YOUR WORLD MODEL IS NOT BLIND, YOUR DETECTOR IS: NATIVE PREDICTION ERRORS FOR REGIME DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A world model that has learned how a system evolves should be the first to notice when that system starts to evolve differently. Yet learned representations are routinely judged blind to regime change: a predictive representation that recovers the hidden state of a dynamical system still lets standard sequential change-point detectors miss most changes. We show that this blindness belongs to the detector, not the representation. While standard detectors assume latent states follow a linear process, we instead query the world model directly-scoring each transition by its one-step predictive error normalized by covariance fitted on nominal (calm) data. This native-residual detector requires no retraining and elevates detection rates from a small minority of shifts to near the exact-likelihood ceiling. We then prove why this gap exists. The standard detector's variance channel can only gather evidence above a threshold fixed by its alternative hypothesis grid, which no grid refinement resolves. Furthermore, we prove an absorption principle: a flexible predictor trained on calm data systematically explains away the component of a regime shift aligned with its learned dynamics, while a learned variance head re-weights the residual back toward nominal behavior. This reveals why standardizing by model uncertainty is inherently deceptive. Across convolutional, transformer, and period-folding backbones, native-residual detection alarms roughly three times faster than classical baselines, fails predictably only when an encoder discards one-step temporal order, and outperforms recent detector families, calm-fitted forecasters, and a zero-shot foundation model.

1 INTRODUCTION

A world model that has learned how a system evolves should be the first to notice when the system starts to evolve differently. This premise underlies the use of learned representations for monitoring. Joint-embedding predictive architectures (JEPAs) predict the embedding of the future from the embedding of the present (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024), and their latent state is increasingly asked to support downstream decisions (Ha & Schmidhuber, 2018; Hafner et al., 2025; Assran et al., 2025). The canonical downstream test is *regime detection*. A stream runs in a calm regime and switches at an unknown time to a stressed one; a detector must raise an alarm soon after the switch while keeping false alarms under control. To compare representations on equal terms, the standard protocol freezes each representation and hands it to one fixed sequential detector: calm whitening, a linear VAR(1) null, and a mixture e-process calibrated on calm data (Shin et al., 2023; Ramdas et al., 2023).

By this test, learned world models look blind. We use a benchmark built so that only nonlinear structure carries the change. On it, a JEPA recovers the hidden state better than any other learned representation, yet the frozen detector catches only 0.231 of the changes at a 5% false-alarm rate, against 0.965 for the exact-likelihood oracle (Table 1). Var-JEPA (Gögl & Yau, 2026), which models predictive uncertainty explicitly, does worse still (0.090). The tempting conclusion is that these representations lack the information. That conclusion cannot hold in general. By the data-processing inequality, no representation creates evidence (Cover & Thomas, 2006); a detector can only fail to read what a representation carries. The real question is *which reading* a learned representation calls for.

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.

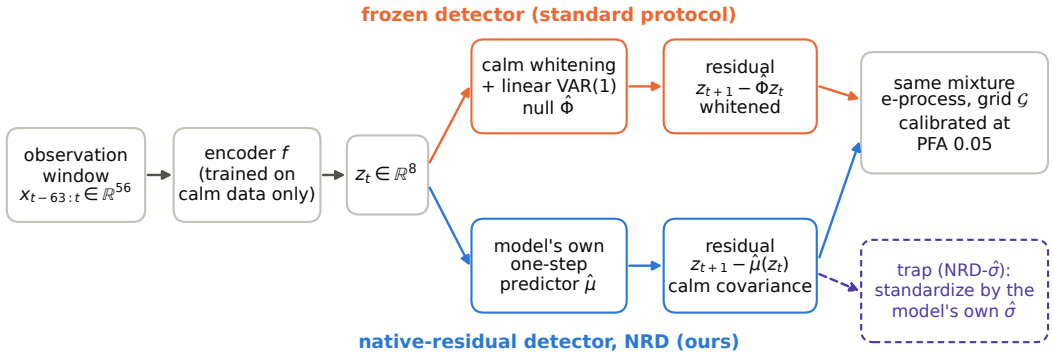

Figure 1: **Two readings of one frozen representation.** Every encoder maps a 64-step window to an 8-dimensional latent and is trained on calm data only. The standard *frozen* detector (orange) scores residuals of a linear VAR(1) null. *Native-residual detection* (blue) scores residuals of the model's own one-step predictor in a calm-fitted metric. Both feed the same mixture e-process, grid and calibration. Standardizing by the model's own predictive scale (violet) is the trap explained by Theorem 3.

Ask the model, not a linear surrogate. A world model comes with its own answer to "what should happen next": its one-step predictor. We score each step by the residual of that predictor, measured in a covariance fitted on calm residuals, and feed it to the *same* e-process with the *same* calibration (Figure 1). We call this *native-residual detection* (NRD). It requires no retraining and adds no parameters. On the same frozen JEPA latent, NRD raises detection from 0.231 to 0.950, within 0.015 of the oracle. The bottleneck was the detector.

Flexibility cuts both ways. Var-JEPA's variance head is the first learned component that opens the frozen detector's *variance channel*, the channel that accumulates evidence over time. Yet NRD gains far less on Var-JEPA ($0.090 \to 0.366$), and standardizing the residual by the model's own predictive scale detects nothing. We explain both facts with two results. First, the frozen detector's variance channel earns evidence only above a threshold fixed by its alternative grid. Second, a flexible calm-trained predictor *absorbs* exactly the part of a regime shift aligned with its own correction, while a variance head re-weights whatever survives back toward calm. A model opens a channel against a linear null and closes it against its own view.

Contributions.

- **A drop-in detector.** NRD reaches near-oracle detection from a frozen JEPA latent, and alarms $3.0\times$ sooner than the frozen detector. It holds across causal convolutional, transformer and period-folding encoders. On the same JEPA latent it is 3-240$\times$ more sensitive than five recent detector families, beats calm-fitted LSTM forecasters on the observations and on the latent, and a 200M-parameter zero-shot foundation model detects nothing (Section 5).
- **Why frozen detectors miss change.** The frozen detector's evidence splits exactly into a variance channel that accumulates and a drift channel that saturates. Over its grid, the variance channel gains evidence only above $\tau_{\mathcal{G}} = 3\log\frac{3}{2} \approx 1.2164$ (Theorem 1), and a denser grid does not rescue it: the restart mixture prices a run faster than a small inflation can pay (Proposition 1).
- **The absorption principle.** The inflation a calm-trained predictor removes relative to a linear null equals twice the alignment between the regime shift and the predictor's correction (Theorem 2). A variance head anchors the null at one and re-weights the remainder (Theorem 3). With the grid threshold, this explains every measured energy ratio without free parameters and yields two design rules: read the raw residual of the *co-adapted* predictor in a calm metric, and never score a residual standardized by the model's own uncertainty against a metric that was not fitted on it.

- **Verified theory and a sharp scope.** All 65 theorems are machine-checked in Lean 4. The method's precondition, that the encoder preserves one-step temporal order, is stated and tested, and so is the regime in which the frozen detector needs no help.

2 Related Work

Predictive architectures and world models. JEPAs predict in embedding space rather than observation space (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025). They rely on anti-collapse heuristics such as EMA targets and variance-covariance penalties (Grill et al., 2020; Bardes et al., 2022), which recent work replaces with principled objectives (Balestriero & LeCun, 2025). Var-JEPA derives the architecture from a single ELBO with an explicit latent generative model and a predictive variance (Gögl & Yau, 2026; Kingma & Welling, 2014). Time-series JEPAs target classification, forecasting and anomaly prediction (Ennadir et al., 2025; Verdenius et al., 2024; He et al., 2026). We add the regime-detection axis and ask which detector channel a representation opens. Hong (2026) report that world-model monitors absorb slow drift below an empirical threshold. Our absorption theorem gives the algebra of such absorption for a calm-trained predictor, and NRD shows how to read what survives.

Sequential change detection. CUSUM and Shiryaev-Roberts procedures are optimal for known pre- and post-change laws (Page, 1954; Lorden, 1971; Moustakides, 1986; Lai, 1998; Tartakovsky et al., 2014). E-values and e-processes give anytime-valid tests (Vovk & Wang, 2021; Ramdas et al., 2023; Howard et al., 2021), and e-detectors aggregate e-processes started at every candidate change time (Shin et al., 2023; Shekhar & Ramdas, 2023; Saha & Ramdas, 2026; Ramdas, 2026). The frozen detector is such a restart mixture over a scaled-Gaussian alternative grid. Prior work found that its restart mixture buys validity rather than power; we show that the *alternative class* decides power.

Deep anomaly and change detection. Forecasting-residual detectors date back at least to LSTM predictors (Malhotra et al., 2015; Hundman et al., 2018). Deep detectors score one-class distances (Ruff et al., 2018), energies (Liu et al., 2020), Mahalanobis distances (Lee et al., 2018), association discrepancies (Xu et al., 2022), or kernel two-sample statistics (Gretton et al., 2012; Li et al., 2019; Chang et al., 2019; Arlot et al., 2019); see Pang et al. (2021); Schmidl et al. (2022); Aminikhanghahi & Cook (2017); Truong et al. (2020) for surveys. Most of these are *pipelines* that fuse a detector head into a model, so we compare with them at the detector level, on the same latents. NRD differs from forecasting-residual detectors in two ways: it reuses a *self-supervised* world model's own predictor without retraining, and it comes with a theory of when that predictor hides the change. To place it against that mainstream alternative on our own protocol, we also fit calm-only LSTM one-step forecasters, on the raw observations and post hoc on the frozen latent, and score their residuals with the same e-process, the same calm-fitted metric and the same calibration (Section 5.3). Our backbones follow TCN, Transformer, PatchTST and TimesNet (Bai et al., 2018; Vaswani et al., 2017; Nie et al., 2023; Wu et al., 2023). Our foundation-model baseline is TimesFM (Das et al., 2024), of the same family as Ansari et al. (2024).

Variance heads. Heteroscedastic heads trained by likelihood are known to trade mean fit for variance fit (Nix & Weigend, 1994; Lakshminarayanan et al., 2017; Seitzer et al., 2022; Stirn et al., 2023). A head of this kind invites likelihood scoring, as probabilistic forecasting pipelines do. Theorem 3 identifies what that costs on the detection side: standardizing a residual by the head's own scale anchors the null at one and discounts exactly the shift that survives the mean. The anchor, not the head, is the problem, and Section 5.3 separates the two by re-fitting the calm metric after standardization.

3 Why Frozen Detectors Miss Change

3.1 Setting

Episodes of length 300 run under a calm law P and, for test episodes, switch at $\nu = 150$ to a stress law Q. An encoder maps a 64-step observation window to a latent $z_t \in \mathbb{R}^8$ and is trained on calm data only. Every representation is scored by the same frozen detector; NRD runs alongside it, never instead. Four synthetic regimes share one generator (Section B):

- SWIRL (the hard case): the rotation angle and innovation scale of a hidden 6-dimensional state depend on its radius, with the marginal law matched exactly.
- ROTOR (the expressible case): a latent rotation changes, and the marginals are provably invariant.
- SECTORS (a diluted case): one of eight sectors rotates.
- SWIRL-LIN (a selectivity control): the rotation changes linearly.

The representations are the privileged true state (TRUE-STATE), PSR, PCA, raw observations, an autoencoder, a contrastive encoder, JEPA and Var-JEPA on a causal TCN, and three further backbones (Section 5.3).

