success has occurred.

## B.2 EXECUTED-WINDOW INVERSION

For $H = 2H'$, concatenate the executed prefixes originating at cycles $j$ and $j + 1$:

$$\widehat{a}_j = \operatorname{vec}\left( \begin{bmatrix} a_{j,1:H'} \\ a_{j+1,1:H'} \end{bmatrix} \right) \in \mathbb{R}^{Hd_a}.$$

We invert this continuation on the flow model's fixed sampler grid, holding $C_j$ constant, to obtain $x_{0,j}$ in Eq. (7). The second prefix is generated after replanning under $C_{j+1}$, so the continuation need not be a sample from the earlier conditional policy. The score $S_{\text{inv},j} = \|x_{0,j}\|_2^2/(2Hd_a)$ is the per-coordinate Gaussian base energy. It omits the change-of-variables term and incurs numerical inversion error; it is not an exact action negative log-likelihood. It can respond to replanning as well as unexpected behavior. Self-scoring uses the deployed policy; cross-scoring uses an independently trained policy.

## B.3 RND-OE ARCHITECTURE AND TRAINING

RND-OE denotes random network distillation on observation embeddings; its loss contains no outlier-exposure term. It takes the context-encoder output $x_j = \text{enc}(C_j) \in \mathbb{R}^{256}$, without actions or future observations. Each coordinate is standardized using the score-fit set to obtain $\widetilde{x}_j$. The fixed random target MLP $f$ has widths $256 \to 1024 \to 2048 \to 4096 \to 128$ and LeakyReLU hidden activations. The predictor $g$ has widths $256 \to 1024 \to 2048 \to 4096 \to 2048 \to 1024 \to 128$, with LeakyReLU in its first three hidden layers and ReLU in its last two. With $N_{\text{fit}}$ score-fit cycles, both training and scoring use the unsquared Euclidean norm:

$$\mathcal{L}_{\text{RND}} = \frac{1}{N_{\text{fit}}} \sum_{n=1}^{N_{\text{fit}}} \|g(\widetilde{x}_n) - f(\widetilde{x}_n)\|_2, \qquad S_{\text{RND},j} = \|g(\widetilde{x}_j) - f(\widetilde{x}_j)\|_2.$$

Architecture and loss follow the released FIPER implementation, but its decision thresholds are not used (Römer et al., 2025). AdamW uses batch size 256, learning rate $10^{-4}$, weight decay

$10^{-5}$, and cosine decay to $10^{-6}$ over at most 250 epochs. Score-fit cycles are split 90/10 for training and validation. Training stops after seven epochs without validation improvement or when the validation-to-training loss ratio exceeds five; the best validation checkpoint is restored.

Rollout-trained variants use all cycles from a 30% subsample of successful policy rollouts; demonstration-trained variants use 100 training demonstrations. In Appendix E, score-fit rollouts are excluded from the analysis pool. For the second confirmation, score fitting uses discovery successes, and both networks and standardization constants are fixed before confirmation collection. These score-fit data are distinct from the split-specific TRAIN successes used to construct RoLCA's progress ranks. RND-OE uses the same rank transform and running average as execution consistency, with the evaluated availability convention in Appendix B.1.

### B.4 Temporal statistics and multichannel decision rules

**Standardized-score baselines.** Within each channel and split, define

$$\zeta_j = \max\left\{0, \frac{S_j - \mu_{b(j)}}{\sigma_{b(j)}}\right\},$$

where $\mu_b$ and $\sigma_b$ are the mean and standard deviation of successful TRAIN scores in progress bin $b$. The standard deviation is floored at $10^{-3}$; bins with fewer than three references use $\mu_b = 0$ and $\sigma_b = 1$. The instantaneous execution baseline calibrates $U_{\max} = \max_{j \in \mathcal{I}_\kappa(R)} \zeta_j$ and alarms at the first eligible $\zeta_j > h$. CUSUM instead uses

$$G_0 = 0, \qquad G_j = \max\{0, G_{j-1} + \zeta_j - d_{b(j)}\}.$$

For each nonempty bin, $d_b$ is the $\min\{n_b, \lceil (n_b+1)(1-\alpha_{\text{ref}})\rceil\}$th order statistic of its $n_b$ standardized references, with $\alpha_{\text{ref}} = 0.10$. This drift remains fixed when the target FPR changes. CUSUM calibrates $\max_{j \in \mathcal{I}_\kappa(R)} G_j$. Missing scores supply no candidate to a maximum and no CUSUM update.

**Joint mean, joint maximum, and OR.** For channel set $\mathcal{C}$, the default mean uses the running averages $A_j^{(c)}$ from Eq. (9). The alternative joint statistics are $\max_c \zeta_j^{(c)}$ and $\max_c A_j^{(c)}$. Each joint rule calibrates its own maximum over eligible records at budget $\alpha$. For positive statistics, calibrating their logarithms is equivalent to calibrating the statistics directly.

The instantaneous OR uses $\zeta_j^{(c)}$ without averaging; the matched running-average OR uses $A_j^{(c)}$, equivalently $\log A_j^{(c)}$. Each calibrates channel-wise rollout maxima at $\alpha/|\mathcal{C}|$ and alarms at the first eligible component crossing. For $m$ calibration successes, the component order-statistic index is $\lceil (m+1)(1-\alpha/|\mathcal{C}|)\rceil$, with threshold $+\infty$ if it exceeds $m$. Missing instantaneous scores contribute no candidate; running averages retain the fixed terminal update described above. OR uses each channel's availability schedule, whereas fusion waits for its required inputs. Thus matched OR fixes the channel statistics, but not a common alarm-opportunity grid or the decision region. Neither rule is guaranteed to have greater power.

For fixed channel rules, Proposition 1 bounds each channel's eligible-alarm probability by $\alpha/|\mathcal{C}|$ on a successful rollout. The union bound therefore controls OR's rollout-level FPR by $\alpha$, without independence across channels.

**Auxiliary channels and product statistics.** The auxiliary set $\mathcal{C}_{\text{aux}} = \{\text{exec}, \text{plan}, \text{vq}\}$ contains execution consistency, the causal prior's negative log-probability of the executed plan token, and the plan-tokenizer quantization residual. These last two scores are demonstration-trained and excluded from the main three-channel monitor, which uses execution, inversion, and RND-OE. The auxiliary product construction is

$$P_j^{(c)} = \prod_{i=1}^{j} e_i^{(c)}, \qquad M_j^{\text{prod}} = \frac{1}{3} \sum_{c \in \mathcal{C}_{\text{aux}}} P_j^{(c)}.$$

Products are computed in the log domain without restarting, and $\max_{j \in \mathcal{I}_\kappa(R)} \log M_j^{\text{prod}}$ is calibrated on successful CAL rollouts. Tables call this rule *mean of three products*. Its validity comes from

outer calibration. In contrast, the product/Ville diagnostic uses execution evidence alone, alarming at the first eligible record with $P_j^{(\mathrm{exec})} = \prod_{i=1}^{j} e_i^{(\mathrm{exec})} > 1/\alpha$. The product is accumulated in the log domain without restarting, and no outer calibration is applied. Marginal validity of the factors would not establish the sequential condition required by Ville's inequality.

### B.5 Computational cost

| Operation | H200 GPU | CPU (4 threads) |
|---|---|---|
| Policy: sample one chunk | 9.2 ms | 11.4 ms |
| Execution consistency | 0.8 ms | 0.6 ms |
| Flow inversion | 9.1 ms | 11.5 ms |
| Auxiliary token scores | 1.6 ms | 1.9 ms |

Table 3: Median of 50 timed calls after warm-up, at batch size one on the lift/B bundle with eight Euler steps. These measurements describe computation, not end-to-end intervention latency.

