CALIBRATION IN THE LOOP: TRAINING VLA FAILURE MONITORS FOR EARLY DETECTION AT A CONDITION-WISE FALSE-ALARM TOLERANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

A failure monitor for a vision-language-action (VLA) policy is a learned score and a threshold that carries the false-alarm guarantee. Under deployment shift the guarantee must hold per operating condition, and the per-condition rule places the deployed threshold at the heaviest successful tail, where a score trained by cross-entropy alarms late, after a retry can no longer succeed. We frame monitor training under deployment shift as a constrained optimization problem: maximize early detection of failures subject to a false-alarm tolerance under every deployment condition, and we solve it with calibration in the loop (CIL), a training framework that computes the calibration rule that will set the deployed threshold inside the objective. The rule is an order statistic of reference successes per condition and the maximum over conditions, exact and differentiable almost everywhere, so the score is trained under the threshold it will be deployed with while dual ascent enforces one constraint per condition; the loss, the constraints and the rule can be exchanged. One trained score serves a marginal and a high-confidence deployment mode whose worst-condition bound holds with probability $1 - \delta$. On held-out banks under the LIBERO-Plus catalog, CIL raises early detection within the tolerance from 39.3% to 59.2% for $\pi_{0.5}$ on LIBERO-Spatial and to 85.5% for OpenVLA-OFT when applied to a SAFE network, against detection baselines thresholded by the same rule; in a reset-and-retry loop the earlier alarms become completed tasks as the reset cost grows; and on a real xArm6 the monitor raises early detection from 34.9% to 44.4% within the tolerance and completes more tasks under recovery than the CoRe and TIDE monitors. Code and data recipes are available at `https://anonymous.4open.science/r/CIL-4340/README.md`.

1 INTRODUCTION

A vision-language-action (VLA) policy is deployed frozen, and a failure monitor decides call by call whether to interrupt execution so that a recovery procedure can run (Gu et al., 2025; Sinha et al., 2022; Zhang et al., 2026b). The monitor is a learned score on the policy's internal state plus a threshold; a split-conformal quantile of the score on successful rollouts bounds the false-alarm rate when calibration and deployment are exchangeable (Papadopoulos et al., 2002; Angelopoulos & Bates, 2023). Both errors of a monitor cost task completions. A false alarm resets an execution that would have succeeded and charges the reset to the execution budget; a late alarm leaves too little budget for a retry. In our closed loop, a calibrated monitor allowed one reset lifts completion under a transient disturbance from 0.708 to 0.886 when a reset costs 20 steps (Section 5.4).

Deployment shifts in cameras, lighting, instructions, initial states or sensor noise invalidate a threshold calibrated on source rollouts alone (Tibshirani et al., 2019; Sinha et al., 2022). Pooling simulated rollouts of the shifted conditions into the calibration set controls only the average false-alarm rate: the monitor stays quiet on easy conditions and alarms freely on the hardest one. Calibrating each condition on its own simulated successes and deploying the largest threshold (CondMax, a maximum rule in the spirit of Yang & Jin (2026)) controls each represented condition, which is the guarantee a deployment needs, but the deployed threshold then sits at the heaviest successful tail. A score trained by cross-entropy separates failures from successes on average; nothing in its training lifts failures above that tail while a retry can still succeed. Figure 1 shows the two failure modes on one task: the

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

calibration-only monitor interrupts a run that would have succeeded and, on a run that fails, alarms after the last point at which a retry could complete the task. In our closed loop it alarms a median of 153 steps after the disturbance, and its completion falls from 0.886 to 0.756 once a reset costs 140 steps (Table 13).

We treat this as a constrained optimization problem: maximize early detection of failures subject to a false-alarm tolerance under every deployment condition. The difficulty is that the threshold in the constraint is not a parameter but the output of the calibration rule. Our observation is that the rule is an order statistic of each condition's reference successes followed by a maximum over conditions, which is exact and differentiable almost everywhere, so the rule can be computed inside the loss. Calibration in the loop (CIL) splits a simulated library into reference and query folds by initial-state regime, sets the threshold on the reference successes, maximizes early detection of the query failures within the actionable horizon under that threshold, and holds each condition's false-alarm surrogate at the tolerance by dual ascent (Figure 2b). Learning the score and recalibrating its threshold therefore happen together, and each improves the other. The framework is not tied to one loss, one constraint set or one rule, and it is not tied to one score family: it starts from SAFE's linear head on the policy's internal features (Gu et al., 2025) or from SAFE's and ActProbe's networks and changes only their training objective.

The trained score is frozen and recalibrated on an independent bank in a marginal mode, whose split-conformal ranks bound each represented condition's expected false-alarm rate, or a high-confidence mode, whose binomial ranks bound the worst-condition rate with probability at least $1 - \delta$ (Figure 2c). Both guarantees are inherited from the calibration step and hold regardless of how the score was trained; CIL changes what the score does at the threshold those ranks select.

We evaluate a frozen $\pi_{0.5}$ and a frozen OpenVLA-OFT on LIBERO-Spatial and LIBERO-10 under the seven-condition LIBERO-Plus catalog, comparing against detection baselines (kNN distance, SAFE-MLP, SAFE-LSTM, ActProbe) thresholded by the same rule on the same calibration successes; in a 1,272-episode closed-loop reset-and-retry protocol against the recovery monitors CoRe and TIDE; and on a UFACTORY xArm6 under three conditions, for detection and for recovery (Section 5).

Our contributions are:

- We frame failure monitoring under deployment shift as a constrained optimization problem and propose calibration-in-the-loop (CIL), a training framework in which the calibration rule that will set the deployed threshold is computed inside the objective. The framework is flexible: the detection loss and the constraints can be exchanged for other objectives and other calibration rules (Section 4).
- We control the false-alarm rate under each deployment condition (conditional coverage) by placing one constraint per possible condition, so that the tolerance holds where the score is hardest to calibrate rather than on average (Section 3).
- We optimize early detection in the loop, so that learning the score and recalibrating its threshold happen together and each improves the other, with the same trained score serving a marginal and a high-confidence deployment mode.
- We validate the framework on two policies and two LIBERO suites in simulation, in a closed loop with reset-and-retry recovery, and on a real robot, against detection baselines under the same calibration rule and recovery baselines under the same reset budget (Section 5).

2 Related work

Calibration-aware training. Stutz et al. (2022) differentiate through a simulated conformal calibration to keep prediction sets small at a target coverage, and Einbinder et al. (2022) train with conformal objectives that improve conditional coverage. We apply the same device to the condition-wise maximum of Equation (1) on held-out reference folds and optimise early detection under per-condition false-alarm surrogates.

Conformal calibration under shift. Split-conformal prediction (Papadopoulos et al., 2002; Lei et al., 2018), its group-conditional variants (Vovk, 2013) and its extensions to shift and PAC-style calibration (Tibshirani et al., 2019; Podkopaev & Ramdas, 2021; Wang et al., 2025; Guan & Tibshirani,

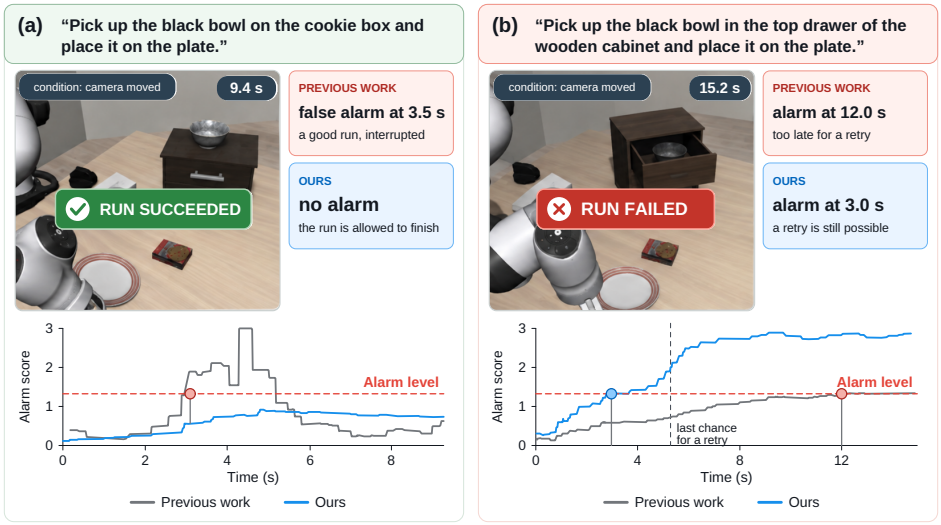


Figure 1: Two $\pi_{0.5}$ runs under a moved camera, monitored by a prior failure monitor (gray) and by ours (blue), both at the same false-alarm tolerance. Left: a run that succeeds; the previous monitor crosses its alarm level at 3.5 s and interrupts it, ours stays silent. Right: a run that fails; the previous monitor alarms at 12 s, after the last point at which a retry could still complete the task, ours at 3 s.

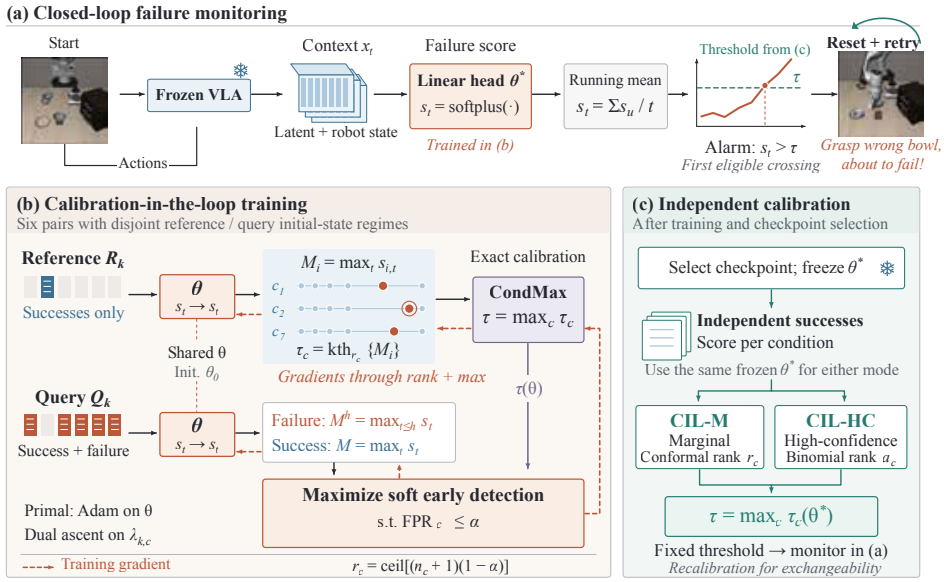


Figure 2: Training the score under its calibration rule moves alarms earlier at the same per-condition tolerance, and the earlier alarms become completed tasks once a reset costs time. (a) The monitor alarms when the running mean of a learned per-call surprise crosses the deployed threshold. (b) CIL computes that threshold inside the loss from reference-fold successes and maximises early detection of query-fold failures under it, subject to per-condition constraints. (c) The frozen score is recalibrated on independent data as CIL-M or CIL-HC.

2022; Si et al., 2023; Bairaktari et al., 2025) calibrate a fixed score. Multi-Distribution Robust Conformal Prediction (Yang & Jin, 2026) takes a maximum over per-distribution p-values for coverage uniform across sources and learns scores that keep aggregated sets small. CondMax is that maximum over per-condition false-alarm thresholds, and CIL trains the score through its finite-sample calibration ranks.

Neyman-Pearson and constrained learning. Neyman-Pearson classification (Scott & Nowak, 2005), binomial-rank thresholds (Tong et al., 2018) and training-conditional conformal validity (Vovk, 2013) underlie Proposition 2. Training a score at a fixed false-positive rate under Lagrangian constraints follows Cotter et al. (2019) and Chamon & Ribeiro (2020); the dual ascent of Equation (6) is that construction with one constraint per condition.

Robot failure detection and recovery. SAFE (Gu et al., 2025) probes a VLA policy's internal features and calibrates the score conformally. SAFECAST (Rajaprakash et al., 2026) trains and calibrates on contrast sets against shift. Others use task progress, multimodal reasoning or fleet supervision (Seligmann et al., 2026; Römer et al., 2025; Liu et al., 2024; Farid et al., 2022; Luo et al., 2024; Sinha et al., 2022). ActProbe (Huang et al., 2026) scores action-chunk consistency and magnitude under a pooled split-conformal threshold; Foresight (Zhang et al., 2026a) scores action-conditioned world-model latents under a functional conformal band. We keep SAFE's score family and change the training objective, then test ActProbe under the same rule and with CIL (Section 5.2) and Foresight's band on our scores (Section E). CoRe (Zhang et al., 2026b) and rewind-based imitation (Zheng et al., 2026) pair a monitor with their own intervention; we hold the recovery fixed and compare monitors against the reset cost (Section 5.4).

3 Problem setting and calibration rule

Monitor. A frozen policy is queried at calls $t = 1, 2, \ldots$, and the monitor observes at each call a context vector $x_t \in \mathbb{R}^d$ built from the policy's internal state and the robot state (Section 5.1). A linear head $\theta \in \mathbb{R}^{d+1}$ gives the per-call surprise $s_t(\theta) = \mathrm{softplus}(\theta^\top [x_t; 1]) = -\log \sigma(-\theta^\top [x_t; 1])$, the negative log-probability of success under a logistic model, and $\bar{s}_t$ is its running mean. The episode statistic is $M(\theta) = \max_{t \in T} \bar{s}_t(\theta)$ over the eligible calls T, and the monitor alarms at the first eligible call with $\bar{s}_t > \tau$. Offline, T is all recorded calls. The early statistic $M^h(\theta) = \max_{t \le h} \bar{s}_t(\theta)$ keeps only the first h calls; the actionable horizon h, the shortest episode recorded for the task (13 to 23 calls on LIBERO-Spatial), is our proxy for the window in which a retry can still succeed. Online, eligibility starts at a fixed call and one reset is allowed (Section 5.4).

Conditions and tolerance. Deployment is a finite set of operating conditions $\mathcal{C}$ (seven LIBERO-Plus perturbation families). Condition c has false-alarm rate $\mathrm{FPR}_c(\tau) = \mathbb{P}(M > \tau \mid \text{success}, c)$ and early detection rate $\mathrm{TPR}_c^h(\tau) = \mathbb{P}(M^h > \tau \mid \text{failure}, c)$. We want a pair (θ, τ) with high early detection subject to $\mathrm{FPR}_c(\tau) \le \alpha$ for each represented c.

Condition-wise calibration (CondMax). With $\mathcal{S}_c = \{M_i\}$ the statistics of the n_c successful calibration rollouts of condition c and $M_{(r)}$ their r-th smallest, CondMax sets

$$\tau_c = M_{(r_c)}, \quad r_c = \lceil (n_c + 1)(1 - \alpha) \rceil, \qquad \tau = \max_{c \in \mathcal{C}} \tau_c, \tag{1}$$

with $\tau_c = +\infty$ when $r_c > n_c$. Calibration uses each condition's successful rollouts and condition labels, and the score is frozen before it sees them. Two guarantees follow.

Proposition 1 (marginal mode)**.** *If, for each condition c, the calibration statistics $\mathcal{S}_c$ and the statistic of a new successful deployment execution of condition c are exchangeable, then $\mathbb{E}[\mathrm{FPR}_c(\tau)] \le \alpha$ for each $c \in \mathcal{C}$, the expectation taken over the calibration draw.*

Intuitively, the deployed threshold is at least each condition's own split-conformal threshold, so no represented condition's expected false-alarm rate exceeds α.

Proposition 2 (high-confidence mode)**.** *Suppose for each c that the calibration statistics $\mathcal{S}_c$ are i.i.d. from the law of the deployment success statistic of condition c. Replace r_c in Equation* (1) *by*

$$a_c = \min\{a \le n_c : \mathbb{P}[\mathrm{Bin}(n_c, 1 - \alpha) \ge a] \le \delta\}, \qquad \tau_c = M_{(a_c)}, \tag{2}$$

with $\tau_c = +\infty$ when no such a exists, and deploy $\tau = \max_c \tau_c$. Then $\mathbb{P}(\max_{c \in \mathcal{C}} \mathrm{FPR}_c(\tau) > \alpha) \le \delta$.

Intuitively, only the condition with the largest population $(1 - \alpha)$-quantile matters: the bad event needs its threshold below that quantile, which the binomial rank allows with probability at most δ, so no union bound over conditions is needed (proofs in Section A). The higher rank raises the threshold and lowers detection.