3.2 THE FROZEN DETECTOR HAS TWO CHANNELS

The frozen detector whitens latents on calm training data and fits a VAR(1) null $\hat{\Phi}$. It forms residuals $r_t = z_{t+1} - \hat{\Phi} z_t$ with calm covariance Σ_0 and $A = \Sigma_0^{-1}$, and aggregates a restart mixture over change times k and the variance grid $\mathcal{G} = \{1, 1.5, 2.5, 4\}$:

$$\log E_t = \log \sum_{k \le t} \sum_{g \in \mathcal{G}} \frac{w_k}{|\mathcal{G}|} e^{V_g(n) + M_g(n)}, \qquad \begin{aligned} V_g(n) &= \tfrac{SS_n}{2}\big(1 - \tfrac{1}{g}\big) - \tfrac{nd}{2} \log g, \\ M_g(n) &= \tfrac{\eta^2 D_n}{2g(1+n\eta^2)} - \tfrac{d}{2} \log(1 + n\eta^2). \end{aligned} \tag{1}$$

Here $n = t - k + 1$, $SS_n = \sum_s r_s^\top A r_s$ is the run's energy, $D_n = S_n^\top A S_n$ with $S_n = \sum_s r_s$ is its squared drift, $w_k \propto 0.95^{t-k}$ and $\eta^2 = 1$. The split into a *variance channel* V_g and a *drift channel* M_g is exact, and V_g is precisely the log-likelihood ratio of $\mathcal{N}(0, g\Sigma_0)$ against $\mathcal{N}(0, \Sigma_0)$. Let $\bar{g}$ denote the post-change energy inflation, $\mathbb{E}[SS_n] = nd\bar{g}$. Its one-step estimate γ_{eff} equals the ratio R_{lin} defined in Section 4.1.

Lemma 1 (Two channels). *(a)* $\mathbb{E}[V_g(n)] = \frac{nd}{2}\varphi(\bar{g}, g)$ *with* $\varphi(\bar{g}, g) = \bar{g}(1 - \frac{1}{g}) - \log g$, *and* $\max_{g>0} \varphi(\bar{g}, g) = d_{\text{IS}}(\bar{g}) = \bar{g} - 1 - \log \bar{g} \ge 0$, *attained at* $g = \bar{g}$ *and zero iff* $\bar{g} = 1$. *The variance channel therefore earns an Itakura-Saito divergence (Févotte et al., 2009) that grows linearly in* n. *(b) If* $\mathbb{E}[D_n] \le n\Omega$, *the drift channel's positive part is at most* $\Omega/(2g)$ *for every* n, *while its price* $\frac{d}{2}\log(1 + n\eta^2)$ *grows without bound.*

Theorem 1 (Grid threshold). *For* $n, d > 0$, $\max_{g \in \mathcal{G}} \mathbb{E}[V_g(n)] > 0$ *if and only if* $\bar{g} > \tau_{\mathcal{G}} := \min_{g \in \mathcal{G},\, g>1} \frac{\log g}{1 - 1/g} = 3 \log \frac{3}{2}$. *Numerically* $\tau_{\mathcal{G}} \in (1.21639,\ 1.21640)$. *The per-point thresholds are* 1.2164 *for* $g = 1.5$, 1.5272 *for* $g = 2.5$, *and* 1.8484 *for* $g = 4$.

Below $\tau_{\mathcal{G}}$, every variance alternative in the grid *loses* evidence in expectation after the change. The channel's defining advantage, linear accumulation, is then unavailable, and the detector must rely on the saturating drift channel (Figure 2a). Learned representations sit exactly there. Figure 2b validates the threshold on the true latent: as a change is diluted over more sectors, the best grid point switches from 1.5 to 1 between inflations 1.248 and 1.213, which bracket $\tau_{\mathcal{G}}$, and detection collapses at that point. Proofs are in Section C. All statements, including the certified digits of $\tau_{\mathcal{G}}$, are machine-checked in Lean 4 with Mathlib (de Moura & Ullrich, 2021; The mathlib Community, 2020) using only the three standard axioms (Section E).

A denser grid does not rescue the channel. The threshold belongs to the grid: $g \mapsto \log g/(1-1/g)$ increases to 1 as $g \downarrow 1$, so alternatives near one push $\tau_{\mathcal{G}}$ toward one. What no grid changes is the rate at which evidence arrives. The best alternative earns at most $\frac{d}{2} d_{\text{IS}}(\bar{g})$ per step, while the restart mixture charges $\log(1/\rho)$ per step to keep a run alive, so the channel accumulates only above $\bar{g} = 1.169$ here (Proposition 1). On SWIRL the measured inflation is $R_{\text{lin}} = 1.089$, far below that, and densifying the grid to 8 and 16 points confirms it: frozen detection moves from 0.231 to 0.224 and 0.221, while NRD is unchanged (Section C.2).

4 NATIVE-RESIDUAL DETECTION AND THE ABSORPTION PRINCIPLE

Definition 1 (Native-residual detection, NRD). Let $\hat{\mu}$ be a model's own one-step predictor: the JEPA head for the one-step target, or Var-JEPA's conditional-prior mean. NRD forms $\hat{r}_t = z_{t+1} - \hat{\mu}(z_t)$, fits the metric $\Sigma_c = \text{Cov}(\hat{r})$ on *calm* calibration residuals, and runs the e-process equation 1 with $A = \Sigma_c^{-1}$ and the same grid and achieved-PFA calibration as the frozen detector.

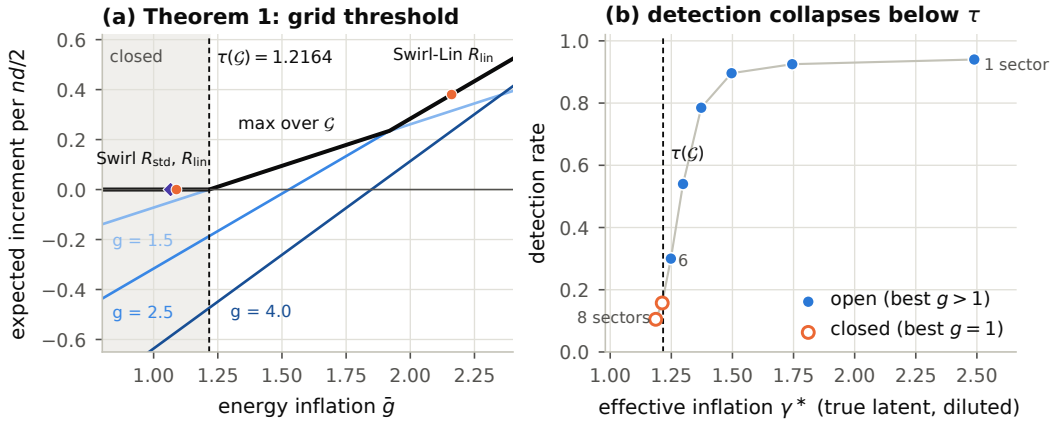

Figure 2: **The grid threshold.** (a) Expected variance-channel increment per unit $nd/2$ for each grid point. The upper envelope is positive only for $\bar{g} > \tau_{\mathcal{G}}$. The measured SWIRL ratios (7-seed means) fall in the closed region; the SWIRL-LIN ratio does not. (b) Dilution curve on the true latent: detection collapses where the best grid point drops to $g = 1$, between the two inflations that bracket $\tau_{\mathcal{G}}$.

Two variants isolate the ingredients. NRD-$\hat{\sigma}$ standardizes $\hat{r}_t$ by the model's own predictive scale $\hat{\sigma}(z_t)$ instead of Σ_c. The *calm-fitted null* replaces $\hat{\mu}$ by an MLP fitted post hoc on calm latents, which is available for every representation, including those without a predictor.

4.1 THE ABSORPTION PRINCIPLE

Let P, Q be the calm and stress laws of consecutive latents (z, z'), let $\mu_c(z) = \mathbb{E}_P[z' \mid z]$, and let $\mu_Q = \mu_c + \delta$, so that δ is the regime mean shift. Write $a = \mu_c - \hat{\mu}$ for the model's calm error and $c = \hat{\mu} - \mu_{\text{lin}}$ for the model's *correction* to the frozen linear null μ_{lin}. The calm-optimal diagonal variance head is $\hat{\sigma}_i^2(z) = \mathbb{E}_P[(z'_i - \hat{\mu}_i(z))^2 \mid z]$. For a predictor f, let $I_f = \mathbb{E}_Q\|z' - f(z)\|^2 - \mathbb{E}_P\|z' - f(z)\|^2$ be its residual inflation, and define

$$R_{\text{lin}} = \tfrac{1}{d}\mathbb{E}_Q\|z' - \mu_{\text{lin}}(z)\|^2_{\Sigma_{\text{lin}}^{-1}}, \qquad R_{\text{raw}} = \frac{\mathbb{E}_Q\|z' - \hat{\mu}(z)\|^2}{\mathbb{E}_P\|z' - \hat{\mu}(z)\|^2}, \qquad R_{\text{std}} = \tfrac{1}{d}\mathbb{E}_Q\left\|\frac{z' - \hat{\mu}(z)}{\hat{\sigma}(z)}\right\|^2, \tag{2}$$

with Σ_{lin} the calm second moment of the linear residual. These are the frozen detector's view, the model's raw view and the model's standardized view. Two assumptions are used, and each is invoked only where needed.

- (M) *Marginal matching*: $Q_z = P_z$. This holds by design for SWIRL and ROTOR, and was verified numerically to within 1%.
- (O) *Calm optimality along c*: $t \mapsto \mathbb{E}_P\|z' - \hat{\mu}(z) - t\,c(z)\|^2$ is minimized at $t = 0$. This holds, for instance, for predictors with a linear output layer and a linear skip path.

Lemma 2 (Calm anchoring). *$R_{\text{raw}}(P) = R_{\text{std}}(P) = 1$, and $\hat{\sigma}_i^2 = a_i^2 + \text{Var}_P(z'_i \mid z)$.*

Theorem 2 (Mean absorption). *Under (M), $I_{\text{lin}} - I_{\text{model}} = 2\,\mathbb{E}_Q\langle c, \delta\rangle$. Hence the model's residual inflates less than the linear null's iff $\mathbb{E}_Q\langle c, \delta\rangle > 0$, and does not inflate at all ($R_{\text{raw}} \le 1$) iff $2\,\mathbb{E}_Q\langle c, \delta\rangle \ge I_{\text{lin}}$. Under (M) and (O), the stress energies satisfy exactly $\mathbb{E}_Q\|z' - \mu_{\text{lin}}\|^2 - \mathbb{E}_Q\|z' - \hat{\mu}\|^2 = \mathbb{E}_Q\|c + \delta\|^2 - \mathbb{E}_Q\|\delta\|^2$.*

Theorem 3 (Variance absorption). *For any state-dependent diagonal head, $R_{\text{std}} = \frac{1}{d}\mathbb{E}_Q\sum_i\big[(a_i + \delta_i)^2 + \text{Var}_Q(z'_i \mid z)\big]/\hat{\sigma}_i^2(z)$. With the calm-optimal head, $R_{\text{std}} = 1 + \frac{1}{d}\mathbb{E}_Q\sum_i\big[\delta_i^2 + 2a_i\delta_i + \text{Var}_Q - \text{Var}_P\big]/\hat{\sigma}_i^2$. Consequently, if $\text{Var}_Q \le \hat{\sigma}^2$, then $R_{\text{std}} \le 1 + \frac{1}{d}\mathbb{E}_Q\|a + \delta\|^2_{\hat{\sigma}^{-2}}$, which reduces to $1 + \frac{1}{d}\mathbb{E}_Q\|\delta\|^2_{\hat{\sigma}^{-2}}$ when the model is calm-exact ($a = 0$).*

Theorem 2 says that a calm-trained predictor absorbs *exactly the part of the shift aligned with its own correction* (Figure 3). Theorem 3 says that a variance head anchors the null at one and divides

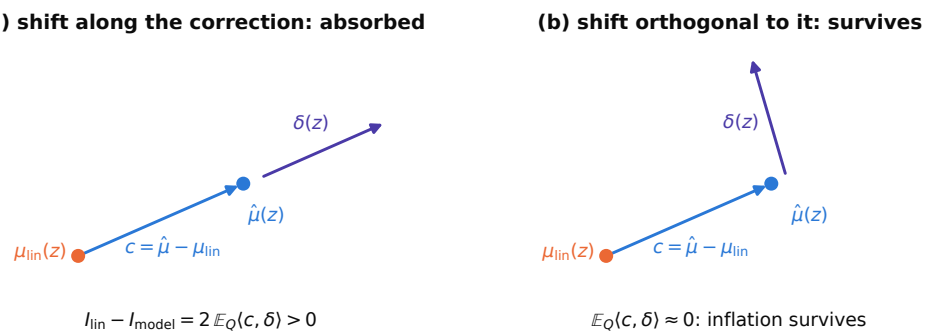

Figure 3: **Mean absorption, geometrically (Theorem 2).** At a state z, the model's correction $c = \hat{\mu} - \mu_{\text{lin}}$ is fixed by calm training. A regime shift δ along c is absorbed, because it lowers the model's inflation by $2\langle c, \delta\rangle$. A shift orthogonal to c survives in the model's residual, where NRD can read it.

whatever shift survives the mean by the calm uncertainty at the same state. The more expressive the model, the larger its correction c and its head's reach, and the more of the change it closes against its own view. It opens a channel against a linear null and closes it against itself: this is the *absorption principle*. Proofs are in Section D. All statements are checked in Lean for finite-support laws, which covers every empirical average we report.

