Who Detects What? Budgeted Detection of Conditional Mechanism Shifts in Frozen-Policy RL Streams

Anonymous authors
Paper under double-blind review

Abstract

Reinforcement learning policies are increasingly deployed as frozen components, yet their environments can change beneath a stable surface: observation marginals remain approximately unchanged while a conditional reward or transition mechanism shifts. We study deployment-time detection: given a frozen-policy stream with a pre-registered episode-level false-alarm budget ($\alpha = 0.05$), which detector signal families actually detect which mechanism changes, and at what minimum injected intensity? In two MiniGrid environments we construct controlled mechanism injections—transition slip (seven intensities, a small-marginal-shift variant of the premise) and reward flip (four probabilities, the stable-marginal condition exactly)—verified by independent probes, and compare five detector families in a frozen 165-run matrix. Three findings emerge: (i) detection is *mechanism-specific*—marginal- and trajectory-statistics detectors are blind to reward flips at all intensities (for the marginal scores entailed by construction, for the joint kNN realized empirically), while the conditional discriminator, joint-most-sensitive on slip, is near-blind to weak flips that the value residual detects perfectly; (ii) an episode-stream circular-shift calibration is required for honest false-alarm control, as window-score-sequence shifting is provably a no-op; (iii) batch-style AUROC rankings do not substitute for budgeted deployment-level detection rates ($\rho = 0.885$; one persisted-score seat). A pre-registered hold-out on two unseen intensity levels passes all frozen criteria, replicating the specificity structure, as does a pre-registered cross-cell confirmation of the flip mechanism on the second environment; we will release all streams, manifests, per-window artifacts (persisted-score seats), and the pre-registered protocol.

1 Introduction

Setting and stakes. Reinforcement learning agents are routinely trained once and deployed frozen, operating on streams they can no longer learn from; temporal quality degradation is the norm even in standard supervised deployments, afflicting 91% of model–dataset pairs over time (Vela et al., 2022). The classical monitoring assumption—that drift will announce itself through changed input statistics—fails for a growing class of environment changes: a transition probability that slips only in rarely visited state–action corners, or a reward that flips sign at successful termination, can leave the distribution of raw observations essentially untouched while systematically corrupting the agent's objective. We call this setting **marginal-stable, conditional-mechanism shift** and study it from the deployment side: the policy is frozen, no drift labels arrive, and the monitor must respect a pre-registered episode-level false-alarm budget. The learning-side counterpart has been studied (Wang et al., 2025); we address the deployment-side detection gap.

Our approach. We build a budgeted detection protocol around *frozen-reference signals*: a family of frozen references—each capturing a different statistical channel (observation marginal, action sequence, transition kernel, conditional discrepancy between (s, a) and (s', r), and value residual)—is calibrated exclusively on a dedicated no-drift stream, with thresholds set under a fixed episode-level FAR budget ($\alpha = 0.05$). Mechanism injections

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.

are verified by independent probes—marginal tests, realized replacement rates, and fixed-policy return impact—so that detection evidence never doubles as its own ground truth. The comparison matrix—(7 slip + 4 flip levels) × 3 seeds × 5 families = 165 runs across two environments—was frozen before data generation, and a pre-registered hold-out on two excluded intensity levels completes the design.

Contributions.

- **C-Protocol.** To our knowledge, no existing protocol combines (i) RL deployment-time streams with frozen-policy reference signals, (ii) controlled mechanism-injection ground truth verified by independent probes, and (iii) pre-registered episode-level false-alarm budgets across a detector family; we provide such a protocol and will release all streams, manifests, and per-window artifacts (persisted for the discriminator seat; §5).
- **C-Specificity.** A controlled demonstration that detector × mechanism specificity is structural in this setting: five signal families × two mechanisms separate into complementary coverage, with pre-registered expected failures realized exactly, and the structure replicated on two held-out intensity levels under a pre-registered gate.
- **C-Calibration.** The episode-stream circular-shift calibration principle for windowed detectors on autocorrelated episode streams, with theoretical grounding in permutation-test error theory and an empirical demonstration that the sequence-level alternative is weak.
- **C-Thresholds.** Detection thresholds $\hat{\varepsilon}$—the minimum injected intensity detectable under a fixed false-alarm budget, with bootstrap CIs and no level interpolation—as a replacement for unconstrained ranking claims.

Scope and honesty. All claims are established in a controlled two-environment, two-mechanism setting with three policy seeds—a boundary set by design, since probe-verified mechanism-level ground truth requires controlled injection; generalization beyond it is future work. Negative and boundary results are reported rather than filtered (§5); we claim no new detector algorithm—the contribution is the protocol, the calibration principle, and the measured specificity structure.

2 Related Work

Drift detection for non-stationary decision processes. Detection here has largely targeted mean reward shifts: PS-RMAB equips restless bandits with mean-shift detectors whose guarantees require a known shift-magnitude lower bound (Li et al., 2026)—a condition marginal-stable, conditional-mechanism shifts violate by construction. An edit-operation study measures Levenshtein/Damerau/DTW distances between baseline and drift episode sequences with a Welch t-test layer (Lee & Shim, 2025)—our B2 signal family, but a single batch-tested detector without mechanism-specificity analysis, marginal-stability control, or a calibrated false-alarm budget. The OOD-RL-Bench suite compares eight detectors under nine anomaly injectors (Mittag et al., 2026), scored by matched-time ranking metrics without numeric-level false-alarm budgets; its top-ranked detector by PEDM exhibits the worst detection latency, hinting that ranking quality and deployment alarm behavior decouple—a decoupling our batch-vs-deployment analysis quantifies directly ($\rho = 0.885$, B5/slip). Pointwise monitors (Zhang et al., 2024; Fang & Zdun, 2024) detect pointwise OOD states or learned-model misfit, a different problem level from stream-level mechanism change under a fixed policy; the DRL-side BehaviorGuard flags trigger-agnostic backdoor traces in the action distribution at deployment (Yu et al., 2026), but without frozen-reference conditional-mechanism nulls, episode-level false-alarm budgets, or marginal-stability controls. An adjacent pre-registered 800-cell factorial over deployed safety classifiers reports detection difficulty dominated by a classifier×shift interaction ($\eta^2 = 0.185$) under empirically calibrated thresholds (Leong, 2026)—mirroring, outside decision streams, our budgeted detector×mechanism specificity. A KDD benchmarking framework standardizes Monte Carlo drift injection and timing-aware metrics across fourteen stream detectors (Cerqueira et al., 2026), but leaves open frozen-policy decision streams, marginal-stability controls, episode-level false-alarm budgets, and mechanism-family attribution.

Conditional drift: learning-side robustness vs. detection-side budgets. Closest in problem class, Wang et al. (2025) study policy *learning* under concept drifts with preserved covariate marginals, establishing that this regime differs fundamentally from joint-distribution shift; on the theory side, convergence analyses of switching non-stationary MDPs show that a single policy trained across hidden-mode switches converges only to the stationary-distribution-weighted *average-environment* optimum (Amiri & Magnússon, 2026)—a structural suboptimality that presumes, but does not provide, the ability to know *when* the mechanism has changed, which is precisely the deployment-time detection question we isolate; WATCH (Prinster et al., 2025) operationalizes the complementary dichotomy online via weighted conformal test martingales. We address the **deployment-time detection** counterpart for frozen-policy decision streams; neither line answers the evaluation question we target—WATCH is a single adaptive monitor rather than a controlled comparison of signal families, and neither work reports at which injected intensity a detector stops being useful under a fixed budget. Outside decision streams, forecasting work separates concept drift from temporal shift (Zhao et al., 2025) and Model-X change-point detection targets conditional change given covariates (Huang et al., 2025); both are non-decision-stream settings and neither budgets deployment-level false alarms. The causal literature formalizes precisely the conditional-vs-marginal separation our setting relies on: the sparse-mechanism-shift hypothesis casts shift as few changing causal conditionals distinct from per-node marginal tests (Perry et al., 2022), a kernel/copula statistic is provably zero when the causal dependence is intact and strictly positive only if it changes, insensitive to joint or marginal variation (Gavioli-Akilagun et al., 2026), and dataset-shift diagnostics test conditional shifts $P_{Y|X}$ unchanged while explicitly allowing covariate marginals P_X to vary (Maia Polo et al., 2023)—anchoring the mechanism-specific detection our comparison targets, though all are defined outside decision streams and lack deployment-level false-alarm budgets, which we address via a budgeted detection-protocol comparison (cf. the budgeted e-detector toolkit of Saha & Ramdas, 2026).

Residual, conformal, and stream-level calibration. Domberg & Schildbach (2026) use world-model prediction residuals for OOD detection in online continual RL; our residual seat (C1) differs in calibration discipline rather than signal family—residuals standardized on a dedicated calibration stream, empirical-null thresholds with episode-stream circular-shift correction, and a pre-registered FAR budget, controls absent from rolling-threshold designs. Katz et al. (2026) calibrate preemption thresholds to bound the *unpreempted-failure* rate; our calibration targets the complementary quantity—false alarms under no-drift—and the two guarantees do not substitute for each other. Anytime-valid tests for streaming shift (Khan & Syed, 2025) rest on exchangeability assumptions that episode streams with overlapping windows violate, which is why our protocol calibrates on the deployment stream itself; fixed-reference conditional conformal test martingales (CCTM) restore anytime-valid type-I control under a held-out calibration set (Shaer et al., 2026)—sharing our frozen-calibration discipline, but as a single proposed monitor rather than a budgeted comparison across signal families; conformal two-sample testing (Bansal et al., 2025) supports our discriminator seat's graceful-degradation design. Our own implementation audit illustrates the failure mode: a pre-repair sequential monitor alarmed on every window of clean streams, and the repaired build still exhibited naive FAR departing from the nominal level (§4); independent measurement likewise reports empirical false-alarm rates of deployed drift detectors fluctuating away from nominal batch levels in continuous monitoring (Singh, 2026); a concurrent conformal-martingale monitor fires on all 135 clean runs of real forecast streams—independent contemporaneous evidence that deployment streams violate exchangeability and demand explicit null calibration (Han & Qu, 2026).