The deployed threshold is thus the order statistic of one condition's successful tail: on our LIBERO-Spatial calibration stream the robot condition sets the CIL-M threshold of the paper configuration (Section 4) at 1.338, and without it the camera condition, the next largest, would set it at 1.016 (Table 9). Cross-entropy training never evaluates the score at that tail or within h calls; Section 4 puts the rule into the loss.

4 CALIBRATION-IN-THE-LOOP TRAINING

Regime-paired folds. Training uses simulated rollouts of a library of catalog variants, each labelled with its condition c and its initial-state regime. We form $K = 6$ leave-one-initial-state-regime-out pairs $(\mathcal{R}_k, \mathcal{Q}_k)$, one per library initial state: the reference fold $\mathcal{R}_k$ holds the successes of the held-out regime, and the query fold $\mathcal{Q}_k$ holds the successes and failures of the other regimes. We compute the threshold on $\mathcal{R}_k$ alone, so each query episode meets a threshold from a calibration it did not enter, as at deployment. The reference folds are library episodes that shape θ through gradients; the deployment calibration bank of Section 3 never enters training or selection.

Exact threshold inside the loss. For a pair k and parameters θ, the per-condition threshold is the order statistic

$$\tau_c(\theta) = \mathrm{kthvalue}_{r_c}\{M_i(\theta) : i \in \mathcal{R}_k^{\mathrm{succ},c}\}, \qquad \tau(\theta) = \max_{c \in \mathcal{C}} \tau_c(\theta), \tag{3}$$

with $\mathcal{R}_k^{\mathrm{succ},c}$ the successes of condition c in the reference fold, $r_c = \lceil (n_c + 1)(1 - \alpha) \rceil$ and $n_c = |\mathcal{R}_k^{\mathrm{succ},c}|$. This is the rule of Equation (1) and is differentiable almost everywhere. The gradient of $\tau(\theta)$ flows through the selected order statistic of the selected condition into the statistics $M_i(\theta)$, and through the maximum and the running mean into the per-call surprises.

Objective. With $\sigma(u/T)$ a sigmoid of temperature T, the soft early detection of condition c' and the soft false-alarm rate of condition c on the query fold of pair k are

$$\widetilde{\mathrm{TPR}}^{\,h}_{c'}(\theta) = \frac{1}{|\mathcal{Q}_k^{\mathrm{fail},c'}|} \sum_{i \in \mathcal{Q}_k^{\mathrm{fail},c'}} \sigma\Big(\tfrac{M_i^h(\theta) - \tau(\theta)}{T}\Big), \quad \widetilde{\mathrm{FPR}}_c(\theta) = \frac{1}{|\mathcal{Q}_k^{\mathrm{succ},c}|} \sum_{i \in \mathcal{Q}_k^{\mathrm{succ},c}} \sigma\Big(\tfrac{M_i(\theta) - \tau(\theta)}{T}\Big), \tag{4}$$

where $\mathcal{Q}_k^{\mathrm{fail},c'}$ and $\mathcal{Q}_k^{\mathrm{succ},c}$ are the query failures of condition c' and the query successes of condition c. With one dual vector λ_k per pair, the loss on pair k is

$$L_k(\theta; \lambda_k) = -\sum_{c' \in \mathcal{C}_{\mathrm{det}}} \nu_{c'} \widetilde{\mathrm{TPR}}^{\,h}_{c'}(\theta) + \sum_{c \in \mathcal{C}} \lambda_{k,c} \Big(\widetilde{\mathrm{FPR}}_c(\theta) - \alpha \Big) + \beta \frac{\|\theta - \theta_0\|^2}{1 + \|\theta_0\|^2}. \tag{5}$$

The detection term runs over $\mathcal{C}_{\mathrm{det}} \subset \mathcal{C}$, the conditions with enough library failures to estimate a detection rate (camera, language, noise, robot; equal weights ν), while the constraints cover all seven conditions. The anchor pulls θ toward the cross-entropy head θ_0 that initialises training. After every Adam step on θ, the pair's dual variables move by projected ascent,

$$\lambda_{k,c} \leftarrow \max\big\{0,\ \lambda_{k,c} + \eta\big(\widetilde{\mathrm{FPR}}_c(\theta) - \alpha\big)\big\}, \tag{6}$$

so a condition over the tolerance gains weight until it returns under it. The sigmoid terms shape θ; the statements of Propositions 1 and 2 come from the final calibration of the frozen score. Temperature rule and hyper-parameters are in Section C.

Selection and deployment. A development split with variants disjoint from the training split provides its own reference and query halves. After every epoch we recalibrate θ on its reference successes with Equation (1), record macro early detection over $\mathcal{C}_{\mathrm{det}}$ and the worst per-condition false-alarm rate on its query half, and select the checkpoint with the highest macro early detection among those whose worst rate is at or below α (Algorithm 1, Section C). We freeze the selected score and recalibrate it on the deployment calibration bank with Equation (1) (CIL-M) or Equation (2) (CIL-HC).

Table 1: Failure detection experiments at $\alpha = 0.10$; all values in percent. Every method is thresholded by the condition-wise CondMax rule on the same calibration successes (CIL-M with the split-conformal rank, CIL-HC with the high-confidence rank, $\delta = 0.05$). Worst FPR is the highest per-condition false-alarm rate. Early detection counts failures flagged within the actionable horizon; ROC-AUC uses the within-horizon statistic; balanced accuracy is $(\text{early} + 100 - \text{FPR})/2$. The thresholds of the baselines are adjusted so that their false-alarm rate stays within the tolerance on every condition.

Method	LIBERO-Spatial					LIBERO-Long				
	Early↑	ROC-AUC↑	Bal. acc.↑	FPR	worst FPR	Early↑	ROC-AUC↑	Bal. acc.↑	FPR	worst FPR
$\pi_{0.5}$										
kNN distance	3.1 ± 1.2	54.8 ± 2.3	50.2 ± 0.6	2.7 ± 0.5	3.4 ± 0.9	3.6 ± 1.0	51.4 ± 1.9	50.2 ± 0.6	3.2 ± 0.5	4.4 ± 1.2
SAFE-MLP	47.6 ± 4.1	**92.9** ± 1.1	72.1 ± 2.1	3.4 ± 0.5	8.6 ± 1.9	20.8 ± 2.6	88.1 ± 1.0	58.0 ± 1.3	4.8 ± 0.7	7.8 ± 1.8
SAFE-LSTM	61.8 ± 3.7	91.4 ± 1.4	79.2 ± 1.9	3.4 ± 0.5	9.6 ± 2.1	39.1 ± 2.9	**88.8** ± 1.1	68.1 ± 1.4	3.0 ± 0.5	6.4 ± 1.5
ActProbe	49.2 ± 3.9	85.7 ± 1.9	72.2 ± 2.0	4.7 ± 0.6	9.6 ± 1.8	29.3 ± 2.8	87.3 ± 1.1	63.9 ± 1.4	1.5 ± 0.4	3.9 ± 1.2
CIL-M (ours)	59.2 ± 3.8	88.0 ± 1.5	77.3 ± 1.9	4.7 ± 0.7	9.7 ± 2.0	**46.5** ± 3.0	84.9 ± 1.3	70.4 ± 1.5	5.7 ± 0.7	9.6 ± 1.8
CIL-HC (ours)	**61.8** ± 3.8	91.8 ± 1.3	**79.2** ± 1.9	3.3 ± 0.5	9.6 ± 2.2	45.8 ± 2.9	84.8 ± 1.5	**71.3** ± 1.4	3.2 ± 0.5	5.0 ± 1.3
OpenVLA-OFT										
kNN distance	10.2 ± 2.3	93.7 ± 1.0	54.9 ± 1.1	0.3 ± 0.2	3.6 ± 1.8	11.8 ± 2.1	90.3 ± 1.3	55.5 ± 1.1	0.8 ± 0.3	4.7 ± 1.6
SAFE-MLP	0.0 ± 0.0	99.0 ± 0.2	49.8 ± 0.1	0.3 ± 0.2	1.8 ± 1.2	0.0 ± 0.0	94.2 ± 0.7	49.2 ± 0.3	1.7 ± 0.4	2.2 ± 1.1
SAFE-LSTM	54.1 ± 3.6	99.0 ± 0.3	76.9 ± 1.8	0.3 ± 0.2	1.8 ± 1.1	34.2 ± 3.4	96.4 ± 0.6	67.0 ± 1.7	0.3 ± 0.2	1.6 ± 0.9
ActProbe	4.3 ± 1.3	98.0 ± 0.4	52.0 ± 0.7	0.3 ± 0.2	0.9 ± 0.7	2.6 ± 1.1	96.6 ± 0.3	51.3 ± 0.6	0.1 ± 0.1	0.5 ± 0.4
CIL-M (ours)	**85.5** ± 2.5	**99.1** ± 0.2	**92.0** ± 1.3	1.5 ± 0.4	5.5 ± 2.4	**67.2** ± 3.7	**97.0** ± 0.8	**82.7** ± 1.9	1.8 ± 0.5	6.0 ± 1.8
CIL-HC (ours)	69.4 ± 3.4	98.8 ± 0.3	84.5 ± 1.7	0.4 ± 0.2	1.8 ± 1.1	51.4 ± 3.2	95.8 ± 1.1	75.5 ± 1.6	0.4 ± 0.2	1.2 ± 0.7

What the gradient does. The detection term lifts failure statistics above $\tau(\theta)$ within the first h calls. The gradient through $\tau(\theta)$ lowers the upper successful tail of the condition that sets the threshold, which lowers the threshold shared by all seven conditions. The duals settle on the noise condition, whose tail is hardest to hold under the tolerance (Figure 5, Section C). The head has $d + 1 = 143$ parameters and trains in under two minutes on a CPU.

5 Experiments

5.1 Setup

Simulation. A frozen $\pi_{0.5}$ (Physical Intelligence et al., 2025) and a frozen OpenVLA-OFT run the LIBERO-Spatial and LIBERO-10 tasks under the LIBERO-Plus catalog (Fei et al., 2025), whose variants fall into seven conditions (background, camera, language, layout, light, noise, robot); the monitor's context is a 142-dimensional projection of the policy's internal features and proprioception. A simulated library trains and selects the monitor, a calibration stream at other initial states sets its thresholds, and a predeclared fresh bank of 1,500 episodes per policy and suite, never used for any decision, is evaluated (Table 5 and Section B). θ_0 is SAFE's logistic head (Gu et al., 2025) refitted on the training split; CIL starts from it and also trains the SAFE and ActProbe networks in place of the linear head, and each ours row of Tables 1 and 3 is the best-performing CIL variant within the tolerance on that bank (Table 17). kNN distance, SAFE-MLP, SAFE-LSTM and ActProbe (Huang et al., 2026) are trained on the same library and thresholded by CondMax on the same calibration successes (Table 18 and Section F). Tolerances are $\alpha = 0.10$ and $\delta = 0.05$; early detection is the fraction of failures flagged within the actionable horizon, a success is a false alarm the first time its statistic crosses the threshold, and paired differences carry root-block bootstrap 95% intervals (Section C).

Real world. We fine-tune $\pi_{0.5}$ on 250 scripted demonstrations of 3 pick-and-place and 2 push tasks on a UFACTORY xArm6 with base and wrist RGB-D cameras. The target objects and destinations vary in shape and size. The arm runs three pick-and-place tasks under three conditions (nominal, distractors, perturbed start pose).

Table 2: Realized false-alarm rates of the $\pi_{0.5}$ monitors when the baselines use their own calibration rules (SAFE-MLP and SAFE-LSTM: the pooled functional band; ActProbe: its running maximum against a pooled split-conformal quantile; kNN: a pooled split-conformal quantile), all at $\alpha = 0.10$ on the same calibration successes, and CIL-M and CIL-HC use CondMax; all values in percent ($\pm$ std). Pooled FPR and the worst per-condition FPR on the fresh banks; a rate above the tolerance is marked in red.

Method	LIBERO-Spatial		LIBERO-Long	
	FPR	worst FPR	FPR	worst FPR
kNN distance	**10.7** $\pm$ 0.9	**13.1** $\pm$ 1.7	**10.2** $\pm$ 1.0	**13.3** $\pm$ 2.1
SAFE-MLP	9.2 $\pm$ 0.9	**20.6** $\pm$ 2.9	**13.1** $\pm$ 1.2	**21.3** $\pm$ 3.1
SAFE-LSTM	**10.5** $\pm$ 1.0	**23.0** $\pm$ 2.8	**13.2** $\pm$ 1.1	**21.5** $\pm$ 2.8
ActProbe	**10.5** $\pm$ 0.9	**23.0** $\pm$ 3.3	**10.6** $\pm$ 1.0	**16.7** $\pm$ 2.5
CIL-M (ours)	4.7 $\pm$ 0.7	9.7 $\pm$ 2.0	5.7 $\pm$ 0.7	9.6 $\pm$ 1.8
CIL-HC (ours)	3.3 $\pm$ 0.5	9.6 $\pm$ 2.2	3.2 $\pm$ 0.5	5.0 $\pm$ 1.3

5.2 Does training under the rule detect failures earlier at the same tolerance?

(a) Main results. Table 1 reports early detection, ROC-AUC, balanced accuracy and false-alarm rates on the four fresh banks, two policies on two suites. The CIL-M and CIL-HC rows are the best CIL variant within the tolerance on each bank, the linear head on the $\pi_{0.5}$ banks and CIL applied to a SAFE network on OpenVLA-OFT; on every bank they raise early detection over the cross-entropy head with each condition within the tolerance (per-condition detections in Figure 6, tests A and B in Table 7, every variant in Table 17).

(b) Ablations. The controls of Table 6 (cross-entropy continuation, stop-gradient, unconstrained, IID pairs) isolate the gradient through the calibration order statistic and the per-condition constraints as the sources of the high-confidence gain (Section C).

(c) Baselines. With their thresholds adjusted to the per-condition tolerance, SAFE-LSTM is the strongest baseline on the $\pi_{0.5}$ banks, tying CIL-HC on LIBERO-Spatial and trailing CIL-M on LIBERO-Long, while kNN distance, SAFE-MLP and ActProbe trail both; under their own pooled rules the SAFE and ActProbe scores exceed the tolerance on the noise condition (Table 18).

5.3 Does the guarantee hold?

(a) Unadjusted thresholds. Table 2 thresholds every $\pi_{0.5}$ monitor by the baselines' original threshold: the baselines exceed the per-condition tolerance on the heaviest-tailed condition of LIBERO-Spatial, while both CIL modes hold it on both banks.

(b) Within-condition splits. On 500 random within-condition splits of the fresh bank's successes, both modes meet their bound on every bank; on tests A and B, calibrated at initial states 0–1 and evaluated at 2–3, the noise condition exceeds the tolerance for every score (Sections D and G).

(c) CIL with functional CP. A condition-wise time-varying band keeps the gain, pooled rules exceed the tolerance on the heaviest-tailed condition, and an unrepresented condition is protected only by a heavier represented tail (Tables 8 and 9 and Section E).

5.4 Do earlier alarms become completed tasks, and at what reset cost?

Protocol. Each of the 1,272 test-B episodes receives one transient disturbance; the monitor may trigger one recovery, a scene reset charged c steps against the execution budget followed by a retry. The comparators are the recovery baselines CoRe and TIDE and a no-monitor reference; detection methods are compared in Table 1, and the calibration rules on our score in Table 13 (Section I).

(a) Cheap resets. At a 20-step reset every monitor completes far more than the no-monitor reference (Figure 3), and CIL-M alarms a median of 81 steps after the disturbance (Figure 2e).

(b) Costly resets. As the reset cost grows, CoRe and TIDE keep their completion only through resets of episodes that would have succeeded (63% and 75%) and fall below the reference first, while the

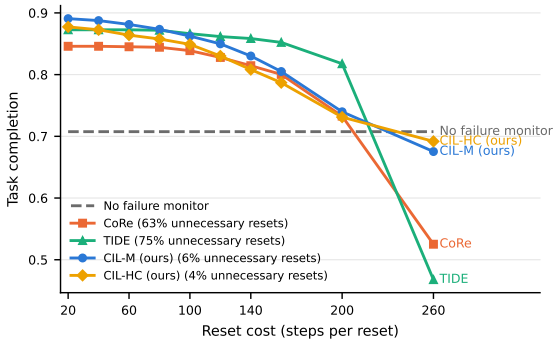


Figure 3: Failure recovery experiments under a controlled reset budget and tolerance FP rate: task completion against the reset cost charged per recovery, with each monitor's unnecessary-reset rate (resets of episodes that would have succeeded). Completion at every cost is recomputed from the recorded step counts; the full table of arms is Table 13.

