SURPRISE IS NOT ENOUGH: DYNAMICS-CONSISTENT OBSERVATION RELIABILITY IN REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning agents deployed on physical systems act on what their sensors report, and some sensor failures produce readings that still look plausible: a channel freezes at its last value, lags behind the true state, or drifts slowly away from it. A policy cannot distinguish such a reading from a real change in state, and robustness training does not tell it which observation to distrust. We introduce dynamics-consistent reliability (DCR), a per-step, per-channel estimate of whether an observation can be trusted, computed inside the RL loop from the agent's own experience. DCR compares each reading with what a learned dynamics model predicts from past observations and actions, and scores the residual two-sided on the model's own clean scale over several window lengths, the classical innovation test of fault diagnosis, so that a residual that is too small or too biased is itself evidence of a fault. The scoring rule has no fitted parameters and each of its components is standardized against a known null distribution. This design addresses a failure shared by the detectors used in practice, from temporal-jump thresholds and reconstruction error to ensemble disagreement and innovation gating: they score an observation by how surprising it is, and a plausible fault is not surprising; the temporal-jump detector ranks a frozen sensor as more reliable than a working one, with per-channel AUROC of 0.004–0.021 on three of four MuJoCo tasks, and the other surprise scores are weakened. We evaluate DCR on nine sensor-fault families with per-step ground truth, in closed loop, with every family held out of tuning when tested. Over five seeds DCR reaches mean held-out AUROC of 0.994 on Hopper, 0.986 on Walker2d, 0.995 on HalfCheetah and 0.792 on Ant, and outperforms PEDM- and DEXTER-style RL out-of-distribution detectors in every task; averaging the PEDM-style score over a window recovers most of that margin, so temporal accumulation is the larger lever and two-sided scoring adds a consistent 0.1–3.6 points. A learned head trained on injected faults is below the fixed statistic on average over held-out families. We further characterize when history-only estimation is impossible, derive a single-channel drift-detectability bound and show that the other channels carry most of the evidence, and build a controller that dead-reckons a flagged channel through the dynamics model and recovers 47–65% of the return lost to a fault on HalfCheetah, while failing on Walker2d, where isolation errs, and on two-channel faults. Code, fault definitions, trained policies and results are released.

1 INTRODUCTION

A reinforcement learning agent deployed on a physical system acts on sensor readings, and some sensor failures leave the reading plausible. A joint encoder freezes at its last value; a velocity estimate arrives several control steps late; an inertial bias creeps upward over a minute. None of these readings looks wrong. Each is a value the sensor could legitimately have produced, and a policy that cannot tell a faulty reading from a genuine state change will act on the fault. Robustness training does not close this gap: a policy trained to tolerate noise still cannot say *which* observations it should not have trusted, and so cannot fall back, re-plan, or ask for help. We study observation reliability as

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.

an explicit, per-step, per-channel estimation target: the probability that what the agent just received reflects the state it is meant to measure.

The detectors practitioners reach for share one mechanism. A temporal-jump threshold, a reconstruction error, an ensemble's disagreement, and an innovation gate all score an observation by how *surprising* it is, and raise an alarm when the surprise is large. This shared mechanism has a shared blind spot (Proposition 1): for any fault whose residual is stochastically no larger than the clean residual, every score that is monotone in the residual magnitude has an AUROC of at most one half. A frozen channel under a temporal-jump detector is the extreme case: its residual is exactly zero, and the detector ranks every faulty step as more reliable than every clean step. The failure is inversion, not degradation, and no threshold fixes a backwards ranking. On three of our four tasks the temporal-jump detector's per-channel AUROC on frozen channels is 0.004–0.021; on Ant it is 0.36.

Two things are needed to escape it, and we find empirically that both are necessary. The first is *dynamics consistency*: the reading must be compared with what a model of the dynamics predicts, not with persistence or with the plausibility of the value alone; a one-step dynamics model does not follow a frozen channel, because the other channels and the executed actions say it should move. The second is *two-sided scoring on the residual's own clean scale*: a residual that is implausibly small, or whose windowed mean or variance is off in either direction, is evidence of a fault. Referred to the model's own one-step uncertainty, this is a standard test with a known null distribution (Gao et al., 2015; Khalastchi & Kalech, 2018). The setting is new.

Perturbed observations in RL have so far been studied as a *robustness* problem: train a policy that keeps performing under adversarial or faulty inputs (Zhang et al., 2020; Gu et al., 2025b; Skand et al., 2024), and out-of-distribution detection in RL scores whole observations or trajectories by surprise using models fit on clean rollouts of a converged policy (Haider et al., 2023; Nasvytis et al., 2024; Haider et al., 2026). Recent work extends both views to sensor dropout, hardware faults, trajectory-scoring benchmarks and sim-to-real monitors (Vogt-Lowell et al., 2026; Schoepp et al., 2024; Mittag et al., 2026; Munn et al., 2026). Runtime monitors for robot policies detect *task* failure from consistency and progress signals (Agia et al., 2024; Gu et al., 2025a), not which sensor to distrust. Per-dimension deviation signals exist in a sim-to-real monitor (Munn et al., 2026), one-sided. To our knowledge no prior work evaluates *which observation the agent should not trust* with per-step, per-channel ground truth, fault families held out of training, an estimator fit inside the RL loop, closed-loop evaluation, and a two-sided test.

The estimator is fit only on the agent's own training stream, with no separate phase of clean rollouts from a converged policy, and is evaluated closed loop on nine per-channel fault families with per-step ground truth, each held out in turn. Under this protocol the dynamics-consistent two-sided test outperforms every surprise-based detector, including PEDM- and DEXTER-style detectors we re-implement, in every task; the largest single lever is temporal accumulation, which also lifts the PEDM-style score most of the way.

Two findings run against our initial design: a learned reliability head trained on injected faults is below the fixed statistic on average over held-out families, and a leave-channel-out predictor adds nothing to detection once two-sided windowed scoring is in place. We further characterize when history-only estimation is impossible (Proposition 2), derive a drift-detectability bound and show why measured detectability exceeds it, and recover part of the lost return by dead-reckoning the flagged channel through the dynamics model.

Contributions. (i) Dynamics-consistent reliability (DCR): we bring the classical two-sided innovation test of fault diagnosis into the RL loop as a per-step, per-channel reliability estimate on a learned dynamics model's own residual scale, with no fitted scoring parameters, and show that it outperforms the detectors RL practice relies on. (ii) A closed-loop, leave-one-family-out evaluation protocol over nine sensor-fault families with per-step, per-channel ground truth, complementing robustness benchmarks that measure return, not detection. (iii) Why the detectors in use fail: any score monotone in the residual inverts on plausibility-preserving faults (Proposition 1); an input ladder isolating the two necessary ingredients; two negative results on learned alternatives. (iv) An impossibility characterization for history-only estimation and a drift detectability bound that the experiments test. (v) A reliability-aware controller and its return recovery. Code, fault definitions and the full breakdown are released.

2 Related Work

Model-based fault detection and isolation. Classical fault detection tests the residual, or innovation, between a model's prediction and the received measurement for changes in its mean, covariance and whiteness, not only its size (Mehra & Peschon, 1971; Willsky, 1976); see Ding (2013); Gao et al. (2015); Khalastchi & Kalech (2018) for surveys and van Wyk et al. (2020) for sensor identification in automated vehicles. We apply such a two-sided test to a learned predictor fit online in the RL loop, evaluate it in closed loop, and show why one-sided scoring fails (Proposition 1).

Out-of-distribution and anomaly detection in RL. OOD detectors for RL score observations or trajectories by their deviation from nominal behaviour, using ensemble dynamics models (Haider et al., 2023), windowed time-series features (Nasvytis et al., 2024) or a learned world model (Haider et al., 2026), fit on rollouts of a trained policy. The preprints of Mittag et al. (2026) and Munn et al. (2026) contribute a benchmark and a sim-to-real monitor; the latter scores each dimension but, like PEDM, alarms only on large prediction errors, which Proposition 1 shows cannot detect plausible faults. None of these detectors is evaluated against per-step, per-channel ground truth on fault families held out of tuning.

Robust RL under observation perturbation. Robust RL trains policies that keep performing under corrupted observations, through adversarial regularization (Zhang et al., 2020), random sensor dropout during training (Skand et al., 2024), or sequence models that infer missing readings from history (Vogt-Lowell et al., 2026), and benchmarks them by return (Gu et al., 2025b); a preprint adapts policies to hardware faults by continued training (Schoepp et al., 2024). Such methods let a policy tolerate a fault but do not tell it which observation to distrust.

Runtime failure monitoring for robot policies. Runtime monitors detect when a learned robot policy is failing its task, using the temporal consistency of its actions and its progress toward the goal (Agia et al., 2024) or the policy's internal features (Gu et al., 2025a). However, they do not identify which input is at fault, particularly when a corrupted reading still looks plausible. A per-channel reliability estimate is complementary: it can explain a failure they flag, or catch a sensor fault before it causes one.

Calibration. Each component of the DCR statistic is standardized against a null distribution the model itself supplies, so its scale needs no fault examples, and the alarm threshold is a quantile of the score on clean data. Learned scores offer no such guarantee: even modern neural classifiers are miscalibrated and need post-hoc correction on held-out data (Guo et al., 2017).

3 When surprise cannot work

Setting. Let $o_t \in \mathbb{R}^D$ be the received observation and $\hat{o}_t = g(o_{<t}, a_{<t})$ any causal prediction of it. Write $r_t = o_t - \hat{o}_t$ for the residual and P_0 for the law of a single residual coordinate $r_{t,d}$ under clean operation. For a temporal-jump detector $g(o_{<t}, a_{<t}) = o_{t-1}$; for a reconstruction detector $\hat{o}_t$ is the autoencoder output; for an innovation gate g is a learned or filtered one-step model. We call a detector *surprise-based* if its per-channel score is $S = h(|r_{t,d}|)$ for some non-decreasing h. The temporal-jump threshold, the one-sided innovation gate and reconstruction error are of this form; ensemble disagreement and ensemble likelihood score surprise in a looser sense (the spread of predictions, or a likelihood that also depends on the predicted scale) and are covered in §6 empirically rather than by the proposition.

Definition 1 (Plausibility-preserving fault)**.** *A fault on channel d is* plausibility-preserving *for predictor g if the residual $r'_{t,d}$ it induces is stochastically no larger in magnitude than the clean residual:* $\Pr(|r'_{t,d}| > x) \le \Pr(|r_{t,d}| > x)$ *for all* $x \ge 0$.

Proposition 1 (Surprise-based scores are at best uninformative)**.** *Let $r \sim P_0$ be a clean residual coordinate with a continuous law and r' the residual of a plausibility-preserving fault, independent of r. For any non-decreasing h, the per-channel score $S = h(|\cdot|)$ satisfies*

$$\mathrm{AUROC}(S) \;=\; \Pr\big(h(|r'|) > h(|r|)\big) + \tfrac{1}{2}\Pr\big(h(|r'|) = h(|r|)\big) \;\le\; \tfrac{1}{2}.$$

If the fault residual is degenerate at zero (a frozen channel under a temporal-jump detector) and h is strictly increasing, $\mathrm{AUROC}(S) = 0$*: the detector ranks every faulty step as more reliable than every clean step.*

Proof idea. Replace r' by a copy coupled to an independent clean draw r'' so that $|r'| \le |r''|$ almost surely (Strassen); a non-decreasing h preserves the order, so the faulty score exceeds the clean score no more often than $h(|r''|)$ does, and by exchangeability that is half the time. Full proof in Appendix F. The statement is elementary; its value is that it applies to any monotone score at once.

Remark 1 (Which faults are plausibility-preserving for which predictor)**.** *Plausibility preservation is a property of the fault* and *the predictor. Under the temporal-jump predictor a frozen channel gives $r' \equiv 0$; after onset a lagged channel gives a clean one-step difference ($r' \overset{d}{=} r$); and a drift of rate ρ adds a per-step shift $\rho\sigma_d$, which is below the residual spread σ_r when the drift is slow. Value-plausibility detectors (reconstruction error, density estimates, feature forests) are weakened by the same faults, since a frozen or lagged value stays inside the marginal distribution, but invert only where the residual is strictly smaller than clean; §6 finds inversion for the temporal jump alone. A predictor that models the dynamics behaves differently: the frozen channel's residual is minus the change the dynamics say should have happened, which is large relative to a good model's residual scale. Dynamics consistency is therefore the first ingredient; two-sided scoring, which addresses faults that shrink or bias the residual without enlarging it, is the second, and §6 confirms both.*

The fix is two-sided. Proposition 1 fails the moment h is allowed to decrease: the windowed variance statistic $(\sum_{i\le W} z_i^2 - W)/\sqrt{2W}$ of the normalized residual $z = r/\sigma$ increases when the residual is implausibly *small*. This requires σ to be the residual's own clean scale; normalized by the spread of the observation, a residual is always small and "too small" is undefined. Two-sided innovation tests and sequential accumulation are standard in model-based fault diagnosis (Page, 1954; Basseville & Nikiforov, 1993; Ding, 2013; Gao et al., 2015; Khalastchi & Kalech, 2018) and in deployed sensor-anomaly systems (van Wyk et al., 2020); their origin is the innovations approach of Mehra & Peschon (1971) and Willsky (1976).