Execution consistency and eight-step inversion together take 9.9 ms on the measured H200 GPU and 12.1 ms on four CPU threads (Table 3). At 20 Hz, a four-action prefix spans 200 ms, so these two monitoring operations take less than 10% of that interval. This excludes policy generation and is distinct from observation availability: inversion first becomes available after step 8. The measurements do not characterize the full RND-OE fusion pipeline or end-to-end intervention latency.

## C Experimental protocol and statistical analysis

### C.1 Populations and analysis status

The primary study contains 30,502 rollouts: 11,752 discovery rollouts and three separate populations of 6,250 for flow-inversion confirmation, published-monitor comparisons, and rollout-trained confirmation. The second confirmation has 5,574 successes and 676 failures (Table 5). All six primary cells use rectified-flow actors; B/D identify BC-RNN/Diffusion Policy *feature backbones*, not different actor families. The external-actor extension is separate (Appendix G).

| Study | Evaluation design |
|---|---|
| Discovery | Splits, metrics, FPR levels, and initial monitors specified before analysis; later amendments added comparisons. |
| Flow-inversion confirmation | Population, monitors, analyses, and decision criteria fixed before seed opening. |
| Published-monitor comparison | Exploratory comparison with implementation and analysis choices specified before each analysis stage. |
| Rollout-trained confirmation | Channel, comparators, population, and decision criteria fixed before seed opening; stopping-convention, matched-OR, timing, and calibration-conditional analyses reported post hoc. |
| External-actor extension | Exploratory evaluation, excluded from primary confirmatory claims. |

Table 4: Roles of the study populations. Protocol qualifications are given in this section and Appendix G.

Within each analysis pool, 50 shared rollout-level splits allocate successes to TRAIN (50%), CAL (25%), and evaluation (25%). Any reserved score-fit rollouts are first removed, as described in Appendix E. Failures are used only for evaluation and appear in every split. Score-model fitting, rank-reference construction, and threshold calibration are distinct operations.

For each confirmation, candidate monitors and decision criteria were fixed before seed opening. The second-confirmation seed was reassigned from a planned robustness study; a rehearsal on earlier populations could cancel the study but could not change its criteria. Its conditional calibration analysis was specified but omitted initially and computed after inspecting outcomes. It is reported as post hoc, together with matched running-average OR, E/S comparisons, the clock, and early endpoints. These analyses reuse the same population and do not constitute an independent confirmation.

### C.2 MONITORING CONVENTIONS AND TIMING ENDPOINTS

The original convention (E) monitors through $T_{\max} = 400$, including observations after success. Convention (S) retains opportunities with $\tau_j \le \min\{T_{\text{succ}}, T_{\max}\}$, including ties with success. Failures remain active through step 400. For (S), every threshold is recalibrated using eligible opportunities on CAL successes; score trajectories and TRAIN references are held fixed. Masking evaluation alone is insufficient.

The clock uses distinct fused-opportunity steps $\mathcal{O} = \{8, 12, \ldots, 400\}$. For calibration success $r$, define

$$q_r = \max\{t \in \mathcal{O} : t \le T_{\text{succ}}(R_r)\}, \qquad \max \varnothing = -\infty.$$

Its (S) threshold $h_{\text{clk}}$ is the $k = \lceil (m+1)(1-\alpha) \rceil$ order statistic of $q_1, \ldots, q_m$, or $+\infty$ if $k = m+1$. The earliest possible clock alarm is

$$t_{\text{clk}} = \min\{t \in \mathcal{O} : t > h_{\text{clk}}\}, \qquad \min \varnothing = +\infty,$$

fitted separately in each cell and split. Under (E), all calibration maxima equal 400 and the clock never crosses its threshold.

Early recall counts failures alarmed by step 100; before-clock recall counts those alarmed strictly before $t_{\text{clk}}$. Misses remain in the denominator, and ties with the clock are excluded. Both E and S use the same (S)-calibrated clock reference. All 676 second-confirmation failures time out at step 400, so these endpoints measure warning before timeout, not detection before independently validated failure onset.

Table 5: Second-confirmation population and clock reference. CAL counts are per split; clock entries are median first-alarm steps under (S). B/D denote visual-feature backbones. Totals: 6,250 rollouts, 5,574 successes, and 676 failures.

| Cell | Rollouts | Successes | Failures | CAL | Clock $t_{\text{clk}}$ |
|---|---|---|---|---|---|
| lift/B | 1800 | 1698 | 102 | 424 | 72 |
| lift/D | 2300 | 2201 | 99 | 550 | 64 |
| can/B | 650 | 567 | 83 | 141 | 144 |
| can/D | 800 | 727 | 73 | 181 | 128 |
| square/B | 350 | 196 | 154 | 49 | 272 |
| square/D | 350 | 185 | 165 | 46 | 284 |

### C.3 AGGREGATION AND PAIRED ESTIMANDS

For monitor $q$, let $r_{i,s}^{(q)}$ be recall in cell $i$ and split $s$. With $N_{\text{cell}}$ cells, macro recall is

$$\widehat{r}^{(q)} = \frac{1}{N_{\text{cell}}} \sum_{i=1}^{N_{\text{cell}}} \text{median}_{s=1,\ldots,50}\, r_{i,s}^{(q)}.$$

Here $N_{\text{cell}} = 6$ for the primary study and 4 for the external-actor extension. FPR is aggregated in the same order. We distinguish

$$\widehat{\Delta}_{a,b} = \frac{1}{N_{\text{cell}}} \sum_{i=1}^{N_{\text{cell}}} \text{median}_{s=1,\ldots,50} \left( r_{i,s}^{(a)} - r_{i,s}^{(b)} \right), \tag{13}$$

$$\widehat{D}_{a,b} = \widehat{r}^{(a)} - \widehat{r}^{(b)}. \tag{14}$$

The paired macro gain $\widehat{\Delta}$ need not equal the macro-recall difference $\widehat{D}$: for joint mean versus execution on the second confirmation, these are $0.132$ and $0.523 - 0.403 = 0.120$. G-C1 and the second-confirmation criteria use $\widehat{\Delta}$; the original G-C2 criterion and published-monitor comparisons use $\widehat{D}$. The first-confirmation G-C2 gain is $0.104$ under $\widehat{D}$, compared with $0.100$ under $\widehat{\Delta}$. The overlapping splits describe sensitivity to partitioning the observed success pool and are not independent replications.

### C.4 BOOTSTRAP UNCERTAINTY AND ANALYSIS CORRECTIONS

We use 2,000 within-cell failure-bootstrap resamples, sharing failure IDs across monitors and splits within each cell. Each resample recomputes the reported statistic, including the cell medians and macro aggregation. Intervals for $\widehat{\Delta}$ use paired within-split differences; intervals for $\widehat{D}$ use differences of separately aggregated macro recalls. The 2.5th and 97.5th percentiles give pointwise 95% intervals conditional on the observed success pool and fixed splits. They do not account for sampling a new success pool or selection of post-hoc comparisons. Split interquartile ranges describe partition sensitivity, not failure-sampling uncertainty.