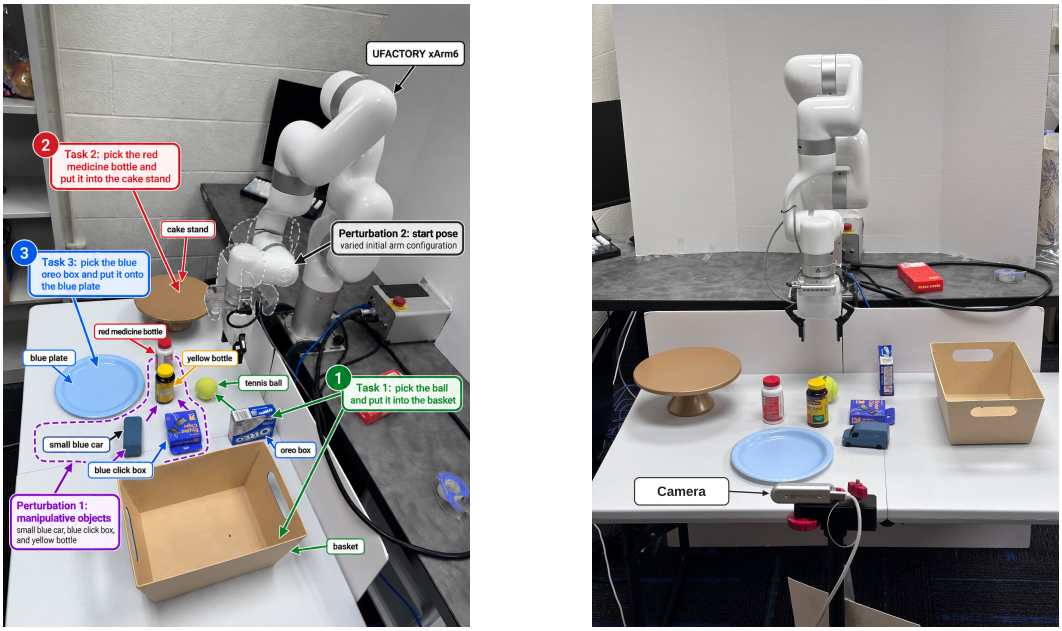


Figure 4: Real-robot setup. Left: the UFACTORY xArm6 workspace with the three pick-and-place tasks (1: tennis ball into the basket; 2: red medicine bottle onto the cake stand; 3: blue Oreo box onto the blue plate) and the two perturbed conditions, distractor objects placed near the objects and the carry path (small blue car, blue click box, yellow bottle) and a varied initial arm configuration. Right: the scene from the table-mounted RealSense base camera; a second camera is mounted on the wrist. Each task is run on six object layouts under the nominal and the two perturbed conditions.

two CIL monitors reset 4 to 6% and degrade last (Figure 3, Table 15); the matched θ_0 controls split the gap between context features and training objective (Section I).

5.5 Does the gain carry to a real robot?

(a) Detection. Over the six folds CIL-M detects the most failures early within the tolerance (Table 3); the baselines reach the tolerance only after their thresholds are raised, which costs them most of their detection, and the paired interval against θ_0 includes zero with 63 failures (Section J). **(b) Recovery.**

Table 3: Failure detection on the real robot (xArm6, six leave-one-layout-out folds, 156 episodes, 63 failures) at $\alpha = 0.10$; all values in percent ($\pm$ std). Every method is thresholded by CondMax with the split-conformal rank on the calibration layouts (the high-confidence rank needs 29 calibration successes per condition, which the folds do not hold). CIL-M is the best-performing CIL variant within the tolerance. The thresholds of the baselines are adjusted so that their false-alarm rate stays within the tolerance on every condition.

Method	Early↑	ROC-AUC↑	Bal. acc.↑	FPR	worst FPR
kNN distance	7.9 ± 3.0	53.1 ± 5.2	51.8 ± 1.1	4.3 ± 1.9	7.4 ± 3.8
SAFE-MLP	33.3 ± 8.7	85.2 ± 4.1	63.4 ± 4.4	6.5 ± 3.1	8.3 ± 4.3
SAFE-LSTM	25.4 ± 9.1	81.8 ± 4.9	60.5 ± 4.1	4.3 ± 2.7	7.4 ± 5.8
CIL (ours)	**44.4** ± 8.4	**86.6** ± 3.5	**70.1** ± 4.0	4.3 ± 2.3	5.6 ± 3.7

Table 4: Failure recovery on the real robot (xArm6) under a controlled reset budget and false-alarm tolerance: task completion per monitor. max 3 retries.

Monitor	Completion↑
No failure monitor	40.5 ± 6.7
CoRe	62.4 ± 3.9
TIDE	69.8 ± 2.8
CIL (ours)	**76.4** ± 3.6

Under the reset-and-retry protocol of Section 5.4 with at most three retries, CIL-M completes the most tasks, above TIDE and CoRe and far above the no-monitor reference (Table 4).

6 Conclusion

A VLA failure monitor's threshold is set by a rule, and the score can be trained for it. Computing the condition-wise maximum threshold inside the loss raised the linear head's early detection at the same nominal tolerance from 0.393 to 0.592 on a held-out LIBERO-Plus bank, from 0.334 to 0.465 on LIBERO-10 and from 0.349 to 0.444 on the xArm6 over 63 failures, and, applied to SAFE's networks, reached 0.855 and 0.694 on OpenVLA-OFT against 0.427 and 0.239 for the cross-entropy head. At a 140-step reset cost earlier alarms became completed tasks, 7.4 points over a calibration-only CondMax monitor (3.1 over the matched cross-entropy control) and 5.2 in the high-confidence mode at the same unnecessary-reset rate.

Reproducibility statement

The objective, exact threshold, selection rule and both calibration modes are in Sections 3 and 4 and Algorithm 1; hyper-parameters, data roles, calibration banks, eligibility and the cost rule are in Sections B, C, 5.1 and 5.4. Code, records and the claims trace will be released.

AI Use Statement

AI assistants (Anthropic Claude Code and OpenAI Codex) were used in this work to write and refactor experiment code, run and monitor experiments, aggregate results into tables and figures, and draft and edit text. All experimental designs, methods and claims were specified by the authors, every reported number was produced by the authors' code on the authors' data, and the authors reviewed and verified all AI-assisted code and text and take full responsibility for the content.

References

Anastasios N. Angelopoulos and Stephen Bates. Conformal Prediction: A Gentle Introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Konstantina Bairaktari, Jiayun Wu, and Zhiwei Steven Wu. Kandinsky Conformal Prediction: Beyond Class- and Covariate-Conditional Coverage. In *International Conference on Machine Learning (ICML)*, 2025.

Luiz F. O. Chamon and Alejandro Ribeiro. Probably Approximately Correct Constrained Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019.

Bat-Sheva Einbinder, Yaniv Romano, Matteo Sesia, and Yanfei Zhou. Training Uncertainty-Aware Classifiers with Conformalized Deep Learning. *arXiv preprint arXiv:2205.05878*, 2022.

Alec Farid, David Snyder, Allen Z. Ren, and Anirudha Majumdar. Failure Prediction with Statistical Guarantees for Vision-Based Robot Control. In *Robotics: Science and Systems XVIII*, 2022. doi: 10.15607/rss.2022.xviii.042.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: In-depth Robustness Analysis of Vision-Language-Action Models. *arXiv preprint arXiv:2510.13626*, 2025.

Qiao Gu, Yuanliang Ju, Shengxiang Sun, Igor Gilitschenski, Haruki Nishimura, Masha Itkina, and Florian Shkurti. SAFE: Multitask Failure Detection for Vision-Language-Action Models. *arXiv preprint arXiv:2506.09937*, 2025. NeurIPS 2025.

Leying Guan and Robert Tibshirani. Prediction and outlier detection in classification problems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):524–546, 2022. doi: 10.1111/rssb.12443.

Bingjia Huang, Xiangyu Li, Xiang Wang, Liang Mi, Zixu Hao, Weijun Wang, Hao Wu, Kun Li, Yunxin Liu, and Ting Cao. Actprobe: Action-space probe for early failure detection of generative robot policies. *arXiv preprint arXiv:2606.08508*, 2026.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Huihan Liu, Yu Zhang, Vaarij Betala, Evan Zhang, James Liu, Crystal Ding, and Yuke Zhu. Multi-Task Interactive Robot Fleet Learning with Visual World Models. *arXiv preprint arXiv:2410.22689*, 2024. CoRL 2024.

Rachel Luo, Shengjia Zhao, Jonathan Kuck, Boris Ivanovic, Silvio Savarese, Edward Schmerling, and Marco Pavone. Sample-efficient safety assurances using conformal prediction. *The International Journal of Robotics Research*, 43(9):1409–1424, 2024. doi: 10.1177/02783649231221580.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive Confidence Machines for Regression. In *Machine Learning: ECML 2002*, pp. 345–356. Springer, 2002. doi: 10.1007/3-540-36755-1_29.

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.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence (UAI)*, 2021.

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

Ralf Römer, Adrian Kobras, Luca Worbis, and Angela P. Schoellig. Failure Prediction at Runtime for Generative Robot Policies. *arXiv preprint arXiv:2510.09459*, 2025. NeurIPS 2025.

Clayton Scott and Robert Nowak. A Neyman-Pearson Approach to Statistical Learning. *IEEE Transactions on Information Theory*, 51(11):3806–3819, 2005. doi: 10.1109/TIT.2005.856955.

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

Wenwen Si, Sangdon Park, Insup Lee, Edgar Dobriban, and Osbert Bastani. PAC Prediction Sets Under Label Shift. *arXiv preprint arXiv:2310.12964*, 2023.

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.

David Stutz, Krishnamurthy Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning Optimal Conformal Classifiers. In *International Conference on Learning Representations (ICLR)*, 2022.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal Prediction Under Covariate Shift. In *Advances in Neural Information Processing Systems*, 2019.

Xin Tong, Yang Feng, and Jingyi Jessica Li. Neyman-Pearson Classification Algorithms and NP Receiver Operating Characteristics. *Science Advances*, 4(2):eaao1659, 2018. doi: 10.1126/sciadv.aao1659.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Machine Learning*, 92(2-3): 349–376, 2013. doi: 10.1007/s10994-013-5355-6.

Nien-Shao Wang, Duygu Nur Yaldiz, Yavuz Faruk Bakman, and Sai Praneeth Karimireddy. Conformal Prediction Adaptive to Unknown Subpopulation Shifts. *arXiv preprint arXiv:2506.05583*, 2025.

Yuqi Yang and Ying Jin. Multi-Distribution Robust Conformal Prediction. *arXiv preprint arXiv:2601.02998*, 2026.

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*, 2026a.

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

Gehan Zheng, Sanjay Seenivasan, Matthew Johnson-Roberson, and Weiming Zhi. Rewind-IL: Online Failure Detection and State Respawning for Imitation Learning. *arXiv preprint arXiv:2604.16683*, 2026.

Table 5: Data roles. Root = catalog variant (a LIBERO-Plus perturbation of a task); roots never overlap between the training/development library and any stream (checked: 0 shared roots); the fresh bank is a random sample of catalog pairs and shares roots with the library only at other initial states. Initial-state indices refer to the LIBERO initial-state file. Selection use: whether the role was used to choose θ (training), a checkpoint (development), or was inspected before the final evaluation.

Role	Source	Episodes	Roots	Initial states	Use
Training	simulated library	3526	652	4–9	trains θ (reference/query regime pairs)
Development	simulated library	1252	232	4–9	checkpoint selection (ref./query halves, disjoint roots)
Final calibration	calibration stream	1272	686	0–1 (layout: 0)	thresholds for tests A/B and the fresh bank; never trained on
Test A	evaluation stream	1272	672	2–3 (layout: 0)	inspected during development (exploratory)
Calibration, disturbed	disturbed calibration stream	1300	700	0–1	thresholds for the closed loop
Test B	disturbed evaluation stream	1272	672	2–3 (layout: 0)	confirmatory offline test; closed-loop episodes
Fresh bank	catalog (variant, initial-state) pairs never previously collected	1500	1064 (406 shared with the library at other initial states)	0–49 (random pairs)	never inspected before the final evaluation
LIBERO-10 (second domain; same design, layout excluded: single initial state)					
Training	simulated library	2136	356	4–9	trains θ (regime pairs)
Development	simulated library	744	124	4–9	checkpoint selection
Final calibration	calibration stream	1200	600	0–1	thresholds for the fresh bank; variants disjoint from the library
Fresh bank	predeclared catalog pairs	1500	1086 (294 episodes on library variants at other initial states)	0–49 (random pairs)	never inspected before the final evaluation

A PROOFS

Proof of Proposition 1. Fix c and let M_{n_c+1} be the statistic of a new successful execution of condition c. Exchangeability of $\{M_1, \ldots, M_{n_c}, M_{n_c+1}\}$ gives $\mathbb{P}(M_{n_c+1} > M_{(r_c)}) \le 1 - r_c/(n_c + 1) \le \alpha$ for $r_c = \lceil (n_c + 1)(1 - \alpha) \rceil$, the split-conformal bound (Papadopoulos et al., 2002; Lei et al., 2018); when $r_c > n_c$ the threshold is $+\infty$ and the probability is 0. Since $\tau = \max_{c'} \tau_{c'} \ge \tau_c$, $\mathbb{P}(M_{n_c+1} > \tau) \le \mathbb{P}(M_{n_c+1} > \tau_c) \le \alpha$. The left-hand side equals $\mathbb{E}[\mathrm{FPR}_c(\tau)]$ because the score is frozen and the new execution is independent of the calibration draw given the condition. □

Proof of Proposition 2. Let $q_c = \inf\{q : \mathbb{P}(M \le q \mid \text{success}, c) \ge 1 - \alpha\}$ be the population $(1 - \alpha)$-quantile of condition c and fix $c^\star \in \arg\max_c q_c$. On the event $\{\tau_{c^\star} \ge q_{c^\star}\}$ we have $\tau \ge \tau_{c^\star} \ge q_{c^\star} \ge q_c$ for each c, hence $\mathrm{FPR}_c(\tau) = \mathbb{P}(M > \tau \mid c) \le \mathbb{P}(M > q_c \mid c) \le \alpha$ for each c, by the definition of q_c and because the frozen score is independent of the calibration data. Therefore $\{\max_c \mathrm{FPR}_c(\tau) > \alpha\} \subseteq \{\tau_{c^\star} < q_{c^\star}\}$: the bad event is contained in the event of one condition, so no union bound over conditions is needed. If no rank qualifies for $c^\star$ then $\tau_{c^\star} = +\infty$ and the event is empty. Otherwise $\tau_{c^\star} = M_{(a_{c^\star})}$ is below $q_{c^\star}$ if and only if at least $a_{c^\star}$ of the $n_{c^\star}$ i.i.d. calibration statistics are below $q_{c^\star}$; each is below with probability $p = \mathbb{P}(M < q_{c^\star} \mid c^\star) \le 1 - \alpha$, so

$$\mathbb{P}(\tau_{c^\star} < q_{c^\star}) = \mathbb{P}[\mathrm{Bin}(n_{c^\star}, p) \ge a_{c^\star}] \le \mathbb{P}[\mathrm{Bin}(n_{c^\star}, 1 - \alpha) \ge a_{c^\star}] \le \delta,$$

by monotonicity of the binomial tail in p and the choice of $a_{c^\star}$ in Equation (2). □

Remark (ranks that do not exist). A condition with too few calibration successes has no qualifying high-confidence rank ($n_c < 29$ at $\alpha = 0.10$, $\delta = 0.05$), so $\tau = +\infty$ and the monitor is silent on all conditions. The simulated calibration banks of this paper hold at least 77 successes per condition (Section D); the real-robot calibration pairs hold 9 to 12 and use the marginal mode (Section J).

B DATA ROLES AND THE FRESH BANK

Table 5 lists each data role. No root is shared between library and stream variants, between the training and development splits, or between the development reference and query halves. Test A (initial states 2–3) was inspected during development; test B and the fresh bank were not.

Fresh bank. The fresh bank is a predeclared random sample of (variant, initial state) pairs from the catalog at initial states 0–49 (seed 20260917, 1,500 pairs), excluding all previously collected pairs, rolled out with the frozen policy and no monitor. It was first evaluated for the frozen comparison of θ_0 and CIL in both modes by a single scripted run on 2026-09-18; the four ablation arms, their re-run with preprocessing pinned to the headline checkpoint (Section C), the non-library subset, the sensitivity grid, the CIL-trained SAFE and ActProbe networks and the OpenVLA-OFT study were evaluated on it afterwards. The ours rows of Table 1 are chosen on this bank among the within-tolerance variants of Table 17; the development-selected linear head is the paper configuration.