4 When history-only estimation is impossible

Proposition 2 (Indistinguishability)**.** *Let $\mathbb{P}_0$ and $\mathbb{P}_f$ be the laws of the observation–action history $(o_{\le t}, a_{<t})$ under clean operation and under a fault process f. Any estimator $\hat{\rho}_t = \phi(o_{\le t}, a_{<t})$ satisfies $\left|\mathrm{AUROC}(\hat{\rho}_t) - \frac{1}{2}\right| \le \mathrm{TV}(\mathbb{P}_0, \mathbb{P}_f)$. In particular, if $\mathbb{P}_f = \mathbb{P}_0$ no history-only estimator beats chance.*

Proof idea. Couple the fault-side score to a clean-side copy so that they differ with probability TV; against an independent clean score the copy has AUROC exactly $\frac{1}{2}$ (Appendix F).

Three fault processes are indistinguishable or nearly so, and no estimator in this paper or elsewhere should be expected to detect them: (i) a constant bias present from the first step on a channel whose offset is not otherwise identified by the dynamics; (ii) a *replay* fault that emits a trajectory consistent with the true dynamics and the executed actions; (iii) the onset step of a frozen channel, whose received value equals the clean value by construction. Case (iii) is why we report detection delay and never relabel the onset step.

A single-channel detectability bound for drift. Let a drift of rate ρ (in units of the channel's spread σ_d per step) act on channel d. Suppose the predictor *propagates* the received value of channel d, so that the normalized residual of channel d acquires only the per-step increment as a shift, $\delta = \rho\sigma_d/\sigma_{r,d}$. If the estimator may use only channel d's own residual, the windowed mean statistic $m_W = \sum_{i\le W} z_i/\sqrt{W}$ is $\mathcal{N}(0,1)$ under clean data and $\mathcal{N}(\delta\sqrt{W}, 1)$ under drift, so

$$\mathrm{AUROC}_{\text{own}}(m_W) \;=\; \Phi\Big(\frac{\rho\,\sigma_d\sqrt{W}}{\sqrt{2}\,\sigma_{r,d}}\Big), \tag{1}$$

which is at chance as $\rho \to 0$ for any finite window. Both hypotheses of Eq. (1) can fail in ways that make the fault *more* detectable, and the experiments separate them. First, a drifted channel enters the dynamics model's input and biases the predictions of every *other* channel, so the remaining channels carry evidence the bound ignores; in closed loop it also changes the actions. Second, to the extent the model predicts channel d from the other channels and the actions rather than from its own last value, the residual on channel d reflects the accumulated bias $\rho\sigma_d(t - t_0)$ rather than the per-step increment. §6 decomposes the statistic into the faulty channel's own term and the remaining channels' term, in closed and open loop, at both checkpoints. The own-channel term sits at the bound only

where the learned model is persistence-like, which we observe early in training on Hopper; once the model predicts the channel from the other channels and the actions, the own-channel term exceeds the bound, and in every task at every rate the remaining channels supply most of the detectability.

5 DYNAMICS-CONSISTENT RELIABILITY ESTIMATION

Problem. An agent receives $o_t = \mathcal{C}_t(s_t, o_{<t}, \zeta_t)$, where s_t is the true state, $\zeta_t \in \{\text{clean}\} \cup \mathcal{F}$ is a latent per-channel fault mode and $\mathcal{C}_t$ is the fault map (identity when clean). We say *fault*, as in the fault detection and isolation literature, for what the machine-learning literature calls an observation corruption: a sensor that enters a bad mode and stays in it. The reliability of channel d at time t is $\rho_{t,d} = \Pr(\zeta_{t,d} = \text{clean} \mid o_{\le t}, a_{<t})$ and the observation's reliability is $\min_d \rho_{t,d}$. We ask for a causal estimator, learned from the agent's own experience without a clean-rollout phase, that ranks faulty steps above clean steps on fault families it was never tuned on.

Predictor. A heteroscedastic one-step model g maps the last k received observations and actions to a Gaussian over o_t with per-channel mean $\mu_{t,d}$ and scale $\sigma_{t,d}$ (we predict the increment $o_t - o_{t-1}$ and add it back), fit by maximum likelihood on the agent's own training stream up to the checkpoint with a memory of the most recent 300k steps, never on rollouts of the final policy. The scale $\sigma_{t,d}$ is the model's own one-step residual scale on clean data, which is what makes an implausibly *small* residual detectable (§3). A *leave-channel-out* variant $g^{\setminus d}$ receives the same window with channel d hidden and predicts the level of every channel from the rest; it plays no role in detection but is what makes fault *isolation* work in the controller below.

Statistic. Let $z_{t,d} = (o_{t,d} - \mu_{t,d})/\sigma_{t,d}$ be the normalized residual, $\mathcal{N}(0,1)$ under a calibrated model on clean data. For a window W ending at t, the standardized windowed mean and variance statistics

$$m_W = \frac{1}{\sqrt{W}} \sum_{i=t-W+1}^{t} z_{i,d}, \qquad v_W = \frac{\sum_i z_{i,d}^2 - W}{\sqrt{2W}}$$

are both $\mathcal{N}(0,1)$ under the null, and a fault moves either in *either* direction: a drift drives m_W away from zero; a frozen channel drives v_W negative under a persistence-like predictor and positive under a dynamics model, since the model keeps predicting motion the channel no longer reports. The DCR score of channel d is $\max_{W \in \{8,32,128\}} \max(|m_W|, |v_W|)$ and the step score is the maximum over channels. This is the two-sided innovation test of model-based fault diagnosis (Gao et al., 2015; Ding, 2013) evaluated at three scales on a learned model; it has no fitted scoring parameters, so there is nothing to overfit to a fault family. The maximum over windows, statistics and channels is not itself standard normal; alarm thresholds are therefore set as quantiles of the score on clean episodes.

Calibration. Each component m_W, v_W is standardized against a null distribution the model itself supplies, so no fault examples are needed to set a scale; the alarm threshold on the aggregate score is a clean-data quantile, chosen without fault examples. The learned head of the next paragraph has neither property.

Learned head (ablation). For comparison we also train a small MLP, shared across channels, that maps a per-channel feature vector (the signed and log-magnitude residuals, the windowed statistics above, the raw windowed difference variance and the temporal jump) to $\Pr(\text{fault})$, using a class-balanced logistic loss on clean training episodes into which faults from the *training* families are injected. It is always evaluated on a held-out family.

Online protocol. Predictors and scales are fit only on the agent's own training stream up to the evaluated checkpoint, including the random-policy warm-up and every intermediate policy. Evaluation rollouts are closed loop: the policy acts on the faulty observation, so faults alter behaviour and the true state as they would on a physical system. Alarm thresholds are per-channel quantiles of the statistic on the clean evaluation episodes, which also serve as the clean set for AUROC, so reported false-alarm rates are in-sample.

Acting on reliability. The controller raises a latching alarm on channel d when its statistic exceeds its threshold on three consecutive steps, isolates one culprit per alarm (the largest exceedance), and thereafter replaces that channel by dead-reckoning: the model's one-step prediction from the *repaired* history. Without isolation the dead-reckoned value drifts, inflates the other channels' residuals and latches them one by one until the policy acts on imagination (Appendix E). We therefore test the remaining channels with the culprit hidden from the predictor's input, using $g^{\setminus d}$, against a threshold calibrated for that test, and cap simultaneous flags at three. An oracle that flags the true mask with the same repair bounds what the mechanism can achieve.

6 EXPERIMENTS

Tasks and agents. Hopper-v5, Walker2d-v5, HalfCheetah-v5 and Ant-v5 from Gymnasium MuJoCo (Towers et al., 2025), each trained with SAC (Haarnoja et al., 2018) at the CleanRL defaults (Huang et al., 2022) for one million steps and five seeds. The agent's entire training stream (observations, actions, and the simulator's true state) is logged. Reliability estimators are evaluated at the 150k and 1M checkpoints and are fit only on the stream up to that checkpoint.

Fault families. Nine per-channel fault families, each with a per-step ground-truth mask (Appendix A). Plausibility-preserving: *freeze*, *lag* (2–10 steps), *drift* (rate 0.003–$0.05\,\sigma_d$ per step). Surprise-visible: *spike* (sparse impulses), *noise* burst, *dropout* to zero, channel *swap*. Mixed: *gain* error and *saturation*. Faults begin at a random onset and persist; the label marks the sensor's state, not the reading, so a spike-faulted channel is labelled faulty between impulses.

Protocol. Leave-one-family-out (LOFO): the head evaluated on family h is trained with faults from the other eight families only, and no hyperparameter is tuned on h. Evaluation uses 40 clean and 30 faulted closed-loop episodes per family, checkpoint and seed. We report pooled per-step AUROC, AUPRC, false-alarm rate at 95% recall, and detection delay (steps from onset to first alarm at a threshold set at 0.1% per-step false-alarm rate on clean steps).

Detectors. Surprise-based: temporal jump; one-sided innovation gate normalized by the observation spread or by the predictor's own scale; autoencoder reconstruction error (Sakurada & Yairi, 2014); probabilistic-ensemble NLL and disagreement in the manner of PEDM (Haider et al., 2023; Chua et al., 2018; Lakshminarayanan et al., 2017) (our variant scores the mean per-member Gaussian log-likelihood rather than the minimum error); and a DEXTER-style windowed-feature isolation forest (Nasvytis et al., 2024; Liu et al., 2008). Classical, all on the same learned predictor so that only the scoring rule differs: the two-sided windowed innovation test (standardized windowed mean and variance of the normalized residual) of model-based fault diagnosis (Gao et al., 2015; Ding, 2013); two-sided CUSUM on the normalized residual (Page, 1954) with reference $k{=}0.5$ and decision limit $h{=}5$; the χ^2 innovation-consistency test (Bar-Shalom et al., 2001) over $W{=}32$; and the PEDM-style score averaged over the same window. Ours: the DCR statistic, its learned-head ablation, and the input ladder. All learned components share the predictors described in §5; all are causal (Appendix B).

6.1 SURPRISE INVERTS ON PLAUSIBILITY-PRESERVING FAULTS

Figure 1 shows one Hopper episode in which a channel freezes. The temporal-jump detector's estimated reliability of that channel rises to 1 within a few steps and stays there, while both dynamics-consistent detectors drop to zero. At the channel level (Appendix Table 3) the temporal-jump detector's AUROC on a frozen channel is 0.004–0.021 on Hopper, Walker2d and HalfCheetah and 0.360 on Ant; on a dropped channel 0.002–0.028; on a saturated channel 0.165–0.253: near-perfect detectors with the sign reversed, as Proposition 1 predicts for a residual that is identically zero. Reconstruction error and the DEXTER-style forest degrade without inverting. At the step level, where the score is a maximum over channels, closed-loop evaluation masks the inversion because the policy's reaction moves the other channels. The one-sided innovation gate on our learned predictor does *not* invert (0.73–0.83 channel-level AUROC on freeze): a one-step dynamics model does not follow a frozen channel. Dynamics consistency is therefore the first ingredient.

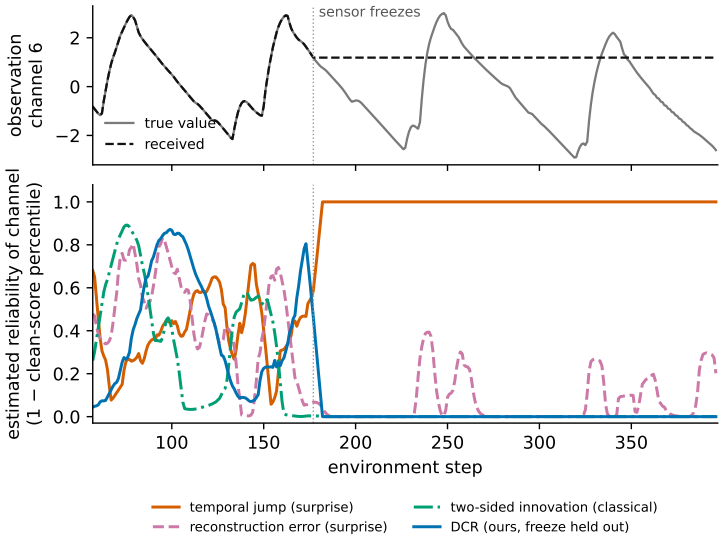


Figure 1: A frozen channel on Hopper-v5. Top: true and received value of the frozen channel. Bottom: per-channel estimated reliability, shown as one minus the percentile of each detector's score within its own clean-episode distribution. Surprise-based detectors report a dead sensor as reliable; dynamics-consistent detectors report it as unreliable within a few steps of onset.