Sequential testing of conditional discrepancy. CJSD estimates a conditional Jensen–Shannon discrepancy as the held-out cross-entropy gap of two discriminators, with a covariate-null theorem, a drift-mass law, and misspecification control (Props. 1–4) (Oda, 2026). We adapt this measurement principle to RL transition tuples, $X = (s, a)$ and $Y = (s', r)$, where three structural differences arise: actions are policy-generated (endogenous covariates); the reward channel enters Y and may change alone—a case our experiments show CJSD detects only weakly at full intensity (0.533 at q=1.0) while value residuals detect it everywhere ($\hat{\varepsilon}$=0.1); and evaluation operates on episode streams under window-level

FAR calibration rather than pooled two-dataset discrimination. CJSD's dimension analysis (kNN plug-in CMI fails beyond d=64; the discriminator route remains stable to d=256) is independently replicated by our ordering. Ensemble composition is an established theme: TSB-AD finds rankings vary across domains (Liu & Paparrizos, 2024); engineering practice combines detectors heuristically, e.g. a simultaneous-alarm AND rule without joint false-alarm calibration (Weiß et al., 2026); and conformal test martingales can be provably blind to arbitrarily large changes when the conformity measure is misaligned with the change type (Hallberg Szabadváry, 2026). Our contribution is not a new combination rule but a **mechanism-level explanation** of why composition is necessary here: each injection mechanism alters a different conditional statistic, so coverage gaps are structural rather than empirical accidents (§5).

3 Problem Setting

3.1 Streams, Policy, and Isolation

Let the deployment environment be an implicit-parameter decision process $\mathcal{M}_e = (\mathcal{S}, \mathcal{A}, P_e, R_e, O_e, G_e, \gamma)$, where the mechanism parameter e modulates reward and transition conditionals. The behavior policy π is **frozen**: a PPO checkpoint (Stable-Baselines3; Schulman et al., 2017; Raffin et al., 2021) whose parameters are fixed before deployment and never updated from the deployment stream, admitted only after passing a pre-registered convergence criterion (per-seat values in the reproducibility package). Streams use deterministic (greedy) rollout of the frozen policy, removing action-sampling noise from the null structure; a companion tension check (§5.6) verifies stochastic robustness.

All evidence is organized around **three isolated streams** per (environment, policy seed): a **calibration stream** ($n_{\text{cal}} = 500$ episodes, dedicated seed range)—the *only* source for observation-marginal histograms, null distributions, residual standardization, and alarm thresholds; a **validation stream** ($n_{\text{val}} = 300$)—a no-drift stream reused with the same thresholds for false-alarm cross-confirmation; and **drift test streams** (300 episodes per cell, disjoint seed range)—mechanism-injected streams on which detection evidence is recorded. Streams are never pooled: calibration thresholds and standardization statistics are frozen before any test stream is scored, and no threshold, window, or drift level is selected on test data.

3.2 Edge-Stability Condition

The setting posits that raw observation marginals remain approximately stable while conditional mechanisms change. We operationalize this as a pre-registered acceptance test: the pooled per-channel observation histogram of the drift test stream must satisfy

$$d_{\text{JSD}}\big(p_{\text{test}}(o),\, p_{\text{cal}}(o)\big) \;\le\; q_{0.95}\big(\mathcal{D}_{\text{null}}\big), \tag{1}$$

where $\mathcal{D}_{\text{null}}$ is constructed by two bootstrap resamples of the calibration stream (sizes n_{cal} and n_{test}), preserving the finite-sample JSD positive bias. The validation stream is scored against the same threshold; failure flags calibration doubt and blocks acceptance.

Identification assumption and failure boundary. The test assumes that histogram-JSD at the registered binning resolution separates marginal changes. If it fires, either the injection perturbed the marginal (design iteration, at most two rounds) or calibration sample size is insufficient (one re-run with larger n_{cal}); both outcomes are logged distinctly, and neither is reported as "mechanism not constructible" without this audit. Edge-stability is *not* a claim that drift is unobservable from trajectories—it only fixes the marginal axis along which our detector families differ.

Realized outcome. The two mechanism axes realize this condition differently, and both outcomes are part of the registered evidence. Reward-flip streams pass: under the frozen deterministic policy the observation stream is structurally invariant to reward rewriting, and the edge-stability probe records $\text{JSD}_{\text{obs}} = 0.0$ against a null q_{95} of 0.0. Transition-slip streams fail at every attempted intensity: across the transition-slip identification ladder ($p \in \{0.3, 0.15, 0.075\}$, 9/9 rows) the observed JSD exceeds the null q_{95} by 2.8×–254×. Slip cells are therefore *not* marginal-stable; we retain them as *small-marginal-shift* cells, asking

whether any detector family can exploit a marginal signal this small under the calibration budget. The marginal-stable premise is realized by the reward-flip axis.

3.3 Mechanism Injection and Independent Probes

Injections modify only the conditional layer, leaving the observation trajectory structurally unchanged. **Transition slip** (DoorKey-5x5, `ConditionalSlipWrapper`): when the agent-centric view shows forward cell and left cell both walls, `forward` is redirected to `left` with probability p; the slip RNG is seeded independently of the reset stream, so mechanism randomness never reuses environment randomness. **Reward flip** (Empty-8x8, `HiddenRewardWrapper`): the successful-termination reward is replaced with -1 with per-episode probability q, drawn from an independent RNG; for $q = 1.0$ no random draws are consumed (bit-compatible with the original edge-stability probe semantics). Both environments are from the MiniGrid library (Chevalier-Boisvert et al., 2023). Whether an injection "took" is decided by **independent probes, not by the detectors under test**: (i) edge stability (§3); (ii) a mechanism probe—offline replay statistics verifying that the realized replacement rate falls within ± 0.10 of the nominal conditional probability; (iii) **fixed-policy return impact**—an independent-sample comparison of pre/post-injection returns with pre-registered acceptance ($p < 0.01$, $|d| > 0.8$). If (i)–(ii) hold but returns are unaffected, the cell is registered as a *non-identifiable case*—a negative result that is kept, not deleted. For transition slip only, a restricted intensity ladder (initial $p = 0.3$; two pre-registered downward revisions $0.15 \to 0.075$, each revision itself pre-registered and fully logged) distinguishes principled iteration from any-intensity scanning: an exhausted ladder terminates the cell as a negative result—no stronger detector is swapped in (pre-registered stopping rule).

4 Budgeted Detection with Frozen-Reference Signals

4.1 Two Frozen Roles

Two distinct components are frozen—conflating them is the central pitfall of residual monitoring: the **frozen behavior policy** π generates the trajectory stream; the **frozen reference predictors** (B5/C1 components) generate residual signals. The reference model is trained only on the calibration stream, never updated at deployment, and architecturally separate from π—otherwise residuals would measure policy mismatch, not environment drift.

4.2 Detector Families

All scores are computed per episode, converted to null distributions on the calibration stream, and compared as window statistics against calibrated thresholds. We deliberately sample five *signal families* rather than five estimators of one signal:

- **B1 (state-marginal JSD, windowed)**: mean per-dimension JSD between window observation marginals and calibration marginals—the observation-marginal channel.
- **B2 (action-sequence edit distance)**: mean normalized Levenshtein distance from each test action sequence to the **deduplicated** reference set of unique action sequences. Deduplication is a correctness fix: the 500-episode DoorKey reference set contains only 48 unique trajectories, under which a min-distance statistic degenerates to a zero-variance null (diagnosed in audit tools to be released with the code).
- **B4 (transition-kernel kNN)**: kNN distance in the joint $(X, Y) = ((s, a), (s', r))$ space—the transition-conditional channel.
- **B5 (conditional discrepancy, CJSD-style)**: $d_{\mathrm{CJS}} = (\text{T1 CE} - \mathrm{CE}_{\mathrm{heldout}}) - \text{T2 CE}$, where T1 is a conditional predictor of Y given X (its held-out calibration error is the reference baseline) and T2 is a joint discriminator on (X, Y); both are lightweight MLPs averaged over a **5-seed ensemble**, fit on calibration tuples only with fixed inference RNG. Sequential separation of covariate and mechanism shift has recently been formalized for supervised data streams (Oda, 2026); we **adapt the conditional-discrepancy measurement principle to RL transition tuples** $X = (s, a)$, $Y = (s', r)$—where the action channel, the reward channel, and episode structure change the hypothesis

structure—and evaluate it under a pre-registered detector pool rather than as a standalone proposal.
- **C1 (value residual)**: residual of a frozen value predictor regressed on $s \to$ return, computed per episode. The reward channel enters this signal through the return–value path even when trajectories are unchanged—the only flip-effective channel in our pool.

Predictive-skill check. Reference predictors must show measurable gains over trivial baselines, else null variance swamps signal; the formal matrix verifies this implicitly (33/33 cells with usable null calibrations; the single B1/one-seed failure is boundary-characterized in §5).

Aggregation. Episode scores → sliding-window mean (window = 100, stride = 50) → window statistics vs. threshold → window alarms; where defined (B2/B4/C1), episode-level rates are reported in parallel. Sequential interfaces (CUSUM / Page-Hinkley) exist in the codebase but were pre-registered *out* of the main matrix (seat B3 demoted to an expected-failure appendix).

Identification assumption and failure boundary. Under edge stability, a conditional mechanism change must alter some conditional predictive relation—*some* family should respond, but which one is decided by its signal channel. The failure boundary is empirical and pre-registered: a mechanism that alters no family's channel is unidentifiable—B1/B2/B4's zero detection at all flip intensities is exactly this boundary realized as an expected-failure result.

4.3 Calibration under a Fixed False-Alarm Budget

Every detector operates under the **same pre-registered false-alarm budget** $\alpha = 0.05$, defined nominally at the episode level and operationalized per test window (thresholds are set on calibration-window statistics and FAR is accounted over the per-cell test windows; §5); comparisons are detection rates at fixed FAR, not ranking metrics. Calibration proceeds in four frozen steps: (i) score standardization on the calibration stream; (ii) naive threshold = q95 of calibration-stream window statistics; (iii) **episode-stream circular-shift calibration**—the null *episode-score sequence* is randomly rotated $n_{\text{circ}} = 200$ times (seed 20260908), window statistics are **recomputed** on each rotated stream and pooled ($200 \times 9 = 1800$ samples) before taking q95; (iv) every cell reports the FAR triple (nominal α / naive empirical / calibrated empirical).