LIBERO-10 roles. The LIBERO-10 roles were drawn once from its catalog before any LIBERO-10 rollout existed (seed 20260922): 80 variants per non-layout condition, stratified over the ten base tasks, at initial states 4–9 for the library; 100 further variants per condition, disjoint from the library,

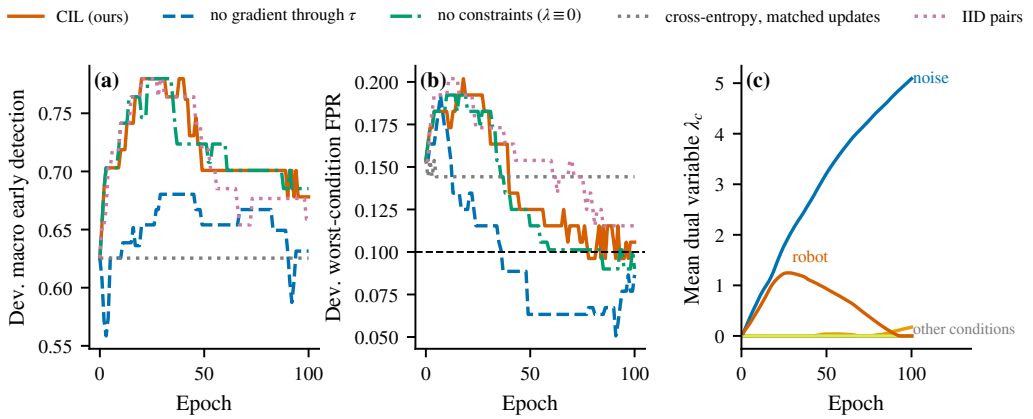


Figure 5: The noise dual dominates training and robot is a distant second. Development curves of CIL and its ablations; checkpoints are selected among epochs whose worst development false-alarm rate is at or below the tolerance (dashed line in (b)). (a) Macro early detection over the four objective conditions. (b) Worst-condition false-alarm rate. (c) Mean dual variable per condition for CIL.

at initial states 0–1 for the calibration stream; and 1,500 random (variant, initial state) pairs at initial states 0–49 over the whole non-layout catalog, excluding the pairs of the other two roles, for the fresh bank. The actionable horizon of a task is its shortest library or calibration episode; in simulation each failure ran to the step cap, so the shortest episode of a task is a success.

Claims trace. A single script generates, from the raw rollout records, the evidence inventory and a claims trace that maps each number in the paper to a field of a result file.

C Training algorithm, ablations, hyper-parameters and compute

Algorithm 1 Calibration-in-the-loop (CIL) training

Require: regime pairs $\{(\mathcal{R}_k, \mathcal{Q}_k)\}_{k=1}^{K}$, tolerance α, horizon h, head θ_0, T, η, β
1: $\theta \leftarrow \theta_0$; $\lambda_{k,c} \leftarrow 0$ for all pairs k and conditions c
2: **for** epoch $= 1, \ldots, E$ **do**
3: **for** each pair k **do**
4: compute $\tau(\theta)$ on $\mathcal{R}_k$ Equation (3)
5: compute $M_i(\theta)$, $M_i^h(\theta)$ on $\mathcal{Q}_k$ and the surrogates $\widetilde{\mathrm{TPR}}_{c'}^{\,h}(\theta)$, $\widetilde{\mathrm{FPR}}_c(\theta)$
6: $\theta \leftarrow \theta - \mathrm{Adam}\big(\nabla_\theta L_k(\theta; \lambda_k)\big)$ with gradient clipping Equation (5)
7: $\lambda_{k,c} \leftarrow \max\{0, \lambda_{k,c} + \eta(\widetilde{\mathrm{FPR}}_c(\theta) - \alpha)\}$ for all c Equation (6)
8: **end for**
9: recalibrate θ on the development reference half; record macro early detection and worst false-alarm rate on its query half
10: **end for**
11: select the checkpoint with the highest macro early detection among those with worst false-alarm rate $\leq \alpha$
Ensure: frozen θ, recalibrated on the deployment calibration bank (CIL-M or CIL-HC)

Algorithm 1 is the training loop of Section 4. When no checkpoint meets the tolerance, the arm takes the checkpoint with the lowest worst rate and is reported as infeasible.

Ablations in detail. Four controls share θ_0, preprocessing, aggregation, selection rule and calibration bank: a cross-entropy continuation with the CIL update budget of 600 updates, the objective with the gradient through $\tau(\theta)$ stopped, the objective without constraints ($\lambda \equiv 0$), and IID variant pairs in place of regime pairs (Table 17; direct contrasts against the paper configuration in Table 6, computed on one common set of root-block resamples per mode). The continuation finds no feasible checkpoint, falls back to epoch 1 (head within 0.004 of θ_0) and matches θ_0's early detection. The stop-gradient arm keeps $+0.115$ over θ_0 in each mode; CIL minus stop-gradient is $+0.063$ [$+0.000$, $+0.122$] in

Table 6: Fresh bank, paired contrasts with root-block bootstrap 95% intervals (the same resamples are used for every contrast within a mode). Left: CIL minus each ablation (pooled early detection and pooled FPR). Right: paired difference vs θ_0 in pooled early detection on the 922-episode non-library subset.

	Marginal: CIL minus row		High-conf.: CIL minus row		Non-library subset, row minus θ_0	
Arm	Δ early det.	Δ FPR	Δ early det.	Δ FPR	marginal	high-conf.
θ_0 (cross-entropy, same architecture)	+0.178 [+0.121, +0.242]	+0.011 [+0.004, +0.019]	+0.199 [+0.140, +0.259]	+0.013 [+0.006, +0.021]	–	–
+ cross-entropy continuation, CIL update budget	+0.178 [+0.121, +0.242]	+0.011 [+0.004, +0.019]	+0.199 [+0.140, +0.259]	+0.014 [+0.007, +0.022]	+0.000 [+0.000, +0.000]	+0.000 [+0.000, +0.000]
CIL without the gradient through τ	+0.063 [+0.000, +0.122]	+0.005 [-0.002, +0.014]	+0.084 [+0.024, +0.141]	+0.007 [+0.000, +0.014]	+0.139 [+0.067, +0.218]	+0.113 [+0.041, +0.188]
CIL without constraints ($\lambda \equiv 0$)	-0.005 [-0.037, +0.025]	+0.001 [-0.005, +0.007]	+0.047 [+0.006, +0.090]	+0.005 [+0.001, +0.009]	+0.191 [+0.114, +0.275]	+0.157 [+0.089, +0.227]
CIL with IID pairs instead of regime pairs	-0.016 [-0.048, +0.012]	-0.015 [-0.023, -0.007]	+0.031 [-0.012, +0.076]	+0.001 [-0.004, +0.005]	+0.200 [+0.122, +0.284]	+0.157 [+0.090, +0.225]
CIL (ours)	–	–	–	–	+0.200 [+0.124, +0.286]	+0.209 [+0.132, +0.289]

the marginal mode and $+0.084$ [$+0.024$, $+0.141$] in the high-confidence mode. The unconstrained arm ties the paper configuration in the marginal mode (-0.005 [-0.037, $+0.025$], worst rate 0.097) and trails it by $+0.047$ [$+0.006$, $+0.090$] in the high-confidence mode. The IID-pair arm reaches $+0.194$ over θ_0 in the marginal mode with the robot condition at 0.122 (pooled rate 0.058 against CIL's 0.044); CIL minus IID-pair is $+0.031$ [-0.012, $+0.076$] in the high-confidence mode.

Selected checkpoints and sensitivity. At the selected epoch the mean dual is 4.37 for noise, 0.31 for robot and at most 0.01 elsewhere (Figure 5c). CIL selects epoch 78 (development macro early detection 0.701, worst development false-alarm rate 0.096, the robot condition at 10 of 104 successes); the stop-gradient arm selects epoch 37 and the unconstrained arm epoch 85. Refitting the randomised PCA from scratch changes the preprocessing at order 10^{-4}, leaves θ_0's results unchanged and moves each CIL development checkpoint above the tolerance (lowest worst rate 12 of 104); the fallback checkpoint, epoch 89, detects 0.513 of the fresh bank's failures early ($+0.120$ [$+0.070$, $+0.171$] over θ_0; high-confidence 0.487, $+0.162$ [$+0.110$, $+0.216$]) with each condition under the tolerance. With the preprocessing pinned, the same code reproduces epoch 78 and the paper-configuration rows of Table 17. Of the sensitivity grid's alternatives under CondMax in the marginal mode, temperature 0.05 (epoch 93) detects 0.592 at a worst rate of 0.097 (camera) and anchor 10^{-4} 0.581 at 0.091, within tolerance; PCA 256, the grid's development choice, detects 0.644 at 0.115 (noise), $C = 0.1$ 0.602 at 0.124 and training $\alpha = 0.20$ 0.586 at 0.120, over the tolerance; PCA 64 collapses to 0.031. In the high-confidence mode no alternative exceeds the paper configuration's 0.524 (anchor 10^{-4} ties it).

Hyper-parameters and compute. All arms use Adam with learning rate 10^{-3} for 100 epochs, gradient clipping at 5, anchor $\beta = 10^{-3}$, dual step $\eta = 5$, temperature $T = 0.064$ (0.1 times the interquartile range of the baseline statistic) and $\alpha = 0.10$ for both the training constraint and the provisional calibration. The baseline logistic head uses balanced class weights and inverse regularisation $C = 0.3$. The temperature-0.05 variant of Table 1 sets $T = 0.05$ and is otherwise the paper configuration. For the networks, the anchor pulls toward the released weights and the anchor and learning rate are chosen on the development split: SAFE-LSTM output layer on LIBERO-Spatial anchor 10^{-3}, learning rate 10^{-3}, epoch 31; SAFE-LSTM on LIBERO-10 anchor 0.1, learning rate 10^{-4}, epoch 22; SAFE-MLP on OpenVLA-OFT anchor 10^{-3}, learning rate 10^{-4}, epoch 51; SAFE-LSTM on OpenVLA-OFT anchor 10^{-2}, learning rate 3×10^{-5}, epoch 25. The head has 143 parameters and each arm trains in under two minutes on a CPU; the closed-loop arms take about three hours each on one GPU with five parallel simulator shards.

D Offline results on tests A and B and the within-condition splits

Tests A and B. Table 7 reports the marginal mode on test A (evaluation stream, initial states 2–3, calibrated on the 1,165 successes of the calibration stream) and test B (disturbed evaluation stream, calibrated on the 915 successes of the disturbed calibration stream) for θ_0, the cross-entropy continuation, the three pinned ablations, the IID-pair arm and CIL. The per-condition calibration sets hold 196, 163, 190, 100, 198, 160 and 158 successes (background, camera, language, layout, light, noise, robot) on the calibration stream and 159, 120, 155, 77, 157, 130 and 117 on the disturbed stream. The robot condition sets the deployed threshold for both scores: 1.754 (A) and 1.859 (B) for θ_0, 1.338 (A) and 1.524 (B) for CIL. In the high-confidence mode the thresholds rise to 2.029 and 2.255 for θ_0 and to 1.442 and 1.850 for CIL; pooled early detection is then 0.363 (θ_0) against 0.615 (CIL) on A and 0.190 against 0.311 on B, with worst rates of 0.064 (robot) and 0.083 (noise) for θ_0

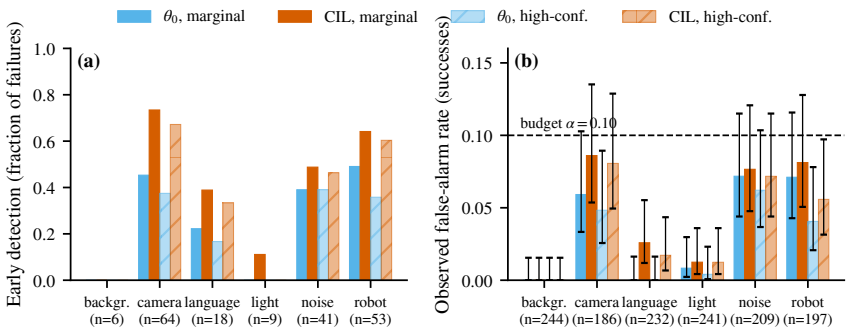


Figure 6: The paper configuration of the linear head (0.571 marginal, 0.524 high-confidence; Table 17) adds camera, robot and language detections, and each condition's false-alarm rate stays under the dashed tolerance. Fresh bank (1,500 held-out episodes), thresholds from the calibration stream at $\alpha = 0.10$. (a) Early detection per condition (fraction of that condition's failures alarmed within the actionable horizon; n = failures). (b) Observed false-alarm rate per condition on successes with Wilson 95% intervals (n = successes); the dashed line is the tolerance. Within each condition the left pair of bars is the marginal mode and the right, lighter pair the high-confidence mode.

Table 7: Offline results on test A (evaluation stream, initial states 2–3, inspected during development) and test B (disturbed evaluation stream, 1,272 episodes, never used for selection), marginal mode, $\alpha = 0.10$. All arms share θ_0, features, aggregation and calibration bank. Paired macro early detection vs θ_0 with root-block bootstrap 95% intervals. CIL's noise condition exceeds the nominal tolerance on both tests and θ_0's on test B (initial-state shift between the calibration bank and the tests, Appendix D).

	Test A				Test B (confirmatory)			
Arm	Early	FPR	Worst FPR	Δ macro [CI]	Early	FPR	Worst FPR	Δ macro [CI]
θ_0	0.467	0.033	0.096 (robot)	–	0.285	0.044	0.129 (noise)	–
Cross-entropy continuation	0.467	0.034	0.103 (robot)	+0.000 [+0.000, +0.000]	0.285	0.042	0.129 (noise)	+0.000 [+0.000, +0.000]
Cross-entropy, CIL update budget	0.467	0.033	0.096 (robot)	+0.000 [+0.000, +0.000]	0.285	0.044	0.129 (noise)	+0.000 [+0.000, +0.000]
CIL without gradient through τ	0.504	0.040	0.116 (noise)	+0.046 [-0.018, +0.113]	0.314	0.041	0.136 (noise)	+0.030 [-0.004, +0.065]
CIL without constraints	0.637	0.049	0.121 (noise)	+0.182 [+0.106, +0.255]	0.375	0.044	0.121 (noise)	+0.101 [+0.055, +0.148]
CIL, IID pairs	0.644	0.056	0.145 (noise)	+0.196 [+0.116, +0.272]	0.350	0.037	0.106 (noise)	+0.069 [+0.026, +0.114]
CIL (ours)	0.637	0.047	0.110 (noise)	+0.182 [+0.102, +0.262]	0.422	0.048	0.136 (noise)	+0.145 [+0.095, +0.195]

and 0.092 (noise) and 0.068 (noise) for CIL. Switching CIL from the marginal to the high-confidence mode changes its pooled early detection by -0.022 $[-0.050, +0.000]$ on A and -0.111 $[-0.144, -0.080]$ on B. On test A θ_0's noise rate is 0.087 and its worst condition is robot at 0.096.

Random within-condition splits. For each of 500 random splits of the fresh bank's successes into calibration and test halves within each condition (calibration halves of 122 background, 93 camera, 116 language, 120 light, 104 noise and 98 robot successes), the table below gives the mean per-condition test false-alarm rate, the mean worst-condition rate and the fraction of splits whose worst-condition rate exceeds $\alpha = 0.10$ ($\delta = 0.05$ in the high-confidence mode).

	mean test FPR per condition						mean worst	splits with worst $> \alpha$
Score / mode	backgr.	camera	language	light	noise	robot		
θ_0, marginal	0.000	0.065	0.001	0.008	0.081	0.078	0.094	38.8%
CIL-M	0.000	0.081	0.018	0.010	0.072	0.056	0.089	35.8%
θ_0, high-confidence	0.000	0.037	0.000	0.003	0.049	0.029	0.054	0.6%
CIL-HC	0.000	0.048	0.002	0.002	0.031	0.013	0.049	3.0%

E THRESHOLD RULES AND UNREPRESENTED CONDITIONS

A time-varying threshold. Foresight (Zhang et al., 2026a) thresholds a per-timestep score with a functional conformal band: with μ_t and σ_t the mean and standard deviation of the calibration successes' scores at call t, the nonconformity of calibration success i is $R_i = \sup_t(s_t^{(i)} - \mu_t)/\sigma_t$, $\hat{q}$ is the $\lceil (n+1)(1-\alpha) \rceil$-th smallest R_i, and the monitor alarms at the first call with $s_t \geq \mu_t + \hat{q}\,\sigma_t$. We

Table 8: Threshold rules on the fresh bank at the same nominal tolerance $\alpha = 0.10$ ($\delta = 0.05$ in the high-confidence mode): pooled early detection and the worst observed per-condition false-alarm rate (bold: above the tolerance) for the two scores, and CIL minus θ_0 with root-block bootstrap 95% intervals on common resamples. The functional band $\delta_t = \mu_t + \hat{q}\,\sigma_t$ (Zhang et al., 2026a) is calibrated on the calibration-stream successes, pooled or per condition with the deployed band the pointwise maximum; it thresholds either the per-call surprise s_t or its running mean $\bar{s}_t$.