6.2 Two-sided scoring on a dynamics-consistent residual

Table 1 summarises held-out AUROC at 1M steps over five seeds; per-family breakdowns and the 150k checkpoint are in Appendix D. The classical two-sided windowed test on the same predictor as the one-sided gate raises mean AUROC from 0.83–0.88 to 0.976–0.991 on Hopper, Walker2d and HalfCheetah, and the multi-scale version (DCR) adds a further 0.5–2.2 points, reaching 0.994, 0.986, 0.995 and 0.792 on Ant (five seeds; seed standard deviation of the mean at most 0.011). On the same residual, mean-shift CUSUM reaches 0.82–0.88 (0.69 on Ant), since plausibility-preserving faults change the residual's spread more than its mean, and the χ^2 innovation test matches the classical two-sided test (0.976–0.992; 0.768 on Ant). The gain over one-sided scoring is largest on spike (23–33 points) and smallest on noise (2–7 points); the spike gain reflects the label semantics, since a spike-faulted channel is labelled faulty between impulses and only an accumulating score can flag those steps. Against the re-implemented RL detectors, the PEDM-style ensemble NLL is the strongest per-step baseline (0.67–0.94) and is below DCR in all 36 task–family cells, significantly in 31 under a seed-paired t-test (Appendix Table 14); ensemble disagreement and the DEXTER-style forest are far behind. Averaging the PEDM-style score over $W{=}32$ steps lifts it to 0.991, 0.969, 0.994 and 0.756: temporal accumulation is the larger lever, and two-sided scoring on the model's own scale adds the remaining 0.1–3.6 points, significantly on Walker2d and Ant and within seed noise on Hopper and HalfCheetah. False alarms at 95% recall follow the same order on the three tasks: 0.4–6.4% of clean steps for DCR, 27–55% for the PEDM-style score, 84–99% for temporal jump (Appendix Table 22). Detection delay (Appendix D) is under ten steps on Hopper and HalfCheetah, 75–200 steps on Walker2d, and Ant's contact-channel faults are mostly missed; temporal jump needs on the order of a hundred steps. DCR does not degrade from 150k to 1M; Walker2d improves from 0.948 to 0.986 as the policy and hence the predictor's training distribution stabilise.

6.3 Input ladder

Appendix Table 4 adds one information source per rung: temporal jump $\rightarrow$ one-sided residual of a dynamics model $\rightarrow$ per-step two-sided $|\log z^2| \rightarrow$ windowed two-sided $\rightarrow$ multi-scale. Two rungs are instructive. Per-step two-sided scoring is *worse* than one-sided (0.78 vs 0.88 on Hopper, 0.57 vs 0.83 on Walker2d): a single small residual happens by chance constantly, so "too small" becomes evidence only when accumulated over a window. And the leave-channel-out residual adds nothing once windowed two-sided scoring is in place (0.79–0.95 alone; combined with the full residual,

Table 1: Held-out AUROC at 1M steps, 5 seeds: mean over the nine fault families (all) and over the three plausibility-preserving families (PP), averaged over seeds. Per-family tables, with two further weak baselines, in Appendix D.

Detector	Hopper		Walker2d		HalfCheetah		Ant	
	PP	all	PP	all	PP	all	PP	all
Temporal jump	0.560	0.649	0.572	0.631	0.380	0.491	0.514	0.538
Reconstruction (AE)	0.841	0.848	0.762	0.752	0.909	0.850	0.667	0.627
DEXTER-style (IF)	0.677	0.735	0.656	0.679	0.879	0.861	0.541	0.544
Innovation / own scale (1-sided)	0.927	0.877	0.860	0.830	0.822	0.837	0.687	0.666
PEDM-style NLL	0.973	0.937	0.905	0.872	0.975	0.943	0.692	0.670
PEDM-style NLL, windowed	**0.992**	0.991	0.977	0.969	**0.994**	0.994	0.751	0.756
CUSUM on residual (k=0.5, h=5)	0.924	0.875	0.849	0.817	0.880	0.880	0.709	0.693
χ^2 innovation test W=32	0.990	0.990	0.981	0.976	0.988	0.992	0.769	0.768
2-sided, windowed W=32	0.990	0.989	0.982	0.976	0.987	0.991	0.770	0.769
DCR (2-sided, multi-scale)	0.992	**0.994**	**0.985**	**0.986**	0.993	**0.995**	**0.794**	**0.792**
Learned head (LOFO)	0.956	0.975	0.944	0.954	0.969	0.979	0.729	0.743

never better). Cross-channel prediction does not detect these faults in a single proprioceptive stream, although §6.5 shows it is what isolates them.

The learned head does not beat the statistic. The "learned head" rows are the estimator we set out to build: a small network mapping the same features, and more, to a fault probability, trained on faults injected from the training families. Its mean held-out AUROC is 0.74–0.98, below the fixed statistic on average in every task and in 32 of 36 task–family cells, worst on the families least like its training set (drift, lag). Trained on all nine families it closes 37–54% of the gap (Appendix D): part of the shortfall is generalization across fault families, part is the head itself.

6.4 Drift: what the bound predicts and what cross-channel consistency adds

Appendix Figure 2 sweeps the drift rate over two orders of magnitude at 1M steps and decomposes the windowed-mean statistic into the faulty channel's own term and the remaining channels' term. In every task the remaining channels carry most of the detectability at the slowest rates: a drifted channel enters the model's input and biases the predictions of every other channel. And open-loop injection into logged clean episodes matches the closed-loop result, so the effect is the model's input being corrupted, not the policy reacting. The faulty channel's own term behaves as Eq. (1) says it should for the model it has: at 150k steps on Hopper, where the learned model is still persistence-like, it sits on the bound (Appendix D); at 1M steps it exceeds the bound in all four tasks (HalfCheetah at both checkpoints). A slowly drifting sensor is found not because its own readings look wrong but because it makes every other reading inconsistent with the dynamics.

6.5 Acting on reliability

Appendix Figure 3 reports return under each fault family with no gating, with the DCR controller of §5, and with an oracle that flags the true mask and applies the same repair. On HalfCheetah the controller recovers 47–65% of the lost return on every family (drift: 3886→7991 of a 10888 ceiling) and costs 5% on clean episodes. On Hopper it recovers 69% on drift and 38–53% on saturation, spike, dropout, lag and noise, but only 11% on gain, where the policy tolerates the fault and any false isolation costs more than it saves, and 3% on swap, whose two faulty channels single-culprit isolation cannot represent. Walker2d is the honest failure: the oracle recovers most of the lost return on every family, so the repair works, but with the same fixed threshold our isolation errs and the controller *lowers* return on five of nine families (noise 4173→3245, dropout 2727→1111). Per-task threshold tuning, which we deliberately did not do, would have hidden it. The controller's own development is in Appendix E: three versions failed, and the cascade in which a repaired channel corrupts the predictions of the others until the policy acts on imagination is the reason the leave-channel-out predictor is in this paper.

6.6 WHERE HISTORY-ONLY ESTIMATION FAILS

Ant is the weak task in every table, and Proposition 2 says why: 78 of its 105 channels are contact forces that are exactly zero whenever the body is off the ground, so a frozen, dropped or saturated contact channel produces a history identical to a clean one for long stretches. Split by the faulted channel's group (Appendix Table 19), the 27 kinematic channels reach 0.85–0.93 held-out AUROC on freeze, dropout, gain and lag at 1M steps, in line with the other tasks, while contact channels sit at 0.54–0.69; families that inject energy (spike, noise, drift) are detected on both groups (≥ 0.91).

7 LIMITATIONS

Synthetic faults. All nine families are injected and modelled on documented failure modes; no real sensor log is used. **Persistent, sparse faults.** Faults persist to the end of the episode and affect one channel (two for swap); the controller's latching alarm and single-culprit isolation encode this assumption, and swap is where they fail. Intermittent faults would require an un-latching rule. **Label semantics.** The label marks the sensor's state, so a spike-faulted channel is faulty between impulses and a frozen contact-force channel that reads zero is faulty while identical to clean; this favours detectors that accumulate evidence, hence the windowed PEDM-style baseline. **In-sample thresholds.** Alarm thresholds and reported false-alarm rates use the same clean evaluation episodes. **Episode length.** Faulted episodes end earlier when the robot falls, weighting early post-onset steps more. **Controller settings.** The latch length, flag cap and hidden-test threshold were chosen on Hopper seed-1 diagnostics and then held fixed; they are not part of the held-out detection claim. **Simulation.** Clean one-step residuals in MuJoCo are small and well calibrated; on hardware the residual scale is larger and the two-sided test correspondingly less powerful. **Online but not adversarial.** The estimator sees the agent's own stream, which we assume clean during training; robust scale estimation when the training stream itself contains faults is not evaluated here. **Indistinguishable cases.** Constant bias from the first step, replay of a dynamically consistent trajectory, and the onset step of a frozen channel are undetectable from history alone (Proposition 2); we report detection delay rather than relabel.

8 CONCLUSION

One axis separates reliability estimators for sensor faults: whether the fault keeps the reading plausible. Surprise-based detectors sit on the wrong side of it for frozen, lagged and drifting sensors, and their ranking reverses rather than degrades. Dynamics prediction, accumulated over time and scored two-sided on the model's own scale, escapes the inversion without fitted parameters and beats every learned alternative on unseen families on average. Faults invisible from history remain undetectable, and acting on the estimate requires isolating the culprit.

AI USE STATEMENT

In this work, LLM is used to polish and aid writing. We use LLM to provide initial paraphrasing for some sentences that need polishing, then manually refine them based on provided suggestions. We also use LLM to format tables and generate initial draft code for the graphs and visualization of the results. We use LLM to help find related work, which we read and verified before citing.

ETHICS STATEMENT

A reliability estimate that is overconfident on a plausible-looking fault is worse than no estimate, because a controller will act on it; this is why we report how thresholds are set, detection delay, and the failure cases (Ant contact channels, Walker2d isolation) rather than ranking alone. The fault families are synthetic and the tasks are simulated; no data about people is involved.

REPRODUCIBILITY STATEMENT

Code, fault-family definitions, trained policy checkpoints and the full per-seed result files are provided as anonymised supplementary material and will be released publicly on acceptance. Environment

versions (Gymnasium MuJoCo v5), SAC hyperparameters, predictor and head architectures, window lengths, seed counts and the evaluation protocol are given in Appendix C; the fault families in Appendix A; the tests in Appendix B. Every table and number in the paper regenerates from the released result files with two scripts.

REFERENCES

Christopher Agia, Rohan Sinha, Jingyun Yang, Zi-ang Cao, Rika Antonova, Marco Pavone, and Jeannette Bohg. Unpacking failure modes of generative policies: Runtime monitoring of consistency and progress. In *CoRL*, 2024.

Yaakov Bar-Shalom, X. Rong Li, and Thiagalingam Kirubarajan. *Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software*. Wiley, 2001.

Michèle Basseville and Igor V Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, 1993.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *NeurIPS*, 2018.

Steven X. Ding. *Model-Based Fault Diagnosis Techniques: Design Schemes, Algorithms and Tools*. Springer, 2nd edition, 2013.

Zhiwei Gao, Carlo Cecati, and Steven X. Ding. A survey of fault diagnosis and fault-tolerant techniques—part I: Fault diagnosis with model-based and signal-based approaches. *IEEE Transactions on Industrial Electronics*, 62(6):3757–3767, 2015.

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

Shangding Gu, Laixi Shi, Muning Wen, Ming Jin, Eric Mazumdar, Yuejie Chi, Adam Wierman, and Costas Spanos. Robust gymnasium: A unified modular benchmark for robust reinforcement learning. In *ICLR*, 2025b.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *ICML*, 2018.

Tom Haider, Karsten Roscher, Felippe Schmoeller da Roza, and Stephan Günnemann. Out-of-distribution detection for reinforcement learning agents with probabilistic dynamics models. In *AAMAS*, pp. 851–859, 2023.

Tom Haider, Karsten Roscher, and Stephan Günnemann. Imagining trajectories for anomaly detection in reinforcement learning from images. *Machine Learning*, 115:205, 2026.

Shengyi Huang, Rousslan Fernand Julien Dossa, Chang Ye, Jeff Braga, Dipam Chakraborty, Kinal Mehta, and João GM Araújo. Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *JMLR*, 23(274), 2022.

Eliahu Khalastchi and Meir Kalech. On fault detection and diagnosis in robotic systems. *ACM Computing Surveys*, 51(1):1–24, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *ICDM*, 2008.

Raman K Mehra and John Peschon. An innovations approach to fault detection and diagnosis in dynamic systems. *Automatica*, 7(5):637–640, 1971.

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.

Humphrey Munn, Brendan Tidd, Peter Bohm, Marcus Gallagher, and David Howard. RAPT: Model-predictive out-of-distribution detection and failure diagnosis for sim-to-real humanoid deployment. *arXiv preprint arXiv:2602.01515*, 2026.

Linas Nasvytis, Kai Sandbrink, Jakob Foerster, Tim Franzmeyer, and Christian Schroeder de Witt. Rethinking out-of-distribution detection for reinforcement learning: Advancing methods for evaluation and detection. In *AAMAS*, 2024.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41:100–115, 1954.

Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *MLSDA Workshop at ACML*, 2014.

Sheila Schoepp, Mehran Taghian, Shotaro Miwa, Yoshihiro Mitsuka, Shadan Golestan, and Osmar Zaïane. Enhancing hardware fault tolerance in machines with reinforcement learning policy gradient algorithms. *arXiv preprint arXiv:2407.15283*, 2024.

Skand, Bikram Pandit, Chanho Kim, Fuxin Li, and Stefan Lee. Simple masked training strategies yield control policies that are robust to sensor failure. In *CoRL*, 2024.

Mark Towers, Ariel Kwiatkowski, John Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Jordan Terry, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025.

Franco van Wyk, Yiyang Wang, Anahita Khojandi, and Neda Masoud. Real-time sensor anomaly detection and identification in automated vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):1264–1276, 2020.

Kevin Vogt-Lowell, Theodoros Tsiligkaridis, Rodney Lafuente-Mercado, Surabhi Ghatti, Shanghua Gao, Marinka Zitnik, and Daniela Rus. When sensors fail: Temporal sequence models for robust PPO under sensor drift. In *ICLR Workshop on Continual and Adaptive Optimization (CAO)*, 2026. arXiv:2603.04648.