Corollary 1 (The measured absorption triple)**.** *On* SWIRL *(Var-JEPA, 7 seeds),* $R_{\text{lin}} = 1.089$*,* $R_{\text{raw}} = 0.959$ *and* $R_{\text{std}} = 1.066$*. Both* R_{lin} *and* R_{std} *lie below* $\tau_{\mathcal{G}}$ *on* every *seed (maxima* 1.125 *and* 1.090*), so no grid alternative can grow evidence from either view (Theorem 1).* $R_{\text{raw}} < 1$ *on every seed, so by Theorem 2 the model's correction is aligned with the shift (*$2\mathbb{E}_Q\langle c, \delta\rangle > I_{\text{lin}}$*) and its raw residual does not inflate at all. None of the three views can use the variance channel. Any gain of NRD must therefore come through the drift channel of the model's raw residual, whose calm metric is sharp because the predictor is co-adapted. Section 5 tests this empirically; it is not a theorem.*

5 EXPERIMENTS

Protocol. Every detector is calibrated the same way, which is what makes this a comparison of detectors. Four disjoint episode sets come from one generator: the encoder trains on calm *fit* episodes, calm *calibration* episodes supply thresholds, a second calm *holdout* set measures the false-alarm rate those thresholds achieve, and *test* episodes carry the change at ν. We sweep 40 thresholds at quantiles of the calibration running maximum and interpolate the (achieved PFA, detection) curve at 0.05; no threshold ever sees a test episode or a label, and the number is undefined when 0.05 lies outside the span the calibration sample reaches (Section F). Detection rate is the fraction of test episodes alarmed after ν, and mean delay the average first alarm time after ν over detected episodes, at the same achieved level. Every detection claim uses 7 seeds with paired t-tests and exact Wilcoxon signed-rank tests (Wilcoxon, 1945). The exact-likelihood oracle scores the true latent path with the true transition densities; nothing may systematically exceed it.

5.1 NRD REACHES THE ORACLE

Table 1 and Figure 4a answer the opening question. Read through its own co-adapted predictor, the JEPA latent that the frozen detector scores at 0.231 detects 0.950 of changes, against an oracle of 0.965 on the same episodes. It never exceeds the per-seed oracle by more than 0.003, which is interpolation noise (Figure 7a). The finding that "JEPA recovers the state but not usably" was therefore a property of the frozen linear detector. Var-JEPA also gains significantly ($0.090 \to 0.366$), and in this run far less than JEPA (NRD on Var-JEPA vs. on JEPA: $t = -7.45$, $p = 0.0003$). We read that gap with care: Var-JEPA's objective samples latent variables during training, so a re-run is a different draw and its detection rate moves while its energy ratios do not (Section H.7). What Theorems 2 and 3 speak about is those ratios, and Section 5.2 measures them.

Table 1: **SWIRL, 7 seeds**: detection at achieved PFA 0.05, with the seed range in brackets. NRD reads the model's own one-step residual in a calm-fitted metric. Tests are paired over seeds.

detector	Var-JEPA latent	JEPA latent
frozen (linear VAR null)	0.090 [0.035, 0.133]	0.231 [0.127, 0.307]
NRD (ours)	**0.366** [0.100, 0.711]	**0.950** [0.920, 0.983]
NRD-$\hat{\sigma}$ (standardized by own $\hat{\sigma}$)	0.000 (3 seeds)	n/a (no variance head)
NRD vs. frozen: paired t	$t = 3.76, p = 0.0095$	$t = 28.5, p = 1.2 \times 10^{-7}$
NRD vs. frozen: Wilcoxon	$p = 0.0156$	$p = 0.0156$
references: exact-likelihood oracle 0.965 [0.953, 0.983] on the same episodes; TRUE-STATE 0.419 (3 seeds)		

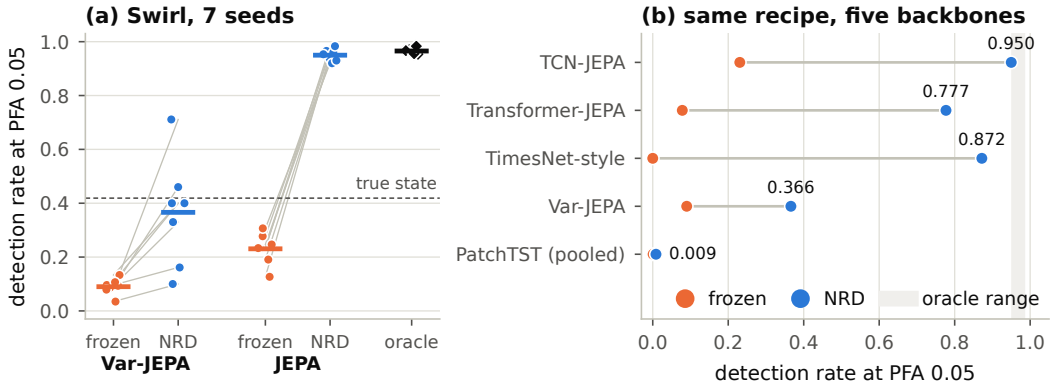

Figure 4: **NRD closes the gap to the oracle.** (a) Per-seed detection on SWIRL; grey lines pair the frozen detector and NRD on the same seed. NRD on JEPA sits at the oracle band; TRUE-STATE is the privileged linear reference. (b) The same recipe across five backbones on SWIRL (7 seeds each). It fails only for PatchTST, whose mean-pooled patch code discards one-step temporal order.

5.2 THE ABSORPTION PRINCIPLE, MEASURED

Table 2 and Figure 5 test the theory in four regimes. On SWIRL and SECTORS the frozen view sits below $\tau_{\mathcal{G}}$ and NRD wins significantly: $0.090 \to 0.366$ and $0.020 \to 0.317$ ($t = 3.65$, $p = 0.0107$, Wilcoxon $p = 0.0156$). The SECTORS result exceeds the privileged TRUE-STATE reference (0.133) while remaining far from the oracle (0.941). On SWIRL-LIN the shift is linear and fully expressible by the calm predictor. The frozen view inflates to 2.161, far above $\tau_{\mathcal{G}}$ on every seed, and detects 0.963. The model's residual instead *deflates* ($R_{\text{raw}} = 0.571$), and NRD collapses to 0.002 ($t = -122.7$). On ROTOR the rotation opens the frozen channel wide ($\gamma_{\text{eff}} \gg \tau_{\mathcal{G}}$), and Var-JEPA's flexible predictor absorbs its own native signal (0.168, seed sd 0.137). The two controls rule out a detector exploit: NRD helps exactly where the change is *not* expressible by the calm predictor, and stops helping where it is.

Opening the channel is not enough. Var-JEPA is the first learned representation whose frozen residual energy inflates on every seed ($\gamma_{\text{eff}} = 1.085$ on average; Section G), robustly across eight architectural variants, where every deterministic and projected representation averages below one and even the true state exceeds one on only two of three seeds. Yet the grid argmax never leaves $g = 1$ and frozen detection falls *below* JEPA's, consistent with Theorem 1: the inflation stays under $\tau_{\mathcal{G}}$ in every variant, so the opened channel cannot accumulate evidence. Standardizing by the model's own scale is worse still. NRD-$\hat{\sigma}$ detects nothing, consistent with Theorem 3, which anchors the standardized null at one ($R_{\text{std}} = 1.066$). The anchor, not the variance head, is what costs the evidence: re-fitting the calm metric of the standardized residual recovers detection (Section 5.3).

5.3 GENERALITY: BACKBONES, NULLS, DETECTOR FAMILIES, FOUNDATION MODELS

Backbones (Table 3, Figure 4b). NRD transfers to a causal transformer ($0.078 \to 0.777$; budget-limited, Section H.1) and to a TimesNet-style encoder. The TimesNet latent is so nonlinear that the

Table 2: **Four regimes, Var-JEPA, 7 seeds.** Frozen vs. NRD detection, with the absorption triple of equation 2. †For ROTOR only the frozen view $\gamma_{\text{eff}} \equiv R_{\text{lin}}$ was measured (mean over seeds).

regime	change	frozen	NRD	R_{lin}	R_{raw}	R_{std}
SWIRL	state-dependent, matched marginals	0.090	**0.366**	1.089	0.959	1.066
SECTORS	rotation of 1 of 8 sectors (diluted)	0.020	**0.317**	1.044	0.977	1.006
ROTOR	latent rotation, invariant marginals	**0.933**	0.168	2.65†	n/a	n/a
SWIRL-LIN	linear rotation (control)	**0.963**	0.002	2.161	0.571	0.648

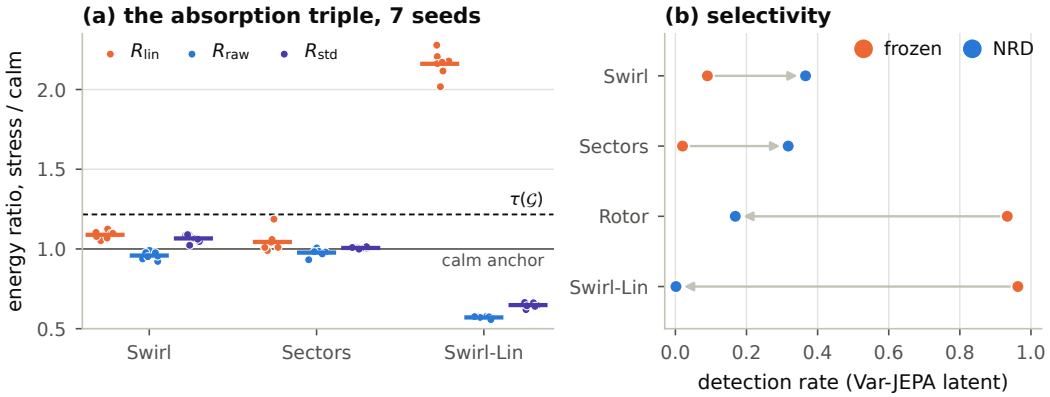

Figure 5: **Absorption, measured.** (a) The triple R_{lin} (frozen view), R_{raw} (model's raw view) and R_{std} (model's $\hat{\sigma}$-standardized view), per seed. Every SWIRL and SECTORS ratio lies below $\tau_{\mathcal{G}}$. On SWIRL-LIN the frozen view is far above it, while the model's views deflate below calm. (b) Frozen→NRD detection for Var-JEPA: the gain is selective.

frozen detector is blind on *both* benchmarks (0.000), yet its own predictor reaches 0.872 and 0.950. This is the cleanest demonstration that the detector, not the representation, was blind. NRD fails for PatchTST on SWIRL (0.009), because mean-pooling patch codes destroys the one-step temporal order that carries the change. On ROTOR, where the change survives pooling, PatchTST reaches the ceiling like the other backbones. This is the method's precondition: *the encoder must preserve one-step temporal structure*.