	Marginal mode					High-confidence mode				
	θ_0		CIL			θ_0		CIL		
Rule	early	worst FPR	early	worst FPR	CIL − θ_0 [CI]	early	worst FPR	early	worst FPR	CIL − θ_0 [CI]
CondMax on the episode statistic (Equation (1); Tables 1–2)	0.393	0.072 (noise)	0.571	0.086 (camera)	+0.178 [+0.121, +0.242]	0.325	0.062 (noise)	0.524	0.081 (camera)	+0.199 [+0.140, +0.259]
Pooled split-conformal threshold on the episode statistic	0.754	**0.376 (robot)**	0.754	**0.299 (robot)**	+0.000 [-0.033, +0.031]	0.717	**0.325 (robot)**	0.733	**0.259 (robot)**	+0.016 [-0.025, +0.053]
Functional band, pooled, per-call score (Foresight's rule)	0.743	**0.239 (noise)**	0.770	**0.199 (camera)**	+0.026 [-0.020, +0.075]	0.723	**0.220 (noise)**	0.749	**0.183 (camera)**	+0.026 [-0.021, +0.077]
Functional band, pooled, running mean	0.780	**0.365 (robot)**	0.764	**0.320 (robot)**	-0.016 [-0.053, +0.020]	0.749	**0.335 (robot)**	0.738	**0.299 (robot)**	-0.010 [-0.054, +0.031]
Functional band, condition-wise maximum, per-call score	0.539	**0.102 (camera)**	0.592	**0.102 (camera)**	+0.052 [-0.006, +0.114]	0.471	0.081 (noise)	0.492	0.086 (camera)	+0.021 [-0.022, +0.064]
Functional band, condition-wise maximum, running mean	0.445	0.077 (noise)	0.597	**0.105 (noise)**	+0.152 [+0.095, +0.215]	0.377	0.057 (noise)	0.524	0.059 (camera)	+0.147 [+0.095, +0.202]

Table 9: Leave-one-condition-out on the fresh bank: thresholds calibrated on the calibration stream without condition c, observed false-alarm rate of c (bold: above the tolerance $\alpha = 0.10$); n = successes / failures of c. CondMax w/o c = the rule of Equation (1) over the remaining conditions; pooled w/o c = one split-conformal quantile over the remaining successes; retrained = CIL trained, selected and calibrated without c (camera and language drops have no development-feasible checkpoint and use the fallback). τ = deployed CIL-M threshold without c (setting condition) against 1.338 (robot) with every condition; early = early detection of c's failures by CIL-M without / with c in calibration.

		θ_0, CondMax w/o c		CIL, CondMax w/o c		CIL-M	CIL retrained w/o c			
Held-out c	n	M	HC	M	HC	pooled w/o c	M	HC	τ w/o c	early w/o / with
background	244 / 6	0.000	0.000	0.000	0.000	0.004	0.000	0.000	1.338 (robot)	0.00 / 0.00
camera	186 / 64	0.059	0.048	0.086	0.081	**0.183**	0.075	0.054	1.338 (robot)	0.73 / 0.73
language	232 / 18	0.000	0.000	0.026	0.017	**0.142**	0.000	0.000	1.338 (robot)	0.39 / 0.39
light	241 / 9	0.008	0.004	0.012	0.012	0.017	0.012	0.012	1.338 (robot)	0.11 / 0.11
noise	209 / 41	0.072	0.062	0.077	0.072	**0.196**	**0.187**	**0.153**	1.338 (robot)	0.49 / 0.49
robot	197 / 53	**0.239**	**0.117**	**0.168**	0.061	**0.335**	**0.193**	**0.102**	1.016 (camera)	0.74 / 0.64

apply the band rather than Foresight's world model because our records keep no video frames. The band thresholds the per-call surprise s_t or its running mean $\bar{s}_t$, pooled over conditions as in Foresight or condition-wise with the deployed band the pointwise maximum over conditions (the analogue of Equation (1)); calls with fewer than ten supporting calibration successes reuse the last supported band, and the high-confidence variant replaces the rank by the binomial rank of Equation (2).

The gain is largest under the rule the score was trained for (Table 8). Each pooled rule, band or single threshold, reaches its higher detection by exceeding the tolerance on the heaviest-tailed condition (worst rate 0.18 to 0.38), and the two scores tie under it. The condition-wise band on the running mean is within 0.03 of CondMax for the paper configuration (0.597 against 0.571, worst rate 0.105) and keeps the gain over θ_0 ($+0.152$ [$+0.095$, $+0.215$]; high-confidence $+0.147$ [$+0.095$, $+0.202$]). On the per-call score it gives 0.539 for θ_0 and 0.592 for CIL (worst rate 0.102 for both), a difference of $+0.052$ [-0.006, $+0.114$]. On test B the condition-wise call band gives CIL 0.350 against 0.422 under CondMax, and CIL minus θ_0 is $+0.093$ [$+0.055$, $+0.130$].

Leaving a condition out of calibration. A condition left out of calibration stays under the tolerance only when a represented condition has a heavier successful tail (Table 9). We calibrate on the calibration stream without condition c and observe c on the fresh bank. Removing a condition changes the deployed threshold only when that condition sets the maximum. Robot sets $\tau = 1.338$ for CIL, and without it camera sets $\tau = 1.016$, so robot's rate rises to 0.168 (CIL-M) and 0.239 (θ_0-M), and to 0.061 and 0.117 in the high-confidence mode; removing any other condition leaves the threshold and the rates unchanged. A pooled threshold without c is worse in each case (robot 0.335, noise 0.196, camera 0.183). Retraining CIL without c in the objective and the constraints does not protect c: trained without noise, its noise rate is 0.187 (early detection 0.66 against 0.49), and without camera or language no development checkpoint meets the tolerance, so the fallback checkpoint is used.

Table 10: ActProbe (Huang et al., 2026) trained on our library with its recipe: pooled early detection and worst observed per-condition false-alarm rate (bold: above $\alpha = 0.10$) on the fresh bank, test A and the disturbed test B, thresholds from the corresponding calibration stream. Rows (a)–(c) use the cross-entropy network selected on the development split, row (d) the same network trained with CIL. For reference, θ_0 / CIL-M detect 0.393 / 0.571 (fresh), 0.467 / 0.637 (A) and 0.285 / 0.422 (B).

	Fresh bank		Test A		Test B (disturbed)	
Row	early	worst FPR	early	worst FPR	early	worst FPR
(a) Its rule: running maximum, pooled split-conformal quantile	0.670	**0.230 (noise)**	0.667	**0.197 (noise)**	0.414	**0.175 (layout)**
(b) Running maximum under CondMax, marginal	0.576	**0.124 (noise)**	0.556	**0.104 (noise)**	0.095	0.050 (robot)
(b) Running maximum under CondMax, high-confidence	0.492	0.096 (noise)	0.459	0.087 (noise)	0.015	0.042 (robot)
(c) Running mean of $\mathrm{softplus}$(logit), CondMax, marginal	0.581	**0.153 (noise)**	0.570	**0.139 (noise)**	0.221	0.067 (robot)
(c) Running mean, CondMax, high-confidence	0.529	**0.115 (noise)**	0.496	0.092 (noise)	0.057	0.050 (robot)
(d) (c) with the network trained by CIL, marginal	0.565	0.097 (camera)	0.556	**0.122 (robot)**	0.054	0.059 (robot)
(d) (c) with the network trained by CIL, high-confidence	0.393	0.054 (camera)	0.385	0.064 (robot)	0.010	0.050 (robot)

Table 11: SAFE's released detectors (Gu et al., 2025) trained on our library with their recipe (SAFE-MLP: two-layer network with cumulative per-call score; SAFE-LSTM: recurrent per-call probability), regularisation and seed selected on the development split: pooled early detection [95% interval] and worst observed per-condition false-alarm rate (bold: above $\alpha = 0.10$) on the fresh bank, test A and the disturbed test B, thresholds from the corresponding calibration stream. Row (a): SAFE's own functional conformal band; rows (b): the same score under CondMax. For reference, θ_0 / CIL-M (paper configuration) detect 0.393 / 0.571 (fresh), 0.467 / 0.637 (A) and 0.285 / 0.422 (B).

	Fresh bank		Test A		Test B	
Row	Early det.	Worst FPR	Early det.	Worst FPR	Early det.	Worst FPR
SAFE-MLP: $\lambda_{\mathrm{reg}} = 0.001$*, seed 1, development early-statistic AUC 0.968*						
(a) SAFE's band, marginal	0.827 [0.770, 0.879]	**0.206 (noise)**	0.800 [0.718, 0.872]	**0.249 (noise)**	0.447 [0.385, 0.505]	**0.235 (noise)**
(b) CondMax, marginal	0.550 [0.467, 0.630]	**0.124 (noise)**	0.578 [0.473, 0.667]	**0.133 (noise)**	0.075 [0.042, 0.109]	0.098 (noise)
(b) CondMax, high-confidence	0.476 [0.392, 0.556]	0.086 (noise)	0.481 [0.366, 0.583]	**0.116 (noise)**	0.000 [0.000, 0.000]	0.045 (noise)
SAFE-LSTM: $\lambda_{\mathrm{reg}} = 0.01$*, seed 0, development early-statistic AUC 0.954*						
(a) SAFE's band, marginal	0.827 [0.767, 0.882]	**0.230 (noise)**	0.785 [0.704, 0.854]	**0.254 (noise)**	0.527 [0.462, 0.583]	**0.250 (noise)**
(b) CondMax, marginal	0.702 [0.628, 0.770]	**0.120 (noise)**	0.719 [0.632, 0.795]	**0.162 (noise)**	0.219 [0.166, 0.267]	0.045 (noise)
(b) CondMax, high-confidence	0.618 [0.542, 0.693]	0.096 (noise)	0.615 [0.517, 0.696]	0.098 (noise)	0.149 [0.106, 0.190]	0.030 (noise)

F ACTPROBE AND SAFE ON OUR BANKS

ActProbe. We reproduce ActProbe (Huang et al., 2026) from its released code on our recorded rollouts. Its two signals are computed per policy call from the stored action chunks as in the release: the magnitude is the root-mean-square of the five executed steps of the current chunk, and the consistency error is the mean squared difference between the previous chunk's unexecuted 45 steps and the first 45 steps of the current one (zero at the first call). With a normalised call index they form the three-dimensional input of the released network, a 32-unit LSTM whose initial state is a 16-dimensional bottleneck of a Qwen3-Embedding-0.6B embedding of the instruction, followed by a skip-connected MLP (22,945 parameters). We embed the 326 distinct instructions of our banks, the rewritten instructions of the language variants included. Training on our training split follows their recipe: binary cross-entropy over valid calls with the episode label broadcast to each call, AdamW (learning rate 10^{-3}, weight decay 10^{-4}), a cosine schedule over 400 epochs, gradient clipping at 1, batch 64, dropout 0.4, and early stopping on the development split's q-AUC at a quarter of the episode; of five seeds we keep the best development seed (q-AUC 0.906 to 0.914). At deployment the rule keeps the largest probability seen so far and compares it with a split-conformal quantile of the calibration successes' episode maxima; that rule and the CondMax rows are calibrated on the calibration stream and scored within the actionable horizon. Under our selection rule the cross-entropy ActProbe checkpoint has a worst development rate of 0.154. For the CIL rows the released network replaces the linear head in Algorithm 1: the per-call surprise is $\mathrm{softplus}$ of its logit, dropout is off during the CIL updates, and the regime pairs, exact threshold, surrogates, duals, anchor to the cross-entropy weights and selection are unchanged. Four protocol choices differ from the release: sequences are indexed by policy call rather than environment step; the development split stands in for their validation split; the calibration stream is the calibration population; and their hidden-state baselines and F1 and time-to-detection metrics are not run. On OpenVLA-OFT the policy executes its whole 8-step chunk, so consecutive chunks do not overlap; the consistency error follows ActProbe's rule for adjacent inferences, the squared difference of the consecutive executed actions across the

chunk boundary, the magnitude is taken over the 8 executed steps, and the seed is chosen on the development split (no grid). On the two $\pi_{0.5}$ banks the rows of Table 1 use the configuration with the highest development q-AUC on a grid over learning rate, weight decay, hidden size and seed (learning rate 3×10^{-3}, weight decay 10^{-4}, 32 units, seed 0 on both), which detects 0.560 and 0.408 on LIBERO-Spatial and 0.470 and 0.278 on LIBERO-10; the released recipe's rows are in Table 18.

On the disturbed test B the ActProbe rows under CondMax with the running maximum, and the CIL-trained rows, are nearly silent (early detection at most 0.10): the disturbance raises the consistency error of successful executions (mean 0.059 against 0.030) and the condition-wise thresholds follow that tail, while the internal-state scores keep 0.285 (θ_0) and 0.422 (CIL). Its own pooled rule keeps 0.414 by exceeding the tolerance on the layout condition (0.175). On test A the CIL-trained network's robot rate is 0.122. ActProbe reads only the action chunks, so its rows are the black-box comparison of this paper.

SAFE. We port SAFE-MLP and SAFE-LSTM from the released `failure_prob` code. SAFE-MLP is a two-layer network on the per-call feature whose outputs are summed over calls, trained with SAFE's loss ($+s$ on successes, $-s$ on failures); SAFE-LSTM is a 256-unit LSTM with a per-call sigmoid trained by per-call cross-entropy with the episode label. Both use class weights inverse to frequency, L2 regularisation, Adam and 1,000 full-batch epochs, and both read SAFE's π_0 feature, the pre-velocity latent (2,048 values, standardised on the training split). The regularisation weight in $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ and then the seed in $\{0, 1, 2\}$ are selected on the development split by the ROC-AUC of the early statistic (SAFE-MLP: 10^{-3}, seed 1, AUC 0.968; SAFE-LSTM: 10^{-2}, seed 0, AUC 0.954). Row (a) of Table 11 applies SAFE's own functional conformal band (a 30/70 split of the calibration successes fits the mean trajectory and its modulation, the width is the $(1-\alpha)$ quantile of the remaining successes' maximal standardised deviation, trajectories edge-padded); rows (b) apply CondMax to the episode maximum of the score. The rows of Table 1 on the two $\pi_{0.5}$ banks use the configuration with the highest development early-statistic AUC on a grid over λ, learning rate, epochs and seed: on LIBERO-Spatial SAFE-MLP $\lambda = 10^{-3}$, learning rate 3×10^{-3}, seed 0 (0.544 and 0.476) and SAFE-LSTM the released configuration (0.702 and 0.618); on LIBERO-10 SAFE-MLP $\lambda = 10^{-3}$, learning rate 10^{-3} (0.267 and 0.193) and SAFE-LSTM $\lambda = 0.1$, learning rate 3×10^{-4}, 300 epochs (0.494 and 0.391). On OpenVLA-OFT no grid was run and the rows use the released recipe with $\lambda = 10^{-2}$ chosen on the development split. For the CIL rows the network replaces the linear head in Algorithm 1 as for ActProbe, anchored to the released weights.

SAFE's band exceeds the tolerance on the noise condition on each bank (0.21 to 0.25). Under CondMax SAFE-MLP's cumulative score falls to 0.075 on test B, because the disturbance inflates the summed score of successful executions and raises the disturbed calibration threshold; SAFE-LSTM keeps 0.219 against CIL's 0.422. On the fresh bank SAFE-LSTM under CondMax is the strongest external row (0.702 marginal at a noise rate of 0.120, 0.618 high-confidence at 0.096) and is above the linear head in both modes (0.592 and 0.524); CIL applied to its output layer ties its 0.618 in the high-confidence mode, and applied to the whole network gives 0.702 at 0.118 (camera) and 0.544 at 0.043 (Table 17).