Second-confirmation intervals use independent random streams across cells and shared resamples across monitors, splits, endpoints, and conventions within a cell, except the marked historical interval in Table 11. After unblinding, 17 earlier paired-gain intervals were corrected to bootstrap $\widehat{\Delta}$ rather than $\widehat{D}$. The second-confirmation stream correction moved interval endpoints by at most $0.0023$ across 20 checked contrasts and changed none of their inclusion-of-zero conclusions. Point estimates and prespecified outcomes were unchanged; no gate used an interval. The timing correction placed two terminal joint-max alarms at step 400 and restored that opportunity under (S), without changing thresholds or reported macro summaries. Detailed correction records accompany the reproducibility materials.

### C.5 FPR ACCEPTANCE CRITERIA AND REFERENCE BANDS

Figure 2 reports cellwise FPR medians and interquartile ranges over splits. Two distinct single-split constructions accompany these summaries.

**Prespecified binomial acceptance bands.** Let $B \sim \mathrm{Binomial}(n_{\mathrm{eval}}, \alpha)$ and let $Q_B$ be its quantile function. The acceptance band is

$$\left[\frac{Q_B(0.025)}{n_{\mathrm{eval}}}, \frac{Q_B(0.975)}{n_{\mathrm{eval}}}\right].$$

It models evaluation-count variation at fixed conditional FPR $\alpha$, excluding calibration randomness. Examples on the second confirmation are $[0.076, 0.125]$ for $n_{\mathrm{eval}} = 551$ and $[0.021, 0.192]$ for $n_{\mathrm{eval}} = 47$. All 30 first-confirmation and 18 second-confirmation checks fall within their prespecified bands.

**Beta-binomial reference bands.** Under conditionally i.i.d. continuous calibration and evaluation statistics, with $k = \lceil (m+1)(1-\alpha) \rceil \le m$,

$$\Pi \sim \mathrm{Beta}(m+1-k, k), \qquad B \mid \Pi \sim \mathrm{Binomial}(n_{\mathrm{eval}}, \Pi).$$

Here $\Pi$ is the calibration-set-dependent conditional FPR. The central 95% range of $B/n_{\mathrm{eval}}$ includes calibration and evaluation variation, with expected FPR $(m+1-k)/(m+1) \le \alpha$. For $k = m+1$, the count is zero. Figure 2 uses these bands for reference, not acceptance testing. Ties need not preserve the exact Beta-binomial law or its central two-sided coverage.

Neither construction gives 95% coverage for medians over overlapping splits or simultaneous coverage for the 48 checks. Wide bands in small cells do not establish tight agreement with $\alpha$ or exclude an FPR of $0.15$. We do not estimate a population-level confidence interval for median FPR over newly sampled success pools. The formal guarantee comes from Proposition 1, independently of these checks. The execution-channel product/Ville diagnostic exceeds the first-confirmation upper acceptance limit in every cell, with median FPRs $0.205$–$0.378$.

### C.6 HUMAN ANNOTATION

Five raters reviewed blinded two-camera videos of the 1,081 simulator-defined discovery failures. The first round covered ambiguous cases and a stratified audit; the second covered the remaining cases after poor agreement with geometric labels was observed. Overlapping cases estimated agreement, majority votes determined categories, and four cases without a majority were adjudicated.

Fleiss' agreement coefficients were $0.451$ and $0.475$ in the two rounds, with 63% pairwise agreement. Human labels support diagnostic categories; they did not change simulator outcomes or geometric event times and do not independently validate failure onset.

| Human category | Count | Percent |
|---|---|---|
| Contact / execution | 456 | 42.2 |
| Stalling / no progress | 239 | 22.1 |
| Planning / action choice | 159 | 14.7 |
| Failed recovery / error propagation | 116 | 10.7 |
| Ambiguous / other | 99 | 9.2 |
| Not a failure / spurious | 12 | 1.1 |
| Total | 1,081 | 100.0 |

Table 6: Human annotation of simulator-defined discovery failures. The geometric labeler agrees with human consensus on 48.8% of auto-confident cases. Four cases without a majority were adjudicated; simulator outcomes remain unchanged.

# D Additional results for the main study

## D.1 Discovery and flow-inversion confirmation

Tables 7–10 report discovery and first-confirmation results under (E). The self-scored inversion gain falls from $+0.023$ $[+0.004, +0.039]$ to $-0.003$ $[-0.026, +0.018]$, failing both G-C1 criteria. G-C1a requires a positive paired macro gain. G-C1b requires positive gains in at least four cells and no cell gain below $-0.02$ (square/B exempt); only three cells improve and square/D loses $0.026$. The execution-only running-average comparator was selected by discovery macro recall among single-channel conformal monitors and fixed across confirmation cells and splits. Product-mean and joint-maximum gains over auxiliary instantaneous OR remain positive, but their mutual difference is inconclusive on confirmation. The product-versus-OR contrast also changes temporal aggregation (Figure 4).

*(a) Macro performance and descriptive timing*

| Monitor | FPR | Recall | Recall$_{\neg\mathrm{amb}}$ | Recall$_{<\mathrm{dl}}$ |
|---|---|---|---|---|
| $U_{\max}$ conformal | 0.092 | 0.392 | 0.397 | 0.331 |
| CUSUM conformal | 0.095 | 0.410 | 0.419 | 0.351 |
| Running-average evidence | 0.093 | 0.425 | 0.431 | 0.389 |
| exec+inv fusion (self) | 0.093 | 0.444 | 0.453 | 0.415 |
| exec+inv fusion (cross) | 0.094 | 0.470 | 0.476 | 0.435 |

*(b) Cellwise recall*

| Monitor | lift/B | lift/D | can/B | can/D | square/B | square/D |
|---|---|---|---|---|---|---|
| $U_{\max}$ conformal | 0.19 | 0.51 | 0.21 | 0.57 | 0.57 | 0.30 |
| CUSUM conformal | 0.20 | 0.48 | 0.34 | 0.57 | 0.62 | 0.26 |
| Running-average evidence | 0.20 | 0.51 | 0.38 | 0.59 | 0.49 | 0.38 |
| exec+inv fusion (self) | 0.27 | 0.59 | 0.44 | 0.59 | 0.39 | 0.39 |
| exec+inv fusion (cross) | 0.31 | 0.56 | 0.45 | 0.60 | 0.52 | 0.38 |

Table 7: Discovery under (E), at $\alpha = 0.10$. Entries are split medians; macro metrics average six cell medians. Diagnostic recalls exclude ambiguous failures ($\neg\mathrm{amb}$) or require alarms before the annotated deadline ($<\mathrm{dl}$), whose relation to failure onset is not independently validated. Self/cross denote deployed/independently trained inversion models; B/D denote feature backbones.