Which null? Co-adaptation. A post-hoc null improves detection only modestly (Figure 10, Section H): on 3 seeds the frozen linear null and a calm-fitted MLP null give 0.126 and 0.194 on JEPA latents, and 0.116 and 0.213 on PSR, while a 200M-parameter zero-shot foundation model (TimesFM-2.5; Das et al., 2024) scored with the same calm metric detects 0.000. The co-adapted 1,096-parameter predictor reaches 0.950: the latent becomes sharply predictable by *that* predictor in calm, and the state dependence of the change is left in its residual. Pretrained generality does not substitute for this.

Recent detector families were run on the same latents at the same achieved PFA (Section H.3): energy scores (Liu et al., 2020), Deep SVDD (Ruff et al., 2018), e-CUSUM (Shin et al., 2023), a kernel-MMD change-point statistic (Gretton et al., 2012; Chang et al., 2019) and a marginal z-score. Marginal families are blind by construction (≤ 0.22 everywhere), since SWIRL matches the marginal law, and on JEPA latents every family is 3-240× below NRD. On Var-JEPA latents the MMD statistic (0.579) beats NRD (0.366), which absorption predicts: what the predictor absorbs, a kernel two-sample test on the latent can still see.

Forecast-then-detect, and how the residual is scored (Section H.6, Section H.4). Fitting a calm-only LSTM one-step forecaster and monitoring its residuals is the mainstream alternative. Under our protocol it reaches 0.047 on the observations and 0.157 post hoc on the frozen latent, against 0.950 for the encoder's own head: the predictor trained *with* the encoder is the one whose residual carries the change. The calm metric barely matters (0.947 to 0.950 for diagonal, shrunk and identity). The anchor of Theorem 3 does: scoring the $\hat{\sigma}$-standardized residual against a fixed metric detects 0.109, while re-fitting the calm metric of that standardized residual recovers 0.878. The rule is therefore

Table 3: **Backbones** trained with the same JEPA objective, scored by the frozen detector and by NRD. Seven seeds, except †three. The transformer uses a larger budget of 400 episodes and 4k steps (Section H.1).

		SWIRL		ROTOR	
backbone	encoder	frozen	NRD	frozen	NRD
TCN-JEPA	causal dilated convolution	0.231	**0.950**	0.957	0.941
Transformer-JEPA	causal self-attention	0.078	**0.777** (sd 0.169)	0.945	0.950
TimesNet-style	FFT periods, 2D convolution	0.000	**0.872** (sd 0.094)	0.000†	**0.950**†
Var-JEPA	TCN + ELBO, variance head	0.090	0.366	0.933	0.168
PatchTST-JEPA	patches, mean pooling	0.001	0.009	0.932	0.929

not "do not standardize" but: never score an anchored residual against a metric that was not fitted on it.

Alarm delay (Figure 12, Section H.4). At the same achieved false-alarm rate, NRD alarms after 21.4 steps against 64.1 for the frozen detector, a 3.0× speed-up: the gain is also in how long the stream runs under the new regime. Figure 12b is the empirical side of Proposition 1: lowering the grid threshold to 1.0050 does not lift frozen detection.

Controls and ablations (Section I). Standardizing by the model's own scale against a fixed metric gives 0.000. Post-hoc nulls are not enough. The gain is selective (SWIRL-LIN). PatchTST is the stated counterexample. The transformer confirms the finding but needs a larger budget. Section G reports the Var-JEPA variants, the pre-registered predictions and the channel readout of every representation.

6 LIMITATIONS

No real-data claim. The near-oracle result (0.950) is on a synthetic benchmark, and we make no real-data detection claim. TCPD (van den Burg & Williams, 2020) has series too short for the episode protocol. On NAB AWS CloudWatch (Lavin & Ahmad, 2015; Ahmad et al., 2017) the frozen detector's energy inflation is far *above* the grid threshold (median $R_{\text{lin}} = 1.91$ against $\tau_{\mathcal{G}} = 1.2164$) and the model's own residual inflates as well, so by Theorems 1 and 2 the frozen detector already has what it needs there and NRD has no room to help (Section F). Finding real data whose change is invisible to the frozen alternative class, as SWIRL is by construction, is the main open task. **Stationary calm windows.** The calm-fitted metric assumes stationary pre-change residuals. Under a mild cycle or gain drift that the calibration episodes also see, the advantage survives; with thresholds calibrated on stationary data and a perturbed stream, the achieved false-alarm rate moves to at most 0.060 against a nominal 0.05 (Section H.5). **Variance and reproducibility.** NRD on Var-JEPA is seed-variable on ROTOR (sd 0.137), and the transformer needs a larger budget than the TCN (sd 0.169). Var-JEPA samples latent variables during training, so a re-run is a different draw: read its detection rates as a distribution over runs, while its energy ratios are stable (Section H.7). The remaining limitations, on implementation fidelity, the reach of the theorems and the provenance of three quantities, are stated in Section J.

7 CONCLUSION

A learned world model can look blind to regime change only because the detector asks it the wrong question. The frozen linear detector asks whether the latent is a linear process, and its variance channel earns evidence only above a grid threshold that no denser grid repairs. The model's own co-adapted predictor asks whether the latent is doing what calm training taught it to expect, and the answer is near-oracle, and three times sooner, whenever the encoder preserves one-step temporal order. The same flexibility that lets a variance head open a channel lets the model absorb what flows through it. We prove this, measure it without free parameters, and turn it into a rule: detect with the model, not against it.

AI USE STATEMENT

In this work, generative AI tools were used to (i) create or modify visual figures based on our experimental results, (ii) edit English grammar, phrasing, and vocabulary, and (iii) assist in producing Lean formalizations for checking selected mathematical equations. The tools were not used to generate data, formulate the research questions or theoretical claims, design or run experiments, analyze or interpret experimental results, or make scientific decisions. They were also not used to replace the authors' judgment about the scientific content.

We reviewed all AI-assisted work. The experimental data and results were checked against the original runs and stored result files; AI-assisted figures were checked against those results; and the Lean formalizations were compiled and checked by the authors to ensure that they corresponded to the equations stated in the paper. AI-assisted language edits were manually reviewed for meaning and accuracy. The authors take responsibility for the final content of this work, including all text, claims, figures, and formalizations produced with AI assistance.

REFERENCES

Subutai Ahmad, Alexander Lavin, Scott Purdy, and Zuha Agha. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 262:134–147, 2017. doi: 10.1016/j.neucom.2017.04.070.

Samaneh Aminikhanghahi and Diane J. Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367, 2017. doi: 10.1007/s10115-016-0987-z.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. arXiv:2403.07815.

Sylvain Arlot, Alain Celisse, and Zaid Harchaoui. A kernel multiple change-point algorithm via model selection. *Journal of Machine Learning Research*, 20(162):1–56, 2019.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. arXiv:2301.08243.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. arXiv:2404.08471.

Alberto D. Cencillo, Leonardo Concepción, Isaac Triguero, and Julián Luengo. VACE: Learning geometrically structured representations for time series anomaly detection. *arXiv preprint arXiv:2605.23504*, 2026.

Wei-Cheng Chang, Chun-Liang Li, Yiming Yang, and Barnabás Póczos. Kernel change-point detection with auxiliary deep generative models. In *International Conference on Learning Representations (ICLR)*, 2019.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006. doi: 10.1002/047174882X.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, 2024.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, volume 12699 of *Lecture Notes in Computer Science*, pp. 625–635, 2021. doi: 10.1007/978-3-030-79876-5_37.

Sofiane Ennadir, Siavash Golkar, and Leopoldo Sarra. Joint embeddings go temporal. *arXiv preprint arXiv:2509.25449*, 2025. NeurIPS 2024 Workshop on Time Series in the Age of Large Models.

Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. Nonnegative matrix factorization with the Itakura–Saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830, 2009. doi: 10.1162/neco.2008.04-08-771.

Moritz Gögl and Christopher Yau. Var-JEPA: A variational formulation of the joint-embedding predictive architecture – bridging predictive and generative self-supervised learning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, volume 306 of *PMLR*, 2026. arXiv:2603.20111.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Yanan He, Yunshi Wen, Xin Wang, and Tengfei Ma. SC-JEPA: Stabilizing latent predictive learning for time-series anomaly prediction. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, 2026. arXiv:2602.04643.

Zhe Hong. The boiling frog threshold: Criticality and blindness in world model-based anomaly detection under gradual drift. *arXiv preprint arXiv:2603.08455*, 2026.

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

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Tze Leung Lai. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information Theory*, 44(7):2917–2929, 1998. doi: 10.1109/18.737522.

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

Alexander Lavin and Subutai Ahmad. Evaluating real-time anomaly detection algorithms – the Numenta anomaly benchmark. In *IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 2015. arXiv:1510.03336.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview preprint, version 0.9.2, 2022. https://openreview.net/pdf?id=BZ5a1r-kVsf.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Shuang Li, Yao Xie, Hanjun Dai, and Le Song. Scan B-statistic for kernel change-point detection. *Sequential Analysis*, 38(4):503–544, 2019. doi: 10.1080/07474946.2019.1686886.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Gary Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, 1971. doi: 10.1214/aoms/1177693055.

Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. Long short term memory networks for anomaly detection in time series. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, pp. 89–94, 2015.

Xudong Mou, Rui Wang, Bo Li, Tianyu Wo, Jie Sun, Hui Wang, and Xudong Liu. RoCA: Robust contrastive one-class time series anomaly detection with contaminated data. *arXiv preprint arXiv:2503.18385*, 2025.

George V. Moustakides. Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14(4):1379–1387, 1986. doi: 10.1214/aos/1176350164.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

David A. Nix and Andreas S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of the IEEE International Conference on Neural Networks (ICNN)*, pp. 55–60, 1994. doi: 10.1109/ICNN.1994.374138.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954. doi: 10.2307/2333009.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2):1–38, 2021. doi: 10.1145/3439950.

Aaditya Ramdas. Universality of e-detectors for ARL control. *arXiv preprint arXiv:2608.12660*, 2026.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4), 2023. doi: 10.1214/23-STS894.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pp. 4393–4402, 2018.

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

Aitor Sánchez-Ferrera, Elisabeth Wetzer, Kristoffer Wickstrøm, Michael Kampffmeyer, and Robert Jenssen. ProtoX-AD: Self-explainable time series anomaly detection and characterization. *arXiv preprint arXiv:2606.13277*, 2026.

Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. Anomaly detection in time series: A comprehensive evaluation. *Proceedings of the VLDB Endowment*, 15(9):1779–1797, 2022. doi: 10.14778/3538598.3538602.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations (ICLR)*, 2022.

Shubhanshu Shekhar and Aaditya Ramdas. Sequential change detection via backward confidence sequences. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, 2023.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. E-detectors: A nonparametric framework for sequential change detection. *The New England Journal of Statistics in Data Science*, pp. 229–260, 2023. doi: 10.51387/23-NEJSDS51.

Andrew Stirn, Hans-Hermann Wessels, Megan Schertzer, Laura Pereira, Neville E. Sanjana, and David A. Knowles. Faithful heteroscedastic regression with neural networks. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *PMLR*, 2023.