G LIBERO-10

Protocol and selection. The LIBERO-10 roles follow Section B (Table 5). Layout variants of LIBERO-Plus have a single initial state and are excluded, leaving six conditions, of which camera, language, noise and robot have at least 50 library failures and enter the detection objective. The actionable horizon is the shortest library or calibration-stream episode of each task (29 to 67 calls). CIL selects epoch 28 (development macro early detection 0.463, worst development rate 0.088, against 0.227 and 0.135 for θ_0). Table 1 gives both modes (Table 12 repeats the linear-head rows and the released-recipe ActProbe rows). Under its own rule the released ActProbe exceeds the tolerance on the camera condition (0.167); under CondMax it detects 0.411 and 0.293, and the grid-selected configuration 0.470 at 0.180 (camera), over the tolerance, and 0.278, against the linear head's 0.465 and 0.347. Training its network with CIL lifts the high-confidence mode of the running-mean statistic from 0.255 to 0.334 and leaves the marginal mode about unchanged, 0.424 to 0.437, with noise at 0.105. SAFE-LSTM's grid-selected score detects 0.494 at 0.096 (noise) and 0.391 (released recipe 0.370 and 0.275); trained with CIL from the released recipe it detects 0.540 at 0.107 (camera), over the tolerance, and 0.458 at 0.050 (robot), $+0.208$ [$+0.152$, $+0.258$] against θ_0 and $+0.183$ [$+0.134$,

Table 12: LIBERO-10 fresh bank (1,500 predeclared episodes, 389 failures; 1206 episodes on variants outside the LIBERO-10 library), thresholds from the LIBERO-10 calibration stream (932 successes) at $\alpha = 0.10$ ($\delta = 0.05$ in the high-confidence mode); columns as in Table 1; bold: worst observed per-condition rate above the tolerance. ActProbe rows: the released detector trained on the LIBERO-10 library with its recipe (Section F).

Score / training	Early det.	Δ vs θ_0 [95% CI]	Pooled FPR	Worst FPR (condition)	Early det., non-library
Marginal mode					
θ_0 (cross-entropy)	0.334	–	0.016	0.044 (robot)	0.329
CIL (ours)	0.465	+0.131 [+0.097, +0.168]	0.057	0.096 (noise)	0.460
ActProbe, its rule	0.586	[0.532, 0.640]	0.106	**0.167 (camera)**	0.573
ActProbe score under CondMax	0.411	[0.353, 0.468]	0.038	0.077 (robot)	0.415
ActProbe network, running-mean statistic, trained with CIL	0.437	[0.381, 0.494]	0.074	**0.105 (noise)**	0.439
High-confidence mode					
θ_0 (cross-entropy)	0.249	–	0.006	0.017 (robot)	0.247
CIL (ours)	0.347	+0.098 [+0.066, +0.130]	0.028	0.046 (noise)	0.345
ActProbe score under CondMax	0.293	[0.239, 0.347]	0.015	0.039 (robot)	0.296
ActProbe network, running-mean statistic, trained with CIL	0.334	[0.280, 0.389]	0.043	0.064 (noise)	0.335

$+0.238$] against the released SAFE-LSTM score under the same rule; CIL on the grid-selected SAFE-LSTM keeps no update at the gate (epoch 0). SAFE-MLP detects 0.267 at 0.107 (camera), over the tolerance, and 0.193 (Tables 17 and 18).

Guarantee check. Over 500 random within-condition splits of the LIBERO-10 fresh bank's successes, the high-confidence worst-condition rate exceeds α in 3.4% of splits for CIL and 2.2% for θ_0, within $\delta = 5\%$. In the marginal mode the mean test rate of each condition is at most 0.102 (CIL, noise), against a split-conformal bound of 0.100 for that condition's calibration half of 109 successes.

H OpenVLA-OFT on LIBERO-Spatial

Policy and features. The second policy is OpenVLA-OFT fine-tuned on LIBERO-Spatial, run as released: an L1-regression action head, a third-person and a wrist image, proprioception input and 8-action chunks executed in full, so the policy is called every eight steps. The per-call feature is the last-layer hidden state of its action tokens, averaged over the action dimensions, at the first and last chunk step (8,192 values); the context is the paper's, a 128-component PCA of that feature, 12 proprioception values, the reset-pose deviation and the instruction flag (142 values), in place of the 2,048-value pre-velocity latent of $\pi_{0.5}$ (Section 5.1).

Banks. The banks reuse the exact (variant, initial state) pairs of the $\pi_{0.5}$ LIBERO-Spatial banks (Table 5): the library holds 4,778 episodes with 1,022 failures (robot 514, language 193, noise 89, light 76, camera 69, background 64, layout 17), the calibration stream 1,272 episodes with 1,068 successes (86 to 194 per condition, 86 for robot), the fresh bank 1,500 episodes with 255 failures (successes/failures per condition: background 229/21, camera 239/11, language 206/44, light 231/19, noise 230/20, robot 110/140; layout absent), and a nominal stream of 500 episodes on the ten base tasks holds 464 successes (92.8%; the policy's release reports 97.6% under its own evaluation loop). All six non-layout conditions enter the detection objective and the constraints, and the actionable horizon is 9 to 14 calls per task, the shortest library or calibration episode.

Pooled rules alarm on the robot condition. Every pooled rule reaches early detection of 0.953 to 0.984 by alarming on 0.373 to 0.545 of the robot-condition successes: θ_0 under a pooled split-conformal quantile 0.973 at 0.545 and under the pooled binomial quantile 0.953 at 0.500, SAFE-MLP and SAFE-LSTM under their band 0.984 and 0.977 at 0.446, ActProbe under its rule 0.969 at 0.373 (Table 18). The robot perturbation shifts every monitor's score, so a pooled calibration is not a valid monitor on this policy. The linear CIL head under the pooled quantile halves θ_0's worst rate (0.933 at 0.191) and stays over the tolerance, and SAFE's band at a nominal level matched post hoc on the fresh bank ($\alpha = 0.125$ for SAFE-MLP, 0.105 for SAFE-LSTM) is over it as well (0.536 and 0.464) and is not deployable.

SAFE without shift data. Trained on the nominal stream alone, with its band calibrated on the nominal initial states 0–9, SAFE-MLP detects 0.996 at a pooled false-alarm rate of 0.454 (worst 0.773, robot) and SAFE-LSTM 0.925 at 0.310 (0.709, robot): every perturbed execution looks anomalous

to a monitor that has seen only nominal ones. At a post-hoc matched level of $\alpha = 0.01$ they detect 0.875 at 0.318 and 0.788 at 0.236, still over the per-condition tolerance, and under CondMax 0.973 at 0.536 and 0.933 at 0.882 (Table 18).

Baselines and CIL under CondMax. SAFE-MLP and SAFE-LSTM use the released recipe with the library's shift data, $\lambda = 10^{-2}$ and the seed chosen on the development split (no sensitivity grid was run on this policy); ActProbe follows Section F, with its consistency error on adjacent inferences because the chunks do not overlap. Under CondMax, θ_0 detects 0.427 at a worst rate of 0.082 (robot) and 0.239 at 0.027; SAFE-MLP 0.361 at 0.064 and 0.000 at 0.018; SAFE-LSTM 0.776 at 0.036 and 0.541 at 0.018; ActProbe 0.286 at 0.045 and 0.043 at 0.009 (Table 1). The linear head in the paper configuration detects 0.733 at 0.054 ($+0.306$ [$+0.237$, $+0.373$] against θ_0) and 0.537 at 0.027 ($+0.298$ [$+0.235$, $+0.363$]), and the development-selected alternative of its grid (learning rate 3×10^{-3}) 0.722 at 0.036 and 0.596 at 0.018 ($+0.353$ [$+0.276$, $+0.436$]). CIL applied to SAFE-MLP (anchor 10^{-3}, learning rate 10^{-4}, epoch 51, chosen on the development split) detects 0.855 at 0.054 in the marginal mode, $+0.494$ [$+0.424$, $+0.568$] against SAFE-MLP's own score under the same rule, and 0.263 in the high-confidence mode; CIL applied to SAFE-LSTM (anchor 10^{-2}, learning rate 3×10^{-5}, epoch 25) detects 0.749 at 0.018 in the marginal mode, -0.027 [-0.076, $+0.019$] against SAFE-LSTM's 0.776, and 0.694 at 0.018 in the high-confidence mode, $+0.153$ [$+0.102$, $+0.206$] against its 0.541. No source pairs the network rows with θ_0. The same linear recipe retrained inside the real-robot kit reaches 0.647 and 0.655 in two runs; the linear-head rows of Table 17 are the suite-pipeline runs.

Per-condition band and posterior mixture. Under SAFE's band computed per condition with the pointwise maximum deployed, θ_0 detects 0.545 at 0.082, SAFE-MLP 0.357 at 0.054 and SAFE-LSTM 0.000: the robot-condition band swallows every score. With the band inside the loop, SAFE-MLP detects 0.804 at 0.064, SAFE-LSTM 0.753 at 0.036 and the linear head 0.690 at 0.064, and the high-confidence widths silence both SAFE networks (0.000). Under the mixture rule θ_0 detects 0.729 at 0.091 (robot), within tolerance, and the paper configuration 0.792 at 0.102 (language), over it ($+0.063$ [$+0.020$, $+0.105$]); the linear heads within tolerance under this rule are the training-tolerance-0.05 head at 0.788 (0.087) and the non-selected PCA-256 alternative at 0.816 (0.097) (Tables 17 and 18).

Training tolerance. The training-tolerance grid (α inside the loop 0.05 and 0.02, deployment at 0.10) was added after the first fresh-bank results and is selected on the development split only. For the linear head, tolerances 0.10, 0.05 and 0.02 give 0.733, 0.761 (worst 0.045) and 0.718 in the marginal mode and 0.537, 0.616 (0.036) and 0.561 in the high-confidence mode, and joint development selection over the whole linear grid still picks the learning-rate-3×10^{-3} head. For SAFE-MLP with CIL, tolerance 0.02 gives 0.835 at 0.054 and 0.584 at 0.009 against 0.855 and 0.263 at tolerance 0.10; for SAFE-LSTM with CIL, 0.765 at 0.027 and 0.349 at 0.004 against 0.749 and 0.694. Joint development selection over the tolerances picks the tolerance-0.02 run for both networks, so the two OpenVLA-OFT rows of ours in Table 1 are the development gate within their own anchor and learning-rate grid at tolerance 0.10 and not the joint development choice. The loop aligns the score at the training quantile, and the development metric, macro early detection under the marginal CondMax rule, does not rank checkpoints for the high-confidence rule; selection under the deployed rule is not implemented.

Ranking is saturated. Early pooled ROC-AUC of the episode statistic within the horizon is 0.983 for θ_0, 0.982 for the paper configuration, 0.981 for the development-selected head, 0.990 for SAFE-MLP and SAFE-LSTM, 0.991 and 0.988 for the two networks with CIL, and 0.980 for ActProbe (0.916 with CIL); the nominal-only SAFE-MLP and SAFE-LSTM rank at 0.965 and 0.932, and the full-episode AUC is at least 0.979 for every monitor because failures run to the step cap. CIL moves the AUC of the linear head and of the SAFE networks by less than 0.01, so the gap between 0.427 and 0.733 or 0.855 within the tolerance is a threshold effect, not a ranking effect. On the $\pi_{0.5}$ LIBERO-Spatial bank the same AUCs are 0.882 (θ_0), 0.883 (CIL), 0.929 (SAFE-MLP), 0.914 (SAFE-LSTM) and 0.857 (ActProbe), where SAFE's networks rank better than the linear head.

Table 13: Failure recovery experiments under controlled reset budget and FP rate.

Monitor	Compl. c=20	Δ vs CondMax	Compl. c=140	Δ vs CondMax	Unnec. resets	Median delay (steps)
No monitor (matched budget)	0.708	–	0.708	–	0	–
Source-CP (earlier score, single threshold)	0.878	-0.008 [-0.025, +0.010]	0.866	+0.109 [+0.083, +0.135]	0.511	21
CondMax (earlier score)	0.886	–	0.756	–	0.037	153
CoRe monitor	0.846	-0.040 [-0.060, -0.021]	0.814	+0.058 [+0.033, +0.082]	0.633	16
TIDE monitor	0.873	-0.013 [-0.030, +0.002]	0.858	+0.102 [+0.078, +0.127]	0.750	7
Posterior mixture	0.914	+0.028 [+0.016, +0.041]	0.851	+0.095 [+0.076, +0.115]	0.097	89
Robust conformal set	0.910	+0.024 [+0.014, +0.035]	0.822	+0.065 [+0.050, +0.082]	0.080	114
Earlier learned score	0.876	-0.010 [-0.025, +0.004]	0.792	+0.035 [+0.019, +0.053]	0.049	110
CondMax, $\alpha = 0.20$	0.914	+0.028 [+0.015, +0.041]	0.855	+0.099 [+0.078, +0.121]	0.124	81
θ_0-M (CIL preprocessing, eligibility)	0.881	-0.005 [-0.020, +0.009]	0.800	+0.043 [+0.027, +0.061]	0.057	98
θ_0-HC	0.872	-0.014 [-0.028, -0.001]	0.784	+0.028 [+0.011, +0.046]	0.037	119
CIL-M (ours)	0.891	+0.005 [-0.009, +0.018]	0.830	+0.074 [+0.054, +0.094]	0.064	81
CIL-HC (ours)	0.877	-0.009 [-0.024, +0.005]	0.808	+0.052 [+0.034, +0.071]	0.037	95

Table 14: Reset-fidelity diagnostic: completion of a retry started by a timed reset at policy call 12 without any disturbance (24 episodes per task, tasks 0, 4, 5), compared with a fresh attempt. The original restore left a stale controller goal and an unrestored gripper command ramp; the corrected restore is used for every arm in this paper. Right: completion of two arms measured with the defective restore, discarded and re-run.

Reset mechanism	Task 0	Task 4	Task 5	All ($n = 72$)
Fresh attempt (no reset)	1.00	0.79	0.67	0.819
Original restore	0.92	0.42	0.12	0.486
True environment reset	0.96	0.83	0.67	0.819
Corrected restore	1.00	0.83	0.79	0.875
Discarded closed-loop arms (defective restore, same 1,272 episodes)				
CondMax: 0.797 (corrected: 0.886)				
CIL-M: 0.779 (corrected: 0.891)				

I CLOSED-LOOP DETAILS

Disturbances. The object displacement moves the target object by a random planar offset of 4 to 7 cm (uniform magnitude and direction) at a step drawn uniformly from 15 to 35, unless the gripper is already in contact with it; the forced gripper opening holds the gripper command open for 10 steps once the gripper has been closed for 8 consecutive steps after step 30. The kind is drawn per episode with equal probability from a per-episode seed shared by all arms, so the arms face the same disturbance schedule.

Reset mechanism and diagnostic. The recovery restore writes the recorded initial joint and object state into the simulator, re-reads the state before resetting the operational-space controller's goal, restores the gripper command ramp and lets the controller settle for five steps; on CPU the restored observation matches a fresh start to 0.01 cm at the end effector with the gripper open. The diagnostic runs 72 undisturbed episodes of tasks 0, 4 and 5 (24 per task) with a timed reset at call 12 (Table 14): a fresh attempt completes 0.819, a true environment reset 0.819 and the restore used for all closed-loop arms 0.875. A restore that resets the goal before re-reading the state and leaves the ramp unrestored completes 0.486 (its settle steps move the arm 2 cm and 0.12 rad); the closed-loop arms of Table 13 use the corrected restore.

Comparators. The three less conservative rules on the earlier score are the posterior mixture, the robust conformal set and CondMax at $\alpha = 0.20$; the earlier learned score is a further row. Source-CP uses one threshold from 45 source successes. The θ_0 controls share CIL's preprocessing, eligibility and calibration bank, and CoRe and TIDE follow their published recipes (Table 13).

Columns and curves. Table 15 gives the full columns at a 20-step reset, completion at each reset cost (an episode with r resets that succeeded after n executed steps succeeds under cost c iff $n \leq 420 - c\,r$) and the paired differences of CIL-M and CIL-HC against their matched controls and against CondMax. Rescued is the fraction of the 372 reference failures completed; rescuable refers to the 270 reference failures whose undisturbed twin succeeded; unnecessary and ruined are the fractions of the 900 reference-success episodes that were reset, and reset and then failed; pre-onset

Table 15: Closed-loop evaluation, full columns (reset cost 20). Δ vs none = paired completion difference vs the matched-budget no-monitor reference with root-block bootstrap 95% intervals; unnecessary and ruined = fraction of the 900 reference-success episodes that were reset / reset and then failed; rescued = fraction of the 372 reference failures completed; rescuable = the 270 reference failures whose undisturbed twin succeeded; pre-onset = alarms before the disturbance onset, as a fraction of the reference failures with a defined onset.