Why the stream-level shift. A tempting cheaper alternative is to circularly shift the *window-score sequence* and reuse precomputed window statistics. This is a no-op calibration: shifting preserves the multiset of window scores, so the pooled quantile inherits the same autocorrelated arrangement and fails to control FAR (Propositions 1–2, Appendix G)—consistent with permutation-test theory on level maintenance under autocorrelated nulls (Romano & Tirlea, 2022) and with surrogate-method studies showing that the surrogate choice materially changes error rates (Cannistra et al., 2025). Rotating at the *episode* level changes episode boundaries and forces window recomputation, breaking multiset invariance; in the formal matrix this moved the most sensitive seat into band.

Failure handling. A cell whose calibrated FAR leaves the band ($|\text{FAR} - \alpha| > 2\alpha$ or $\text{FAR} > 3\alpha$) is flagged `calibration_failed` and boundary-characterized, never silently dropped (7/165 rows, one seed of one detector; Appendix B). If calibration sample size is insufficient, n_{cal} is increased and recalibration is logged—never a lowered budget.

4.4 Comparison Protocol and Detection Thresholds

The comparison matrix was frozen before data generation: (7 slip levels—including two extrapolation levels, $p \in \{0.0375, 0.01875\}$, outside the identification ladder—+ 4 flip levels) $\times$ 3 policy seeds $\times$ 5 detector families = 165 runs across the two environments; 300 drift episodes per cell; 500 calibration episodes per (experiment, seed). Beyond matched-time ranking metrics (Mittag et al., 2026), this protocol enforces a **fixed deployment-level FAR budget** and pre-registered detection-threshold reporting—turning ranking claims into budgeted detection claims.

Detection threshold $\hat{\varepsilon}$. For each (detector, mechanism): the lowest intensity at which the pooled window-alarm rate (3 seeds × 5 test windows per level) significantly exceeds the empirical calibrated null FAR (one-sided binomial test, $p < 0.05$; bootstrap-1000 CI; **no interpolation between levels**). Levels whose seed-level effects fall below the pre-registered MDE ($|d| \geq 0.3$, $\alpha = 0.01$, power ≥ 0.8) are registered as detection limits without significance claims (§5 details the screening-versus-MDE discipline). Descending-scan tolerates non-monotonicity by returning the smallest significant level; non-significant levels are registered as boundaries.

5 Experiments

5.1 Detection-Rate Matrix

Table 1 reports window-level calibrated detection rates (mean ± SD, 3 seeds). Under transition slip, B1, B2, and B5 are perfect from $p = 0.075$ upward and C1 from $p = 0.15$, while B4 never reaches perfect detection below $p = 0.7$; at the two weakest levels B1/B5 lead, B2 follows, and B4/C1 swap ranks. Under reward flip, the matrix separates structurally: B1, B2, and B4 register **zero detection at all four intensities, including** $q = 1.0$, exactly matching the pre-registered expected-failure hypothesis; B5 detects only the strongest level ($q = 1.0$: 0.533 ± 0.094) and is silent at $q \leq 0.5$; C1 alone detects at every level (1.000 ± 0.000, including the weakest $q = 0.1$). Episode-level rates, where defined, are directionally consistent (B1/B5 have no episode-level definition by construction); the full 165-run table with complete FAR triples will be released as CSV.

Table 1: Window-level calibrated detection rate (mean ± SD, $n = 3$ seeds); FAR(cal) = mean calibrated FAR on drift cells (including the seven `calibration_failed` runs); $\hat{\varepsilon}$ = detection threshold, bootstrap-95% CI (n/a = no significant level). Slip shows five of seven pre-registered levels ($p \in \{0.3, 0.5\}$ omitted; full 165-run table to be released with the code); the §3 ladder refers to the initial identification phase.

(a) Transition slip (DoorKey-5x5).

Detector	p=0.7	p=0.15	p=0.075	p=0.0375	p=0.01875	FAR(cal)	$\hat{\varepsilon}$ [CI95]
B1 (JSD)	1.000±0.000	1.000±0.000	1.000±0.000	0.733±0.189	0.467±0.189	0.074±0.105	0.01875 [.019,.038]
B2 (edit)	1.000±0.000	1.000±0.000	1.000±0.000	0.667±0.249	0.400±0.163	0.111±0.000	0.01875 [.019,.038]
B4 (kNN)	1.000±0.000	0.867±0.189	0.600±0.163	0.400±0.000	0.133±0.189	0.111±0.000	0.0375 [.019,.15]
B5 (CJSD)	1.000±0.000	1.000±0.000	1.000±0.000	0.733±0.249	0.467±0.094	0.074±0.052	0.01875 [.019,.038]
C1 (resid.)	1.000±0.000	1.000±0.000	0.600±0.163	0.333±0.094	0.267±0.189	0.111±0.000	0.0375 [.019,.075]

(b) Reward flip (Empty-8x8).

Detector	q=1.0	q=0.5	q=0.25	q=0.1	FAR(cal)	$\hat{\varepsilon}$ [CI95]
B1 (JSD)	0.000	0.000	0.000	0.000	0.000±0.000	n/a
B2 (edit)	0.000	0.000	0.000	0.000	0.000±0.000	n/a
B4 (kNN)	0.000	0.000	0.000	0.000	0.000±0.000	n/a
B5 (CJSD)	0.533±0.094	0.000	0.000	0.000	0.000±0.000	1.0 [1.0,1.0]
C1 (resid.)	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	0.000±0.000	0.1 [0.1,0.1]

5.2 Detection Thresholds $\hat{\varepsilon}$ and Mechanism Specificity

On slip, B1/B2/B5 are jointly most sensitive ($\hat{\varepsilon} = 0.01875$), with B4 and C1 at 0.0375. On flip, $\hat{\varepsilon}$ is undefined (n/a, no significant level) for B1/B2/B4; B5 is significant only at $q = 1.0$; C1 attains $\hat{\varepsilon} = 0.1$ with the CI collapsed to $[0.1, 0.1]$.

$\hat{\varepsilon}$ quantifies budgeted detection capability (§4): the smallest injected intensity at which a detector remains useful under a fixed false-alarm budget. The pooled binomial test (15 window indicators per level: 3 seeds × 5 test windows) treats overlapping indicators as exchangeable and is therefore a screening device, not a valid significance test; the pre-registered seed-level MDE gate is the primary discipline. Seed-level results at the 0.01875 slip level that fall below the MDE are registered as detection limits (B4: 0.133 ± 0.189; C1: 0.267 ± 0.189, $p = 0.076$)—no seed-level significance claims for these cells.

The matrix exhibits a structural specificity pattern (Figure 2; App. F), organized as three evidence chains:

1. **Observation- and trajectory-statistics detectors are blind to reward-channel changes in our setting.** For B1/B2 the blindness is entailed by score construction: reward flip does not alter the observation marginals or the (s, a, s') trajectories their scores depend on. For B4, whose kNN input does include r, blindness is empirical: the perturbation touches only terminal rewards of successful episodes, diluted in the joint distance by the state dimensions.
2. **The conditional-discrepancy detector is mechanism-asymmetric.** B5 is joint-most-sensitive on slip ($\hat{\varepsilon} = 0.01875$) yet nearly blind to reward flip (only $q = 1.0$: 0.533): the discriminator couples $X = (s, a)$ to $Y = (s', r)$, and a flip that alters r only at successful terminations perturbs the joint statistics too weakly at $q \leq 0.5$.
3. **The value-residual detector is complementary.** C1 attains $\hat{\varepsilon} = 0.1$ on flip (full detection at all levels) but is secondary on slip ($\hat{\varepsilon} = 0.0375$): the reward channel enters residuals through the return–value path even when trajectories are unchanged.

Together: **no single detector family dominates across mechanism types** in this controlled setting—a property of these five families under two mechanisms, not a universal theorem. Mechanism and environment are not crossed in the original matrix, so specificity is entangled with environment identity; a pre-registered cross-cell confirmation (flip on DoorKey, zero recalibration) replicated the structure (C1 at the same $\hat{\varepsilon} = 0.1$ floor, trajectory-channel seats blind; Appendix D).

5.3 Oracle Combination: B5 ∪ C1

We evaluate the pre-registered union rule (a window alarms if either detector's own calibrated threshold fires) on the B5–C1 pair, the two mechanisms' winners. The union shows **zero complementarity loss** in all 33 cells—equal to max(single-detector) everywhere: full flip recovery where B5 is blind ($q \leq 0.5$: union 1.000) and 1.000 held on slip at $p \geq 0.075$. **Honest FAR accounting**: the union reuses each detector's independently calibrated threshold without a paired joint null, so $\text{FAR}_{\text{union}}$ is reported only as a Bonferroni upper bound (slip: $0.185 = 0.074 + 0.111$; flip: 0.000), which exceeds $\alpha = 0.05$ on slip and is therefore *not* used for ranking or gate decisions; an empirically paired joint null is now established on the flip×DoorKey cell in a formal pre-registered experiment (Appendix D.8). Recomputation reproduces all 33 formal values exactly (a fail-closed determinism check). These provide direct constructive evidence for mechanism-conditional composition, subject to the §5 scope limitation.

5.4 Calibration Fidelity and Batch-vs-Deployment Consistency

Under episode-stream circular-shift calibration, B5—the most sensitive seat—achieves FAR(cal) 0.074 ± 0.052 on DoorKey with **21/21 cells inside the $\alpha \pm 2\alpha$ band** (0.000 on Empty); B2/B4/C1 are in-band. Seven of 165 runs trigger the pre-registered `calibration_failed` flag (all B1/DoorKey/seed_1): boundary-characterized, excluded from ranking, $\hat{\varepsilon}$ unchanged when that seed is dropped (App. B). For B5, the only seat with persisted per-window scores, batch AUROC ranks align strongly but imperfectly with deployment detection rates: $\rho = 0.885$ (17/21 concordant) on slip, $\rho = 0.775$ (3/6) on flip; other seats: `n/a` (scores not persisted).

5.5 Negative Results, Boundary Cases, and Limitations

The honesty ledger comprises five items: (i) optimistic summary-layer $\hat{\varepsilon}$ CIs (per-window scores persisted for B5 only; cross-cell persists all seats); (ii) the 1/9 naive-FAR resolution floor; (iii) the B1 seed_1 calibration failure; (iv) scope limited to 2 environments × 2 mechanisms × 5 families × 3 seeds, with off-diagonal cells and a third environment family crossed in pre-registered confirmations; and (v) pre-registered expected-failure baselines (B3 demoted to a non-run attachment). The complete ledger is in App. F.1; no registered finding is omitted.