Alexander Tartakovsky, Igor Nikiforov, and Michèle Basseville. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. CRC Press, 2014. doi: 10.1201/b17279.

The mathlib Community. The Lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP)*, 2020. arXiv:1910.09336.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

Gerrit J. J. van den Burg and Christopher K. I. Williams. An evaluation of change point detection algorithms. *arXiv preprint arXiv:2003.06222*, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Stijn Verdenius, Andrea Zerio, and Roy L. M. Wang. LaT-PFN: A joint embedding predictive architecture for in-context time-series forecasting. *arXiv preprint arXiv:2405.10093*, 2024.

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

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. doi: 10.2307/3001968.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations (ICLR)*, 2023.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations (ICLR)*, 2022.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. TS2Vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

APPENDIX

Contents. A Notation B Benchmarks and representations C Two-channel theory: proofs D Absorption theorem: proofs E Lean formalization F Detector and evaluation protocols G Channel readout, variants and pre-registration H Additional results I Ablations and controls J Extended limitations K Extended discussion

A NOTATION

Table 4 collects the symbols used throughout.

Table 4: Symbols used in the paper.

symbol	meaning
$P,\ Q$	calm and stress laws of consecutive latents (z, z'); the change happens at $\nu = 150$
$z_t \in \mathbb{R}^d$, $d = 8$	latent produced by an representation from a 64-step observation window
$\hat{\Phi}$, r_t, Σ_0, A	frozen VAR(1) null, its residual, calm residual covariance, $A = \Sigma_0^{-1}$
SS_n, D_n	run energy $\sum r^\top A r$ and squared drift $S_n^\top A S_n$, $S_n = \sum r$
$V_g(n)$, $M_g(n)$	variance and drift channels of the e-process at grid point g
$\mathcal{G} = \{1, 1.5, 2.5, 4\}$	variance-alternative grid; $\eta^2 = 1$ drift prior; restart weights $\propto 0.95^{t-k}$
$\bar{g}$, γ_{eff}	post-change energy inflation; its one-step estimate (equals R_{lin})
$\varphi(\bar{g}, g)$, d_{IS}	$\bar{g}(1 - 1/g) - \log g$; Itakura-Saito divergence $x - 1 - \log x$
$\tau_{\mathcal{G}} = 3 \log \frac{3}{2}$	grid threshold (Theorem 1), ≈ 1.2164
μ_c, μ_Q, δ	calm and stress conditional means of z' given z; $\delta = \mu_Q - \mu_c$
$\hat{\mu}$, μ_{lin}	the model's own calm-trained one-step predictor; the frozen linear null
a, c	calm error $\mu_c - \hat{\mu}$; the model's correction $\hat{\mu} - \mu_{\text{lin}}$
$\hat{\sigma}^2$	calm-optimal diagonal variance head $\mathbb{E}_P[(z' - \hat{\mu})^2 \mid z]$
I_f	inflation $\mathbb{E}_Q\|z' - f\|^2 - \mathbb{E}_P\|z' - f\|^2$ of predictor f
R_{lin}, R_{raw}, R_{std}	frozen, raw-model and $\hat{\sigma}$-standardized energy ratios, Equation (2)

B BENCHMARKS AND REPRESENTATIONS

B.1 THE FOUR SYNTHETIC SETTINGS

All synthetic settings share the geometry of the prior benchmark. Result files use the original identifiers: SWIRL is `nonlinear_swirl`, SWIRL-LIN is `linear_rotation`, ROTOR is `nonlinear_latent`, SECTORS is `object_structured`, and the privileged reference TRUE-STATE is `state_linear`. A latent state of dimension 6, made of three rotation planes, is observed through a fixed random two-layer $\tanh$ network into 24 channels with noise of standard deviation 0.15, and 32 pure-noise channels of higher variance are appended, so $d_x = 56$. Episodes have length 300, a burn-in of 200 steps and a change at $\nu = 150$. Constants: $\rho = 0.9$, innovation scale $\sigma = 0.4$, base angle $\theta_0 = 0.60$, $\kappa = 0.17$, $c = \sigma^2/(1 - \rho^2)$, $m = 2c$.

SWIRL. In each rotation plane with state $s \in \mathbb{R}^2$,

$$\begin{aligned} &\text{calm:} && s_{t+1} = \rho R(\theta_0) s_t + \varepsilon_t, && \varepsilon_t \sim \mathcal{N}(0, \sigma^2 I), \\ &\text{stress:} && s_{t+1} = \rho_s R(\theta(s_t)) s_t + \varepsilon_t, && \theta(s) = \alpha + \kappa(\|s\|^2 - m), \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_s^2 I), \end{aligned}$$

with $\rho_s = \rho(1 + (\kappa m)^2)$, $\sigma_s^2 = c(1 - \rho_s^2)$ and $\alpha = \theta_0 + \kappa m - 2\arctan(\kappa m)$. These constants match the marginal law exactly and compensate the lag-1 autocovariance to 0.8%, so the change lives in the *state dependence* of the conditional law. In observation coordinates the compensation closes the variance channel. In state coordinates $\kappa(\|s\|^2 - m)$ is a genuine variance modulation, which is why the true state can reopen it.

SWIRL-LIN (control). The stress angle is $\theta_0 + 1.0$ in every plane, with the same ρ and σ. The transition itself changes linearly, so a linear null can express the change.

ROTOR (Benchmark B of the prior study). A stable VAR(1) latent with isotropic innovations whose rotation angle is the only thing the regime changes. Its stationary covariance cI solves

$\Sigma = \rho^2 R\Sigma R^\top + \sigma^2 I$ for *every* orthogonal R, so the marginal of the latent, and hence of the observations, is identical across regimes. The change lives purely in temporal dependence.

SECTORS (dilution). Eight exchangeable sectors, each with a 4-dimensional rotating latent loaded by six observed series ($d_x = 48$). At the change exactly one sector's rotation angle changes. The marginals are invariant by the same rotation argument, but the lag-1 autocovariance is not matched. The difficulty is that the change lives in one eighth of the channels. The privileged TRUE-STATE reference sees all sectors' true latents; over the prior study's five seeds it detects 0.133 (seed 0: 0.109), and the exact-likelihood oracle detects 0.941.

B.2 MODELS, BACKBONES AND BUDGETS

All learned models use $d_z = 8$, a 64-step context, target windows that *end* at $t + \Delta$ for $\Delta \in \{1, 2, 4, 8\}$, an EMA target encoder (decay 0.996), batch 256, AdamW, and checkpoint selection under a fixed feasibility constraint (median per-dimension standard deviation at least 0.5, effective rank at least 4) that is never consulted against detection labels. The small tier uses encoder depth 3 and width 64. Unless noted, models train for 2000 steps on 150 calm episodes.

JEPA. Causal dilated TCN encoder (Bai et al., 2018). For each Δ a two-layer GELU MLP head predicts the normalized EMA target code, with VICReg-style variance and covariance penalties (Bardes et al., 2022). The NRD uses the $\Delta = 1$ head.

Var-JEPA. The same TCN trunk emits $[\mu_{s_x}; \log \sigma^2_{s_x}]$. The objective follows the Var-JEPA ELBO (Gögl & Yau, 2026): full-window Gaussian reconstruction of the context and target windows, a conditioning latent with a unit-Gaussian prior, and a learned conditional prior $p(s_y \mid s_x, z) = \mathcal{N}(m, \operatorname{diag} v)$, which is the predictor with a variance head. The KL weights are $\alpha_{s_x} = \alpha_z = 0.2$ and $\alpha_{s_y} = 1$. The KL weight is 0 for the first 30% of steps and then ramps linearly over 40%. Log-variances are clamped to $[-8, 4]$, and no VICReg or other anti-collapse penalty is used. The scored latent is the posterior mean μ_{s_x}; the variance head modulates the latent and is never appended to it, because appending dilutes the channel (Section G.1). The NRD uses the conditional-prior mean.

Transformer-JEPA. Causal transformer encoder (depth 2, width 64, 8 heads, learned positions) under the JEPA objective (Vaswani et al., 2017). The final configuration trains for 4000 steps on 400 calm episodes at learning rate 5×10^{-4} (Section H.1).

PatchTST-JEPA. Non-overlapping patches of length 8, a linear patch embedding, a bidirectional transformer over patches, and mean pooling, as in PatchTST (Nie et al., 2023), trained with the JEPA objective.

TimesNet-style JEPA. Top-3 FFT periods; each period folds the window into a 2-D tensor processed by two 3×3 convolutions and mean-pooled; period features are averaged (Wu et al., 2023). This is our in-context implementation of the encoder, not the full TimesNet stack.

Reference and projected representations. TRUE-STATE (true state), PSR (spectral subspace identification), PCA (top eight components), raw observations, an autoencoder, and a contrastive encoder (van den Oord et al., 2018), of the family that TS2Vec also belongs to (Yue et al., 2022); details as in the prior study.

Parameter counts (small tier, measured by instantiating the modules): JEPA encoder $41{,}608$, one-step predictor head $1{,}096$, whole model $45{,}992$. The zero-shot foundation model used in Section 5.3 has about 2×10^8 parameters.

C TWO-CHANNEL THEORY: PROOFS

The frozen detector's per-run log-evidence at grid point g is $V_g(n) + M_g(n)$ as in Equation (1); the split reproduces the implemented log-evidence to $|\Delta| \le 3.6 \times 10^{-15}$ (prior verification script, 23/23 checks).

Gaussian origin of V_g. For $\mathcal{N}(0, g\Sigma_0)$ against $\mathcal{N}(0, \Sigma_0)$ and $s = r^\top \Sigma_0^{-1} r$, we have $\log\det(g\Sigma_0) = d \log g + \log\det \Sigma_0$ (`det_scaled_cov`) and $r^\top (g\Sigma_0)^{-1} r = s/g$ (`quad_scaled_cov`). The log-density difference is therefore $\frac{s}{2}(1 - \frac{1}{g}) - \frac{d}{2}\log g$ (`gaussian_scale_llr`), and summing over a run gives $V_g(n)$.