**Illustrative trajectory.** Figure 5 illustrates an eligible threshold crossing; it does not establish failure-onset timing.

| Monitor | FPR | Recall | l/B | l/D | c/B | c/D | s/B | s/D |
|---|---|---|---|---|---|---|---|---|
| $U_{\max}$ conformal | 0.090 | 0.348 | 0.17 | 0.52 | 0.15 | 0.47 | 0.52 | 0.26 |
| CUSUM conformal | 0.089 | 0.364 | 0.17 | 0.50 | 0.28 | 0.47 | 0.49 | 0.27 |
| Execution alone | 0.093 | 0.396 | 0.19 | 0.54 | 0.41 | 0.51 | 0.46 | 0.27 |
| exec+inv fusion (self) | 0.092 | 0.385 | 0.20 | 0.56 | 0.45 | 0.49 | 0.37 | 0.24 |
| exec+inv fusion (cross) | 0.090 | 0.412 | 0.22 | 0.58 | 0.49 | 0.49 | 0.46 | 0.23 |

Table 8: First confirmation under (E), at $\alpha = 0.10$: 6,250 rollouts and 50 paired splits. Execution alone is the running-average comparator fixed from discovery. l/c/s denote lift/can/square; B/D denote feature backbones. Both G-C1 criteria fail, as detailed in the text.

| Monitor | Discovery | First confirmation |
|---|---|---|
| $U_{\max}$ conformal | .392 [.365, .418] | .348 [.313, .385] |
| CUSUM conformal | .410 [.381, .436] | .364 [.326, .402] |
| Running-average evidence | .425 [.398, .453] | .396 [.357, .435] |
| Mean of three products | .424 [.397, .452] | .370 [.332, .407] |
| Joint maximum of three $\zeta$ scores | .380 [.354, .406] | .355 [.319, .392] |
| OR budget splitting | .301 [.276, .323] | .266 [.231, .296] |
| exec+inv mean (self) | .444 [.415, .473] | .385 [.345, .424] |
| exec+inv mean (cross) | .470 [.441, .499] | .412 [.371, .450] |

Table 9: Macro recall under (E), at $\alpha = 0.10$, with pointwise 95% failure-bootstrap intervals conditional on the success pool and splits.

| Comparison | Discovery | First confirmation |
|---|---|---|
| exec+inv (self) − execution | $+.023$ $[+.004, +.039]$ | $-.003$ $[-.026, +.018]$ |
| exec+inv (cross) − execution | $+.043$ $[+.027, +.061]$ | $+.018$ $[-.005, +.039]$ |
| Mean of products − OR | $+.121$ $[+.093, +.150]$ | $+.100$ $[+.063, +.138]$ |
| Joint $\zeta$ max − OR | $+.073$ $[+.058, +.090]$ | $+.079$ $[+.059, +.098]$ |
| Mean of products − joint $\zeta$ max | $+.045$ $[+.016, +.071]$ | $+.022$ $[-.021, +.058]$ |
| Mean of products − execution | $-.009$ $[-.033, +.016]$ | $-.031$ $[-.069, +.006]$ |
| Joint $\zeta$ max − $U_{\max}$ | $-.010$ $[-.021, +.001]$ | $+.005$ $[-.008, +.020]$ |

Table 10: Paired macro gains $\widehat{\Delta}$ with pointwise 95% failure-bootstrap intervals under (E), at $\alpha = 0.10$. Execution denotes the fixed execution-only running-average comparator. Product, joint $\zeta$ maximum, and OR use the auxiliary channel set; OR thresholds its instantaneous standardized scores.

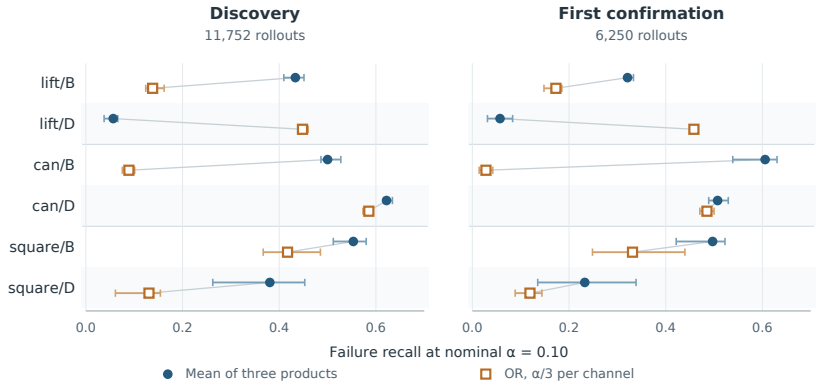



Figure 4: Auxiliary-channel mean of products versus instantaneous OR ($\alpha/3$ per channel) under (E), at $\alpha = 0.10$. Both use execution, plan, and quantization-residual scores. Points and bars show medians and split interquartile ranges, not confidence intervals. The comparison changes temporal aggregation and the decision rule; paired-gain intervals appear in Table 10.

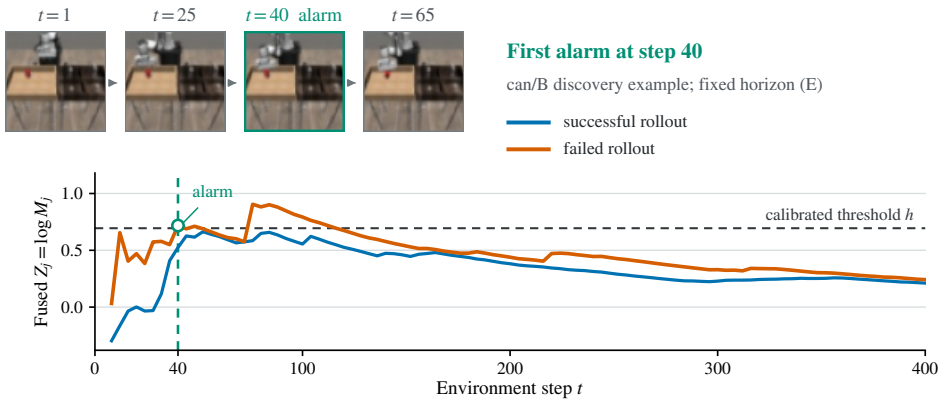



Figure 5: Illustrative discovery rollouts for can/B under fixed-horizon monitoring (E). The failed rollout first crosses the calibrated threshold at step 40; the successful rollout stays below it. B denotes the BC-RNN feature backbone. Frames show the failed rollout at the indicated steps. Failure-onset timing is not independently established.

### D.2 Rollout-trained confirmation

Under (E), execution, execution plus inversion, RND-OE, and joint mean reach macro recalls $0.403$, $0.398$, $0.507$, and $0.523$. Joint mean passes the prespecified macro-gain thresholds against the first two comparators but fails the per-cell criterion: five cells improve, while square/D loses $0.064$, beyond the permitted $0.02$. All 18 FPR checks pass their acceptance bands; the prespecified three-monitor suite has macro FPRs $0.090$–$0.092$. RND-OE accounts for most of the full-horizon increase: joint mean's paired gain over RND-OE is $+0.012$ $[-0.016, +0.039]$. RND-OE uses successful policy rollouts for score-model training, whereas the original scores use demonstration-trained models; all use successful TRAIN rollouts for rank references. These comparisons therefore do not isolate a score-family effect.

| Contrast under (E) | $\widehat{\Delta}$ | 95% interval |
|---|---|---|
| *Prespecified comparisons* | | |
| Joint mean − execution | $+0.132$ | $[+0.100, +0.164]$ |
| Joint mean − (execution + inversion) | $+0.127$ | $[+0.098, +0.155]^{\dagger}$ |
| Joint mean − instantaneous OR | $+0.118$ | $[+0.087, +0.148]$ |
| *Post-hoc comparisons* | | |
| Joint mean − matched running-average OR | $+0.076$ | $[+0.057, +0.093]$ |
| Matched running-average OR − instantaneous OR | $+0.054$ | $[+0.023, +0.081]$ |

Table 11: Second-confirmation paired contrasts under (E). The first two rows exceed the prespecified macro-gain threshold $0.05$; the separate per-cell criterion fails. The three OR contrasts are separate, nonadditive estimands. Intervals use independent cell streams except †, which retains the original interval outside that audit.