5.6 Rollout-Mode and Hold-Out Checks

Rollout-mode robustness. The main matrix uses deterministic (greedy) rollout; a pre-registered tension check verifies that the edge-JSD discrimination premise survives stochas-

tic deployment. Crossing three converged checkpoints with rollout mode {deterministic, stochastic} and mechanism {none, slip 0.3, goal-reward negative control}, all six units satisfy the pre-registered H1 rule ($G = \text{drift}_{q05} - \text{null}_{q95} > 0$; stochastic G from 3.6×10^{-5} to 3.2×10^{-3}) and the negative control stays clean. The check covers the marginal-JSD signal only; an exploration-track boundary case under an *unconverged* checkpoint did not reproduce (not upgraded; App. C); protocol in App. F.2.

Hold-out generalization. A pre-registered hold-out tests whether the specificity structure transfers to two unseen intensities (slip $p = 0.1$, flip $q = 0.75$; outside the registered ladders) on fresh streams (30 units, 300 episodes each; criteria frozen before data generation). The verdict is **PASS** (30/30): on the reward axis, C1 detects on all 3 seeds while B1/B2/B4 remain blind on every seed and B5 is silent at $q = 0.75$; on the transition axis, all four trajectory-window families detect at $p = 0.1$ with at least 2/3 seeds each (B1's missing seed is the characterized B1/DoorKey/seed_1 boundary unit). The realized slip-trigger rate (0.0996) matches the nominal level; the structure reads as intensity-general within our controlled setting (protocol and seed-level results in App. F.3).

6 Conclusion

For a frozen RL policy on a marginal-stable stream undergoing conditional-mechanism change, *which* detector families detect *which* mechanism under a pre-registered false-alarm budget? Our protocol (§3–4) yields three answers. First, **detection is mechanism-specific**. Observation- and trajectory-statistics detectors register zero detection on reward flips at every intensity; the CJSD-style discriminator is jointly most sensitive on slip ($\hat{\varepsilon} = 0.01875$) yet nearly blind to flips that the value residual detects everywhere ($\hat{\varepsilon} = 0.1$). This entails mechanism-conditional selection or composition: the pre-registered B5∪C1 union recovers detection power on both mechanisms with zero complementarity loss, replicated on the hold-out and both off-diagonal pairs (§5.3, App. D & D.8; union FAR is a Bonferroni-only bound, over budget on slip). Second, **calibration is not interchangeable**: episode-stream circular-shift calibration brings the most sensitive seat into the $\alpha \pm 2\alpha$ FAR band in all 21 cells; the cheaper window-score-sequence shift is provably a no-op. Third, **batch rankings are not deployment detections**: AUROC ordering correlates with, but does not substitute for, budgeted detection rates ($\rho = 0.885$ slip, 0.775 flip; single persisted-score seat)—arguing for detection-threshold reporting over unconstrained ranking claims.

Limitations and future work. Five registered limitations are boundary-treated in §5 and App. B–F. Three extensions: an anytime-valid counterpart of the B5∪C1 paired joint null (window scores are not generally e-processes (Dandapanthula & Ramdas, 2025; Halme & Koivunen, 2026)), persisting per-window scores for all seats, and richer environments, non-frozen policies, LLM-agent streams (Oderinwale, 2026; Wang, 2026).

Reproducibility Statement

All experiments follow a pre-registered protocol: hypotheses, mechanisms, seeds, detectors, statistical units, and stopping rules were fixed before data generation, and every result table maps to an experiment ID with a machine-readable manifest (configuration, seeds, code revision, checkpoint checksums, and wall-clock). Mechanism injections are verified by trigger-rate probes before any statistic is interpreted, and all detection thresholds are calibrated on independent null streams with episode-level false-alarm budgets. Failed and invalid runs are retained and reported. Code, episode streams, per-window scores (persisted for the discriminator seat; the remaining seats are re-derivable from the episode streams), and manifests will be released. The main text reports the three training seeds required by our protocol as mean ± SD over seeds; per-seed verdicts appear in the appendix (per-seed detection counts in the hold-out and calibration audits, App. F–B), and the complete per-seed result table (5 detectors × 11 levels × 3 seeds) will be released as CSV with the code.

References

Mohsen Amiri and Sindri Magnússon. Reinforcement learning in switching non-stationary MDPs: Algorithms and convergence analysis. In *IEEE Conference on Decision and Control (CDC)*, 2026.

Vansh Bansal, Tianyu Chen, and James G. Scott. Conformal C2ST: Turning weak classifiers into strong two-sample tests. *arXiv preprint arXiv:2507.17026*, 2025.

C. Cannistra, L. Hoang, A. E. Yuan, and W. Shou. Subtle methodological variations substantially impact correlation test results in ecological time series. *eLife Reviewed Preprint*, 2025. doi: 10.7554/eLife.108198.1. v1 2025-10-21.

Vitor Cerqueira, Heitor Murilo Gomes, Marco Heyden, Bernhard Pfahringer, and Albert Bifet. A framework for evaluating and benchmarking concept drift detection methods. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '26), Volume 2*, pp. 8671–8682, 2026. doi: 10.1145/3770855.3819070.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), New Orleans, LA, USA*, December 2023. Version 3.1.0 used.

Sanjit Dandapanthula and Aaditya Ramdas. Multiple testing in multi-stream sequential change detection. *arXiv preprint arXiv:2501.04130*, 2025.

Fabian Domberg and Georg Schildbach. Self-adapting robotic agents through online continual reinforcement learning with world model feedback. *arXiv preprint arXiv:2603.04029*, 2026.

Zhizhou Fang and Uwe Zdun. Detecting environment drift in reinforcement learning using a Gaussian process. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 992–999, 2024. doi: 10.1109/ICTAI62512.2024.00142.

Shakeel Gavioli-Akilagun, Kieran Wood, and Francesco Quinzan. Detecting changes in causal dependence with kernels and copulas. *arXiv preprint arXiv:2605.05809*, 2026.

Johan Hallberg Szabadváry. Conformal blindness: A note on a-cryptic change-points. *arXiv preprint arXiv:2601.01147*, 2026.

Topi Halme and Visa Koivunen. Multi-stream quickest change detection: Foundations and recent advances. *arXiv preprint arXiv:2604.18008*, 2026.

Weijia Han and Lisha Qu. When the martingale never stops firing: Anytime-valid gating on real forecast streams. In *E-Values: From Statistics to ML, Workshop at NeurIPS 2026*, 2026.

Yiwen Huang, Yan Dong, Mengying Yan, Ziye Tian, Chuan Hong, Doudou Zhou, and Molei Liu. Model-X change-point detection of conditional distribution. *arXiv preprint arXiv:2505.12023*, 2025.

Garrett Ethan Katz, Adebayo Braimah, Qinru Qiu, and Simon Khan. Conformal preemption of failures in sequential decision-making agents. *Reinforcement Learning Journal*, 7, August 2026. to appear.

Behraj Khan and Tahir Qasim Syed. E-SHIFT: Anytime-valid sequential hypothesis testing for distribution shift in streaming learning systems. *arXiv preprint arXiv:2510.03839*, 2025.

Chang-Hwan Lee and Alexander Shim. Detecting model drifts in non-stationary environment using edit operation measures. *arXiv preprint arXiv:2509.11367*, 2025.

Jun Wen Leong. Online shift detection and conformal adaptation for deployed safety classifiers. *arXiv preprint arXiv:2606.11949*, 2026.

Kuan-Ta Li, Chia-Chun Lin, Ping-Chun Hsieh, and Yu-Chih Huang. A modularized framework for piecewise-stationary restless bandits. *arXiv preprint arXiv:2604.10177*, 2026.

Qinghua Liu and John Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*, 2024.

Felipe Maia Polo, Rafael Izbicki, Evanildo Gomes Lacerda Jr, Juan Pablo Ibieta-Jimenez, and Renato Vicente. A unified framework for dataset shift diagnostics. *Information Sciences*, 2023.

Emil Mittag, Richard Dazeley, and Peter Vamplew. OOD-RL-Bench: A benchmark framework for out-of-distribution detection in reinforcement learning. *arXiv preprint arXiv:2607.12523*, 2026.

Kentaro Oda. Separating covariate shift from mechanism change with two discriminators: CJSD, a conditional discrepancy with an exact covariate-concept decomposition. *arXiv preprint arXiv:2608.19885*, 2026.

Hamidah Oderinwale. Agent trajectories as programs: fingerprinting and programming coding-agent behavior. *arXiv preprint arXiv:2606.16988*, 2026.

Ronan Perry, Julius von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022.

Drew Prinster, Xing Han, Anqi Liu, and Suchi Saria. WATCH: Adaptive monitoring for AI deployments via weighted-conformal martingales. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, 2025.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

Joseph P. Romano and Maria Tirlea. Permutation testing for dependence in time series. *Journal of Time Series Analysis*, 43(5):781–807, 2022. doi: 10.1111/jtsa.12638.

Aytijhya Saha and Aaditya Ramdas. Non-partitioned e-detectors for nonparametric sequential change detection. *arXiv preprint arXiv:2607.28322*, 2026.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shalev Shaer, Yarin Bar, Drew Prinster, and Yaniv Romano. Testing for distribution shifts with conditional conformal test martingales. In *Forty-third International Conference on Machine Learning*, 2026.

Raj Shekhar Singh. When drift detectors cry wolf: False alarm rates in continuous ML monitoring. In *ICLR 2026 Workshop on Continual Adaptive Optimization (CAO)*, 2026.

Daniel Vela, Andrew Sharp, Richard Zhang, Trang Nguyen, An Hoang, and Oleg S. Pianykh. Temporal quality degradation in AI models. *Scientific Reports*, 12:11654, 2022. doi: 10.1038/s41598-022-15245-z.

Chunxiao Wang. Nautilus compass: Black-box persona drift detection for production LLM agents. *arXiv preprint arXiv:2605.09863*, 2026.

Jingyuan Wang, Zhimei Ren, Ruohan Zhan, and Zhengyuan Zhou. Distributionally robust policy learning under concept drifts. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, pp. 64244–64281, 2025.

Matthias Weiß, Athreya Hosahalli Prakash, Maurice Artelt, Falk Dettinger, Nasser Jazdi, and Michael Weyrich. Self-adaptive anomaly detection with reinforcement learning and human feedback in connected vehicles. 2026.