Proof of Lemma 1. (a) By linearity, $\mathbb{E}[V_g(n)] = \frac{nd\bar{g}}{2}(1 - \frac{1}{g}) - \frac{nd}{2}\log g = \frac{nd}{2}\varphi(\bar{g}, g)$ (`expected_llr`). Apply $\log x \le x - 1$ at $x = \bar{g}/g$: $\log \bar{g} - \log g \le \bar{g}/g - 1$, i.e. $\varphi(\bar{g}, g) = \bar{g} - \bar{g}/g - \log g \le \bar{g} - 1 - \log \bar{g} = d_{\mathrm{IS}}(\bar{g})$ (`varShape_le_dIS`). Equality holds at $g = \bar{g}$ (`varShape_self`). Finally $d_{\mathrm{IS}}(x) \ge 0$ with equality iff $x = 1$, since $\log x < x - 1$ for $x \ne 1$ (`dIS_nonneg`, `dIS_eq_zero_iff`). (b) For a zero-mean stationary residual process with autocovariances $\Gamma(h)$, $\mathbb{E}[D_n] = \mathrm{tr}(A\Omega_n)$ with $\Omega_n = \sum_{|h|<n}(n - |h|)\Gamma(h)$, which grows like $n\Omega$. If $\mathbb{E}[D_n] \le n\Omega$ then $\frac{\eta^2 \mathbb{E}[D_n]}{2g(1+n\eta^2)} \le \frac{\Omega}{2g} \cdot \frac{n\eta^2}{1+n\eta^2} \le \frac{\Omega}{2g}$ (`drift_positive_part_le`). The price $\frac{d}{2}\log(1 + n\eta^2)$ is non-decreasing (`drift_price_mono`) and exceeds any bound M once $n \ge e^{2M/d}/\eta^2$ (`drift_price_unbounded`). □

Proof of Theorem 1. At $g = 1$ the increment is 0 (`varShape_one`). For $g > 1$, $1 - 1/g > 0$, so $\varphi(\bar{g}, g) > 0 \iff \bar{g} > \log g/(1 - 1/g)$ (`varShape_pos_iff`). The three thresholds are $3\log\frac{3}{2}$, $\frac{5}{3}\log\frac{5}{2}$ and $\frac{4}{3}\log 4$ (`thr_three_halves`, `thr_five_halves`, `thr_four`). The first is the smallest: $9\log\frac{3}{2} < 5\log\frac{5}{2}$ because $(3/2)^9 = 19683/512 < 3125/32 = (5/2)^5$, and $9\log\frac{3}{2} < 4\log 4$ because $(3/2)^9 < 4^4$ (`thr_lt_five_halves`, `thr_lt_four`). Hence some $g \in \mathcal{G}$ has a positive expected increment iff $\bar{g} > \tau_{\mathcal{G}}$ (`grid_closure`); multiplying by $nd/2 > 0$ gives the detector-scale statement (`grid_closure_scaled`); the contrapositive is `grid_dead`. For the digits, $\log\frac{3}{2} = -\log(1 - \frac{1}{3})$ is bracketed by 16 terms of the Mercator series with remainder at most $(1/3)^{17}/(2/3)$, which gives $0.40546505 < \log\frac{3}{2} < 0.40546515$ and therefore $1.21639 < \tau_{\mathcal{G}} < 1.21640$ (`log_three_halves_bounds`, `tauG_bounds`). Likewise $\log\frac{5}{4}$ (10 terms) and $\log 2$ give $1.52715 < \frac{5}{3}\log\frac{5}{2} < 1.52716$ and $1.84839 < \frac{4}{3}\log 4 < 1.84840$. □

Dilution. Spreading a fixed change c over more latent dimensions lowers the effective inflation $1 + c/d$ and strictly lowers d_{IS} (`dIS_mono`, `dilution_decreases_divergence`). This is the direction of the prior study's dilution law $\bar{g} = 1 + k(v - 1)/d$. Table 5 reproduces the prior study's measured dilution curve; its predicted growth column is exactly $\frac{d}{2}\max_{g\in\mathcal{G}}\varphi(\gamma^*, g)$ from Lemma 1 and Theorem 1. The best grid point switches from 1.5 to 1 between $\gamma^* = 1.248$ and 1.213, which bracket $\tau_{\mathcal{G}}$ (`dilution_switch`).

Table 5: Dilution curve from the prior study (true latent of the object benchmark; only distractor sectors are added, so the information about the change is identical in every row).

sectors	d	γ^*	best grid g	predicted growth / step	detection
1	4	2.489	2.5	1.155	0.940
2	8	1.745	1.5	0.704	0.925
3	12	1.496	1.5	0.560	0.896
4	16	1.372	1.5	0.416	0.785
5	20	1.298	1.5	0.272	0.540
6	24	1.248	1.5	0.127	0.300
7	28	1.213	1.0	0.000	0.158
8	32	1.186	1.0	0.000	0.105

C.1 What Theorem 1 does and does not say about the diagnostic g^*

The channel diagnostic reports g^*, the grid point that maximizes a standardized post-versus-calm contrast $[\mathbb{E}_{\mathrm{post}}\,\mathrm{term} - \mathbb{E}_{\mathrm{calm}}\,\mathrm{term}]/\mathrm{SD}_{\mathrm{calm}}\,\mathrm{term}$ over (g, n). In this contrast the price $\frac{nd}{2}\log g$ cancels, so $g^* > 1$ is possible even when $\bar{g} < \tau_{\mathcal{G}}$. For example, TRUE-STATE on seed 0 has $\gamma_{\mathrm{eff}} = 1.112$ and $g^* = 1.5$. Theorem 1 is a statement about *evidence growth* under the post-change law. That $g^* = 1$ on every Var-JEPA seed and variant is *consistent with* Theorem 1, but it is not a consequence of it.

C.2 Denser grids, continuous alternatives and the evidence ceiling

Theorem 1 is a statement about one grid, and the obvious objection is that a better grid would open the channel. It would, and it would not help.

The threshold is only a property of the grid. Write $h(g) = \log g/(1 - 1/g)$ for the per-point threshold of Theorem 1. On $g > 1$ we have $h(g) > 1$, and adding the alternative $1 + t$ brings the threshold down to at most $1 + t$ (`grid_threshold_gt_one`, `grid_threshold_le_one_add`), so $\tau_{\mathcal{G}} = \min_{g \in \mathcal{G}, g>1} h(g)$ falls toward one as alternatives are added near one, and no grid can push it below one. Table 16 lists $\tau_{\mathcal{G}}$ for the grids we run.

Proposition 1 (Evidence ceiling)**.** *Let $\bar{g} \geq 1$ and let ρ be the restart-weight ratio. For* every *grid, including the continuum $g > 1$, the variance channel's expected log-evidence for a run of length n is at most $\frac{nd}{2} d_{\mathrm{IS}}(\bar{g})$, and the run's mixture weight costs $n \log(1/\rho)$. The channel therefore grows without bound only if $d_{\mathrm{IS}}(\bar{g}) > \frac{2}{d} \log \frac{1}{\rho}$, which for $d = 8$ and $\rho = 0.95$ means $\bar{g} > 1.169$ ($\bar{g} > 1.255$ for the grid $\mathcal{G}$).*

What the grid cannot change. By Lemma 1(a) the expected variance-channel evidence of a run of length n is at most $\frac{nd}{2} d_{\mathrm{IS}}(\bar{g})$ whatever the grid, including the continuum, with equality at $g = \bar{g}$. The restart mixture prices a run of length n at $w_k \propto \rho^{n-1}$, so the term's expected log-weight falls by $\log(1/\rho)$ per step. The channel accumulates without bound only if

$$\tfrac{d}{2}\, d_{\mathrm{IS}}(\bar{g}) \;>\; \log \tfrac{1}{\rho}, \tag{3}$$

which for $d = 8$ and $\rho = 0.95$ requires $\bar{g} > 1.169$, against $\bar{g} > 1.255$ for the grid $\mathcal{G}$ and $\bar{g} > \tau_{\mathcal{G}}$ for growth before the restart price. The measured SWIRL inflations, $R_{\mathrm{lin}} = 1.089$ on average and 1.125 at worst, are below all three. The Lean development certifies the worst case: for every $\bar{g} \leq 1.125$ and every g, the expected growth rate stays under the restart price (`evidence_ceiling_swirl`), using $d_{\mathrm{IS}}(x) \leq (x - 1) - 2(1 - 1/\sqrt{x})$ (`dIS_sqrt_bound`).

Even ignoring the restart price, the whole post-change window is short. With $n = 150$ steps and $d = 8$, the best continuous alternative supplies at most $\frac{nd}{2} d_{\mathrm{IS}}(\bar{g})$ nats in total, which on five of the seven seeds is below the anytime-valid level $\log(1/\alpha) \approx 3.00$ (Table 17).

Measured. Table 16 runs the frozen detector and NRD on SWIRL under the three grids, on the ablation sweep of Section H.4. Densifying the grid moves frozen detection by at most 0.027 in absolute terms and does not change the conclusion; NRD is likewise unchanged, since its evidence does not come from the variance channel.

D ABSORPTION THEOREM: PROOFS

We write the proofs for general laws. The Lean formalization states them for finitely supported laws: a marginal p over states z and conditional weights over finitely many outcomes, as in the structure `Law`. This covers every empirical average reported in the paper. Assumption (O) enters Lean as the hypothesis `hopt`: $\forall t,\ \mathbb{E}_P\|z' - \hat{\mu}\|^2 \leq \mathbb{E}_P\|z' - (\hat{\mu} + t\, b)\|^2$ with $b = \mu_{\mathrm{lin}} - \hat{\mu} = -c$.

Lemma 3 (Bias-variance)**.** *For any predictor f, $\mathbb{E}_Q\|z' - f(z)\|^2 = \mathbb{E}_Q\|\mu_Q(z) - f(z)\|^2 + \mathbb{E}_Q \operatorname{tr} \Sigma_Q(z)$, where $\Sigma_Q(z) = \mathrm{Cov}_Q(z' \mid z)$.*

Proof. Write $z' = \mu_Q(z) + \epsilon$ with $\mathbb{E}_Q[\epsilon \mid z] = 0$ and expand; the cross term vanishes. In Lean this is `bv_scalar` coordinatewise and `risk_decomp` for vectors. □

Identity (B1). Since $\mu_Q - \mu_{\mathrm{lin}} = a + c + \delta$ and $\mu_Q - \hat{\mu} = a + \delta$, Lemma 3 gives $\mathbb{E}_Q\|z' - \mu_{\mathrm{lin}}\|^2 - \mathbb{E}_Q\|z' - \hat{\mu}\|^2 = \mathbb{E}_Q\big[\|c\|^2 + 2\langle c, a\rangle + 2\langle c, \delta\rangle\big]$ (`stress_gap`, written there with $b = -c$ as $\mathbb{E}_Q[\|b\|^2 - 2\langle b, \delta\rangle - 2\langle a, b\rangle]$). With $\delta = 0$ this gives the calm gap $\mathbb{E}_P[\|c\|^2 + 2\langle c, a\rangle]$ (`calm_gap`).

Lemma 4 (Cross term)**.** *Under (O), $\mathbb{E}_P\langle a, c\rangle = 0$; under (M) as well, $\mathbb{E}_Q\langle a, c\rangle = 0$.*

Proof. By Lemma 3, $J(t) = \mathbb{E}_P\|z' - \hat{\mu} - t\, c\|^2 = J(0) - 2t\, \mathbb{E}_P\langle a, c\rangle + t^2\, \mathbb{E}_P\|c\|^2$. If $J(0) \leq J(t)$ for all t, then $2tA \leq t^2 B$ for all t with $A = \mathbb{E}_P\langle a, c\rangle$ and $B \geq 0$. Taking $t = A/(B + 1)$ gives $A^2(B + 2) \leq 0$, so $A = 0$ (`quad_nonneg_forces_zero`, `calm_orthogonality`). The function $\langle a(z), c(z)\rangle$ depends on z only, so under $Q_z = P_z$ its Q-expectation equals its P-expectation (`lemmaH`, `cross_term_zero`). □

Proof of Theorem 2. Inflation identity. Subtracting the calm gap from (B1), $I_{\text{lin}} - I_{\text{model}} = \big(\mathbb{E}_Q - \mathbb{E}_P\big)\big[\|c\|^2 + 2\langle c, a\rangle\big] + 2\mathbb{E}_Q\langle c, \delta\rangle$. The first bracket is a function of z only, so it has equal expectations under (M). Hence $I_{\text{lin}} - I_{\text{model}} = 2\mathbb{E}_Q\langle c, \delta\rangle$ (`inflation_identity`). The two criteria are rearrangements (`absorption_iff`, `total_absorption_iff`); $R_{\text{raw}} \le 1 \iff I_{\text{model}} \le 0$. *Stress energies.* Under (M) and (O), Lemma 4 removes $2\mathbb{E}_Q\langle c, a\rangle$ from (B1), and $\|c\|^2 + 2\langle c, \delta\rangle = \|c + \delta\|^2 - \|\delta\|^2$ (`complete_square`, `propC`, `propC_criterion`). □

Table 6: Corollary 1: the measured triple (SWIRL, Var-JEPA, 7 seeds) and what each number instantiates.

quantity	measured	comparison	reading
R_{lin}	1.089 (max 1.125)	$< \tau_{\mathcal{G}}$ on every seed	no grid alternative grows evidence (Theorem 1)
R_{raw}	0.959 (max 0.991)	< 1 on every seed	total mean absorption (Theorem 2)
R_{std}	1.066 (max 1.090)	$< \tau_{\mathcal{G}}$ on every seed	the head anchors at one and re-weights (Theorem 3)
NRD-$\hat{\sigma}$ detection	0.000 (3 seeds)	$R_{\text{std}} < \tau_{\mathcal{G}}$	consistent: nothing the grid can grow
NRD detection	0.950 (JEPA), 0.366 (Var-JEPA)	calm metric	the surviving view, read through the drift channel

Reading Theorem 2. (i) If $\delta \perp c$ in $L^2(Q)$, the model's inflation equals the linear null's: nothing is absorbed, and the model's *smaller* calm energy then makes R_{raw} larger than the Euclidean frozen ratio. (ii) Absorption requires the shift to point along the model's own correction. (iii) Total absorption, $R_{\text{raw}} \le 1$, happens iff $2\mathbb{E}_Q\langle c, \delta\rangle \ge I_{\text{lin}}$; this is what the swirl measurements show on every seed. (iv) On SWIRL-LIN marginal matching fails, because the window law of the latent changes with the transition. The measured deflation $R_{\text{raw}} = 0.571$ is therefore an empirical fact outside the proposition's hypotheses, and we report it as such.