Monitor	Completion	Δ vs none	Unnec.	Ruined	Rescued	Rescued (rescuable)	Pre-onset alarms
Source-CP (earlier score, single threshold)	0.878	+0.171 [+0.143, +0.196]	0.511	0.043	0.688	0.874	0.233
CondMax (earlier score)	0.886	+0.178 [+0.156, +0.202]	0.037	0.008	0.629	0.789	0.011
CoRe monitor	0.846	+0.138 [+0.110, +0.166]	0.633	0.060	0.618	0.763	0.412
TIDE monitor	0.873	+0.165 [+0.138, +0.192]	0.750	0.047	0.677	0.856	0.157
Posterior mixture	0.914	+0.207 [+0.180, +0.232]	0.097	0.009	0.728	0.915	0.019
Robust conformal set	0.910	+0.203 [+0.177, +0.228]	0.080	0.010	0.718	0.896	0.014
Earlier learned score	0.876	+0.168 [+0.145, +0.192]	0.049	0.009	0.597	0.752	0.089
CondMax, $\alpha = 0.20$	0.914	+0.206 [+0.180, +0.232]	0.124	0.016	0.742	0.937	0.024
θ_0-M (CIL preprocessing, eligibility)	0.881	+0.174 [+0.149, +0.199]	0.057	0.018	0.637	0.789	0.033
θ_0-HC	0.872	+0.164 [+0.140, +0.189]	0.037	0.014	0.597	0.730	0.030
CIL-M (ours)	0.891	+0.183 [+0.158, +0.207]	0.064	0.018	0.669	0.819	0.062
CIL-HC (ours)	0.877	+0.170 [+0.145, +0.193]	0.037	0.011	0.608	0.730	0.035

Completion vs synthetic reset cost c (exact recomputation: an episode with r resets that succeeded after n executed steps succeeds under cost c iff $n \le 420 - cr$)

Monitor	c=20	c=40	c=60	c=80	c=100	c=120	c=140	c=160	c=200	c=260
No monitor (matched budget)	0.708	0.708	0.708	0.708	0.708	0.708	0.708	0.708	0.708	0.708
Source-CP (earlier score, single threshold)	0.878	0.877	0.877	0.873	0.873	0.872	0.866	0.850	0.789	0.561
CondMax (earlier score)	0.886	0.873	0.859	0.837	0.818	0.789	0.756	0.722	0.691	0.682
CoRe monitor	0.846	0.846	0.845	0.844	0.839	0.828	0.814	0.800	0.733	0.525
TIDE monitor	0.873	0.873	0.873	0.872	0.866	0.862	0.858	0.852	0.818	0.468
Posterior mixture	0.914	0.913	0.907	0.897	0.889	0.876	0.851	0.825	0.764	0.653
Robust conformal set	0.910	0.904	0.895	0.881	0.870	0.851	0.822	0.796	0.731	0.658
Earlier learned score	0.876	0.868	0.862	0.851	0.840	0.814	0.792	0.762	0.722	0.686
CondMax, $\alpha = 0.20$	0.914	0.908	0.906	0.899	0.893	0.879	0.855	0.829	0.766	0.644
θ_0-M (CIL preprocessing, eligibility)	0.881	0.873	0.863	0.855	0.847	0.829	0.800	0.779	0.720	0.682
θ_0-HC	0.872	0.864	0.857	0.847	0.834	0.809	0.784	0.757	0.718	0.687
CIL-M (ours)	0.891	0.888	0.881	0.873	0.862	0.850	0.830	0.805	0.740	0.675
CIL-HC (ours)	0.877	0.873	0.864	0.858	0.849	0.830	0.808	0.787	0.731	0.692
*Paired completion difference vs the matched control (point estimate; interval excludes zero where marked *)*										
CIL-M $-$ θ_0-M	+0.009*	+0.014*	+0.018*	+0.019*	+0.016*	+0.021*	+0.031*	+0.026*	+0.020*	-0.007*
CIL-HC $-$ θ_0-HC	+0.006	+0.009	+0.007	+0.010	+0.015*	+0.021*	+0.024*	+0.030*	+0.013*	+0.005

Paired completion difference vs CondMax at other reset costs (steps)

Monitor	c=60	c=80	c=100	c=160	c=200	c=260
Source-CP (earlier score, single threshold)	+0.018 [-0.002, +0.038]	+0.036 [+0.014, +0.058]	+0.056 [+0.032, +0.079]	+0.128 [+0.101, +0.155]	+0.098 [+0.071, +0.125]	-0.121 [-0.152, -0.093]
θ_0-M (CIL preprocessing, eligibility)	+0.004 [-0.013, +0.021]	+0.017 [-0.002, +0.035]	+0.029 [+0.010, +0.048]	+0.057 [+0.041, +0.076]	+0.029 [+0.017, +0.042]	+0.001 [-0.009, +0.010]
θ_0-HC	-0.002 [-0.019, +0.014]	+0.010 [-0.009, +0.028]	+0.017 [-0.002, +0.035]	+0.035 [+0.021, +0.051]	+0.027 [+0.016, +0.038]	+0.006 [-0.002, +0.014]
CIL-M (ours)	+0.022 [+0.006, +0.038]	+0.036 [+0.019, +0.053]	+0.045 [+0.025, +0.063]	+0.083 [+0.064, +0.103]	+0.049 [+0.034, +0.064]	-0.006 [-0.017, +0.004]
CIL-HC (ours)	+0.005 [-0.011, +0.021]	+0.020 [+0.002, +0.038]	+0.031 [+0.012, +0.050]	+0.065 [+0.048, +0.084]	+0.040 [+0.027, +0.053]	+0.010 [+0.001, +0.020]

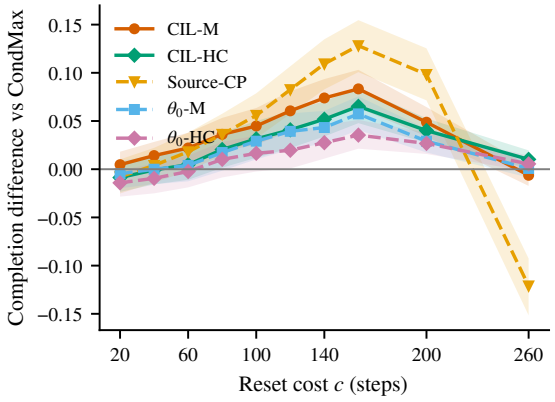


Figure 7: CIL-M's paired gain over CondMax is positive for reset costs of 60 to 200 steps and largest among the tabulated costs at 160 ($+0.083$ [$+0.064$, $+0.103$]); at 260 the interval includes zero (-0.006 [-0.017, $+0.004$]). Paired completion difference against CondMax as a function of the synthetic reset cost, with root-block bootstrap 95% bands (672 catalog variants; pointwise).

alarms are fractions of the 369 failures with a defined onset. Retries after an unnecessary reset complete 0.788 of the time for CondMax, 0.724 for CIL-M and 0.697 for CIL-HC.

Table 16: Real robot, leave-one-layout-out folds: for each test layout, the library and calibration layouts assigned from counts alone, the test episodes (successes / failures), the selected CIL epoch, and the false alarms among successes and early detections among failures of θ_0 and CIL-M under that fold's thresholds.

Test layout	Library layouts	Calibration layouts	Test succ. / fail.	Horizon (calls)	CIL epoch	False alarms / successes		Early detections / failures	
						θ_0	CIL-M	θ_0	CIL-M
1	4, 5, 6	2, 3	12 / 9	35, 37, 13	74	1/12	1/12	2/9	3/9
2	1, 3, 6	4, 5	17 / 4	35, 37, 18	68	0/17	0/17	0/4	0/4
3	4, 5, 6	1, 2	11 / 10	35, 37, 13	74	0/11	0/11	5/10	8/10
4	1, 5, 6	2, 3	12 / 9	39, 37, 13	40	1/12	2/12	5/9	7/9
5	1, 4, 6	2, 3	19 / 17	35, 38, 13	58	1/19	1/19	4/17	7/17
6	1, 4, 5	2, 3	22 / 14	35, 37, 13	18	1/22	0/22	6/14	3/14
pooled, nominal			36 / 12			1/36	2/36	0/12	2/12
pooled, distractors			30 / 18			1/30	1/30	5/18	5/18
pooled, start pose			27 / 33			2/27	1/27	17/33	21/33

J REAL-ROBOT STUDY

Robot and policy. A UFACTORY xArm6 with its gripper and two RGB-D cameras, one fixed at the base and one on the wrist, runs a $\pi_{0.5}$ fine-tuned from the base checkpoint on scripted demonstrations of tabletop pick-and-place tasks; we use three of its tasks (a tennis ball into a basket, a medicine bottle onto a cake stand, a cookie box onto a plate). The observation holds both camera images, the six joint angles and the gripper. Each call returns ten absolute joint-and-gripper targets at 10 Hz, of which eight are executed while the next chunk is requested two ticks early; an episode ends at the operator's verdict, at a stop of the robot's safety layer or after 600 control ticks. The policy runs with ten flow steps. Its sampler carries the input of the action expert's output projection through the denoising loop and returns it with the chunk at each flow step and chunk position, from the same forward pass that produced the executed actions; the returned actions are identical to those of the unmodified sampler, and a call takes 69 ms with or without the latent. The latent is reduced as in simulation (first, last and mean over flow steps and over chunk positions). No monitor runs on the robot; each monitor is trained and evaluated offline on the logged episodes.

Conditions and pilot. Nominal runs the training scene. The distractor condition places three objects, drawn per episode from a fixed pool of three (a tape roll, a clip box, a toy car), within about 10 cm of the object, the destination or the carry path, without touching them. The start-pose condition perturbs each joint of the initial configuration by a uniform draw within ± 0.08 rad per episode, rejecting draws that bring the tool within 180 mm of the base axis or 60 mm of the table. The magnitudes were fixed after an 18-episode pilot on a separate layout, before any main episode: nominal succeeded in 5 of 6 pilot episodes, two distractors far from the path in 4 of 4 and three near it in 1 of 2, a start-pose perturbation of 0.15 rad in 0 of 4 and of 0.08 rad in 1 of 2; the settings with a failure rate between 20 and 60% were kept. A camera-displacement condition was not run because the calibrated base camera cannot be moved. Before the main collection the base camera was displaced and restored to within about 2 cm of its calibrated position at the table; that pose was then frozen and re-checked against a reference frame after each layout, so all three conditions were collected under one camera pose.

Layouts and labels. Object layouts are staged from marked positions and restored for each episode. Ten layouts per task were planned; six (156 episodes) were collected, and roles are assigned leave-one-layout-out from counts alone (Section 5.5, Table 16). Within a layout the episodes of each task run in one launch with the conditions grouped and the repeats shuffled. Eight main-run episodes with a staging fault or an operator intervention were re-run under the written rule. The operator labels success at the gripper's release of the object in or on the destination. Successful episodes are cut five ticks after the release so that, as in simulation, a success ends at completion.

Context, horizon, training and calibration. The context vector has the 128 PCA components of the latent (fitted on library episodes), the joint angles, tool position and gripper, the reset-pose deviation and the instruction flag (constant here, since each instruction is the original), standardised on the library. The objective conditions are those with at least six library failures (distractors and start pose); nominal carries a constraint only. The actionable horizon of a task is the length in calls of

its shortest successful library or calibration episode (35 to 39, 37 to 38 and 13 to 18 calls across the folds); on the robot the shortest episodes are safety stops, which would set the horizon at nine calls for each task. A fold's library holds three layouts (78 to 93 episodes), too few for a development split, so the checkpoint is selected by leave-one-layout-out cross-validation inside the library: for each library layout the CondMax threshold is computed on the successes of the other two, the held-out layout is scored under it, the outcomes are pooled over the three folds, and the selected epoch has the highest macro early detection among those whose pooled worst-condition false-alarm rate is at or below α, as in Section 4. When a reference fold holds fewer than nine successes of a condition, so that its rank does not exist, that condition is left out of that pair's training threshold. Both scores are calibrated in the marginal mode on the fold's calibration pair, which holds 9 to 12 successes per condition, so the rank of Equation (1) is the largest success statistic of each condition and no high-confidence rank exists. Each episode is a test episode in exactly one fold, and the pooled outcomes carry the statistics of Section 5.1 with task-by-layout blocks. SAFE-MLP and SAFE-LSTM are trained on the same library with their released recipe, their regularisation selected by leave-one-layout-out AUC, and evaluated under CondMax (Table 3) and under their functional band (Table 18); trained with CIL on the same folds, SAFE-MLP detects 0.492 at 0.111 (start pose), over the tolerance, with four folds keeping the released network, and SAFE-LSTM 0.365 at 0.037; the other linear-head variants (development-selected checkpoint 0.397, CondMax in the loop with a marginal constraint 0.381, training B 0.302) are within tolerance and below the paper configuration (Table 17). No ActProbe run exists on the robot.

K ALL CIL VARIANTS AND ALL BASELINE RULES

Variants. Table 17 lists every variant of ours on the four held-out banks, per decision rule, and shades the row reported in Table 1 or Table 3, the best-performing variant within the tolerance on that bank; the development-selected linear head is the paper configuration. Training A is Algorithm 1 as written, the CondMax threshold and the per-condition constraints inside the loop, and its sensitivity grid changes one factor at a time: PCA dimension 64 or 256, inverse regularisation C 0.1 or 1.0, temperature 0.05, anchor 10^{-4} or 10^{-2}, learning rate 3×10^{-3} or 3×10^{-4} over 200 epochs, dual step 20, and the training tolerance 0.15, 0.20 and 0.30 on the $\pi_{0.5}$ banks and 0.05 and 0.02 on OpenVLA-OFT, deployment staying at 0.10; each run keeps the development gate of Algorithm 1, and gate infeasible marks a run with no development checkpoint within the tolerance, reported at its fallback. The three ablation arms (no gradient through τ, no constraints, IID pairs) are described in Section C. Training B puts the pooled threshold in the loop (on LIBERO-Spatial no checkpoint meets the gate and the row is θ_0; Bm selects the checkpoint at the matched false-alarm rate) and A-marg keeps the CondMax threshold with a marginal constraint; band-loop puts SAFE's functional band, pooled or condition-wise, inside the loop; net-CIL trains the SAFE-MLP, SAFE-LSTM or ActProbe network with the loop, anchored to its released weights (for SAFE-LSTM on LIBERO-Spatial also its output layer alone), and a run whose gate keeps no update (epoch 0) is the base network itself and is not counted as a variant, as the truncated placeholder runs on OpenVLA-OFT are not; kit re-trains the recipe inside the real-robot kit, whose horizons differ by one to two calls on some tasks; on the robot the selection is the pre-registered leave-one-layout-out pooled rule (Section J), and the high-confidence rows are 0.000 for every variant because no high-confidence rank exists there (Section A).

Rules. Table 18 lists θ_0, SAFE-MLP, SAFE-LSTM and ActProbe under every rule: CondMax in both modes (Equations (1) and (2)), the pooled split-conformal quantile and the pooled binomial quantile at $\delta = 0.05$ and 0.10, the posterior mixture rule (Wang et al., 2025), SAFE's pooled functional band and its condition-wise version with the pointwise maximum deployed, each also with the high-confidence width, and ActProbe's running maximum against a pooled quantile. Rows are grouped by trained configuration: the released recipe, the configuration selected on the development split of the sensitivity grid and its runner-up (never selected, listed for the range), source-only or nominal-only training and calibration without shift data, the kit's re-evaluation of the same checkpoint, refits of θ_0 that share a preprocessing with a grid alternative, and bands at a nominal level matched post hoc on the fresh bank, which are not deployable. Under the pooled split-conformal quantile and under SAFE's pooled band every baseline exceeds the tolerance on its heaviest-tailed condition on every bank (θ_0 0.376, robot, on LIBERO-Spatial, 0.545 on OpenVLA-OFT and 0.111 on the xArm6), ActProbe's rule exceeds it on the three simulated banks (0.230, 0.167 and 0.373), the mixture rule

exceeds it for θ_0's paper checkpoint and for every SAFE and ActProbe row on the $\pi_{0.5}$ banks and the robot and keeps θ_0 within it on OpenVLA-OFT (0.729 at 0.091, robot), and the condition-wise band holds the tolerance at a lower detection (Sections E and H).

Table 17: Every CIL variant on the three fresh banks under the two CondMax rules (marginal = Maxcond, high-confidence = Maxcond-HC), same metrics as Table 1. Variants: the linear head under the paper training (A), the pooled/marginal training (Bm), the dev-selected configuration grid (PCA size, learning rate, epochs, dual step, anchor, constraint weight, temperature, training tolerance) and CIL applied to the SAFE-MLP / SAFE-LSTM networks (full network, output layer only, training-tolerance grid). The row used in Table 1 is marked ⋆; a variant whose score equals its base network (selection returned epoch 0) is omitted. Bold worst FPR exceeds the tolerance.