Alan S Willsky. A survey of design methods for failure detection in dynamic systems. *Automatica*, 12(6):601–611, 1976.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. In *NeurIPS*, 2020.

A FAULT FAMILIES

All faults act per channel on the received observation, begin at a random onset step in $[20, 200]$ (or by the middle of a short episode) and persist to the end of the episode. The per-step ground truth marks the sensor's mode, so a channel is labelled faulty from its onset regardless of whether the reported value happens to coincide with the clean value. Magnitudes are relative to the clean per-channel spread σ_d estimated from the training stream.

Family	Group	Received value	Parameters (sampled per episode)
freeze	PP	value held at onset	–
lag	PP	clean value k steps earlier	$k \sim \mathcal{U}\{2, \dots, 10\}$
drift	PP	clean $+ s\,\rho\,\sigma_d\,(t - t_0 + 1)$	$\rho \sim \log\text{-}\mathcal{U}[0.003, 0.05]$, $s \in \{\pm 1\}$
spike	SV	clean $\pm\, m\sigma_d$ w.p. p per step	$p \sim \mathcal{U}[0.05, 0.3]$, $m \sim \mathcal{U}[3, 8]$
noise	SV	clean $+ \mathcal{N}(0, (c\sigma_d)^2)$	$c \sim \mathcal{U}[0.5, 2]$
dropout	SV	0	–
swap	SV	two channels exchanged	channels sampled without replacement
gain	mixed	g·clean	$g \in \{0.5, 0.7, 1.4, 2.0\}$
saturation	mixed	clip(clean, $\pm f\sigma_d$)	$f \sim \mathcal{U}[0.3, 0.8]$

One implementation serves offline injection into logged episodes (head training and open-loop controls) and online injection inside a Gym wrapper (closed-loop evaluation and control). Unit tests check, for every family, that nothing is altered before onset, that untouched channels are bit-identical

Table 2: Held-out AUROC. Hopper-v5, 1M steps, 5 seeds. Held-out fault family in columns; PP = plausibility-preserving, SV = surprise-visible. Mean±std over seeds.

Detector	freeze PP	lag PP	drift PP	spike SV	noise SV	dropout SV	swap SV	gain mix	saturation mix	PP mean	all mean
Temporal jump	0.535±0.035	0.589±0.028	0.558±0.032	0.713±0.019	0.914±0.021	0.566±0.039	0.835±0.015	0.567±0.023	0.561±0.029	0.560	0.649
Reconstruction (AE)	0.922±0.043	0.747±0.052	0.853±0.039	0.766±0.032	0.870±0.034	0.908±0.044	0.978±0.009	0.773±0.060	0.813±0.060	0.841	0.848
DEXTER-style (IF)	0.815±0.039	0.575±0.046	0.641±0.065	0.814±0.042	0.777±0.047	0.831±0.048	0.866±0.043	0.620±0.036	0.678±0.041	0.677	0.735
Innovation / obs. spread	0.784±0.052	0.694±0.043	0.712±0.058	0.666±0.027	0.952±0.011	0.793±0.036	0.945±0.018	0.723±0.013	0.703±0.041	0.730	0.775
Innovation / own scale (1-sided)	0.933±0.045	0.948±0.019	0.900±0.047	0.756±0.024	0.978±0.007	0.904±0.025	0.786±0.053	0.861±0.093	0.823±0.020	0.927	0.877
PEDM-style disagreement	0.761±0.032	0.606±0.025	0.720±0.040	0.662±0.039	0.722±0.038	0.746±0.038	0.905±0.024	0.655±0.048	0.649±0.018	0.696	0.714
PEDM-style NLL	0.985±0.011	0.962±0.021	0.973±0.009	0.734±0.039	0.997±0.003	0.972±0.024	0.994±0.010	0.961±0.017	0.858±0.017	0.973	0.937
PEDM-style NLL, windowed	**0.992±0.006**	0.997±0.003	0.987±0.005	0.977±0.004	**1.000±0.000**	0.997±0.004	**0.998±0.002**	0.994±0.005	0.981±0.009	0.992	0.991
CUSUM on residual (k=0.5, h=5)	0.931±0.045	0.941±0.017	0.898±0.046	0.721±0.034	0.968±0.010	0.917±0.019	0.807±0.036	0.863±0.090	0.831±0.022	0.924	0.875
χ^2 innovation test W=32	0.990±0.007	0.997±0.002	0.984±0.006	0.986±0.004	1.000±0.000	0.993±0.004	0.995±0.003	0.987±0.010	0.980±0.005	0.990	0.990
2-sided, windowed W=32	0.990±0.006	0.998±0.002	0.984±0.008	0.985±0.004	1.000±0.000	0.993±0.002	0.994±0.003	0.983±0.014	0.979±0.006	0.990	0.989
DCR (2-sided, multi-scale)	0.989±0.010	**0.999±0.001**	**0.988±0.007**	**0.989±0.004**	1.000±0.000	**0.998±0.002**	0.998±0.003	**0.997±0.004**	**0.989±0.007**	0.992	0.994
Learned head (LOFO)	0.991±0.003	0.955±0.015	0.921±0.021	0.982±0.003	1.000±0.000	0.997±0.001	0.996±0.005	0.974±0.014	0.957±0.017	0.956	0.975

Table 3: Channel-level held-out AUROC (the faulted channel's own score vs. its label). Hopper-v5, 1M steps, 5 seeds. Held-out fault family in columns; PP = plausibility-preserving, SV = surprise-visible. Mean±std over seeds.

Detector	freeze PP	lag PP	drift PP	spike SV	noise SV	dropout SV	swap SV	gain mix	saturation mix	PP mean	all mean
Temporal jump	0.014±0.004	0.588±0.024	0.556±0.018	0.671±0.018	0.936±0.008	0.023±0.005	0.606±0.025	0.534±0.015	0.165±0.025	0.386	0.455
Reconstruction (AE)	0.720±0.039	0.651±0.041	0.706±0.020	0.591±0.012	0.752±0.026	0.739±0.050	0.878±0.018	0.679±0.044	0.681±0.031	0.692	0.711
Innovation / own scale (1-sided)	0.825±0.015	0.866±0.019	0.754±0.034	0.652±0.008	0.935±0.013	0.723±0.042	0.657±0.051	0.790±0.038	0.700±0.025	0.815	0.767
2-sided, windowed W=32	0.915±0.012	0.961±0.011	0.862±0.040	0.965±0.009	0.988±0.003	0.947±0.004	0.968±0.013	0.933±0.008	0.929±0.013	0.913	0.941
DCR (2-sided, multi-scale)	**0.933±0.011**	**0.968±0.008**	**0.883±0.030**	**0.976±0.009**	**0.992±0.002**	**0.967±0.008**	**0.974±0.009**	**0.947±0.008**	**0.951±0.013**	0.928	0.955

to the clean observation, and that the mask equals the set of faulted channels from onset onward, plus family-specific invariants (a frozen channel is constant, a lagged channel equals the clean sequence shifted by k, a drift is linear with slope $\rho\sigma_d$).

B CAUSALITY AND OTHER TESTS

Every detector is tested for causality by perturbing observations and actions after step t and asserting that all scores at or before t are unchanged, at the step level and, where available, at the channel level. The leave-channel-out predictor is tested by perturbing channel d in every input block and asserting that its prediction of channel d is unchanged while its predictions of other channels are not. A regression test covers the frozen-channel case in which the windowed variance of the raw one-step difference became slightly negative through floating-point cancellation and its logarithm undefined.

C TRAINING AND ESTIMATOR DETAILS

SAC follows the CleanRL defaults: two 256-unit hidden layers for actor and critics, batch 256, learning rates 3×10^{-4} (actor) and 10^{-3} (critics and temperature), $\gamma = 0.99$, $\tau = 0.005$, 5k random warm-up steps, automatic temperature tuning, replay of 10^6. Predictors are two-layer 256-unit SiLU MLPs with Gaussian heads, trained for 30 epochs with Adam and a cosine schedule on the full stream up to the checkpoint. Windows are $W \in \{8, 32, 128\}$, fixed before any fault experiment and not revisited. Heads (ablations only) are two-layer 64-unit MLPs trained with a class-balanced logistic loss on 80 recent clean episodes into each of which one fault per training family is injected.

D ADDITIONAL RESULTS

Seed-paired significance. Table 14 gives DCR minus the PEDM-style NLL score, per-step and windowed, per task and family, with a paired t-test over the five seeds. Against the per-step score DCR is ahead in all 36 cells and significantly so in 31. Against the windowed score it is ahead in 25 cells, significantly in 15, and behind in 11, none significantly; the largest deficit is 0.8 points (HalfCheetah drift). The gain from two-sided scoring over windowing alone is therefore concentrated on Walker2d and Ant, and on spike in every task; on Hopper and HalfCheetah the other differences are within seed noise.

Table 4: Input ladder, held-out AUROC. Hopper-v5, 1M steps, 5 seeds. Held-out fault family in columns; PP = plausibility-preserving, SV = surprise-visible. Mean±std over seeds.

Detector	freeze PP	lag PP	drift PP	spike SV	noise SV	dropout SV	swap SV	gain mix	saturation mix	PP mean	all mean
Temporal jump	0.535±0.035	0.589±0.028	0.558±0.032	0.713±0.019	0.914±0.021	0.566±0.039	0.835±0.015	0.567±0.023	0.561±0.029	0.560	0.649
Innovation / own scale (1-sided)	0.933±0.045	0.948±0.019	0.900±0.047	0.756±0.024	0.978±0.007	0.904±0.025	0.786±0.053	0.861±0.093	0.823±0.020	0.927	0.877
2-sided, per step	0.808±0.038	0.767±0.045	0.749±0.031	0.609±0.007	0.894±0.018	0.783±0.082	0.906±0.011	0.754±0.034	0.708±0.019	0.774	0.775
PEDM-style NLL, windowed	0.992±0.006	0.997±0.003	0.987±0.005	0.977±0.004	**1.000±0.000**	0.997±0.004	**0.998±0.002**	0.994±0.005	0.981±0.009	0.992	0.991
2-sided, windowed $W=32$	0.990±0.006	0.998±0.002	0.984±0.008	0.985±0.004	1.000±0.000	0.993±0.002	0.994±0.003	0.983±0.014	0.979±0.006	0.990	0.989
DCR (2-sided, multi-scale)	0.989±0.010	**0.999±0.001**	**0.988±0.007**	0.989±0.004	1.000±0.000	**0.998±0.002**	0.998±0.003	0.997±0.004	**0.989±0.007**	0.992	0.994
DCR on leave-channel-out residual	0.859±0.034	0.646±0.049	0.756±0.068	0.940±0.020	0.829±0.045	0.793±0.077	0.830±0.031	0.723±0.039	0.745±0.093	0.754	0.791
DCR, both residuals	0.987±0.010	0.997±0.003	0.982±0.007	**0.989±0.003**	1.000±0.000	0.998±0.002	0.997±0.003	**0.998±0.001**	0.986±0.008	0.989	0.993
Learned head (LOFO)	0.991±0.003	0.955±0.015	0.921±0.021	0.982±0.003	1.000±0.000	0.997±0.001	0.996±0.005	0.974±0.014	0.957±0.017	0.956	0.975
Learned head (all families)	**0.993±0.002**	0.982±0.007	0.955±0.004	0.982±0.004	1.000±0.000	0.998±0.001	0.998±0.001	0.984±0.007	0.975±0.013	0.977	0.985

Table 5: Channel-level held-out AUROC (the faulted channel's own score vs. its label). Walker2d-v5, 1M steps, 5 seeds. Held-out fault family in columns; PP = plausibility-preserving, SV = surprise-visible. Mean±std over seeds.

Detector	freeze PP	lag PP	drift PP	spike SV	noise SV	dropout SV	swap SV	gain mix	saturation mix	PP mean	all mean
Temporal jump	0.021±0.004	0.508±0.023	0.550±0.030	0.658±0.016	0.918±0.012	0.028±0.006	0.535±0.021	0.497±0.023	0.201±0.017	0.360	0.435
Reconstruction (AE)	0.693±0.037	0.612±0.039	0.776±0.027	0.591±0.009	0.767±0.021	0.701±0.045	0.820±0.018	0.625±0.045	0.631±0.030	0.694	0.691
Innovation / own scale (1-sided)	0.726±0.022	0.787±0.017	0.654±0.025	0.631±0.007	0.899±0.008	0.638±0.044	0.658±0.036	0.699±0.043	0.586±0.036	0.723	0.698
2-sided, windowed $W=32$	0.880±0.036	0.934±0.012	0.781±0.026	0.965±0.009	0.982±0.004	0.888±0.058	0.899±0.009	0.881±0.044	0.808±0.034	0.865	0.891
DCR (2-sided, multi-scale)	**0.919±0.022**	**0.959±0.008**	**0.819±0.037**	**0.983±0.005**	**0.988±0.002**	**0.930±0.038**	**0.930±0.014**	**0.933±0.025**	**0.872±0.030**	0.899	0.926

Detection delay. At a 0.1% per-step false-alarm threshold on Hopper (Table 15) the one-sided gate, the PEDM-style scores, the classical test and DCR alarm within a median of 0–4 steps of onset and detect at least 99% of faulted episodes; the learned head needs 11–35 steps on the plausibility-preserving families; temporal jump needs 97–138 steps on freeze, lag and drift and detects 35–54%. HalfCheetah behaves like Hopper (Table 16). On Walker2d (Table 17) DCR needs 75–200 steps on every family but swap and detects 73–91% of episodes at this threshold, and on Ant (Table 18) it detects only 13–65% of frozen, lagged, drifting, dropped, gain- and saturation-faulted episodes. First-alarm detection of a persistent fault is easy where per-step AUROC is high and not otherwise; per-step AUROC measures the harder target a controller needs.