Proof of Lemma 2. $R_{\text{raw}}(P) = 1$ trivially (`lemmaA_raw`). By the tower property, $\mathbb{E}_P[(z'_i - \hat{\mu}_i)^2/\hat{\sigma}_i^2] = \mathbb{E}_P\big[\mathbb{E}_P[(z'_i - \hat{\mu}_i)^2 \mid z]/\hat{\sigma}_i^2\big] = 1$ for each coordinate, and averaging over d coordinates gives $R_{\text{std}}(P) = 1$ (`lemmaA_std`). Lemma 3 coordinatewise gives $\hat{\sigma}_i^2 = a_i^2 + \text{Var}_P(z'_i \mid z)$ (`sigHat_decomp`). □

Proof of Theorem 3. Coordinatewise, $\mathbb{E}_Q[(z'_i - \hat{\mu}_i)^2 \mid z] = (\mu_{Q,i} - \hat{\mu}_i)^2 + \text{Var}_Q(z'_i \mid z) = (a_i + \delta_i)^2 + \text{Var}_Q$; dividing by the state-dependent $\hat{\sigma}_i^2(z)$ and averaging gives the exact form (`propD_exact`). Substituting $\hat{\sigma}^2 = a^2 + \text{Var}_P$ gives $\frac{(a_i+\delta_i)^2+\text{Var}_Q}{\hat{\sigma}_i^2} = 1 + \frac{\delta_i^2+2a_i\delta_i+\text{Var}_Q-\text{Var}_P}{\hat{\sigma}_i^2}$ (`propD_decomp`). If $\text{Var}_Q \le \hat{\sigma}^2$ then $\text{Var}_Q/\hat{\sigma}^2 \le 1$, which gives the bound (`propD_bound`); with $\hat{\mu} = \mu_c$ we have $a = 0$ (`propD_bound_wellspec`). □

Corollary 1 in full. Table 6 lists the measured triple with the statement each number instantiates. On seed 0 the triple is $R_{\text{lin}} = 1.1249$, $R_{\text{raw}} = 0.9910$, $R_{\text{std}} = 1.0903$. The experiment log's seed-0 bookkeeping of the absorbed mean shift is: frozen inflation $+0.1195$, model inflation -0.0136, absorbed $+0.1331$ in the log's normalization (`absorbed_seed0` checks the arithmetic).

E LEAN FORMALIZATION

The supplementary folder `lean/` is a Lake project (Lean `v4.33.1`, Mathlib pinned to the same revision) with two modules. Each module ends with a `#print axioms` audit, and `check.sh` compiles them and fails on any `sorryAx`. All 65 theorems depend only on `propext`, `Classical.choice` and `Quot.sound`. Table 7 maps each Lean name to the statement it certifies.

Table 7: Lean theorems and the statements they certify.

Lean name (module `Channels`)	paper statement
`gaussian_scale_llr`, `det_scaled_cov`, `quad_scaled_cov`, `expected_llr`	V_g is a Gaussian LLR; its expectation
`dIS_nonneg`, `dIS_eq_zero_iff`, `varShape_le_dIS`, `varShape_self`, `varShape_one`, `varShape_dead_at_one`	Lemma 1(a)
`drift_positive_part_le`, `drift_price_mono`, `drift_price_unbounded`	Lemma 1(b)
`varShape_pos_iff`, `varShape_nonpos_iff`, `thr_three_halves`, `thr_five_halves`, `thr_four`, `thr_lt_five_halves`, `thr_lt_four`, `grid_closure`, `grid_closure_scaled`, `grid_dead`	Theorem 1
`log_three_halves_bounds`, `tauG_bounds`, `log_five_fourths_bounds`, `thr_five_halves_bounds`, `thr_four_bounds`	certified digits 1.2164, 1.5272, 1.8484
`dIS_mono`, `dilution_decreases_divergence`	dilution
`grid_threshold_gt_one`, `grid_threshold_le_one_add`, `dIS_sqrt_bound`, `ceiling_criterion`, `restart_price_ge`, `evidence_ceiling_swirl`	Proposition 1
`swirl_Rlin_every_seed`, `swirl_Rstd_every_seed`, `object_Rlin_every_seed`, `linrot_Rlin_every_seed`, `nonlinear_latent_every_seed`, `varjepa_variants_every_one`, `dilution_switch`, `absorbed_seed0`, `wilcoxon_all_positive_n7`	measured instances (per seed)
Lean name (module `Absorption`)	**paper statement**
`bv_scalar`, `risk_decomp`	Lemma 3
`B1_pointwise`, `stress_gap`, `calm_gap`, `complete_square`, `nsq_sub_smul`	identity (B1)
`quad_nonneg_forces_zero`, `calm_orthogonality`, `lemmaH`, `cross_term_zero`	Lemma 4
`propC`, `propC_criterion`, `inflation_identity`, `absorption_iff`, `total_absorption_iff`	Theorem 2
`sigHat_decomp`, `lemmaA_raw`, `lemmaA_std`	Lemma 2
`propD_exact`, `propD_decomp`, `propD_bound`, `propD_bound_wellspec`	Theorem 3

F DETECTOR AND EVALUATION PROTOCOLS

Splits. For seed s, calm training episodes use generator seed $s + 1000$ (150 episodes), validation $s+1500$, calibration $s+2000$ (200), calm holdout $s+2500$ (200), and test episodes $s+3000$ (150, change at ν). Every detector sees the same episodes.

Frozen detector (track 1). A mean/covariance whitener fitted on calm training latents (shrinkage 0.1) is applied to every split. A ridge VAR(1) null (10^{-4}) is fitted on whitened calm training latents. The restart-mixture e-process equation 1 runs with burn-in 12, $\eta^2 = 1$ and geometric restart weights with ratio 0.95.

Achieved-PFA calibration. For each of 40 thresholds, set at the $(1-\alpha)$ quantiles of the calibration episodes' running maximum over $[12, \nu)$, the achieved PFA is the fraction of *holdout* calm episodes that alarm. The detection rate is the fraction of test episodes whose first alarm falls at or after ν. The reported rate is the linear interpolation of the (achieved PFA, detection) curve at 0.05; it is undefined when 0.05 lies outside the achieved span. This machinery produces every detection number in the paper, including the oracle's.

Native-residual detection (NRD). $\hat{r}_t = z_{t+1} - \hat{\mu}(z_t)$ on the unwhitened latent, with $\hat{\mu}$ the $\Delta = 1$ JEPA head or Var-JEPA's conditional-prior mean given the posterior-mean conditioning code. The metric is $\Sigma_c = \mathrm{Cov}(\hat{r}) + 10^{-4} I$ over calm calibration residuals after burn-in, followed by the same e-process and calibration. **NRD-$\hat{\sigma}$** replaces the metric by the model's own per-state predictive scale. The **calm-fitted MLP null** is the prior study's track-2 homoskedastic head: a two-hidden-layer GELU MLP fitted by AdamW with early stopping on calm whitened latents, followed by the same e-process.

Recent detector families (3-seed latent dumps). Energy score (Liu et al., 2020), Deep SVDD (Ruff et al., 2018), a kernel-MMD change-point statistic (Gretton et al., 2012; Chang et al., 2019) and a marginal z-score each use their own running-maximum statistic with a calm-quantile threshold at achieved PFA 0.05, without an e-process. e-CUSUM (Shin et al., 2023) uses its e-detector statistic. All are fitted on the same calm-whitened latents.

Foundation model. TimesFM-2.5 ($\sim$200M parameters, zero-shot, no fine-tuning) produces one-step forecasts, and its residuals are scored with the same calm-fitted metric and e-process.

NAB, and why we report no result on it. We ran the full episode protocol on the 17 AWS CloudWatch series of NAB (Lavin & Ahmad, 2015; Ahmad et al., 2017): z-score each series, cut episodes of length 300 with stride 30, call an episode anomalous if it overlaps a labelled window, split the normal episodes 60/20/20 into fit, calibration and holdout, and score the anomalous ones. Two things made us withdraw the comparison. First, it is fragile to a detail that has nothing to do with the question: fitting the frozen detector's linear null and its residual covariance on the split being scored, rather than on calm training episodes, normalises away the energy inflation the detector exists to see, and moves its per-series detection from 0.475 to 0.136 on the Var-JEPA series set. Only the calm-fitted version is consistent with every synthetic experiment here, and under it the frozen detector is not outperformed. Second, the channel readout explains why, and is the useful part: the frozen view's inflation on these series has median $R_{\mathrm{lin}} = 1.91$, above $\tau_{\mathcal{G}}$ on 9 of 10 series, and the model's own residual inflates too ($R_{\mathrm{raw}} > 1$ on 9 of them). By Theorems 1 and 2 this is the regime where the frozen detector already has what it needs and NRD has nothing to add. NAB is therefore not a test of the claim this paper makes.

TCPD. With $T = 150$ and stride 10 on the TCPD series (van den Burg & Williams, 2020), which have 468-816 observations, only one series yields at least 30 null and 3 anomalous episodes. A one-series pool (21 null, 15 anomalous episodes) gives an uninterpretable 0.000 for both detectors. TCPD needs a per-point scan evaluation, which is left to future work.

Statistics. Paired t-tests and exact two-sided Wilcoxon signed-rank tests (`scipy`), recomputed from per-seed rows by our analysis script. With 7 pairs all of the same sign the exact Wilcoxon p is $2 \cdot 2^{-7} = 0.015625$ (`wilcoxon_all_positive_n7`). Seed standard deviations are population standard deviations.

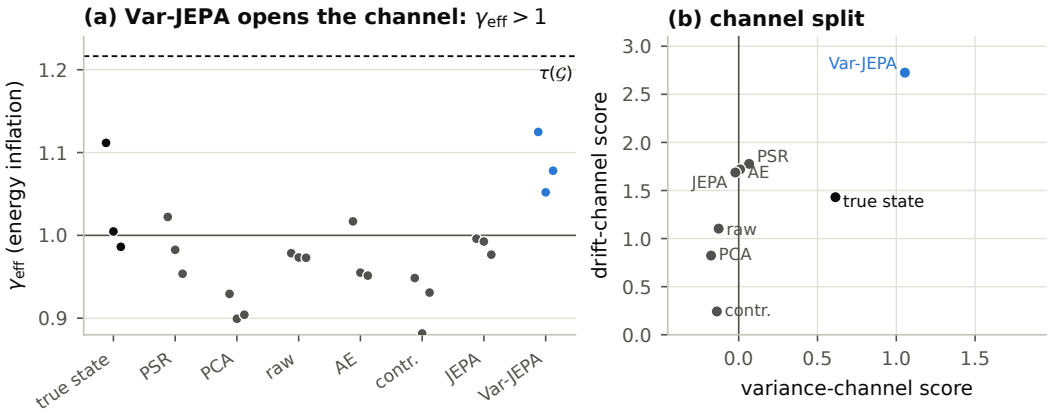

Figure 6: **Channel readout on SWIRL** (3 seeds, frozen view). (a) Energy inflation γ_{eff} per seed: Var-JEPA is the only learned representation above 1 on every seed, and every representation stays below $\tau_{\mathcal{G}}$. (b) Variance- vs. drift-channel scores (seed means).