Variant	Rule	Early det. ↑ pooled / worst	ROC-AUC ↑ pooled / worst	Bal. acc. ↑ pooled / worst	FPR ↓ pooled / worst
$\pi_{0.5}$, *LIBERO-Spatial*					
θ_0 (cross-entropy head, reference)	M	0.393 / 0.000	0.882 / 0.731	0.680 / 0.496	0.032 / 0.072
θ_0 (cross-entropy head, reference)	HC	0.325 / 0.000	0.882 / 0.731	0.650 / 0.498	0.024 / 0.062
CIL on SAFE-LSTM	M	0.702 / 0.111	0.916 / 0.644	0.825 / 0.551	0.051 / **0.118**
CIL on SAFE-LSTM	HC	0.545 / 0.111	0.916 / 0.644	0.763 / 0.551	0.019 / 0.043
CIL on SAFE-LSTM (head only)	M	0.696 / 0.111	0.918 / 0.771	0.821 / 0.549	0.055 / **0.139**
CIL on SAFE-LSTM (head only) ⋆	HC	0.618 / 0.000	0.918 / 0.771	0.792 / 0.494	0.033 / 0.096
CIL linear, pca256	M	0.644 / 0.167	0.910 / 0.749	0.797 / 0.583	0.050 / **0.115**
CIL linear, pca256	HC	0.408 / 0.000	0.910 / 0.749	0.693 / 0.494	0.021 / 0.065
CIL on SAFE-LSTM (training tolerance grid)	M	0.634 / 0.111	0.924 / 0.832	0.796 / 0.549	0.042 / **0.105**
CIL on SAFE-LSTM (training tolerance grid)	HC	0.524 / 0.111	0.924 / 0.832	0.747 / 0.549	0.029 / 0.059
CIL linear, C0.1	M	0.602 / 0.000	0.882 / 0.660	0.773 / 0.500	0.056 / **0.124**
CIL linear, C0.1	HC	0.461 / 0.000	0.882 / 0.660	0.714 / 0.496	0.034 / 0.081
CIL on SAFE-MLP	M	0.602 / 0.000	0.926 / 0.729	0.780 / 0.494	0.042 / **0.105**
CIL on SAFE-MLP	HC	0.487 / 0.000	0.926 / 0.729	0.726 / 0.494	0.036 / 0.091
CIL linear, temp0.05 ⋆	M	0.592 / 0.000	0.880 / 0.666	0.773 / 0.500	0.047 / 0.097
CIL linear, temp0.05	HC	0.466 / 0.000	0.880 / 0.666	0.716 / 0.498	0.034 / 0.081
CIL linear, ta0.20	M	0.586 / 0.167	0.898 / 0.723	0.768 / 0.583	0.050 / **0.120**
CIL linear, ta0.20	HC	0.440 / 0.000	0.898 / 0.723	0.703 / 0.494	0.033 / 0.081
CIL linear, anchor1e-4	M	0.581 / 0.000	0.883 / 0.677	0.767 / 0.500	0.047 / 0.091
CIL linear, anchor1e-4	HC	0.524 / 0.000	0.883 / 0.677	0.743 / 0.494	0.037 / 0.081
CIL linear, training Bm	M	0.576 / 0.167	0.897 / 0.705	0.766 / 0.583	0.044 / **0.100**
CIL linear, training Bm	HC	0.455 / 0.000	0.897 / 0.705	0.709 / 0.494	0.037 / 0.091
CIL linear, lr3e-3	M	0.576 / 0.000	0.879 / 0.632	0.764 / 0.500	0.049 / **0.102**
CIL linear, lr3e-3	HC	0.518 / 0.000	0.879 / 0.632	0.739 / 0.500	0.040 / 0.086
CIL linear, paper config (A)	M	0.571 / 0.000	0.883 / 0.676	0.764 / 0.500	0.044 / 0.086
CIL linear, paper config (A)	HC	0.524 / 0.000	0.883 / 0.676	0.743 / 0.494	0.037 / 0.081
CIL linear, base	M	0.571 / 0.000	0.883 / 0.676	0.764 / 0.500	0.044 / 0.086
CIL linear, base	HC	0.524 / 0.000	0.883 / 0.676	0.743 / 0.494	0.037 / 0.081
CIL linear, anchor1e-2	M	0.555 / 0.000	0.877 / 0.644	0.753 / 0.494	0.050 / **0.102**
CIL linear, anchor1e-2	HC	0.482 / 0.000	0.877 / 0.644	0.722 / 0.496	0.037 / 0.086
CIL linear, C1.0	M	0.534 / 0.000	0.880 / 0.681	0.744 / 0.494	0.045 / 0.091
CIL linear, C1.0	HC	0.471 / 0.000	0.880 / 0.681	0.718 / 0.498	0.036 / 0.081
CIL linear, ta0.15	M	0.518 / 0.000	0.900 / 0.738	0.741 / 0.494	0.037 / 0.091
CIL linear, ta0.15	HC	0.403 / 0.000	0.900 / 0.738	0.686 / 0.494	0.031 / 0.075
CIL linear, ta0.30	M	0.513 / 0.111	0.889 / 0.718	0.735 / 0.549	0.043 / 0.096
CIL linear, ta0.30	HC	0.461 / 0.000	0.889 / 0.718	0.712 / 0.494	0.037 / 0.091
CIL linear, duallr20	M	0.503 / 0.000	0.881 / 0.657	0.730 / 0.494	0.043 / 0.081
CIL linear, duallr20	HC	0.455 / 0.000	0.881 / 0.657	0.712 / 0.498	0.032 / 0.075
CIL linear, lr3e-4_ep200	M	0.503 / 0.000	0.889 / 0.710	0.730 / 0.494	0.042 / 0.096
CIL linear, lr3e-4_ep200	HC	0.471 / 0.000	0.889 / 0.710	0.717 / 0.494	0.037 / 0.081
CIL linear, pca64	M	0.031 / 0.000	0.650 / 0.381	0.511 / 0.500	0.009 / 0.056
CIL linear, pca64	HC	0.026 / 0.000	0.650 / 0.381	0.511 / 0.500	0.005 / 0.030
$\pi_{0.5}$, *LIBERO-10*					
θ_0 (cross-entropy head, reference)	M	0.334 / 0.000	0.846 / 0.642	0.659 / 0.496	0.016 / 0.044
θ_0 (cross-entropy head, reference)	HC	0.249 / 0.000	0.846 / 0.642	0.622 / 0.500	0.006 / 0.017
CIL on SAFE-LSTM	M	0.540 / 0.154	0.848 / 0.648	0.739 / 0.577	0.061 / **0.107**
CIL on SAFE-LSTM ⋆	HC	0.458 / 0.077	0.848 / 0.648	0.713 / 0.538	0.032 / 0.050
CIL linear, ta0.15	M	0.494 / 0.200	0.853 / 0.654	0.717 / 0.568	0.060 / **0.100**
CIL linear, ta0.15	HC	0.393 / 0.077	0.853 / 0.654	0.676 / 0.528	0.041 / 0.087
CIL linear, ta0.30	M	0.483 / 0.231	0.854 / 0.661	0.711 / 0.601	0.060 / **0.110**
CIL linear, ta0.30	HC	0.321 / 0.024	0.854 / 0.661	0.652 / 0.506	0.018 / 0.032
CIL linear, paper config (A) ⋆	M	0.465 / 0.200	0.849 / 0.645	0.704 / 0.570	0.057 / 0.096
CIL linear, paper config (A)	HC	0.347 / 0.146	0.849 / 0.645	0.660 / 0.560	0.028 / 0.046
CIL linear, anchor1e-4	M	0.465 / 0.200	0.850 / 0.645	0.704 / 0.570	0.057 / 0.096
CIL linear, anchor1e-4	HC	0.347 / 0.146	0.850 / 0.645	0.660 / 0.560	0.028 / 0.046
CIL linear, base	M	0.465 / 0.200	0.849 / 0.645	0.704 / 0.570	0.057 / 0.096
CIL linear, base	HC	0.347 / 0.146	0.849 / 0.645	0.660 / 0.560	0.028 / 0.046
CIL linear, training Bm	M	0.416 / 0.000	0.851 / 0.650	0.692 / 0.496	0.032 / 0.066
CIL linear, training Bm	HC	0.308 / 0.000	0.851 / 0.650	0.647 / 0.496	0.014 / 0.032
CIL linear, ta0.20	M	0.404 / 0.195	0.841 / 0.636	0.678 / 0.580	0.047 / 0.078
CIL linear, ta0.20	HC	0.298 / 0.098	0.841 / 0.636	0.637 / 0.538	0.023 / 0.041
CIL on SAFE-MLP	M	0.380 / 0.239	0.844 / 0.763	0.657 / 0.594	0.066 / 0.096
CIL on SAFE-MLP	HC	0.229 / 0.148	0.844 / 0.763	0.596 / 0.548	0.037 / 0.064
CIL linear, pca256	M	0.339 / 0.000	0.858 / 0.716	0.662 / 0.500	0.015 / 0.059
CIL linear, pca256	HC	0.254 / 0.000	0.858 / 0.716	0.625 / 0.500	0.005 / 0.018
CIL linear, pca256_anchor1e-4	M	0.339 / 0.000	0.858 / 0.716	0.662 / 0.500	0.015 / 0.059
CIL linear, pca256_anchor1e-4	HC	0.254 / 0.000	0.858 / 0.716	0.625 / 0.500	0.005 / 0.018
OpenVLA-OFT, LIBERO-Spatial					
θ_0 (cross-entropy head, reference)	M	0.427 / 0.000	0.983 / 0.908	0.710 / 0.500	0.008 / 0.082
θ_0 (cross-entropy head, reference)	HC	0.239 / 0.000	0.983 / 0.908	0.618 / 0.500	0.002 / 0.027
CIL on SAFE-MLP ⋆	M	0.855 / 0.727	0.991 / 0.972	0.920 / 0.857	0.015 / 0.055
CIL on SAFE-MLP	HC	0.263 / 0.000	0.991 / 0.972	0.630 / 0.496	0.002 / 0.009
CIL on SAFE-MLP, tolerance 0.02	M	0.835 / 0.600	0.992 / 0.973	0.912 / 0.800	0.010 / 0.055
CIL on SAFE-MLP, tolerance 0.02	HC	0.584 / 0.350	0.992 / 0.973	0.791 / 0.675	0.002 / 0.009
CIL on SAFE-LSTM, tolerance 0.02	M	0.765 / 0.500	0.993 / 0.973	0.880 / 0.750	0.006 / 0.027
CIL on SAFE-LSTM, tolerance 0.02	HC	0.349 / 0.105	0.993 / 0.973	0.674 / 0.550	0.002 / 0.004
CIL linear, pca256	M	0.761 / 0.550	0.986 / 0.957	0.873 / 0.773	0.015 / 0.045
CIL linear, pca256	HC	0.678 / 0.450	0.986 / 0.957	0.835 / 0.725	0.009 / 0.036
CIL linear, training tolerance 0.05	M	0.761 / 0.650	0.978 / 0.933	0.871 / 0.802	0.019 / 0.045
CIL linear, training tolerance 0.05	HC	0.616 / 0.507	0.978 / 0.933	0.803 / 0.735	0.010 / 0.036
CIL on SAFE-LSTM	M	0.749 / 0.550	0.988 / 0.967	0.872 / 0.775	0.005 / 0.018
CIL on SAFE-LSTM ⋆	HC	0.694 / 0.450	0.988 / 0.967	0.845 / 0.725	0.004 / 0.018
CIL linear, paper config (A)	M	0.733 / 0.671	0.982 / 0.938	0.859 / 0.808	0.014 / 0.055
CIL linear, paper config (A)	HC	0.537 / 0.250	0.982 / 0.938	0.766 / 0.625	0.005 / 0.027

Table 18: Baselines on the three fresh banks under CondMax (marginal and high-confidence) and, where the rule exists, under their own rule (SAFE: pooled functional band at $\alpha = 0.10$; ActProbe: running maximum against a pooled split-conformal quantile). Same metrics as Table 1; own-rule rows report early detection, pooled FPR and worst FPR only. Bold worst FPR exceeds the tolerance.

Variant	Rule	Early det. ↑ pooled / worst	ROC-AUC ↑ pooled / worst	Bal. acc. ↑ pooled / worst	FPR ↓ pooled / worst
$\pi_{0.5}$, *LIBERO-Spatial*					
SAFE-MLP	M	0.550 / 0.000	0.929 / 0.737	0.752 / 0.494	0.045 / **0.124**
SAFE-MLP	HC	0.476 / 0.000	0.929 / 0.737	0.721 / 0.496	0.034 / 0.086
SAFE-LSTM	M	0.702 / 0.222	0.914 / 0.663	0.825 / 0.605	0.051 / **0.120**
SAFE-LSTM	HC	0.618 / 0.000	0.914 / 0.663	0.792 / 0.498	0.034 / 0.096
ActProbe	M	0.576 / 0.333	0.857 / 0.704	0.756 / 0.650	0.063 / **0.124**
ActProbe	HC	0.492 / 0.333	0.857 / 0.704	0.722 / 0.647	0.047 / 0.096
kNN distance	M	0.079 / 0.000	0.548 / 0.485	0.520 / 0.475	0.038 / 0.049
kNN distance	HC	0.031 / 0.000	0.548 / 0.485	0.502 / 0.484	0.027 / 0.034
SAFE-MLP	own rule (band (pooled))	0.827 / –	–	–	0.092 / **0.206**
SAFE-LSTM	own rule (band (pooled))	0.827 / –	–	–	0.105 / **0.230**
ActProbe	own rule (native)	0.686 / –	–	–	0.140 / **0.263**
ActProbe	own rule (band (pooled))	0.665 / –	–	–	0.112 / **0.239**
ActProbe	own rule (band (pooled) HC)	0.654 / –	–	–	0.097 / **0.215**
SAFE-MLP	own rule (band (pooled) HC)	0.796 / –	–	–	0.077 / **0.182**
SAFE-LSTM	own rule (band (pooled) HC)	0.780 / –	–	–	0.077 / **0.196**
$\pi_{0.5}$, *LIBERO-10*					
SAFE-MLP	M	0.224 / 0.136	0.881 / 0.812	0.581 / 0.530	0.061 / 0.093
SAFE-MLP	HC	0.208 / 0.114	0.881 / 0.812	0.580 / 0.529	0.048 / 0.078
SAFE-LSTM	M	0.494 / 0.077	0.888 / 0.752	0.721 / 0.521	0.052 / 0.096
SAFE-LSTM	HC	0.391 / 0.077	0.888 / 0.752	0.681 / 0.528	0.030 / 0.064
ActProbe	M	0.411 / 0.000	0.873 / 0.721	0.687 / 0.500	0.038 / 0.077
ActProbe	HC	0.293 / 0.000	0.873 / 0.721	0.639 / 0.500	0.015 / 0.039
kNN distance	M	0.057 / 0.023	0.514 / 0.479	0.500 / 0.475	0.056 / 0.081
kNN distance	HC	0.036 / 0.000	0.514 / 0.479	0.502 / 0.484	0.032 / 0.044
SAFE-MLP	own rule (band (pooled))	0.686 / –	–	–	0.131 / **0.213**
SAFE-LSTM	own rule (band (pooled))	0.602 / –	–	–	0.116 / **0.201**
ActProbe	own rule (native)	0.586 / –	–	–	0.106 / **0.167**
OpenVLA-OFT, LIBERO-Spatial					
SAFE-MLP	M	0.361 / 0.000	0.990 / 0.942	0.675 / 0.496	0.011 / 0.064
SAFE-MLP	HC	0.000 / 0.000	0.990 / 0.942	0.498 / 0.491	0.003 / 0.018
SAFE-LSTM	M	0.776 / 0.600	0.990 / 0.962	0.885 / 0.800	0.007 / 0.036
SAFE-LSTM	HC	0.541 / 0.316	0.990 / 0.962	0.769 / 0.658	0.003 / 0.018
ActProbe	M	0.286 / 0.000	0.980 / 0.911	0.639 / 0.498	0.009 / 0.045
ActProbe	HC	0.043 / 0.000	0.980 / 0.911	0.520 / 0.498	0.003 / 0.009
kNN distance	M	0.314 / 0.000	0.937 / 0.782	0.654 / 0.500	0.006 / 0.064
kNN distance	HC	0.102 / 0.000	0.937 / 0.782	0.549 / 0.500	0.003 / 0.036
SAFE-MLP	own rule (band (pooled))	0.984 / –	–	–	0.076 / **0.446**
SAFE-LSTM	own rule (band (pooled))	0.977 / –	–	–	0.094 / **0.446**
ActProbe	own rule (native)	0.969 / –	–	–	0.097 / **0.373**