Yinbo Yu, Xueyu Yin, Jiadai Wang, Chunwei Tian, Sai Xu, Qi Zhu, and Daoqiang Zhang. BehaviorGuard: Online backdoor defense for deep reinforcement learning. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, 2026.

Hongming Zhang, Ke Sun, Bo Xu, Linglong Kong, and Martin Müller. A distance-based anomaly detection framework for deep reinforcement learning. *Transactions on Machine Learning Research*, 2024.

Zhiyuan Zhao, Haoxin Liu, and B. Aditya Prakash. Tackling time-series forecasting generalization via mitigating concept drift. *arXiv preprint arXiv:2510.14814*, 2025.

AI Use Statement

Generative AI tools were used in the preparation of this work as follows. AI coding assistants assisted with implementing experimental infrastructure (mechanism-injection wrappers, stream-generation and detector pipelines), always under human-specified protocol designs, and with routine code documentation. AI research assistants supported literature search and citation metadata verification; every citation was independently checked against the original source before inclusion. AI writing assistance was used for copy-editing and LaTeX assembly; all scientific claims, hypotheses, experimental designs, and interpretations were specified by the authors, and all results derive from pre-registered protocols (§3) with frozen configurations, seeds, and decision rules. AI tools did not generate experimental data, did not select thresholds, seeds, or test streams, and were not used to write any result reported in this paper without corresponding machine-verifiable artifacts (streams, manifests, and per-window scores to be released with the code). The authors reviewed all AI-assisted contributions and take full responsibility for the integrity of the work and the final content.

A Expected-Failure Attachments

The comparison protocol pre-registered three detector families as *expected failures*—signal families whose score construction predicts blindness to one or both injected mechanisms—and committed to running (or deliberately demoting) them with their negative expectations registered before data generation. This appendix documents the registrations and their realization status, as required by the protocol's confirmation-bias controls.

A.1 Marginal and trajectory statistics on reward flips (B1/B2/B4). The pre-registered hypothesis chain (protocol §3.3) predicted that detectors scoring observation marginals (B1, state JSD) or trajectory statistics (B2, action-sequence edit distance; B4, transition-kernel kNN) cannot detect reward flips. For B1 and B2 this is entailed by score construction: a reward flip that leaves (s, a, s') trajectories unchanged alters none of the statistics their scores depend on. For B4 the registered expectation was empirical: the flip perturbs only terminal rewards, which the joint kNN distance was predicted to dilute below detectability against the state dimensions. Realization: all three detectors scored 0.000 ± 0.000 detection across all four flip intensities (Table 1b), exactly as registered.

A.2 CUSUM-family sequential monitoring (B3, demoted). A CUSUM-style sequential seat (B3) was demoted from the formal matrix to this attachment, per the pre-registered demotion rule, on the strength of a registered prior expectation: on an external benchmark (OOD-RL-Bench, Mittag et al., 2026), the CUSUM-family detector ranked at chance level (aggregate AUROC 0.495), and an ITRM-style residual detector showed weak performance in prior audits. B3 was therefore not run in the 165-run matrix; its demotion is registered here rather than silently omitted. No B3 numbers are reported from our own streams, and none are claimed.

A.3 Residual-detector weak expectation. The residual seat (C1) carried a registered prior expectation of *weak* performance on transition-type mechanisms (residuals are dominated by the return–value path, which slip perturbs only indirectly; the expectation follows the weak ITRM-style residual results registered in prior audits). Realization on the registered side: C1 was secondary on slip ($\hat{\varepsilon} = 0.0375$, vs. 0.01875 for B1/B2/B5), consistent with a weak-but-not-blind residual channel. The strong flip result (C1, $\hat{\varepsilon} = 0.1$) was *not* part of the expected-failure registration and stands as a positive finding rather than a confirmation; we keep the two directions separate to avoid post hoc reframing.

B Calibration-Failed Boundary Characterization

Of the 165 matrix runs, 7 triggered the pre-registered `calibration_failed` flag (the pre-registered rule: $|\text{FAR} - \alpha| > 2\alpha$ or $\text{FAR} > 3\alpha$). Per the frozen decision rules, these rows are boundary-characterized and excluded from detection-rate ranking; this appendix documents the failure structure and the robustness audit.

B.1 Failure structure. All 7 failed rows concentrate in a single seat–environment–seed cell family: **B1 / DoorKey-5x5 / seed_1**, across all 7 slip levels, with uniformly $\text{FAR}_{\text{cal}} = 0.222$. The uniformity across intensity levels identifies a *seed-level* calibration failure—a mismatch between B1's JSD score distribution on the seed_1 null stream and the pooled calibration threshold—rather than a level-dependent or mechanism-dependent artifact. Under the same calibration machinery, B5 reaches at most 0.111 (in-band) and B2/B4/C1 are entirely in-band (0.000 or 0.111) on the same null streams, which locates the failure in B1's score distribution on that particular seed rather than in the episode-stream circular-shift calibration itself.

B.2 Robustness audit. Dropping seed_1 leaves the B1 DoorKey conclusions unchanged: detection rates at $p = 0.01875$ are 0.600/0.600 (seeds 0/2) and 0.600/1.000 at $p = 0.0375$; the detection threshold $\hat{\varepsilon} = 0.01875$ is unaffected. We nevertheless retain the boundary characterization because, within-sample, we cannot exclude heavy-tailed JSD null distributions as a systematic (rather than seed-idiosyncratic) failure mode of marginal-statistics calibration.

B.3 In-band but conservative cells. On Empty-8x8, all detectors report $\text{FAR}_{\text{cal}} = 0.000$. This is in-band under the pre-registered rule ($|0 - 0.05| = 0.05 \leq 2\alpha = 0.1$) but conservative: the pooled calibrated threshold rests on $200 \times 9 = 1800$ shifted-window samples, so a zero-count empirical FAR under a nominal $\alpha = 0.05$ budget indicates no observed false alarms rather than a verified rate of exactly zero. On DoorKey, the 0.111 cells sit at the 1/9 naive granularity floor inherited from the 9 null-window calibration structure; rankings and boundary statements in the main text use the calibrated scale, whose effective sample size is the pooled 1800. Figure 1 visualizes the full FAR triple per detector.

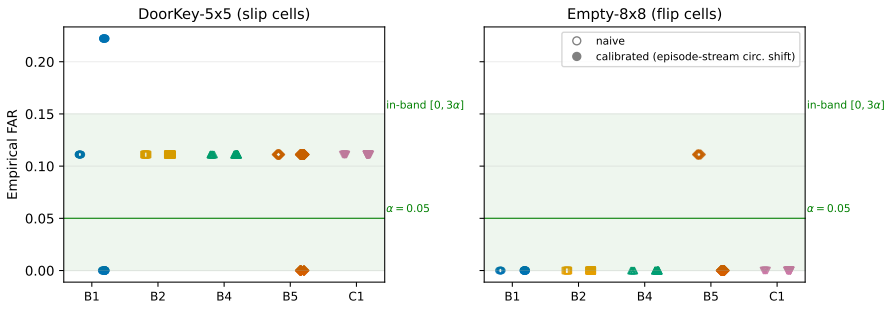

Figure 1: FAR triple per detector: nominal α vs. naive empirical (pinned at the 1/9 granularity floor) vs. episode-stream circular-shift calibrated. Shaded region: pre-registered in-band $[0, 3\alpha]$ (equivalently $|\text{FAR} - \alpha| \leq 2\alpha$ for $\text{FAR} \geq 0$); points above it are the seven B1/DoorKey/seed_1 boundary cells of §B.1, all sharing $\text{FAR}_{\text{cal}} = 0.222$ (one per slip level) and therefore overplotted as a single visible point.

C Entropy-Regularization Ladder: A Convergence-Confounded Boundary

This appendix boundary-characterizes an exploratory extension that asked whether the exploration-track observation—an entropy-regularized *unconverged* checkpoint compressing the marginal-JSD gap (Appendix F.2)—survives as a *training-side* effect: does raising the PPO entropy coefficient monotonically compress the discrimination gap G of the resulting policies? The verdict is **inconclusive** (convergence confound); per the frozen decision rules the data are reported as boundary characterization and do not enter the main evidence chain.

C.1 Design and verdict. Twelve seats: entropy coefficient $\in \{0, 0.005, 0.01, 0.02\} \times$ 3 seeds, each trained 768k steps on DoorKey-5x5 (the transition-slip identification ladder environment); the ladder protocol (rollout plan and analysis criteria) was frozen before any rollout data generation. Under the pre-registered convergence criterion, own-null (last-5-batch success-rate range $\leq 2\times$ the null $q95$ from fresh re-rollouts), only 4/12 seats converge, and 3 of the 4 are seed_1: convergence is dominated by policy identity (seed) rather than by the regularization level. The pre-registered ladder reading requires converged seats at every level; under this confound, no causal reading of the coefficient $\rightarrow G$ relation is licensed.

C.2 Criterion-validity observation. The training-time criterion (last-5 range < 0.05) and the own-null criterion disagree on seat identity in both directions: the only training-time pass (coefficient 0/seed_2, range 0.0278) is rejected by own-null ($2\times$ null $q95 = 0.0048$, a $\sim 10\times$ stricter effective scale), while four training-time failures (ranges 0.059–0.105) pass own-null. Both criteria are reported unchanged; the pre-registered verdict uses own-null. The disagreement is itself evidence on convergence-criterion validity in this regime, echoing the two-criterion tension documented since the exploration track.

C.3 Boundary data under the confound. The registered diagnostics are reported without upgrade. Mean policy entropy is strictly monotone in the coefficient ($0.017 \rightarrow 0.021 \rightarrow 0.026 \rightarrow 0.129$ nats; per-seed Spearman $\rho = [0.8, 1.0, 0.8]$), and the mean stochastic-mode gap decreases monotonically ($7.3 \times 10^{-4} \rightarrow 4.7 \times 10^{-4} \rightarrow 8.6 \times 10^{-5} \rightarrow -3.1 \times 10^{-4}$; $\rho = -1.0$). However, within-level seed dispersion spans one to two orders of magnitude (at coefficient 0: $[5.3 \times 10^{-5},\ 2.1 \times 10^{-3},\ 3.9 \times 10^{-5}]$; at 0.02: two seats at $\approx 5 \times 10^{-6}$ and one at -9.5×10^{-4}), so seed variance swallows the gradient signal and the monotone mean is outlier-driven. Quality control is otherwise clean: no degenerate seats (mean rollout success rate ≥ 0.99 per level; every seat ≥ 0.86), and the goal-reward negative control passes the registered criterion ($G_{\text{goal}} \leq 0.5\, G_{\text{slip}}$) on all 12 seats—with the caveat that the single coefficient 0.02/seed_2 seat shows a negative-control gap (-9.8×10^{-4}) at the same magnitude as its own (negative) slip gap, consistent with its convergence failure and with the criterion's reduced discriminating power once the slip gap itself is negative.