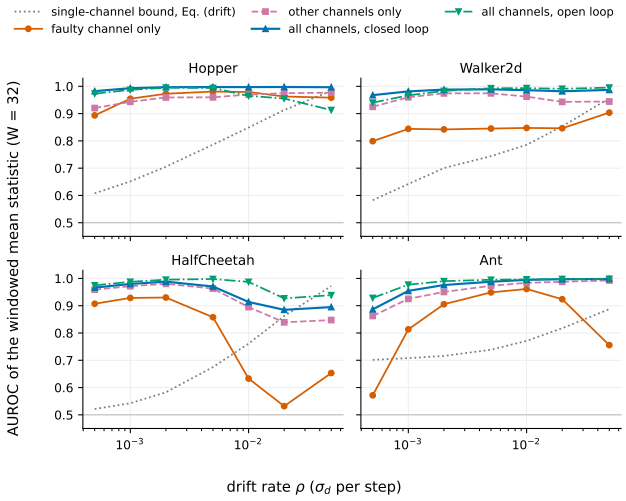


Figure 2: Drift detectability vs. rate at 1M steps ($W=32$). Dotted: single-channel bound Eq. (1). The other channels supply most of the detectability at slow rates, in open loop as in closed loop; the faulty channel's own term exceeds the bound because the 1M model predicts the channel from the dynamics rather than propagating it (150k in Appendix D).

Table 6: Channel-level held-out AUROC (the faulted channel's own score vs. its label). HalfCheetah-v5, 1M steps, 5 seeds. Held-out fault family in columns; PP = plausibility-preserving, SV = surprise-visible. Mean±std over seeds.

Detector	freeze PP	lag PP	drift PP	spike SV	noise SV	dropout SV	swap SV	gain mix	saturation mix	PP mean	all mean
Temporal jump	0.004±0.002	0.491±0.020	0.581±0.032	0.639±0.005	0.706±0.020	0.002±0.001	0.490±0.012	0.485±0.021	0.253±0.026	0.359	0.406
Reconstruction (AE)	0.756±0.019	0.730±0.027	**0.859±0.017**	0.566±0.005	0.756±0.013	0.708±0.046	0.683±0.016	0.693±0.025	0.642±0.039	0.782	0.710
Innovation / own scale (1-sided)	0.832±0.013	0.809±0.012	0.708±0.021	0.579±0.016	0.833±0.019	0.843±0.024	0.708±0.013	0.818±0.012	0.773±0.049	0.783	0.767
2-sided, windowed $W=32$	0.952±0.019	0.948±0.021	0.685±0.047	0.962±0.011	0.975±0.009	0.967±0.035	0.840±0.031	0.967±0.017	0.947±0.049	0.861	0.916
DCR (2-sided, multi-scale)	**0.963±0.014**	**0.957±0.018**	0.707±0.043	**0.979±0.011**	**0.983±0.008**	**0.975±0.030**	**0.864±0.029**	**0.974±0.015**	**0.954±0.044**	0.876	0.929

Table 7: Channel-level held-out AUROC (the faulted channel's own score vs. its label). Ant-v5, 1M steps, 5 seeds. Held-out fault family in columns; PP = plausibility-preserving, SV = surprise-visible. Mean±std over seeds.

Detector	freeze PP	lag PP	drift PP	spike SV	noise SV	dropout SV	swap SV	gain mix	saturation mix	PP mean	all mean
Temporal jump	0.360±0.003	0.510±0.031	0.801±0.007	0.630±0.029	0.922±0.002	0.361±0.001	0.499±0.041	0.511±0.028	0.437±0.016	0.557	0.559
Reconstruction (AE)	0.524±0.027	0.508±0.049	**0.914±0.013**	0.542±0.020	0.842±0.037	0.528±0.032	0.604±0.055	0.535±0.024	0.517±0.016	0.649	0.613
Innovation / own scale (1-sided)	0.503±0.026	0.521±0.035	0.756±0.072	0.600±0.027	0.945±0.027	0.494±0.025	0.544±0.026	0.523±0.027	0.492±0.034	0.594	0.598
2-sided, windowed $W=32$	0.674±0.049	0.810±0.033	0.776±0.060	0.979±0.019	0.989±0.021	0.673±0.060	0.761±0.047	0.671±0.029	0.657±0.041	0.753	0.777
DCR (2-sided, multi-scale)	**0.707±0.047**	**0.839±0.029**	0.790±0.060	**0.988±0.013**	**0.991±0.015**	**0.710±0.052**	**0.786±0.039**	**0.686±0.033**	**0.687±0.037**	0.779	0.798

Drift sweep details. The sweep uses a positive drift sign, so the windowed mean is compared with the bound directly; the slowest rate, 5×10^{-4}, lies below the family's training range (0.003–0.05) by design.

Drift at 150k steps. Table 24 gives the sweep at both checkpoints. At 150k steps on Hopper the faulty channel's own windowed-mean AUROC at the slowest rate ($\rho = 5 \times 10^{-4}$, $W = 32$) is 0.661 against a bound of 0.606, i.e. on the bound, while on Walker2d and Ant it falls below chance (0.403 and 0.212): an immature predictor neither propagates nor pins the channel. The remaining channels' term is above 0.80 on Hopper and Walker2d at that rate, so the decomposition's conclusion holds at both checkpoints.

E CONTROLLER DEVELOPMENT

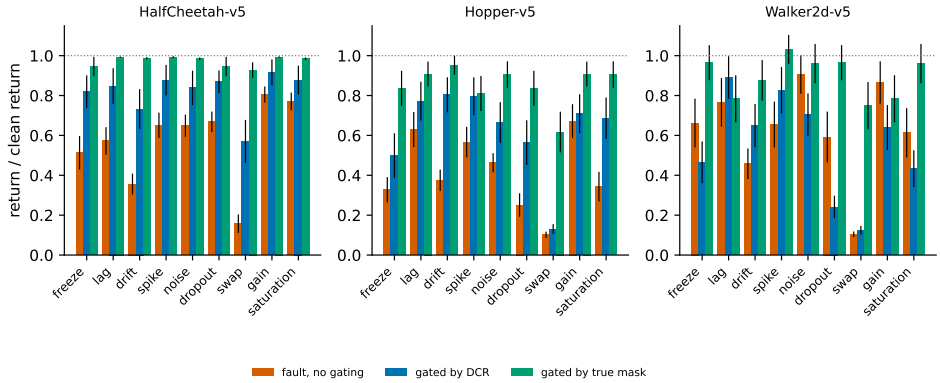


Figure 3: Return relative to clean under each fault family, without gating, gated by DCR, and gated by the true fault mask with the same dead-reckoning repair. Error bars are standard errors over episodes.

Version 1 replaced a flagged channel by the leave-channel-out prediction at a fixed probability threshold with no latching; false alarms replaced good channels with weak guesses and return fell below the ungated policy on most families. Version 2 dead-reckoned through the full predictor with a latching alarm; return still fell, because every channel was eventually latched. Diagnosis showed the first latch was correct in nearly every episode, followed by a cascade: the dead-reckoned value drifts, the other channels' predictions are computed from it, their residuals inflate, they are latched in turn. Version 3 isolated one culprit per alarm and tested the remaining channels with the culprit hidden

Table 8: Input ladder, held-out AUROC. Walker2d-v5, 1M steps, 5 seeds. Held-out fault family in columns; PP = plausibility-preserving, SV = surprise-visible. Mean±std over seeds.

Detector	freeze PP	lag PP	drift PP	spike SV	noise SV	dropout SV	swap SV	gain mix	saturation mix	PP mean	all mean
Temporal jump	0.541±0.019	0.564±0.021	0.612±0.028	0.721±0.013	0.858±0.026	0.540±0.015	0.764±0.032	0.544±0.016	0.534±0.007	0.572	0.631
Innovation / own scale (1-sided)	0.862±0.045	0.890±0.026	0.827±0.026	0.709±0.029	0.954±0.008	0.844±0.031	0.870±0.049	0.808±0.041	0.703±0.020	0.860	0.830
2-sided, per step	0.541±0.055	0.545±0.035	0.574±0.032	0.574±0.015	0.669±0.033	0.531±0.025	0.693±0.013	0.521±0.039	0.521±0.024	0.553	0.574
PEDM-style NLL, windowed	0.979±0.011	0.982±0.010	0.971±0.006	0.980±0.003	0.999±0.001	0.978±0.011	**0.998±0.002**	0.942±0.028	0.892±0.031	0.977	0.969
2-sided, windowed $W=32$	0.985±0.007	0.993±0.005	0.967±0.008	0.984±0.002	0.999±0.001	0.983±0.009	0.992±0.006	0.970±0.007	0.907±0.026	0.982	0.976
DCR (2-sided, multi-scale)	0.988±0.007	**0.995±0.005**	**0.972±0.008**	0.993±0.003	0.998±0.002	0.990±0.006	0.998±0.002	**0.984±0.008**	**0.953±0.022**	0.985	0.986
DCR on leave-channel-out residual	0.817±0.064	0.735±0.051	0.883±0.051	0.962±0.019	0.926±0.049	0.814±0.052	0.950±0.026	0.715±0.080	0.696±0.071	0.812	0.833
DCR, both residuals	0.983±0.010	0.993±0.006	0.965±0.009	**0.995±0.002**	0.998±0.002	0.991±0.005	0.997±0.002	0.974±0.016	0.948±0.033	0.980	0.983
Learned head (LOFO)	**0.993±0.002**	0.913±0.037	0.927±0.011	0.982±0.003	**0.999±0.001**	0.996±0.001	0.992±0.007	0.898±0.032	0.888±0.031	0.944	0.954
Learned head (all families)	0.993±0.002	0.969±0.005	0.935±0.007	0.984±0.002	0.999±0.001	**0.996±0.001**	0.995±0.002	0.929±0.022	0.893±0.033	0.966	0.966

Table 9: Input ladder, held-out AUROC. HalfCheetah-v5, 1M steps, 5 seeds. Held-out fault family in columns; PP = plausibility-preserving, SV = surprise-visible. Mean±std over seeds.

Detector	freeze PP	lag PP	drift PP	spike SV	noise SV	dropout SV	swap SV	gain mix	saturation mix	PP mean	all mean
Temporal jump	0.418±0.063	0.468±0.092	0.252±0.095	0.611±0.047	0.615±0.046	0.445±0.072	0.663±0.057	0.502±0.027	0.442±0.025	0.380	0.491
Innovation / own scale (1-sided)	0.899±0.024	0.875±0.044	0.692±0.100	0.672±0.025	0.927±0.022	0.930±0.033	0.764±0.104	0.907±0.017	0.863±0.083	0.822	0.837
2-sided, per step	0.634±0.018	0.644±0.021	0.726±0.029	0.598±0.023	0.631±0.023	0.606±0.010	0.753±0.036	0.569±0.022	0.576±0.006	0.668	0.638
PEDM-style NLL, windowed	0.997±0.004	0.994±0.013	**0.991±0.002**	0.985±0.003	**1.000±0.000**	**1.000±0.000**	0.999±0.001	**1.000±0.000**	0.983±0.022	0.994	0.994
2-sided, windowed $W=32$	0.997±0.001	0.996±0.003	0.969±0.027	0.987±0.003	0.999±0.002	0.999±0.001	0.989±0.015	0.998±0.001	0.981±0.026	0.987	0.991
DCR (2-sided, multi-scale)	**0.998±0.001**	**0.998±0.002**	0.983±0.014	**0.996±0.001**	1.000±0.000	1.000±0.000	0.998±0.001	0.999±0.000	**0.985±0.020**	0.993	0.995
DCR on leave-channel-out residual	0.968±0.032	0.937±0.027	0.931±0.068	0.983±0.018	0.975±0.022	0.953±0.047	0.960±0.038	0.918±0.050	0.931±0.057	0.945	0.951
DCR, both residuals	0.997±0.004	0.982±0.010	0.962±0.042	0.992±0.012	0.996±0.008	1.000±0.000	0.974±0.037	0.999±0.001	0.977±0.030	0.980	0.987
Learned head (LOFO)	0.997±0.001	0.957±0.022	0.953±0.012	0.979±0.007	0.993±0.002	0.998±0.001	0.990±0.013	0.971±0.010	0.976±0.017	0.969	0.979
Learned head (all families)	0.997±0.001	0.979±0.014	0.966±0.024	0.982±0.005	0.994±0.002	0.998±0.000	**1.000±0.000**	0.983±0.007	0.980±0.019	0.981	0.986

from the predictor input, but against a threshold calibrated for the full-input test; version 4 calibrated a separate threshold for the hidden-input test and capped simultaneous flags at three. Only version 4 recovers return.