### D.3 Monitoring conventions, decision rules, and timing

Tables 1 and 2 report the E/S comparison. Joint mean improves early recall over RND-OE under both conventions, but its full-horizon gains over execution and matched OR are inconclusive under (S). Early-recall differences measure coverage by a cutoff, not paired timing on the same detected failures. The clock's complete eventual recall reflects this population's timeout failures. Figure 6 complements the main (S) curves with (E) timing results.

**Interpreting the OR contrasts.** In Table 11, joint mean versus matched OR changes the multi-channel decision rule and associated opportunities while fixing channel statistics. Matched versus instantaneous OR changes the channel statistic within an OR rule; joint mean versus instantaneous

OR changes both. Their paired gains are not additive because medians of paired differences are not additive. The $+0.076$ gain is therefore neither a fixed fraction of $+0.118$ nor a pure calibration effect.

A common nominal budget does not match realized FPR. Instantaneous OR has FPR $0.075$ under (E); discrete calibration ranks and overlapping component alarms can leave budget unused. For square/B and square/D, $m = 49$ and 46 give component indices equal to $m$ at $\alpha/3$, so the marginal rank bounds are $1/50$ and $1/47$. These are probability bounds, not measured component FPRs. Under (S), instantaneous OR reaches recall $0.666$ versus joint mean's $0.559$; this ordering does not imply general superiority of instantaneous scoring.

**Cellwise heterogeneity.** Table 12 shows that square/D favors execution at the displayed endpoints under both conventions. Lift/B's full-horizon contrast reverses from $+0.029$ under (E) to $-0.343$ under (S), while can/B and lift/D retain positive differences.

Table 12: Cellwise paired recall gains of joint mean over execution. Full-horizon entries include pointwise 95% failure-bootstrap intervals; step-100 entries are point estimates. Each estimate is the median within-split difference over 50 shared splits.

| Cell | Full (E) | Full (S) | $\leq 100$ (E) | $\leq 100$ (S) |
|---|---|---|---|---|
| lift/B | $+0.029$ $[-0.059, +0.118]$ | $-0.343$ $[-0.441, -0.235]$ | $+0.078$ | $-0.265$ |
| lift/D | $+0.414$ $[+0.323, +0.515]$ | $+0.152$ $[+0.081, +0.222]$ | $+0.222$ | $+0.020$ |
| can/B | $+0.217$ $[+0.120, +0.313]$ | $+0.235$ $[+0.133, +0.325]$ | $+0.133$ | $+0.145$ |
| can/D | $+0.116$ $[+0.055, +0.192]$ | $+0.110$ $[+0.041, +0.178]$ | 0.000 | 0.000 |
| square/B | $+0.078$ $[+0.016, +0.120]$ | $+0.036$ $[-0.019, +0.091]$ | $-0.006$ | $-0.029$ |
| square/D | $-0.064$ $[-0.112, -0.012]$ | $-0.100$ $[-0.155, -0.052]$ | $-0.061$ | $-0.082$ |

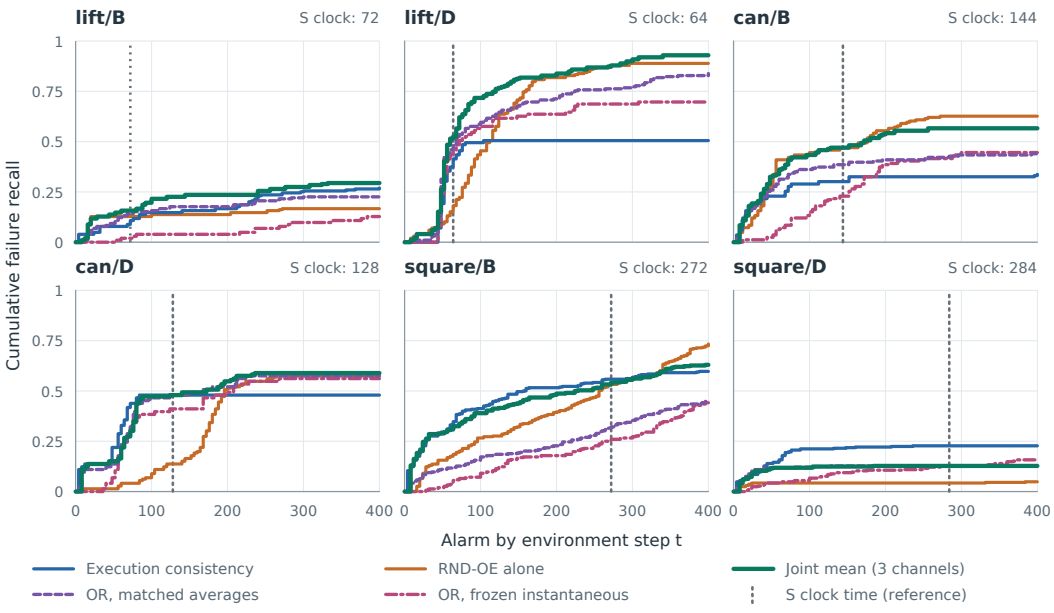



Figure 6: Post-hoc cumulative failure recall under (E) on the second confirmation. Curves show cellwise medians over 50 splits, retaining misses in the denominator. Gray markers show median alarm times of the (S)-calibrated clock; the (E) clock never alarms. All failures time out at step 400, so earlier alarms indicate warning before timeout. B/D denote feature backbones.

# E COMPARISONS WITH PUBLISHED MONITORS

## E.1 EVALUATION DESIGN

A separate 6,250-rollout population supplies 128 sampled chunks per cycle for Sentinel and FIPER statistics (Agia et al., 2025; Römer et al., 2025). Reserving 30% of successes for learned-score fitting leaves an analysis pool of 3,863 successes and 733 failures, evaluated with 50 common splits. Scores include FAIL-Detect logpZO, observation-only RND-OE, action-augmented RND-A, FIPER action-chunk entropy (ACE), and Sentinel temporal consistency (STAC). STAC uses RBF-MMD between overlapping segments of consecutive plan batches. Learned-score training uses either demonstrations or reserved successful policy rollouts.

Comparisons use the evaluated implementations and the fixed-horizon convention (E); they have not been repeated under (S). Score comparisons under a common RoLCA rule (Table 13) are distinct from decision-rule and channel-addition comparisons (Table 14).

*(a) Macro performance*

| Score | Training | FPR | Recall [95% interval] | AUROC |
|---|---|---|---|---|
| $S_{\text{exec}}$ | Demo | .082 | .402 [.37,.44] | .75 |
| $S_{\text{inv}}$ (self) | Demo | .088 | .332 [.30,.37] | .67 |
| $S_{\text{inv}}$ (cross) | Demo | .081 | .374 [.34,.41] | .67 |
| logpZO | Demo | .084 | .400 [.37,.43] | .74 |
| RND-A | Demo | .087 | .224 [.19,.25] | .60 |
| RND-OE | Demo | .083 | .206 [.18,.23] | .58 |
| STAC | – | .090 | .273 [.24,.31] | .57 |
| ACE | – | .077 | .162 [.13,.19] | .60 |
| logpZO | Rollout | .085 | .499 [.46,.53] | .81 |
| RND-A | Rollout | .079 | .436 [.40,.47] | .81 |
| RND-OE | Rollout | .092 | .532 [.50,.56] | .81 |