C.4 Reading against the deployment-side entropy manipulation. The deployment-side temperature sweep (H-null verdict) manipulated policy entropy at deployment time over 3.5 orders of magnitude and left the gap stable ($G \in [3.6 \times 10^{-3}, 5.2 \times 10^{-3}]$); the training-side ladder above is inconclusive. The two results form complementary boundaries of the same question: randomization *at deployment* does not degrade marginal-JSD discrimination on converged policies, while randomization *baked in during training* could not be separated from convergence failure at this scale (3 seeds $\times$ 768k steps). The training-side channel is thereby recorded as a limitation rather than a pending claim.

D Cross-Cell Confirmations

D.1 Motivation and Pre-registration

The original detection matrix (§5) crossed five detector families with two mechanisms but not with environments: transition slip was injected only on DoorKey-5x5 and reward flip only on Empty-8x8. Mechanism specificity and environment identity are therefore confounded at the matrix level—a limitation we registered as ledger item (iv). This appendix reports two pre-registered confirmation experiments, each frozen before any of its data

generation. The first cross-cell confirmation injects the reward-flip mechanism on DoorKey-5x5, crossing the previously unobserved cell; the flip wrapper semantics, intensity ladder ($q \in \{1.0, 0.5, 0.25, 0.1\}$), and three frozen checkpoints are identical to the original flip axis, and thresholds and calibration streams are reused from the DoorKey side of the main matrix with *zero recalibration*. The second cross-cell confirmation (§D.5) injects transition slip on Empty-8x8, crossing the remaining cell and completing the 2×2 mechanism–environment matrix.

D.2 Design and QC

Each of the 12 test streams (4 intensities $\times$ 3 policy seeds) contains 300 episodes; reset seeds are drawn from a fresh contiguous range $[700000, 700400)$, mechanically verified disjoint from all existing seed registries (verification recorded in the manifest). All five detector seats are scored with the byte-identical formal pipeline. QC before interpretation: (i) checkpoint SHA-256 identity for all three seats; (ii) mechanism probes—fixed-policy return impact ($p < 0.01$, $|d| > 0.8$) and per-episode termination-reward audit (flipped fraction $\approx q$, all 12 cells within 3 SE of nominal; q=1.0 exact); (iii) calibration-stream reuse legality (SHA-256 match with formal-pipeline artifacts); (iv) FAR reported as usual, calibration failures boundary-characterized. Unlike the original matrix, *all five seats persist per-window score arrays*, closing the seat-coverage gap noted in ledger item (i).

D.3 Results

The pre-registered structure replicates in full. **C1 (value residual)** detects the flip at every intensity: window-level calibrated detection 1.000 ± 0.000 at all four levels, $\hat{\varepsilon} = 0.1$ (bootstrap CI $[0.1, 0.1]$), matching the Empty-8x8 floor exactly; its episode-level rate tracks the injected flip fraction closely (0.981 ± 0.027, 0.501 ± 0.017, 0.266 ± 0.011, 0.152 ± 0.007 at $q = 1.0/0.5/0.25/0.1$), indicating that the residual alarm localizes to flipped episodes rather than firing indiscriminately. **B1/B2/B4 (observation-, trajectory-, and transition-channel seats)** register no significant detection at any intensity (B1: 0.200 ± 0.163, pooled $p = 0.095$, with 4/12 calibration-failure cells boundary-characterized as in the main matrix; B2: 0.000 ± 0.000; B4: 0.067 ± 0.094, $p = 0.83$)—the predicted constructive blindness. **B5 (conditional discrepancy)** fires only at the strongest intensity (0.267 ± 0.249 at q=1.0, $p = 0.021$; 0.200 ± 0.163, n.s., at $q \leq 0.5$), $\hat{\varepsilon} = 1.0$, the weak-detection pattern predicted from the Empty side. Mean calibrated FAR on drift cells: 0.074 (B1, B5), 0.111 (B2, B4, C1) at nominal $\alpha = 0.05$, consistent with the main-matrix DoorKey calibration.

D.4 Interpretation

The mechanism-specificity structure replicates across environments: the value-residual seat detects the flip at the same intensity floor as on Empty-8x8, while the trajectory-channel seats remain blind—supporting the environment-invariant reading of the specificity claim. Secondary observations retain their pre-registered status: the weak B5 response at q=1.0 is weaker in magnitude than on Empty-8x8 (0.267 vs. 0.533 window rate), and B1's nonzero but non-significant window rate accompanies its calibration-failure boundary cells; neither alters the pre-registered verdict, and both are logged in per-cell artifacts to be released with the code.

D.5 Second Cross-Cell: Transition Slip on Empty-8x8

The second confirmation injects transition slip on Empty-8x8 (`open_forward` trigger, parameters identical to the main-matrix slip axis), completing the cross. The registered seven-level slip ladder ($p \in \{0.7, 0.5, 0.3, 0.15, 0.075, 0.0375, 0.01875\}$) is reused verbatim, so DoorKey-vs-Empty floors are directly comparable on the same ladder; the design comprises 105 units (5 seats $\times$ 7 levels $\times$ 3 seeds) over 300-episode streams, with reset seeds drawn from a fresh range $[710000, 710300)$ (registered zone $[710000, 710400)$, mechanically verified disjoint from all seed registries). Thresholds and calibration streams are reused from the formal Empty side with *zero recalibration*. Because slip admits no flip-rate-style direct observable, mechanism verification uses a pre-registered *first-divergence replay probe*: for each level, a clean replay stream (same reset seeds, same frozen policy, mechanism disabled) is compared episode-wise to the drift stream, the first observation divergence identifies the slipped step,

and the resulting empirical slip rate $\hat{p}$ must fall within 3 SE of nominal at every level (all 7 pass), alongside a return-impact hard gate at the anchor level p=0.7 ($p < 0.01$, $|d| > 0.8$; pass). An eight-item mechanical acceptance worksheet (unit completeness, score-persistence schema, seed disjointness, checkpoint SHA-256 identity, probe verdicts, calibration-failure counts, wall-clock records, one-run rule) is executed before any paper wiring: all items pass, with *zero* `calibration_failed` triggers across all 105 units.

D.6 Results: slip on Empty-8x8

The pre-registered hypothesis (the slip mechanism is detectable by all seats on any environment family) is confirmed, with a strengthening twist. **Window-level calibrated detection is** 1.000 ± 0.000 **for all five seats at all seven intensities**—including the ladder floor $p = 0.01875$—and every seat attains $\hat{\varepsilon} = 0.01875$ (bootstrap CI $[0.01875, 0.01875]$). Episode-level rates are monotone in p for B2 and B4 (1.000 at $p \geq 0.5$, falling to 0.149 at the floor) and 1.000 ± 0.000 at every level for C1. Calibrated FAR is 0.000 on all drift cells; B5's naive window FAR of 0.111 is brought into the band by the episode-stream shift. Table 2 compares same-ladder detection floors across environments.

Table 2: Same-ladder detection floors ($\hat{\varepsilon}$, transition slip) across environments: DoorKey-5x5 (main matrix) vs. Empty-8x8 (second cross-cell confirmation), zero recalibration on the Empty side.

Detector seat	DoorKey-5x5	Empty-8x8
B1 (JSD)	0.01875	0.01875
B2 (edit)	0.01875	0.01875
B4 (kNN)	0.0375	0.01875
B5 (CJSD)	0.01875	0.01875
C1 (resid.)	0.0375	0.01875

D.7 Interpretation: Complete 2×2 Matrix

With both off-diagonal cells crossed under pre-registered, zero-recalibration protocols, the mechanism-specificity structure reads as environment-invariant within our controlled setting in both directions: reward flips leave trajectory-channel seats blind and are carried by the value residual at the same floor on both environments (the first cross-cell confirmation), and transition slips are detected by all five seats on both environments, with the Empty side one ladder step *stronger* for B4/C1 ($0.0375 \rightarrow 0.01875$; the pre-registered reading treats a downward shift as strengthening evidence, which we register as such). One secondary observation is logged as environment-conditioned *sensitivity magnitude* rather than a structural change: C1's episode-level slip detection is full at every level on Empty-8x8 versus $0.609 \rightarrow 0.038$ across the DoorKey ladder—consistent with the pre-registered competing-explanation analysis (higher `open_forward` trigger coverage on the open grid offsets shorter episodes). We additionally register a second, more parsimonious competing explanation surfaced by the paired-joint-null audit: the Empty-side value network fits its reference residuals to a constant (zero variance per seed), so C1's Empty calibration is degenerate—its threshold equals the constant reference value, leaving zero noise margin, under which any positive offset alarms. The value-residual flip claim is independently carried by the DoorKey side (the first cross-cell confirmation), whose reference residuals have genuine variance and whose episode-level detection tracks the injected flip rate; the environment-invariance reading should therefore weight the DoorKey-side evidence more strongly than the degenerate Empty-side calibration.

D.8 Union under a Paired Joint Null

The oracle analysis (§5.3) reports the union FAR only as a Bonferroni upper bound and registers a paired joint null as future work. A pre-registered formal experiment (the paired-joint-null experiment; protocol and scoring script frozen before any scoring run) tests that registration on the two off-diagonal cells of the mechanism–environment matrix: reward flip on DoorKey-5x5 (12 cells) and transition slip on Empty-8x8 (21 cells), 33 units in total. All constants are inherited from the formal pipeline with zero new degrees of freedom (window

100, stride 50, 200 circular rotations, $\alpha = 0.05$; threshold-set seed 20260908, fresh FAR-set seed 20260910; CPU-only). The paired construction feeds the same rotation index to both seats—B5 via the same window components as the formal pipeline, C1 via same-order rotated episode residuals—standardizes each seat by its own threshold-set moments, and combines them under two pre-registered rules: **R-max** ($T = \max(z_{\mathrm{B5}}, z_{\mathrm{C1}})$, pooled 95th-percentile threshold, single budget) and a **Bonferroni baseline** (per-seat $q_{1-\alpha/2}$, alarm on either).