F Additional remarks on the theory

Proof of Proposition 2. Let $X \sim \phi_{\#}\mathbb{P}_f$ and $Y \sim \phi_{\#}\mathbb{P}_0$ be independent scores. Take a maximal coupling $(X, \tilde{X})$ of $\phi_{\#}\mathbb{P}_f$ and $\phi_{\#}\mathbb{P}_0$, so $\Pr(X \neq \tilde{X}) = \mathrm{TV}(\phi_{\#}\mathbb{P}_f, \phi_{\#}\mathbb{P}_0) \leq \mathrm{TV}(\mathbb{P}_f, \mathbb{P}_0)$. Then $\Pr(X > Y) + \frac{1}{2}\Pr(X = Y)$ differs from $\Pr(\tilde{X} > Y) + \frac{1}{2}\Pr(\tilde{X} = Y) = \frac{1}{2}$ by at most $\Pr(X \neq \tilde{X})$.

Proof of Proposition 1. Let $r'' \sim P_0$ be independent of r. Stochastic dominance gives a coupling $(\tilde{r}', \tilde{r}'')$ with $\tilde{r}' \overset{d}{=} r'$, $\tilde{r}'' \overset{d}{=} r''$ and $|\tilde{r}'| \leq |\tilde{r}''|$ almost surely (Strassen). Since h is non-decreasing, $h(|\tilde{r}'|) \leq h(|\tilde{r}''|)$, hence $\Pr(h(|r'|) > h(|r|)) + \frac{1}{2}\Pr(h(|r'|) = h(|r|)) \leq \Pr(h(|r''|) > h(|r|)) + \frac{1}{2}\Pr(h(|r''|) = h(|r|))$, and the right side equals $\frac{1}{2}$ by exchangeability of (r'', r). If $r' \equiv 0$ and h is strictly increasing, $\Pr(h(0) > h(|r|)) = 0$ and $\Pr(h(0) = h(|r|)) = \Pr(|r| = 0) = 0$ by continuity, so the AUROC is 0.

Remark 2 (Aggregation does not rescue it in open loop)**.** *If a step-level score is the maximum over channels and the other channels keep their clean distribution, a frozen channel contributes $h(0)$ to the maximum, so the faulty-step score is the maximum over $D-1$ clean channels, which is stochastically smaller than the clean-step maximum over D channels, and the step-level AUROC is also below $\frac{1}{2}$. In closed loop the premise fails: the policy reacts to the frozen reading and moves the other channels, which can mask the inversion at the step level while leaving it intact at the channel level (§6.1).*

Remark 3 (Adversarial perturbations are the designed limit)**.** *An adversarial observation perturbation is constructed to be small and to look like a valid observation. To the extent that it keeps the residual within its clean range it is a plausibility-preserving fault, and Proposition 1 applies to every surprise-based score; and by Proposition 2 an attack that is to remain undetectable to* any *history-only estimator must produce a history whose law is close to that of clean operation, that is, one consistent with the dynamics and the executed actions. How much a consistency-constrained attacker can still move the policy is a question we do not address here.*

Table 10: Input ladder, held-out AUROC. Ant-v5, 1M steps, 5 seeds. Held-out fault family in columns; PP = plausibility-preserving, SV = surprise-visible. Mean±std over seeds.

Detector	freeze PP	lag PP	drift PP	spike SV	noise SV	dropout SV	swap SV	gain mix	saturation mix	PP mean	all mean
Temporal jump	0.510±0.007	0.508±0.012	0.524±0.013	0.618±0.015	0.584±0.022	0.506±0.011	0.589±0.025	0.498±0.008	0.501±0.007	0.514	0.538
Innovation / own scale (1-sided)	0.562±0.014	0.572±0.025	0.925±0.031	0.655±0.012	0.945±0.018	0.548±0.021	0.676±0.026	0.560±0.013	0.547±0.005	0.687	0.666
2-sided, per step	0.498±0.008	0.502±0.007	0.415±0.025	0.550±0.011	0.523±0.024	0.513±0.012	0.583±0.029	0.508±0.026	0.506±0.017	0.471	0.511
PEDM-style NLL, windowed	0.607±0.020	0.679±0.027	0.966±0.010	0.986±0.004	**0.999±0.001**	0.598±0.029	0.790±0.044	0.593±0.019	0.586±0.017	0.751	0.756
2-sided, windowed $W=32$	0.623±0.027	0.712±0.018	0.975±0.015	0.989±0.004	0.996±0.005	0.610±0.029	0.826±0.031	0.599±0.022	0.594±0.012	0.770	0.769
DCR (2-sided, multi-scale)	**0.654±0.031**	**0.756±0.015**	0.971±0.014	**0.995±0.001**	0.996±0.005	0.640±0.036	**0.854±0.034**	**0.625±0.027**	**0.633±0.017**	0.794	0.792
DCR on leave-channel-out residual	0.552±0.032	0.669±0.022	0.974±0.018	0.989±0.007	0.955±0.031	0.547±0.057	0.823±0.048	0.543±0.028	0.524±0.048	0.732	0.731
DCR, both residuals	0.569±0.027	0.695±0.017	**0.977±0.016**	0.995±0.004	0.980±0.027	0.560±0.050	0.829±0.048	0.555±0.021	0.546±0.035	0.747	0.745
Learned head (LOFO)	0.611±0.015	0.658±0.011	0.919±0.036	0.953±0.017	0.982±0.010	0.619±0.017	0.829±0.047	0.566±0.019	0.548±0.022	0.729	0.743
Learned head (all families)	0.647±0.026	0.674±0.012	0.960±0.018	0.972±0.013	0.988±0.009	**0.656±0.020**	0.822±0.045	0.573±0.019	0.595±0.035	0.760	0.765

Table 11: Held-out AUROC. Walker2d-v5, 1M steps, 5 seeds. Held-out fault family in columns; PP = plausibility-preserving, SV = surprise-visible. Mean±std over seeds.

Detector	freeze PP	lag PP	drift PP	spike SV	noise SV	dropout SV	swap SV	gain mix	saturation mix	PP mean	all mean
Temporal jump	0.541±0.019	0.564±0.021	0.612±0.028	0.721±0.013	0.858±0.026	0.540±0.015	0.764±0.032	0.544±0.016	0.534±0.007	0.572	0.631
Reconstruction (AE)	0.773±0.026	0.640±0.049	0.873±0.026	0.734±0.028	0.793±0.042	0.732±0.030	0.930±0.052	0.627±0.050	0.661±0.044	0.762	0.752
DEXTER-style (IF)	0.738±0.059	0.509±0.034	0.722±0.032	0.766±0.038	0.706±0.053	0.714±0.051	0.827±0.045	0.540±0.045	0.586±0.068	0.656	0.679
Innovation / obs. spread	0.691±0.009	0.611±0.031	0.773±0.013	0.717±0.015	0.885±0.023	0.685±0.024	0.878±0.029	0.590±0.036	0.612±0.010	0.692	0.716
Innovation / own scale (1-sided)	0.862±0.045	0.890±0.026	0.827±0.026	0.709±0.029	0.954±0.008	0.844±0.031	0.870±0.049	0.808±0.041	0.703±0.020	0.860	0.830
PEDM-style disagreement	0.667±0.013	0.577±0.036	0.775±0.026	0.677±0.029	0.696±0.050	0.637±0.022	0.855±0.041	0.578±0.043	0.581±0.018	0.673	0.672
PEDM-style NLL	0.910±0.023	0.880±0.037	0.925±0.018	0.738±0.019	0.975±0.010	0.880±0.020	0.978±0.007	0.832±0.036	0.732±0.024	0.905	0.872
PEDM-style NLL, windowed	0.979±0.011	0.982±0.010	0.971±0.006	0.980±0.003	0.999±0.001	0.978±0.011	**0.998±0.002**	0.942±0.028	0.892±0.031	0.977	0.969
CUSUM on residual (k=0.5, h=5)	0.853±0.048	0.863±0.032	0.830±0.029	0.681±0.044	0.921±0.016	0.838±0.034	0.863±0.046	0.791±0.041	0.709±0.023	0.849	0.817
χ^2 innovation test $W=32$	0.985±0.007	0.993±0.005	0.965±0.007	0.984±0.001	0.999±0.001	0.984±0.007	0.993±0.005	0.969±0.007	0.909±0.025	0.981	0.976
2-sided, windowed $W=32$	0.985±0.007	0.993±0.005	0.967±0.008	0.984±0.002	0.999±0.001	0.983±0.009	0.992±0.006	0.970±0.007	0.907±0.026	0.982	0.976
DCR (2-sided, multi-scale)	0.988±0.007	**0.995±0.005**	**0.972±0.008**	**0.993±0.003**	0.998±0.002	0.990±0.006	0.998±0.002	**0.984±0.008**	**0.953±0.022**	0.985	0.986
Learned head (LOFO)	**0.993±0.002**	0.913±0.037	0.927±0.011	0.982±0.003	**0.999±0.001**	**0.996±0.001**	0.992±0.007	0.898±0.032	0.888±0.031	0.944	0.954

Table 12: Held-out AUROC. HalfCheetah-v5, 1M steps, 5 seeds. Held-out fault family in columns; PP = plausibility-preserving, SV = surprise-visible. Mean±std over seeds.

Detector	freeze PP	lag PP	drift PP	spike SV	noise SV	dropout SV	swap SV	gain mix	saturation mix	PP mean	all mean
Temporal jump	0.418±0.063	0.468±0.092	0.252±0.095	0.611±0.047	0.615±0.046	0.445±0.072	0.663±0.057	0.502±0.027	0.442±0.025	0.380	0.491
Reconstruction (AE)	0.889±0.041	0.881±0.047	0.956±0.011	0.767±0.060	0.870±0.033	0.872±0.056	0.975±0.020	0.722±0.061	0.714±0.077	0.909	0.850
DEXTER-style (IF)	0.902±0.017	0.811±0.052	0.923±0.014	0.846±0.035	0.842±0.026	0.885±0.028	0.956±0.021	0.776±0.034	0.807±0.032	0.879	0.861
Innovation / obs. spread	0.908±0.043	0.899±0.066	0.928±0.022	0.715±0.040	0.914±0.041	0.888±0.064	0.968±0.028	0.772±0.079	0.763±0.099	0.912	0.862
Innovation / own scale (1-sided)	0.899±0.024	0.875±0.044	0.692±0.100	0.672±0.025	0.927±0.022	0.930±0.033	0.764±0.104	0.907±0.017	0.863±0.083	0.822	0.837
PEDM-style disagreement	0.781±0.084	0.785±0.100	0.917±0.020	0.729±0.070	0.773±0.075	0.737±0.118	0.932±0.058	0.643±0.058	0.633±0.076	0.828	0.770
PEDM-style NLL	0.980±0.009	0.977±0.014	0.968±0.008	0.754±0.039	0.983±0.003	0.981±0.007	0.994±0.004	0.955±0.009	0.900±0.050	0.975	0.943
PEDM-style NLL, windowed	0.997±0.004	0.994±0.013	**0.991±0.002**	0.985±0.003	**1.000±0.000**	**1.000±0.000**	**0.999±0.001**	**1.000±0.000**	0.983±0.022	0.994	0.994
CUSUM on residual (k=0.5, h=5)	0.930±0.024	0.922±0.043	0.790±0.086	0.741±0.047	0.944±0.022	0.947±0.019	0.818±0.106	0.921±0.017	0.907±0.056	0.880	0.880
χ^2 innovation test $W=32$	0.997±0.002	0.995±0.005	0.972±0.026	0.988±0.004	0.999±0.000	0.999±0.001	0.995±0.003	0.999±0.001	0.981±0.026	0.988	0.992
2-sided, windowed $W=32$	0.997±0.001	0.996±0.003	0.969±0.027	0.987±0.003	0.999±0.002	0.999±0.001	0.989±0.015	0.998±0.001	0.981±0.026	0.987	0.991
DCR (2-sided, multi-scale)	**0.998±0.001**	**0.998±0.002**	0.983±0.014	**0.996±0.001**	1.000±0.000	1.000±0.000	0.998±0.001	0.999±0.000	**0.985±0.020**	0.993	0.995
Learned head (LOFO)	0.997±0.001	0.957±0.022	0.953±0.012	0.979±0.007	0.993±0.002	0.998±0.001	0.990±0.013	0.971±0.010	0.976±0.017	0.969	0.979

Table 13: Held-out AUROC. Ant-v5, 1M steps, 5 seeds. Held-out fault family in columns; PP = plausibility-preserving, SV = surprise-visible. Mean±std over seeds.