*(b) Cellwise recall*

| Score | Training | lift/B | lift/D | can/B | can/D | square/B | square/D |
|---|---|---|---|---|---|---|---|
| $S_{\text{exec}}$ | Demo | .23 | .50 | .29 | .75 | .29 | .35 |
| $S_{\text{inv}}$ (self) | Demo | .27 | .54 | .42 | .34 | .19 | .24 |
| $S_{\text{inv}}$ (cross) | Demo | .30 | .56 | .37 | .38 | .47 | .16 |
| logpZO | Demo | .30 | .57 | .39 | .82 | .12 | .20 |
| RND-A | Demo | .09 | .06 | .36 | .57 | .20 | .08 |
| RND-OE | Demo | .03 | .09 | .38 | .56 | .08 | .09 |
| STAC | – | .20 | .53 | .25 | .32 | .17 | .17 |
| ACE | – | .20 | .14 | .18 | .18 | .18 | .09 |
| logpZO | Rollout | .26 | .15 | .58 | .68 | .71 | .60 |
| RND-A | Rollout | .15 | .21 | .71 | .67 | .59 | .28 |
| RND-OE | Rollout | .15 | .59 | .72 | .76 | .89 | .09 |

Table 13: Published scores under running-average RoLCA calibration and (E), at $\alpha = 0.10$. Demo/Rollout denote score-model training on demonstrations/reserved successful policy rollouts; "–" denotes no learned score model. AUROC ranks $\max_j \log A_j$ over the full horizon, one value per rollout, with evaluation failures positive and successes negative; ties receive half credit. Macro metrics average cellwise medians over 50 splits. Intervals condition on the success pool and fixed splits; B/D denote feature backbones.

**Thresholds and decision rules.** Let $q_{0.9}$ denote the empirical quantile with linear interpolation (`numpy.quantile`). The ported rules below use this quantile; RoLCA uses the finite-sample order statistic of Proposition 1. Thresholds use only the split's calibration successes, and alarms require strict exceedance. For raw scores $s_j$, FIPER-MQ thresholds $s_j$ at $q_{0.9}\{\max_j s_j\}$ over calibration rollouts, without windowing. The per-step band instead uses $q_{0.9}\{s_j\}$ separately at each record and provides no simultaneous guarantee. Missing test scores cannot trigger either rule; missing per-step thresholds retain the last finite value.

FIPER AND/OR use demonstration-trained RND-OE and ACE with window length one, both choices of our evaluation. Each channel uses its own $q_{0.9}$ of calibration-rollout maxima at the full $\alpha = 0.10$. At each record, the minimum (AND) or maximum (OR) of the two threshold-normalized scores must exceed one; alarms are not retained across records for AND. Missing scores contribute zero to windowed sums. All evaluated component thresholds are positive. This OR rule uses no budget splitting and differs from RoLCA's matched OR.

**Functional band.** FD-band partitions the $m$ calibration sequences in split order: the longest finite-prefix sequence and the first $\lfloor (m-1)/2 \rfloor$ remaining sequences form set 1, of size $N_1 = 1 + \lfloor (m-1)/2 \rfloor$; the rest form set 2. For fitting, missing tails are removed and interior gaps retain the preceding score (zero at the first record). Let $\mu_j$ be the set-1 mean at record $j$ and $L$ the longest sequence length. Retain set-1 sequences whose maximum absolute residual from $\mu_j$ does not exceed the $0.9$-quantile of these maxima; retain all sequences if $(N_1+1)0.9 > N_1$. The modulation $a_j$ is the largest absolute residual at record $j$ among retained sequences, with fallback $1/\max\{j-1, 1\}$ if none reaches that record. Values below $10^{-7}$ are replaced by $1/L$. The threshold $h$ is the $q_{0.9}$ quantile of the set-2 maxima $\max_j\{(s_j - \mu_j)/a_j\}$, giving the upper band $\mu_j + h a_j$, extended by its last finite value. This upward-deviation version reverses the sign of the set-2 residuals in the ported implementation; both signs were evaluated. No finite-sample quantile correction is applied.

**Sentinel component and channel scope.** STAC uses the biased RBF-MMD$^2$ estimate with a median-distance bandwidth between 128 predicted end-effector segments from consecutive chunks, aligned to the same four environment steps. Candidate zero is executed. The evaluated estimator is nonnegative, so its cumulative sum is monotone; the cumulative baseline thresholds this sum at the $q_{0.9}$ of final calibration sums, with missing increments set to zero. Sentinel's VLM progress module is excluded. The ten-channel fusion combines execution, self-scored inversion, STAC, ACE, and each of logpZO, RND-A, and RND-OE in both demonstration- and rollout-trained variants. Each channel contributes its running average to a joint mean or maximum, calibrated once at $\alpha$.

### E.2 Scores, decision rules, and training access

Running-average execution evidence reaches recall $0.402$ $[0.367, 0.441]$, and execution–inversion fusion reaches $0.459$. The two-channel mean exceeds OR on the same channels' instantaneous standardized scores by $0.089$ when self-scored and $0.061$ when cross-scored; these comparisons change both temporal aggregation and the decision rule. Replacing the mean by a maximum changes macro recall by $+0.005$ and $-0.010$, respectively. These contrasts use $\widehat{D}$, not $\widehat{\Delta}$.

Demonstration-trained logpZO differs from execution by $-0.002$ $[-0.040, +0.037]$ in macro recall. At fixed execution score, RoLCA differs from the corrected FAIL-Detect band by $-0.019$ $[-0.045, +0.015]$, with FPRs $0.082$ and $0.112$. Neither contrast establishes a recall advantage. For STAC, RoLCA and the evaluated Sentinel cumulative rule have FPRs $0.090$ and $0.105$. The evaluated FIPER per-step band has macro rollout FPR $0.92$–$1.00$ across scores (cell values from $0.79$). The evaluated OR has macro FPR $0.204$, with cell values $0.164$–$0.250$. Comparisons must account for these different realized false-alarm rates.

Training RND-OE on successful policy rollouts raises recall to about $0.53$, whereas adding ACE to execution–inversion fusion lowers it by $0.135$. These findings concern the evaluated training access and channel combinations and do not establish general superiority over published monitors.

## F Ablations and sensitivity analyses

### F.1 Temporal profiles and auxiliary scores

Figure 7 contrasts repeated execution-score elevations in can/B with a brief early spike in lift/D. These profiles illustrate heterogeneous temporal structure; averages across rollouts do not determine the power of a temporal statistic. The outer-calibrated mean of products and the uncalibrated product/Ville diagnostic are defined separately in Appendix B.4.

The AUROCs here rank $\max_j \zeta_j$, the rollout maximum of a standardized instantaneous score under (E), with evaluation failures positive and successes negative. They use one value per rollout and

*(a) Fixed score: RoLCA (a) versus rule b*

| Score | Rule $b$ | FPR $(a/b)$ | $\widehat{D}_{a,b}$ [95% interval] |
|---|---|---|---|
| $S_{\text{exec}}$ | FIPER-MQ | 0.082 / 0.104 | $+0.248$ $[+0.207, +0.290]$ |
| $S_{\text{exec}}$ | FD-band | 0.082 / 0.112 | $-0.019$ $[-0.045, +0.015]$ |
| $S_{\text{inv}}$ (self) | FIPER-MQ | 0.088 / 0.103 | $+0.057$ $[+0.021, +0.092]$ |
| $S_{\text{inv}}$ (self) | FD-band | 0.088 / 0.117 | $-0.024$ $[-0.053, +0.009]$ |
| logpZO (Demo) | FIPER-MQ | 0.084 / 0.111 | $+0.308$ $[+0.269, +0.346]$ |
| logpZO (Demo) | FD-band | 0.084 / 0.106 | $-0.013$ $[-0.042, +0.020]$ |
| logpZO (Rollout) | FIPER-MQ | 0.085 / 0.113 | $+0.134$ $[+0.097, +0.169]$ |
| logpZO (Rollout) | FD-band | 0.085 / 0.110 | $+0.060$ $[+0.033, +0.089]$ |
| STAC | Sentinel cumulative | 0.090 / 0.105 | $+0.093$ $[+0.058, +0.128]$ |