Results. On the flip×DoorKey cell the paired joint null is established: R-max union FAR is 0.0472/0.0489/0.0467 across the three policy seeds—all within the pre-registered band $|\mathrm{FAR} - \alpha| \leq 2\alpha$—and the union detection rate matches the best single seat (1.000) in all 12 cells, i.e., zero complementarity loss; the Bonferroni baseline agrees (0.0489/0.0483/0.0478, 12/12). On the slip×Empty cell the experiment is *uninformative rather than negative*: the Empty-side value network fits its reference residuals to a constant (zero variance per seed), so the C1 leg's z-standardization divides by zero and its scores become NaN, under which every comparison is silently false—the R-max leg never alarms (a NaN artifact we register as an implementation gap, not a calibration result), and the Bonferroni baseline degenerates to the B5 seat alone, which by itself reproduces the official detection (21/21; FAR 0.0283–0.0289, within band). Verdict under the pre-registered mapping: R-max = MIXED (flip×DoorKey passes both gates; slip×Empty untested), R-bonf = PASS with the single-seat qualification. Diagonal cells (slip×DoorKey, flip×Empty) are excluded by protocol because formal-matrix per-window scores were not persisted and B5 retraining determinism is unverified; they remain a full-matrix union experiment candidate, and a fail-closed zero-variance assertion plus a degenerate-leg policy are registered as the union-experiment prerequisites. Within the tested flip cell, the Bonferroni upper bound is conservatively loose in practice and can be tightened to a single calibrated budget by the paired joint null.

D.9 Full-Matrix Coverage and a Registered Power Boundary

A second pre-registered experiment (the full-matrix union experiment) extends the paired joint null to the full 2×2 mechanism–environment matrix: the two diagonal cells (transition slip on DoorKey-5x5, reward flip on Empty-8x8), a re-run of slip×Empty under a repaired C1 leg, and a re-scoring of flip×DoorKey as a pipeline-consistency check (66 cells, 12 units; constants inherited unchanged, CPU-only). B5 retraining determinism—the reason the paired-joint-null experiment excluded the diagonal cells—is verified directly: two same-seed retrainings produce bit-identical window scores, thresholds within 0.1×IQR of the official values, and identical alarm indicators (12/12 units). Recomputed C1 window alarms match the official `result.json` fields verbatim on all 66 cells. Out-of-sample union FAR stays within the pre-registered band on all 12 units under both combination rules, and the Empty-side units run in the pre-registered degenerate mode whose FAR is carried by the B5 leg, reproducing the paired-joint-null single-seat values verbatim; the flip×DoorKey re-score reproduces the paired-joint-null FAR values verbatim (pipeline continuity).

Registered boundary. On the 54 main judgment cells the union detection rate matches or exceeds the best single seat in 52 under R-max (53 under the Bonferroni baseline); by a pre-disclosed ceiling effect, only 5 low-intensity cells are discriminative, and the R-max union *falls below* the best single seat on 2 of them (seed 1: 0.2 vs. 0.4 at slip rate 0.01875 and 0.6 vs. 0.8 at 0.0375). The loss interacts with the combination rule: the Bonferroni baseline fails only the 0.0375 cell—at 0.01875 it matches the best seat ($0.4 = 0.4$)—so the Bonferroni failure set is a strict subset of the R-max failure set. This is a genuine complementarity loss, not an artifact: FAR calibration, retraining determinism, and recomputation-to-official agreement all pass, and the failures fall exactly inside the pre-disclosed discriminative subset. We therefore register that the paired joint null delivers a single calibrated budget across the full matrix, but its zero-complementarity-loss property is mechanism-dependent: it holds where the two seats' signals are complementary (reward flip) and can fail where they overlap (low-intensity transition slip).

E A Third Environment Family: MiniHack (MH)

E.1 Motivation and Pre-registration

All results in the main matrix and the cross-cell confirmations (§5, Appendix D) live on two MiniGrid environments. To probe whether the mechanism-specificity structure extends beyond a single simulator family, a pre-registered third-environment confirmation (protocol frozen before any of its data generation) injects the reward-flip mechanism on `MiniHack-Room-5x5-v0` (NetHack engine): a third environment family with an 8-action movement space, a 9×9 character-crop partial observation (81 dims, same partial-observability class as MiniGrid's 7×7 image), a constant success reward of 1.0, and non-success terminated endings—verified by a read-only pre-registration probe. The flip wrapper semantics and intensity ladder ($q \in \{1.0, 0.5, 0.25, 0.1\}$) are inherited verbatim; the mechanism predicate (*terminated* with positive reward) transfers exactly, and the flipped-episode audit predicate moves from the episode return to the terminal-step reward (a pre-registered adaptation, mirror of a pre-registered precision fix; a post-hoc per-episode audit of the calibration streams showed successful episodes carry an exact total return of 1.0—only non-success endings end with a small negative terminal step—so both predicates are exact in this task, and the terminal-step form was applied as pre-registered).

E.2 Design

Unlike the cross-cell confirmations, no frozen artifacts exist for a new environment family, so the experiment trains its own behavior policies: three PPO seeds on the raw (mechanism-free) environment under the protocol-default convergence criterion, fresh calibration streams (500 episodes per seed) for threshold bootstrapping, and 12 test streams (4 intensities $\times$ 3 seeds $\times$ 300 episodes) drawn from pre-registered seed zones ($[810000, 810299]$ test, $[820000, 820499]$ calibration) mechanically verified disjoint from all existing registries. Thresholds follow the main-matrix constants ($\alpha = 0.05$, 200 circular rotations, window 100, stride 50) with calibration and test isolation unchanged; a pre-registered marginal-stability probe verifies the edge-stability premise on the new observation representation before any detection claim. All five detector seats are scored through the byte-identical formal pipeline.

E.3 Results

The pre-registered structure replicates: C1 detects the flip at all 4 of 4 intensities (window-level calibrated rate 1.0, i.e., 15/15 pooled test windows across seeds, at every intensity; $\hat{\varepsilon}$ at the lowest rung q=0.1), while B1/B2/B4 register no significant detection at any intensity (episode- and window-level calibrated rates identically 0 in all 12 cells per seat) and B5 fires only at q=1.0 or not at all (episode alarms all null—the registered zero-variance degenerate lineage; window-level calibrated rate 0 at every intensity: fully blind). At the episode level, C1 flags exactly the flipped episodes: its detection rate equals the realized flip rate in all 12 cells ($1.00, 0.453, 0.240, 0.070$ across the ladder) with zero episode-level false alarms. Calibrated FAR on drift cells: 0.000 at both levels for every seat (episode level 12 cells/seat, window level 20 cells/seat; no unit outside the pre-registered $|\text{FAR} - \alpha| \leq 2\alpha$ band, all deviations conservative), nominal $\alpha = 0.05$. The marginal-stability probe on the new representation: per-dimension marginal JSD = 0 for all three seeds (q=1.0 test vs. calibration)—the flip acts on the reward channel only, so the observation marginals are invariant by construction and the edge-stability premise holds in its strongest form.

E.4 Interpretation

The mechanism-specificity structure extends across simulator families: the value-residual seat carries the flip while the trajectory-channel seats remain blind on a third environment family, extending the environment-invariant reading of the specificity claim beyond MiniGrid.

F Full Details: Honesty Ledger, Rollout Check, and Hold-Out Evaluation

F.1 Negative Results, Boundary Cases, and Limitations (Full Ledger)

The protocol's honesty ledger records the following items:

1. **Summary-layer indicator approximation.** $\hat{\varepsilon}$ confidence intervals bootstrap an alarm-indicator array reconstructed as det $\times$ $n_{\text{test_windows}}$ (position information discarded); CIs are therefore optimistic and serve as level-boundary references only. Precise per-window indicators were not persisted for B1–C1 (only B5), a schema lesson we record for future runs.
2. **FAR resolution floor.** Naive-scale empirical FAR cannot approach $\alpha = 0.05$ at $n_{\text{windows}} = 9$ (granularity $1/9 \approx 0.111$); we pre-registered this and interpret naive rows accordingly.
3. **B1 seed_1 calibration failure.** A seed-level JSD score-distribution mismatch (1 of 3 seeds); within-sample we cannot exclude heavy-tailed JSD nulls as systematic, hence "boundary characterization" rather than exclusion-silently (full structure and robustness audit in Appendix B).
4. **Scope.** All findings are established on 2 environments $\times$ 2 mechanisms $\times$ 5 detector families with 3 policy seeds; generalization to other environments, mechanism types, or policy classes is future work.
5. **Registered prior expectations.** CUSUM-family weak performance (AUROC 0.495 on an external benchmark) and residual-detector weak performance were pre-registered as expected-failure baselines; B3 was accordingly demoted to a non-run attachment.

F.2 Rollout-Mode Robustness Check (Full Protocol)

The main matrix evaluates frozen policies with deterministic (greedy) rollout; the following check verifies that the edge-JSD discrimination premise is not an artifact of that choice and survives stochastic deployment. A pre-registered tension-discrimination check crosses the three converged checkpoints (pre-registered convergence criterion) with rollout mode {deterministic, stochastic} and mechanism {none, slip 0.3, goal-reward +1.3 negative control} (200 paired-reset episodes per unit, bootstrap $B = 200$). The pre-registered rule declares H1 when the discrimination gap $G = \text{drift}_{q05} - \text{null}_{q95} > 0$. All six units satisfy H1 (stochastic-mode G from 3.6×10^{-5} to 3.2×10^{-3}), and the goal-reward negative control stays below its null $q95$ in every unit (JSD = 0 in half of them), leaving the false-alarm channel clean. An exploration-track boundary case (gap compression under an entropy-regularized *unconverged* checkpoint) did not reproduce on the converged seats and is recorded as an unconverged-checkpoint boundary phenomenon, not upgraded. The training-side follow-up—an entropy-coefficient ladder over 12 seats—is convergence-confounded and boundary-characterized in Appendix C. Three registration boundaries apply: this check covers the marginal-JSD signal only and says nothing about the conditional detectors (B5/C1) under stochastic deployment; G spans two orders of magnitude across seats (1.2×10^{-5}–3.2×10^{-3}),

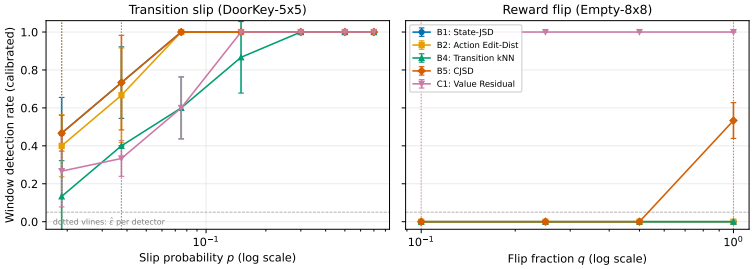

Figure 2: Detection-rate–intensity curves (5 detectors $\times$ 2 mechanisms, faceted; per-window calibrated detection rate summarized over 3 seeds; error bars = seed SD; $\hat{\varepsilon}$ annotated. Figure-internal seat labels: "B1: State-JSD" = B1 (JSD) and "B2: Action-Seq. Dist." = B2 (edit) in Table 1).

so no seat-independence is claimed; and H1 follows the pre-registered $G > 0$ rule without an additional significance test.