Detector	freeze PP	lag PP	drift PP	spike SV	noise SV	dropout SV	swap SV	gain mix	saturation mix	PP mean	all mean
Temporal jump	0.510±0.007	0.508±0.012	0.524±0.013	0.618±0.015	0.584±0.022	0.506±0.011	0.589±0.025	0.498±0.008	0.501±0.007	0.514	0.538
Reconstruction (AE)	0.536±0.012	0.520±0.011	0.944±0.017	0.583±0.012	0.809±0.038	0.537±0.021	0.660±0.025	0.534±0.034	0.520±0.012	0.667	0.627
DEXTER-style (IF)	0.522±0.020	0.487±0.013	0.615±0.046	0.555±0.026	0.559±0.026	0.527±0.021	0.616±0.014	0.506±0.024	0.507±0.014	0.541	0.544
Innovation / obs. spread	0.512±0.010	0.518±0.003	0.675±0.032	0.623±0.013	0.639±0.013	0.527±0.015	0.637±0.021	0.520±0.030	0.518±0.015	0.568	0.574
Innovation / own scale (1-sided)	0.562±0.014	0.572±0.025	0.925±0.031	0.655±0.012	0.945±0.018	0.548±0.021	0.676±0.026	0.560±0.013	0.547±0.005	0.687	0.666
PEDM-style disagreement	0.520±0.012	0.511±0.005	0.826±0.026	0.569±0.014	0.618±0.023	0.532±0.020	0.640±0.023	0.522±0.030	0.515±0.012	0.619	0.584
PEDM-style NLL	0.564±0.012	0.568±0.026	0.944±0.011	0.653±0.012	0.941±0.010	0.559±0.026	0.692±0.032	0.561±0.014	0.546±0.010	0.692	0.670
PEDM-style NLL, windowed	0.607±0.020	0.679±0.027	0.966±0.010	0.986±0.004	**0.999±0.001**	0.598±0.029	0.790±0.044	0.593±0.019	0.586±0.017	0.751	0.756
CUSUM on residual (k=0.5, h=5)	0.576±0.013	0.587±0.022	0.964±0.017	0.762±0.019	0.956±0.012	0.568±0.020	0.700±0.024	0.572±0.009	0.555±0.009	0.709	0.693
χ^2 innovation test $W=32$	0.623±0.027	0.712±0.018	0.972±0.015	0.989±0.004	0.995±0.008	0.609±0.031	0.824±0.031	0.599±0.022	0.593±0.014	0.769	0.768
2-sided, windowed $W=32$	0.623±0.027	0.712±0.018	**0.975±0.015**	0.989±0.004	0.996±0.005	0.610±0.029	0.826±0.031	0.599±0.022	0.594±0.012	0.770	0.769
DCR (2-sided, multi-scale)	**0.654±0.031**	**0.756±0.015**	0.971±0.014	**0.995±0.001**	0.996±0.005	**0.640±0.036**	**0.854±0.034**	**0.625±0.027**	**0.633±0.017**	0.794	0.792
Learned head (LOFO)	0.611±0.015	0.658±0.011	0.919±0.036	0.953±0.017	0.982±0.010	0.619±0.017	0.829±0.047	0.566±0.019	0.548±0.022	0.729	0.743

Table 14: DCR minus the PEDM-style NLL baseline, per-step and windowed (W=32), in AUROC points at 1M steps: mean over five seeds with a 95% confidence interval from a paired t-test over seeds; * marks $p < 0.05$ (two-sided). A positive value favours DCR.

Family	Hopper		Walker2d		HalfCheetah		Ant	
	per-step	windowed	per-step	windowed	per-step	windowed	per-step	windowed
freeze	+0.4* [+0.1, +0.6]	-0.3 [-0.9, +0.3]	+7.9* [+5.3, +10.4]	+1.0* [+0.0, +1.9]	+1.9* [+0.7, +3.0]	+0.1 [-0.4, +0.6]	+9.0* [+6.1, +11.9]	+4.6* [+2.5, +6.8]
lag	+3.7* [+1.1, +6.3]	+0.2 [-0.2, +0.5]	+11.5* [+7.1, +15.8]	+1.3* [+0.4, +2.3]	+2.1* [+0.6, +3.7]	+0.4 [-0.9, +1.8]	+18.8* [+15.1, +22.6]	+7.7* [+5.0, +10.4]
drift	+1.5* [+0.7, +2.4]	+0.1 [-0.8, +1.1]	+4.7* [+1.6, +7.8]	+0.1 [-1.3, +1.5]	+1.6 [-0.5, +3.6]	-0.8 [-2.6, +0.9]	+2.7* [+1.0, +4.4]	+0.5 [-1.5, +2.4]
spike	+25.5* [+21.0, +30.0]	+1.1* [+0.8, +1.4]	+25.6* [+22.9, +28.2]	+1.3* [+0.8, +1.9]	+24.3* [+19.5, +29.0]	+1.2* [+0.9, +1.5]	+34.2* [+32.9, +35.5]	+0.9* [+0.5, +1.3]
noise	+0.3 [-0.1, +0.7]	-0.0 [-0.0, +0.0]	+2.3* [+1.1, +3.6]	-0.0 [-0.2, +0.1]	+1.7* [+1.3, +2.1]	-0.0 [-0.0, +0.0]	+5.5* [+4.2, +6.7]	-0.3 [-0.8, +0.3]
dropout	+2.6 [-0.4, +5.6]	+0.1 [-0.3, +0.6]	+11.0* [+8.6, +13.5]	+1.2* [+0.2, +2.2]	+1.9* [+0.9, +2.8]	-0.0 [-0.0, +0.0]	+8.1* [+5.9, +10.4]	+4.2* [+2.2, +6.3]
swap	+0.3 [-0.6, +1.2]	-0.1 [-0.3, +0.1]	+2.0* [+1.1, +2.9]	-0.0 [-0.2, +0.1]	+0.4 [-0.1, +0.8]	-0.1 [-0.3, +0.0]	+16.2* [+12.8, +19.5]	+6.5* [+3.8, +9.1]
gain	+3.6* [+1.7, +5.4]	+0.2 [-0.3, +0.8]	+15.2* [+10.2, +20.3]	+4.3* [+0.4, +8.2]	+4.4* [+3.3, +5.5]	-0.0 [-0.1, +0.0]	+6.5* [+2.4, +10.5]	+3.3* [+1.3, +5.3]
saturation	+13.1* [+10.7, +15.5]	+0.8 [-0.7, +2.4]	+22.1* [+19.0, +25.2]	+6.1* [+4.2, +8.0]	+8.5* [+4.6, +12.5]	+0.2 [-0.1, +0.5]	+8.8* [+7.4, +10.1]	+4.8* [+3.2, +6.4]

Table 15: Median detection delay in steps at 0.1% per-step FAR (lower is better; median within a seed, mean and std across seeds). Hopper-v5, 1M steps, 5 seeds. Held-out fault family in columns; PP = plausibility-preserving, SV = surprise-visible. Mean±std over seeds.

Detector	freeze PP	lag PP	drift PP	spike SV	noise SV	dropout SV	swap SV	gain mix	saturation mix	PP mean	all mean
Temporal jump	96.6±18.9	133.8±42.3	137.7±29.9	**2.8±0.6**	0.8±0.4	5.6±7.3	**0.0±0.0**	87.3±76.3	49.7±39.0	122.7	57.1
Reconstruction (AE)	22.0±5.7	66.6±44.3	80.9±27.3	2.9±0.5	2.7±2.0	8.6±12.4	**0.0±0.0**	85.5±92.1	32.6±34.0	56.5	33.5
DEXTER-style (IF)	140.8±38.8	270.4±82.5	176.3±17.1	237.3±23.1	228.4±29.2	133.9±74.6	21.0±6.0	390.4±248.8	153.1±38.0	195.8	194.6
Innovation / obs. spread	98.4±49.0	175.6±101.5	119.9±41.8	**2.8±0.6**	1.4±0.5	16.2±12.5	**0.0±0.0**	164.4±143.8	82.7±26.0	131.3	73.5
Innovation / own scale (1-sided)	1.3±0.7	**0.0±0.0**	**1.6±1.6**	**2.8±0.6**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	1.0	0.6
PEDM-style disagreement	87.7±51.5	178.1±85.4	120.9±27.4	5.8±2.0	39.2±30.1	65.9±56.2	1.3±0.4	204.6±125.4	98.7±55.2	128.9	89.1
PEDM-style NLL	**1.2±0.4**	**0.0±0.0**	1.7±1.6	**2.8±0.6**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	1.0	0.6
PEDM-style NLL, windowed	1.5±1.1	**0.0±0.0**	2.8±1.8	3.0±0.8	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	1.4	0.8
CUSUM on residual (k=0.5, h=5)	1.3±0.7	**0.0±0.0**	2.3±2.2	2.9±0.4	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	1.2	0.7
χ^2 innovation test W=32	1.6±1.3	**0.0±0.0**	3.8±2.2	3.1±0.7	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	1.8	0.9
2-sided, windowed W=32	1.6±1.3	**0.0±0.0**	3.8±2.2	3.1±0.7	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	1.8	0.9
DCR (2-sided, multi-scale)	1.6±1.3	**0.0±0.0**	3.0±1.9	2.9±0.4	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	1.5	0.8
Learned head (LOFO)	11.7±7.2	11.3±9.4	35.3±13.8	**2.8±0.6**	**0.0±0.0**	1.4±2.9	**0.0±0.0**	7.8±13.7	10.7±12.9	19.4	9.0

Table 16: Median detection delay in steps at 0.1% per-step FAR (lower is better; median within a seed, mean and std across seeds). HalfCheetah-v5, 1M steps, 5 seeds. Held-out fault family in columns; PP = plausibility-preserving, SV = surprise-visible. Mean±std over seeds.

Detector	freeze PP	lag PP	drift PP	spike SV	noise SV	dropout SV	swap SV	gain mix	saturation mix	PP mean	all mean
Temporal jump	200.8±349.9	200.0±339.1	283.5±307.9	5.3±3.3	18.7±23.4	106.0±144.2	**0.0±0.0**	186.3±73.6	364.0±266.1	228.1	151.6
Reconstruction (AE)	176.6±378.1	166.4±363.4	92.9±57.0	4.9±3.5	34.5±66.9	185.5±395.3	0.3±0.4	313.5±314.9	245.7±363.7	145.3	135.6
DEXTER-style (IF)	890.6±7.5	885.2±17.0	890.8±27.7	891.4±5.2	900.5±8.4	892.0±7.2	881.0±11.9	886.7±16.4	898.9±10.6	888.9	890.8
Innovation / obs. spread	18.4±12.4	7.4±6.2	101.1±23.9	3.5±1.4	6.8±3.4	54.8±55.3	0.1±0.2	249.2±267.3	221.5±173.7	42.3	73.6
Innovation / own scale (1-sided)	**1.0±0.0**	**0.0±0.0**	**4.7±2.5**	**3.3±1.2**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.5±0.5**	1.9	1.1
PEDM-style disagreement	74.3±90.5	59.6±70.1	122.2±37.2	9.7±4.7	29.4±19.9	179.1±251.6	1.9±1.0	343.1±206.2	314.4±308.2	85.4	126.0
PEDM-style NLL	**1.0±0.0**	**0.0±0.0**	5.2±0.4	3.5±1.4	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	0.6±0.5	2.1	1.1
PEDM-style NLL, windowed	1.8±0.4	0.2±0.4	11.0±2.1	3.7±1.2	0.4±0.5	0.1±0.2	**0.0±0.0**	0.8±0.6	1.2±1.1	4.3	2.1
CUSUM on residual (k=0.5, h=5)	**1.0±0.0**	**0.0±0.0**	5.2±0.8	3.6±1.3	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	0.6±0.5	2.1	1.2
χ^2 innovation test W=32	1.6±0.5	0.2±0.4	10.5±3.9	3.7±1.2	0.4±0.5	0.1±0.2	**0.0±0.0**	0.7±0.7	1.2±1.1	4.1	2.0
2-sided, windowed W=32	1.6±0.5	0.2±0.4	10.5±3.9	3.7±1.2	0.4±0.5	0.1±0.2	**0.0±0.0**	0.7±0.7	1.2±1.1	4.1	2.0
DCR (2-sided, multi-scale)	1.6±0.5	0.2±0.4	9.1±3.4	3.8±1.2	0.2±0.4	**0.0±0.0**	**0.0±0.0**	0.5±0.5	1.0±0.7	3.6	1.8
Learned head (LOFO)	7.0±3.7	12.5±15.2	93.8±58.6	5.6±1.6	3.8±2.3	6.4±3.4	0.1±0.2	14.0±5.4	11.1±6.6	37.8	17.1

Table 17: Median detection delay in steps at 0.1% per-step FAR (lower is better; median within a seed, mean and std across seeds). Walker2d-v5, 1M steps, 5 seeds. Held-out fault family in columns; PP = plausibility-preserving, SV = surprise-visible. Mean±std over seeds.