*(b) Self-scored exec+inv fusion (a) versus published rule b*

| Rule $b$ | FPR $(a/b)$ | $\widehat{D}_{a,b}$ [95% interval] |
|---|---|---|
| FD logpZO + band (Demo) | 0.090 / 0.106 | $+0.047$ $[+0.005, +0.091]$ |
| FD logpZO + band (Rollout) | 0.090 / 0.110 | $+0.020$ $[-0.018, +0.061]$ |
| FIPER AND (RND-OE, ACE) | 0.090 / 0.000 | $+0.459$ $[+0.423, +0.496]$ |
| FIPER OR (evaluated) | 0.090 / 0.204 | $+0.266$ $[+0.225, +0.310]$ |
| Sentinel STAC, cumulative | 0.090 / 0.105 | $+0.280$ $[+0.236, +0.322]$ |

*(c) Added channels (a) versus self-scored exec+inv fusion (b)*

| Added score or fusion | FPR $(a/b)$ | $\widehat{D}_{a,b}$ [95% interval] |
|---|---|---|
| logpZO (Demo) | 0.085 / 0.090 | $+0.037$ $[+0.009, +0.065]$ |
| RND-OE (Demo) | 0.080 / 0.090 | $-0.016$ $[-0.039, +0.008]$ |
| STAC | 0.091 / 0.090 | $+0.016$ $[-0.002, +0.038]$ |
| ACE | 0.081 / 0.090 | $-0.135$ $[-0.170, -0.101]$ |
| RND-OE (Rollout) | 0.084 / 0.090 | $+0.106$ $[+0.077, +0.138]$ |
| All channels (ten), mean | 0.083 / 0.090 | $+0.069$ $[+0.031, +0.103]$ |
| All channels (ten), joint maximum | 0.076 / 0.090 | $-0.112$ $[-0.152, -0.067]$ |

Table 14: Monitoring-rule and channel comparisons under (E), at $\alpha = 0.10$. Contrasts are differences of macro recalls, $\widehat{D}$, with paired failure-bootstrap intervals. FIPER-MQ is the raw-score max-quantile rule; FD-band uses upward deviations. FIPER AND/OR use demonstration-trained RND-OE and ACE at the same record, window one, and full budget per channel. Demo/Rollout indicate score-training data. Panel (c) uses mean fusion unless marked otherwise; "all channels" contains ten scores. Realized FPRs differ across decision rules.

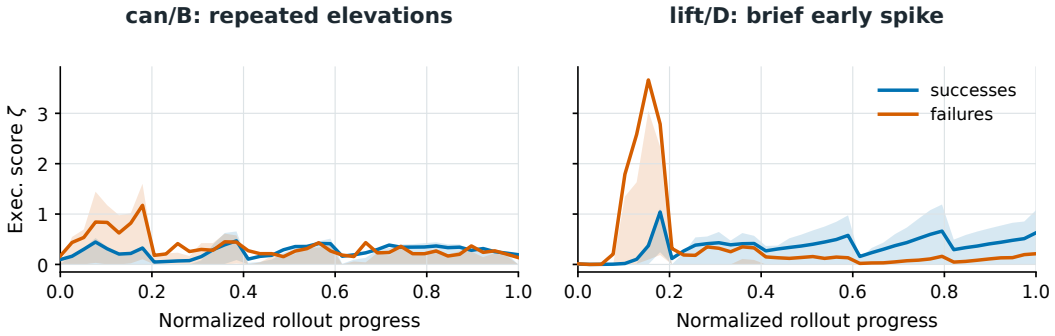



Figure 7: Temporal profiles of standardized execution-consistency scores under (E). Curves show success/failure means and shading shows interquartile ranges. Can/B exhibits repeated elevations; lift/D has a brief early spike. These descriptive profiles do not establish a ranking of detection methods.

cellwise medians over 50 splits; macro values average the six cells. This differs from Table 13's running-average statistic. Cross-scored round-1 plan novelty has cell AUROCs $0.437$–$0.548$. The self-scored variant has AUROC $0.500$ in three cells and $0.439$, $0.403$, and $0.684$ in the others, with the highest value in can/D. Task-trained, ImageNet, and random features give macro prediction-score AUROCs $0.577$, $0.579$, and $0.573$ (cell values $0.300$–$0.681$). For each feature choice, three-channel macro recall remains below that of execution–inversion fusion. An apparent 12.5-percentage-point advantage on weaker actors shrinks to 1–1.5 points after retraining, with a severe cell-level loss. These exploratory diagnostics do not establish a universal aggregation or feature preference.

### F.2 Fusion operators

| | Discovery | | First confirmation | |
|---|---|---|---|---|
| Monitor | FPR | Recall | FPR | Recall |
| OR on three standardized scores | .070 | .301 | .063 | .266 |
| Joint maximum of three $\zeta$ scores | .089 | .380 | .089 | .355 |
| Mean of three products | .093 | .424 | .092 | .370 |
| $U_{\max}$ conformal | .092 | .392 | .090 | .348 |
| Running-average evidence | .093 | .425 | .093 | .396 |
| Mean of three running averages | .096 | .420 | .085 | .351 |
| Maximum of three running averages | .096 | .397 | .088 | .353 |
| exec+inv mean (self) | .093 | .444 | .092 | .385 |
| exec+inv maximum of $\zeta$ (self) | .092 | .403 | .092 | .373 |
| exec+inv maximum of $A$ (self) | .094 | .428 | .091 | .393 |
| exec+inv mean (cross) | .094 | .470 | .090 | .412 |
| exec+inv maximum of $\zeta$ (cross) | .092 | .403 | .096 | .367 |
| exec+inv maximum of $A$ (cross) | .094 | .448 | .090 | .404 |

Table 15: Fusion ablations under (E), at $\alpha = 0.10$. Three-channel rows use execution and the auxiliary token scores, not RND-OE. Maxima of standardized scores and of running averages are distinct constructions.

With auxiliary standardized scores fixed, joint maximum gains $0.073/0.079$ over OR on discovery/first confirmation. The mean of products gains $0.121/0.100$, but also changes transformation and temporal aggregation. With running-average inputs fixed, replacing mean by maximum changes paired recall by $-0.021/+0.005$ for self-scored execution–inversion and $-0.024/-0.010$ for cross-scoring. Only comparisons with identical inputs and temporal statistics isolate the fusion operator. All these results use (E).

### F.3 Finite calibration and alternative budgets

For a fixed statistic, Table 16 compares conformal calibration with an oracle threshold fitted to evaluation successes at the empirical FPR budget. A second, pooled-success reference uses the conformal quantile of all successes. Both references reuse evaluation data and are nondeployable comparisons.

The paired oracle-minus-conformal gap is $0.006$–$0.016$ on discovery and $0.015$–$0.025$ on confirmation. Pooling successes brings the remaining oracle gap within $\pm 0.007$ except for one row. Larger cellwise losses occur with 42–46 calibration successes. Recall need not increase monotonically with calibration size: for square/B it is $0.536$ at $m = 20$ and $0.491$ at $m = 45$, with rank-based FPR bounds $2/21 \approx 0.095$ and $4/46 \approx 0.087$. Rounding loss in the bound vanishes when $(m + 1)\alpha$ is integral; ties and sampling variation can still affect realized performance.