F.3 Hold-Out Generalization on Unseen Intensity Levels (Full Protocol)

The specificity claims in the main text were frozen on the registered intensity ladders; the following pre-registered hold-out evaluation tests whether the *same* specificity structure transfers to two unseen intensity levels—slip $p = 0.1$ (between the registered 0.075 and 0.15) and flip $q = 0.75$ (between 0.5 and 1.0)—on freshly generated streams (5 detectors $\times$ 2 mechanisms $\times$ 3 policy seeds = 30 units, 300 episodes per stream, judgment executed mechanically from criteria frozen before data generation).

The hold-out verdict is **PASS** (30/30 units scored; per-unit bootstrap CI95 lower bound > empirical FAR). On the reward axis, the flip structure replicates exactly: C1 detects on all 3 seeds, B1/B2/B4 remain blind on every seed, and B5—which required $q = 1.0$ to fire in the main matrix—is silent at $q = 0.75$ (0/3 seeds), confirming that its near-blindness extends below the strongest registered level. On the transition axis, all four trajectory-window families (B1/B2/B4/B5) detect at $p = 0.1$ with at least 2/3 seeds each (B2/B4/B5: 3/3; B1: 2/3, with the missing seed being the same B1/DoorKey/seed_1 unit already boundary-characterized in Appendix B); C1's slip performance is recorded descriptively (2/3) without entering the axis criterion. The realized slip-trigger rate (0.0996) matches the nominal $p = 0.1$, and the FAR calibration-fidelity check passes in all units except the single boundary unit above (empirical FAR 0.222, outside $\alpha \pm 2\alpha$, retained and listed per protocol rather than silently excluded). These results support reading the detector–mechanism specificity structure as intensity-general within our controlled setting, rather than an artifact of the registered ladder points.

G Calibration Theory: The Window-Score-Sequence Shift Is Provably a No-Op

This appendix formalizes the claim of §4 that the window-score-sequence circular shift is *provably a no-op*: it is not an approximation to a valid calibration but a relabeling of the naive threshold (Proposition 1), whereas the adopted episode-stream shift provably pools window statistics over circular start positions far beyond the stride grid (Proposition 2). All statements are mechanical or combinatorial; no new experimental data are used, and the empirical echo in §G is the registered diagnosis.

Setup. Let the calibration stream consist of n episodes with episode scores $e_0, \ldots, e_{n-1}$ (null hypothesis $\mathcal{H}_0$). The window map $W_{L,S}$ places m windows of length L at start positions on the stride grid $G = \{0, S, 2S, \ldots, (m-1)S\}$ (with $(m-1)S + L \leq n$), and the window statistic is the mean of the episode scores within the window; the pre-registered values are $n = 500$, $L = 100$, $S = 50$, $m = 9$. A deployment test window alarms when its window statistic exceeds a threshold τ, and a calibration variant is a rule producing τ as an empirical $(1-\alpha)$-quantile of a pooled null sample. Two variants are compared; they differ *only* in how null samples are generated:

- **(A) Episode-stream circular shift** (adopted): rotate the episode sequence by a random offset $k \in \{0, \ldots, n-1\}$, *recompute* all m window statistics on the rotated sequence, and pool over $R = 200$ rotations ($mR = 1800$ samples).
- **(B) Window-score-sequence shift**: compute the window-score sequence $w = (w_0, \ldots, w_{m-1})$ once, cyclically shift *that sequence*, and reuse the precomputed window statistics as null samples.

Proposition 1 (Circularity)**.** In variant (B), for every $R \geq 1$ and every sequence of shift offsets, the pooled null multiset equals $\{w_0, \ldots, w_{m-1}\}$ (each value with multiplicity R); hence the pooled threshold satisfies $\tau_B^{(R)} = q_{1-\alpha}(\{w_0, \ldots, w_{m-1}\}) = \tau_{\text{naive}}$ for all R.

Proof. A cyclic shift is a bijection of sequence positions. Variant (B) generates its null samples by relabeling the m precomputed window statistics—no window statistic is ever recomputed from episode data. The pooled multiset over R shifts is therefore $\{w_j\}$ with each value repeated R times, whose empirical distribution coincides with that of $\{w_j\}$; any

empirical quantile of the pooled multiset equals the corresponding empirical quantile of $\{w_j\}$, which is by definition the naive threshold. □

Corollary 1 (resolution floor). *The variant-(B) threshold takes values in* $\{w_0, \ldots, w_{m-1}\}$ *for all R; with the pre-registered* $(m, \alpha) = (9, 0.05)$ *and the ceiling convention* $\lceil 0.95 \times 9 \rceil = 9$, $\tau_B = \tau_{\text{naive}}$ *is the sample maximum of the 9 stride-grid window statistics. Variant (B) therefore cannot refine the naive threshold's granularity or reduce its dependence on a single autocorrelated arrangement of 9 overlapping windows—formalizing the pre-registered "*1/9 *naive-FAR resolution floor" ledger item (ii) of §5.*

What the episode-level shift pools.

Proposition 2 (Episodic pooling)**.** Consider variant (A) with rotation offsets ranging exhaustively over $\{0, \ldots, n-1\}$. (i) The window multiset of the stream rotated by k equals the multiset of windows of the original sequence whose start positions lie on the shifted grid $G - k \pmod n$. (ii) As k ranges over all n offsets, each length-L circular window $[i, i+L) \pmod n$, $i \in \{0, \ldots, n-1\}$, appears exactly m times in the pooled multiset. (iii) The pooled quantile is therefore the empirical $(1-\alpha)$-quantile of the window statistic evaluated at *all* n circular start positions—rather than only the m stride-grid positions available to variant (B)—and with the registered $R = 200$ random offsets the pooled samples are uniform draws over the same n circular windows ($mR = 1800$ samples).

Proof. (i) Rotating the episode sequence by k and reading windows at fixed grid G is the composition of two rigid operations on the circular arrangement; reading the rotated sequence at start position g is identical to reading the original circular sequence at start position $(g - k) \bmod n$, so the rotated window multiset is $\{W(g-k) : g \in G\}$. (ii) For a fixed grid position g, the map $k \mapsto (g - k) \bmod n$ is a bijection from $\{0, \ldots, n-1\}$ to the n circular start positions; summing over the m grid positions, every circular window is generated exactly m times. (iii) Immediate from (i)–(ii); with random offsets the pooled samples are i.i.d. uniform over the n circular windows by the same bijection. □

Validity remark. Proposition 2 is mechanical; its statistical reading is that variant (A) estimates the null distribution of the window statistic over *window placements*, not merely over the single stride-grid arrangement. Under the null assumption that episode scores are exchangeable (stationary under $\mathcal{H}_0$), every length-L circular window is an exchangeable copy of a fresh test window, so the pooled threshold is a valid finite-sample $(1-\alpha)$ quantile for the test statistic. This is the finite-sample justification for permuting whole episodes rather than derived statistics under autocorrelation (Romano & Tirlea, 2022); that the surrogate choice materially changes error rates is documented by Cannistra et al. (2025). Variant (B) enjoys no analogous guarantee: it never generates a window outside the stride grid, and its null samples inherit a single autocorrelated arrangement.

Empirical echo (registered; no new data). Figure 3 illustrates the two null geometries schematically. The registered diagnosis realized Proposition 1 exactly: the sequence-level variant's "calibrated" FAR was pinned at the naive granularity floor (1/9 per cell, B5)—the predicted identity $\tau_B = \tau_{\text{naive}}$—while the adopted episode-stream shift moved the same seat to $\text{FAR}_{\text{cal}} = 0.074 \pm 0.052$ with 21/21 cells inside the $\alpha \pm 2\alpha$ band (Appendix B, Figure 1). The formal matrix accordingly adopted variant (A) with $n_{\text{circ}} = 200$ (§4), and no result in this paper rests on variant (B).

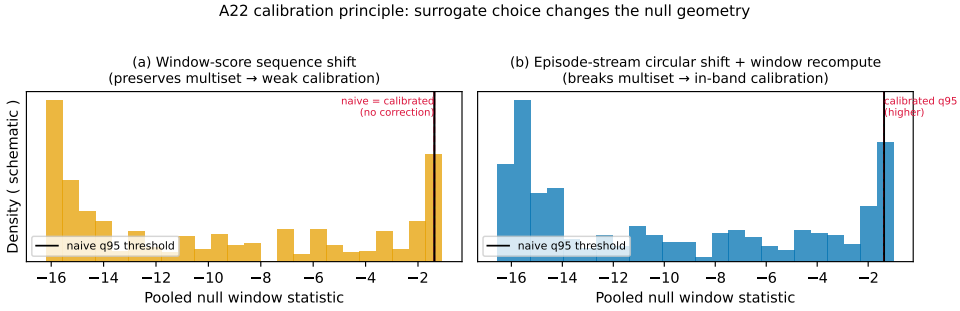

Figure 3: Schematic of the two surrogate geometries (Propositions 1–2; registered diagnosis, no new data; panel-header tags are internal schematic labels). (a) Shifting the window-score sequence leaves the pooled null multiset invariant, so the "calibrated" threshold coincides with the naive stride-grid maximum. (b) The episode-stream circular shift recomputes window statistics over shifted boundaries, pooling over circular start positions and moving the pooled quantile.