Detector	freeze PP	lag PP	drift PP	spike SV	noise SV	dropout SV	swap SV	gain mix	saturation mix	PP mean	all mean
Temporal jump	412.2±331.3	500.9±331.8	227.2±86.8	5.4±2.1	8.0±6.9	264.0±245.8	**0.0±0.0**	472.0±349.0	536.9±396.0	380.1	269.6
Reconstruction (AE)	218.1±126.2	546.5±264.1	169.8±40.9	27.8±38.9	291.1±372.5	249.7±227.3	7.4±8.0	678.2±251.5	448.4±342.7	311.5	293.0
DEXTER-style (IF)	569.9±259.6	792.3±128.3	312.6±19.0	736.7±128.7	857.1±27.8	430.4±205.4	61.1±12.2	853.0±25.5	719.5±187.6	558.3	592.5
Innovation / obs. spread	390.7±365.0	440.2±378.8	206.6±80.8	62.6±103.9	343.2±462.0	281.5±231.9	13.2±18.3	535.9±345.9	428.2±381.3	345.8	300.2
Innovation / own scale (1-sided)	**12.0±10.0**	**5.0±9.1**	34.8±15.7	**3.7±0.8**	0.4±0.9	3.9±8.7	**0.0±0.0**	14.1±19.4	**18.2±22.1**	17.3	10.2
PEDM-style disagreement	344.8±286.4	505.2±311.8	200.2±64.8	63.4±51.7	513.7±440.9	251.1±203.3	13.9±10.7	467.5±360.1	608.5±324.4	350.1	329.8
PEDM-style NLL	12.9±12.0	8.7±12.3	37.1±25.0	3.8±0.9	0.4±0.9	5.1±11.4	**0.0±0.0**	22.1±43.3	24.4±40.1	19.6	12.7
PEDM-style NLL, windowed	65.5±114.2	76.3±139.5	77.3±83.9	16.6±27.3	32.5±69.9	74.4±164.4	0.3±0.7	188.6±355.3	169.3±333.0	73.0	77.9
CUSUM on residual (k=0.5, h=5)	12.4±10.1	6.3±8.8	**33.0±15.9**	**3.7±0.8**	**0.3±0.7**	**3.0±6.7**	**0.0±0.0**	**10.9±15.4**	18.3±22.2	17.2	9.8
χ^2 innovation test W=32	86.4±155.1	125.8±245.7	85.6±107.8	75.9±159.9	77.6±171.0	130.5±284.6	2.6±5.8	198.1±359.4	184.4±356.0	99.3	107.4
2-sided, windowed W=32	86.4±155.1	125.8±245.7	85.6±107.8	75.9±159.9	77.6±171.0	130.5±284.6	2.6±5.8	198.1±359.4	184.4±356.0	99.3	107.4
DCR (2-sided, multi-scale)	84.3±156.3	126.0±245.9	85.7±108.4	75.4±160.2	87.2±192.8	128.7±285.5	2.6±5.8	197.5±360.0	183.0±356.9	98.7	107.8
Learned head (LOFO)	26.5±6.8	122.3±217.8	99.5±71.1	4.1±1.3	2.5±2.6	17.8±12.6	**0.0±0.0**	227.9±340.1	104.3±124.6	82.8	67.2

Table 18: Median detection delay in steps at 0.1% per-step FAR (lower is better; median within a seed, mean and std across seeds). Ant-v5, 1M steps, 5 seeds. Held-out fault family in columns; PP = plausibility-preserving, SV = surprise-visible. Mean±std over seeds.

Detector	freeze PP	lag PP	drift PP	spike SV	noise SV	dropout SV	swap SV	gain mix	saturation mix	PP mean	all mean
Temporal jump	867.8±11.6	872.6±18.4	857.7±42.1	688.1±347.3	877.8±17.1	862.3±20.1	166.0±109.7	863.9±24.0	877.9±21.0	866.0	770.5
Reconstruction (AE)	799.2±124.8	726.8±208.3	336.0±289.3	305.6±417.2	465.2±380.6	842.2±33.5	39.9±36.0	780.5±108.9	844.4±56.4	620.7	571.1
DEXTER-style (IF)	876.8±10.7	880.7±23.6	870.1±31.8	868.6±19.4	883.2±15.1	874.5±18.1	847.1±20.1	877.5±23.0	885.6±17.3	875.9	873.8
Innovation / obs. spread	763.6±102.0	756.9±136.1	591.2±268.3	531.3±334.4	734.3±214.8	720.1±146.6	166.9±147.6	725.2±176.8	803.2±99.7	703.9	643.6
Innovation / own scale (1-sided)	285.5±72.8	133.3±129.2	97.4±54.4	3.7±0.8	**0.0±0.0**	253.5±72.5	**5.1±2.6**	**483.5±181.8**	276.7±85.0	172.1	171.0
PEDM-style disagreement	702.4±159.0	775.5±102.2	665.1±201.8	688.1±179.6	635.6±257.7	662.8±235.2	335.5±300.1	658.9±216.4	742.0±173.3	714.3	651.8
PEDM-style NLL	459.6±255.0	**77.1±43.8**	**63.7±44.2**	**3.5±0.9**	**0.0±0.0**	442.8±269.6	6.2±3.1	551.9±209.7	408.9±169.1	200.1	223.7
PEDM-style NLL, windowed	853.7±19.8	318.1±294.0	150.5±138.1	4.5±1.2	0.2±0.4	858.6±18.6	11.9±10.2	860.8±21.8	867.8±24.5	440.8	436.2
CUSUM on residual (k=0.5, h=5)	**280.4±73.3**	133.3±129.2	90.8±60.2	3.7±0.8	**0.0±0.0**	**251.4±119.6**	**5.1±2.6**	507.8±194.7	**263.7±87.4**	168.2	170.7
χ^2 innovation test W=32	861.2±18.6	574.0±303.1	356.8±279.2	4.4±1.4	0.8±1.8	859.3±20.2	12.6±9.9	864.7±19.9	870.6±26.5	597.3	489.4
2-sided, windowed W=32	861.2±18.6	574.0±303.1	356.8±279.2	4.4±1.4	0.8±1.8	859.3±20.2	12.6±9.9	864.7±19.9	870.6±26.5	597.3	489.4
DCR (2-sided, multi-scale)	859.7±20.8	557.9±291.7	367.2±273.4	4.4±1.4	0.8±1.8	856.2±18.9	12.4±10.0	864.7±19.9	870.6±26.5	594.9	488.2
Learned head (LOFO)	810.3±45.1	375.7±224.2	430.0±239.1	25.5±26.6	19.2±27.0	834.8±28.6	16.7±17.2	829.1±72.7	813.8±67.6	538.7	461.7

Table 19: Ant-v5 at 1M steps, 5 seeds: held-out AUROC of the DCR statistic by the group of the faulted channel (27 kinematic channels vs. 78 contact-force channels).

Channel group	freeze	lag	drift	spike	noise	dropout	swap	gain	saturation
Kinematic	0.851±0.074	0.931±0.047	0.914±0.030	0.992±0.002	0.986±0.015	0.859±0.066	0.990±0.019	0.862±0.067	0.833±0.070
Contact	0.581±0.017	0.689±0.046	0.987±0.005	0.996±0.002	1.000±0.000	0.567±0.025	0.816±0.042	0.542±0.017	0.567±0.012

Table 20: Learned head trained on all nine families (not held out) vs. leave-one-family-out vs. the fixed statistic. AUROC at 1M steps, 5 seeds: mean over the nine fault families (all) and over the three plausibility-preserving families (PP), averaged over seeds. Per-family tables, with two further weak baselines, in Appendix D.

Detector	Hopper		Walker2d		HalfCheetah		Ant	
	PP	all	PP	all	PP	all	PP	all
DCR (2-sided, multi-scale)	**0.992**	**0.994**	**0.985**	**0.986**	**0.993**	**0.995**	**0.794**	**0.792**
Learned head (LOFO)	0.956	0.975	0.944	0.954	0.969	0.979	0.729	0.743
Learned head (all families)	0.977	0.985	0.966	0.966	0.981	0.986	0.760	0.765

Table 21: Held-out AUPRC (faulty steps positive), averaged over the nine held-out fault families and five seeds, 1M steps.

Detector	Hopper	Walker2d	HalfCheetah	Ant
Temporal jump	0.343	0.427	0.481	0.423
Reconstruction (AE)	0.717	0.602	0.829	0.552
DEXTER-style (IF)	0.289	0.371	0.732	0.399
Innovation / obs. spread	0.449	0.516	0.814	0.450
Innovation / own scale (1-sided)	0.831	0.778	0.862	0.613
PEDM-style disagreement	0.414	0.471	0.722	0.460
PEDM-style NLL	0.886	0.815	0.939	0.620
PEDM-style NLL, windowed	0.985	0.957	**0.995**	0.729
CUSUM on residual (k=0.5, h=5)	0.811	0.751	0.884	0.635
χ^2 innovation test W=32	0.980	0.957	0.990	0.731
2-sided, windowed W=32	0.978	0.957	0.989	0.732
DCR (2-sided, multi-scale)	**0.990**	**0.971**	0.995	**0.749**
Learned head (LOFO)	0.929	0.924	0.977	0.688

Table 22: False-alarm rate at 95% recall (lower is better), averaged over the nine held-out fault families and five seeds, 1M steps.

Detector	Hopper	Walker2d	HalfCheetah	Ant
Temporal jump	0.841	0.872	0.985	0.947
Reconstruction (AE)	0.628	0.764	0.625	0.862
DEXTER-style (IF)	0.670	0.732	0.454	0.921
Innovation / obs. spread	0.584	0.753	0.502	0.898
Innovation / own scale (1-sided)	0.687	0.783	0.819	0.835
PEDM-style disagreement	0.746	0.830	0.742	0.890
PEDM-style NLL	0.298	0.545	0.268	0.826
PEDM-style NLL, windowed	0.022	0.178	**0.015**	0.648
CUSUM on residual (k=0.5, h=5)	0.672	0.779	0.643	0.768
χ^2 innovation test W=32	0.034	0.139	0.033	0.616
2-sided, windowed W=32	0.045	0.145	0.035	0.613
DCR (2-sided, multi-scale)	**0.004**	**0.064**	0.018	**0.566**
Learned head (LOFO)	0.114	0.240	0.108	0.692

Table 23: Held-out AUROC at 150k steps, 5 seeds: mean over the nine fault families (all) and over the three plausibility-preserving families (PP), averaged over seeds. Per-family tables, with two further weak baselines, in Appendix D.

Detector	Hopper		Walker2d		HalfCheetah		Ant	
	PP	all	PP	all	PP	all	PP	all
Temporal jump	0.532	0.609	0.567	0.633	0.361	0.489	0.505	0.526
Reconstruction (AE)	0.734	0.759	0.656	0.674	0.861	0.790	0.638	0.591
DEXTER-style (IF)	0.671	0.743	0.543	0.588	0.827	0.812	0.537	0.533
Innovation / own scale (1-sided)	0.939	0.889	0.802	0.799	0.872	0.874	0.621	0.621
PEDM-style NLL	0.966	0.923	0.807	0.804	0.964	0.925	0.604	0.607
PEDM-style NLL, windowed	0.989	0.988	0.876	0.908	**0.990**	0.990	0.648	0.684
CUSUM on residual (k=0.5, h=5)	0.928	0.872	0.768	0.766	0.925	0.909	0.642	0.636
χ^2 innovation test W=32	0.989	0.988	0.907	0.929	0.980	0.987	0.708	0.728
2-sided, windowed W=32	0.989	0.989	0.909	0.930	0.983	0.988	0.712	0.734
DCR (2-sided, multi-scale)	**0.991**	**0.994**	**0.918**	**0.948**	0.989	**0.993**	**0.730**	**0.752**
Learned head (LOFO)	0.956	0.975	0.874	0.911	0.968	0.979	0.664	0.695

Table 24: Drift sweep, AUROC of the windowed-mean statistic ($W=32$) at both checkpoints, seed 1, closed loop. For each task and drift rate ρ (in σ_d per step): the single-channel bound of Eq. (1), the faulty channel's own term, the remaining channels' term, and all channels. At 150k steps the own term sits on the bound on Hopper and below chance on Walker2d and Ant; the remaining channels carry the detectability at both checkpoints.

		150k steps				1M steps			
Task	ρ	bound	own	other	all	bound	own	other	all
Hopper	0.0005	0.606	0.661	0.834	0.927	0.608	0.893	0.920	0.982
	0.001	0.651	0.753	0.858	0.953	0.651	0.954	0.943	0.993
	0.002	0.701	0.890	0.887	0.977	0.706	0.973	0.959	0.997
	0.005	0.789	0.960	0.923	0.988	0.786	0.981	0.960	0.997
	0.01	0.843	0.973	0.942	0.991	0.849	0.979	0.970	0.997
	0.02	0.915	0.974	0.953	0.993	0.913	0.963	0.975	0.998
	0.05	0.992	0.955	0.957	0.991	0.983	0.958	0.977	0.997
Walker2d	0.0005	0.536	0.403	0.806	0.826	0.583	0.799	0.925	0.968
	0.001	0.565	0.580	0.867	0.901	0.642	0.844	0.960	0.981
	0.002	0.605	0.689	0.880	0.942	0.700	0.842	0.974	0.988
	0.005	0.674	0.814	0.905	0.976	0.744	0.845	0.974	0.989
	0.01	0.736	0.794	0.912	0.983	0.786	0.847	0.962	0.986
	0.02	0.801	0.834	0.932	0.986	0.853	0.846	0.943	0.982
	0.05	0.898	0.820	0.953	0.984	0.956	0.904	0.944	0.987
HalfCheetah	0.0005	0.520	0.929	0.963	0.973	0.521	0.907	0.959	0.966
	0.001	0.540	0.952	0.973	0.980	0.542	0.929	0.972	0.980
	0.002	0.578	0.920	0.980	0.987	0.583	0.930	0.980	0.988
	0.005	0.660	0.962	0.991	0.994	0.675	0.858	0.963	0.970
	0.01	0.735	0.898	0.992	0.996	0.761	0.633	0.895	0.914
	0.02	0.816	0.748	0.989	0.993	0.862	0.532	0.839	0.885
	0.05	0.928	0.677	0.987	0.996	0.972	0.653	0.847	0.895
Ant	0.0005	0.547	0.212	0.564	0.599	0.701	0.572	0.862	0.887
	0.001	0.556	0.404	0.634	0.705	0.708	0.813	0.925	0.954
	0.002	0.562	0.542	0.648	0.737	0.716	0.905	0.950	0.976
	0.005	0.579	0.461	0.674	0.761	0.739	0.949	0.973	0.988
	0.01	0.608	0.598	0.781	0.839	0.771	0.961	0.984	0.995
	0.02	0.659	0.524	0.852	0.895	0.817	0.924	0.988	0.997
	0.05	0.758	0.555	0.927	0.946	0.887	0.755	0.992	0.998