At $\alpha = 0.05$, joint standardized-score maximum matches or exceeds the mean of products, while the latter leads at $0.20$ (Table 17). Avoiding the automatic infinite threshold requires

$$k \le m \quad \Longleftrightarrow \quad \alpha \ge \frac{1}{m+1} \quad \Longleftrightarrow \quad m \ge \left\lceil \frac{1}{\alpha} \right\rceil - 1.$$

Thus $\alpha = 0.02$ requires at least 49 calibration successes; meeting this condition does not ensure useful power.

*(a) Discovery*

| Monitor | Conformal | Oracle | $\widehat{\Delta}$ | Unused FPR |
|---|---|---|---|---|
| $U_{\max}$ conformal | .392 | .406 | +.014 | .008 |
| CUSUM conformal | .410 | .414 | +.006 | .005 |
| Running-average evidence | .425 | .446 | +.016 | .007 |
| Mean of three products | .424 | .435 | +.009 | .007 |
| exec+inv mean (self) | .444 | .458 | +.012 | .007 |
| exec+inv mean (cross) | .470 | .487 | +.015 | .006 |

*(b) First confirmation*

| Monitor | Conformal | Oracle | $\widehat{\Delta}$ | Unused FPR |
|---|---|---|---|---|
| $U_{\max}$ conformal | .348 | .359 | +.015 | .010 |
| CUSUM conformal | .364 | .376 | +.015 | .011 |
| Running-average evidence | .396 | .418 | +.021 | .007 |
| Mean of three products | .370 | .398 | +.023 | .008 |
| exec+inv mean (self) | .385 | .419 | +.025 | .008 |
| exec+inv mean (cross) | .412 | .436 | +.023 | .010 |

Table 16: Finite-calibration comparison at $\alpha = 0.10$. Conformal and oracle columns are macro recalls; $\widehat{\Delta}$ is the paired oracle-minus-conformal gain and need not equal their difference. Unused FPR is $\alpha$ minus macro FPR. Oracle thresholds reuse evaluation successes and are not deployable.

*(a)* $\alpha = 0.05$

| Monitor | Disc. FPR | Disc. recall | Conf. FPR | Conf. recall |
|---|---|---|---|---|
| $U_{\max}$ conformal | .039 | .293 | .040 | .270 |
| CUSUM conformal | .040 | .332 | .041 | .265 |
| Running-average evidence | .044 | .336 | .046 | .298 |
| Mean of three products | .043 | .277 | .044 | .241 |
| Joint maximum of three $\zeta$ scores | .040 | .290 | .041 | .267 |
| OR budget splitting | .026 | .176 | .024 | .169 |
| exec+inv mean (self) | .046 | .334 | .041 | .322 |
| exec+inv mean (cross) | .044 | .365 | .042 | .333 |

*(b)* $\alpha = 0.20$

| Monitor | Disc. FPR | Disc. recall | Conf. FPR | Conf. recall |
|---|---|---|---|---|
| $U_{\max}$ conformal | .196 | .499 | .196 | .455 |
| CUSUM conformal | .192 | .530 | .192 | .475 |
| Running-average evidence | .191 | .561 | .197 | .549 |
| Mean of three products | .192 | .582 | .193 | .588 |
| Joint maximum of three $\zeta$ scores | .191 | .493 | .194 | .459 |
| OR budget splitting | .139 | .416 | .140 | .397 |
| exec+inv mean (self) | .188 | .569 | .191 | .522 |
| exec+inv mean (cross) | .194 | .594 | .195 | .545 |

Table 17: Alternative nominal FPR budgets under (E), using the same discovery and first-confirmation populations and splits. Three-channel product, joint-maximum, and OR rows use the auxiliary score set.

### F.4 CALIBRATION-CONDITIONAL THRESHOLDS

| Monitor | Marginal | Conditional $\delta = 0.10$ | Conditional $\delta = 0.05$ |
|---|---|---|---|
| Running-average $S_{\text{exec}}$ | 0.403 | 0.296 | 0.288 |
| exec+inv self-monitor | 0.398 | 0.326 | 0.315 |
| exec+inv+RND-OE | 0.523 | 0.441 | 0.428 |

Table 18: Macro recall under marginal and calibration-conditional thresholds under (E), at target FPR $\alpha = 0.10$.

Applying Eq. (10) under (E) reduces macro FPR to $0.052$–$0.062$ and costs approximately $0.07$–$0.12$ recall relative to marginal calibration. Joint mean retains the highest recall among these three monitors. This analysis was performed after initial confirmation outcomes were inspected and is excluded from prespecified gate outcomes.

## G EXPLORATORY EVALUATION ON EXTERNAL ACTORS

Reference BC-RNN and Diffusion Policy actors are evaluated on ToolHang and TwoArmTransport with raw state observations and $T_{\max} = 700$. Success rates are $0.30/0.40$ on ToolHang and $0.20/0.50$ on TwoArmTransport (BC-RNN/Diffusion Policy). A demonstration-trained monitoring bundle scores uninterrupted rollouts, with cross-scored inversion. The 1,510 rollouts comprise 473 successes and 1,037 failures and are excluded from the primary study.

| | FPR / recall | | | | |
|---|---|---|---|---|---|
| Monitor | Macro | TH/RNN | TH/DP | TR/RNN | TR/DP |
| $U_{\max}$ conformal | .068 / .348 | .088 / .07 | .065 / .25 | .038 / .34 | .081 / .73 |
| CUSUM conformal | .090 / .541 | .088 / .06 | .097 / .31 | .077 / .90 | .097 / .89 |
| Running-average evidence | .072 / .526 | .088 / .15 | .065 / .15 | .038 / .89 | .097 / .92 |
| Mean of three products | .082 / .549 | .088 / .12 | .097 / .25 | .077 / .89 | .065 / .93 |
| OR budget splitting | .051 / .221 | .059 / .11 | .065 / .28 | .000 / .00 | .081 / .50 |
| exec+inv mean (cross) | .080 / .551 | .088 / .17 | .097 / .17 | .038 / .98 | .097 / .89 |
| Product with Ville threshold | .575 / .892 | .824 / .82 | .839 / .85 | .346 / .95 | .290 / .95 |

Table 19: Exploratory external-actor evaluation at $\alpha = 0.10$. Each entry reports FPR / recall. TH/TR denote ToolHang/TwoArmTransport; RNN/DP denote actor families. Cell entries are medians over 50 splits; macro entries average four cell medians. Only product/Ville omits outer calibration. Each cell has 26–34 evaluation successes per split.

Calibrated FPR summaries fall within the reported reference bands, whereas the uncalibrated product/Ville diagnostic reaches macro FPR $0.575$. With only 26–34 evaluation successes per cell and split, these bands are wide and do not give confidence coverage for split medians. The mean of products substantially exceeds OR on TwoArmTransport, while their cellwise recalls are close on ToolHang (Table 19). Cross-scored execution–inversion improves macro recall by approximately $0.03$ over running-average execution. Different actors, tasks, and scoring distributions prevent direct comparison of absolute recall with the main study.

This extension is exploratory: a reset defect was corrected and rollouts recollected after preliminary monitor results were inspected. Checkpoints violating intended holdout isolation were excluded before collection. No primary confirmatory claim depends on these results.