Table 8: Frozen-detector channel readout on SWIRL (3 seeds): γ_{eff}, selected grid point g^* and window n^* per seed, and seed-mean channel scores.

representation	γ_{eff} (seeds 0/1/2)	g^*	n^*	variance ch.	drift ch.
`state_linear`	1.112 / 1.005 / 0.986	1.5 / 1.0 / 1.0	137 / 11 / 12	+0.61	+1.43
PSR	1.022 / 0.983 / 0.954	1.0 / 1.0 / 1.0	11 / 11 / 12	+0.07	+1.78
PCA	0.929 / 0.899 / 0.904	1.0 / 1.0 / 1.0	10 / 11 / 11	-0.18	+0.82
raw	0.978 / 0.973 / 0.973	1.0 / 1.0 / 1.0	67 / 106 / 136	-0.13	+1.10
AE	1.017 / 0.955 / 0.951	1.0 / 1.0 / 1.0	11 / 10 / 10	+0.01	+1.72
contrastive	0.948 / 0.882 / 0.931	1.0 / 1.0 / 1.0	8 / 137 / 8	-0.14	+0.24
JEPA	0.996 / 0.993 / 0.977	1.0 / 1.0 / 1.0	12 / 12 / 11	-0.02	+1.69
Var-JEPA	1.125 / 1.052 / 1.078	1.0 / 1.0 / 1.0	10 / 10 / 10	+1.05	+2.73

G CHANNEL READOUT, VAR-JEPA VARIANTS AND PRE-REGISTRATION

Table 8 and Figure 6 give the per-seed channel readout of every representation, and Table 10 the Var-JEPA architectural variants.

Same-seed detection (3 seeds; frozen track 1, the prior study's nonlinear track 2, and the calm VAR fit R^2 of each latent): Table 9. Var-JEPA's latent is the most linearly predictable in calm ($R^2 = 0.949$), yet the frozen detector reads the least from it. The channel is open, but the representation is less aligned with the frozen alternative class than JEPA's.

Table 9: Same-seed pilot detection on SWIRL (3 seeds).

representation	track 1 (frozen)	track 2 (MLP null)	calm VAR R^2
`state_linear`	0.419	0.389	0.810
PSR	0.178	0.377	0.674
JEPA	0.138	0.056	0.670
Var-JEPA	0.073	0.076	0.949
PCA	0.060	0.084	0.305
raw	0.032	0.074	0.100
AE	0.031	0.019	0.889
contrastive	0.037 (2 seeds)	0.048	0.901

Var-JEPA variants (seed 0, 2000 steps; experiment log). Every small-tier variant opens the channel, and g^* never leaves 1. The medium tier closes it: more capacity absorbs the change even in the mean path.

Table 10: Var-JEPA architectural variants (seed 0; values from the experiment log).

variant	γ_{eff}	variance ch.	drift ch.	g^*	n^*
baseline ($\alpha_{s_y} = 1$)	1.1249	$+1.54$	$+2.76$	1.0	10
$\alpha_{s_y} = 3$	1.0996	$+1.26$	$+2.59$	1.0	10
$\alpha_{s_y} = 5$	1.0881	$+1.14$	$+2.49$	1.0	10
mean-prediction weight 0.3	1.0492	$+0.77$	$+2.37$	1.0	10
mean-prediction weight 1.0	1.0336	$+0.57$	$+1.86$	1.0	9
$\alpha_{s_y} = 3$, mean weight 0.3	1.0165	$+0.27$	$+2.25$	1.0	10
append $\log \sigma^2$ ($d = 16$)	1.0598	$+1.05$	$+2.10$	1.0	9
target offsets $\{1, 2, 4, 8, 16\}$	1.1267	$+1.59$	$+2.73$	1.0	9
medium tier	0.9903	-0.02	$+0.66$	1.0	8

NRD on the 3-seed pilot (experiment log): frozen $0.035/0.087/0.096$; NRD-$\hat{\sigma}$ $0.000/0.000/0.000$; NRD $0.100/0.460/0.161$. In the seed-0 diagnosis the model's standardized residual energy is 0.86 in calm and 0.94 after the change, so the head absorbs the change. The raw predictive-residual energy ratio is 0.98, against 1.125 for the frozen VAR, so the nonlinear mean absorbs as well.

Pre-registered predictions (fixed before training; verdicts on the 3-seed pilot). **P1** $\gamma_{\text{eff}} > 1$ on at least 2 of 3 seeds: *pass* (3 of 3: $1.125/1.052/1.078$). **P2** $g^* > 1$ and $n^* \geq 30$: *fail* ($g^* = 1$, $n^* = 10$). **P3** frozen detection moves toward the TRUE-STATE bracket: *fail* (0.073 vs. JEPA 0.138). **P4** selectivity on SWIRL-LIN: *consistent* (no advantage over the reference). **P5** the ELBO alone prevents collapse: *consistent* (no checkpoint rejected, no anti-collapse penalty).

G.1 WHY THE CHANNEL CAN BE OPENED AT $d_z = 8$

On the exact frozen pipeline (seed 0; experiment log), the true state gives $\gamma_{\text{eff}} = 1.1117$, $g^* = 1.5$ and variance-channel score $+1.83$. A random orthogonal map of the state gives identical values, as an invertible linear re-coordinatization should. The true state plus two neutral noise coordinates ($d_z = 8$) still opens the channel (1.0846, $g^* = 1.5$). Saturating or normalizing warps close it: $\tanh(s)$ gives 0.854, $s/\|s\|$ gives 0.806, and $s/(1 + |s|)$ gives 0.869, all with $g^* = 1$. Appending quadratic coordinates dilutes it: $[s;\ r^2$ of two planes$]$ gives 0.930 and all quadratic features give 0.976. A variance head must therefore *modulate* the latent rather than append coordinates.

H ADDITIONAL RESULTS

Tables 11 to 13 list every per-seed number behind the means of Section 5, and Figure 8 shows ROTOR per seed.

Table 11: SWIRL per seed: frozen and NRD detection for both representations, the per-seed oracle, and the absorption triple (Var-JEPA).

seed	frozen VJ	NRD VJ	frozen J	NRD J	oracle	R_{lin}	R_{raw}	R_{std}
0	0.035	0.100	0.191	0.937	0.967	1.1249	0.9910	1.0903
1	0.087	0.460	0.277	0.966	0.962	1.0520	0.9218	1.0229
2	0.096	0.161	0.233	0.960	0.970	1.0781	0.9376	1.0468
3	0.079	0.711	0.307	0.920	0.953	1.1042	0.9554	1.0865
4	0.093	0.330	0.247	0.930	0.967	1.0961	0.9773	1.0644
5	0.133	0.400	0.127	0.953	0.955	1.0982	0.9512	1.0613
6	0.107	0.400	0.233	0.983	0.983	1.0683	0.9765	1.0900

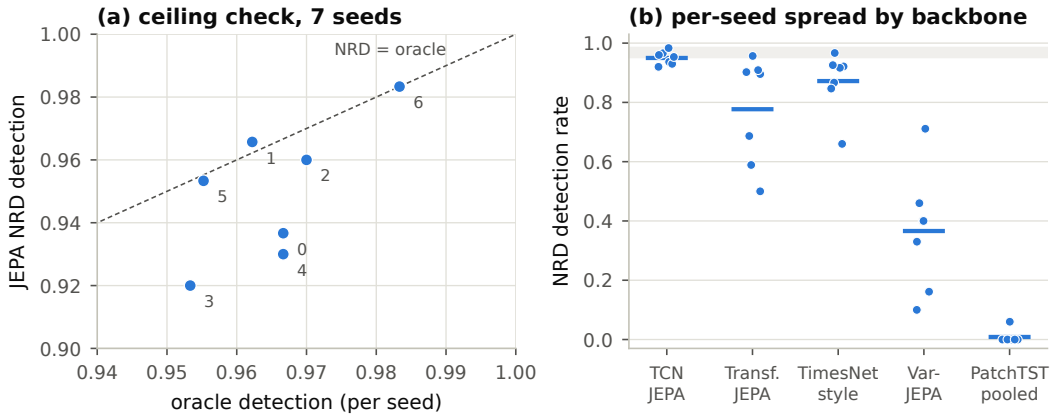

Figure 7: (a) Ceiling check: JEPA's NRD detection never exceeds the per-seed oracle beyond interpolation noise (≤ 0.003); the mean shortfall is 0.015 and the largest is 0.037. (b) Per-seed spread of NRD detection by backbone on SWIRL; the grey band is the oracle range.

Table 12: ROTOR per seed, frozen/NRD. †TimesNet-style was run on 3 seeds.

seed	JEPA	Var-JEPA	PatchTST	Transformer	TimesNet†
0	0.973/0.947	0.927/0.013	0.967/0.927	0.920/0.973	0.000/0.943
1	0.962/0.977	0.902/0.371	0.896/0.944	0.972/0.963	0.000/0.973
2	0.953/0.960	0.957/0.152	0.953/0.953	0.916/0.927	0.000/0.934
3	0.940/0.933	0.913/0.367	0.900/0.970	0.953/0.913	–
4	0.947/0.927	0.959/0.013	0.912/0.927	0.957/0.960	–
5	0.957/0.880	0.920/0.113	0.982/0.920	0.975/0.953	–
6	0.969/0.960	0.956/0.150	0.913/0.863	0.920/0.959	–

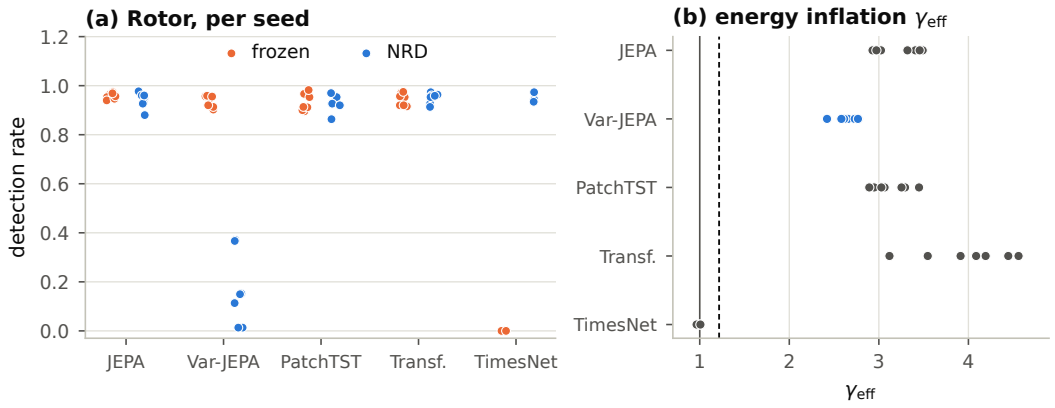

Figure 8: ROTOR: (a) per-seed detection; Var-JEPA's NRD absorbs its own signal while every deterministic backbone reaches the ceiling. (b) Energy inflation γ_{eff}: the channel is wide open for every JEPA-family representation ($\gamma_{\text{eff}} \in [2.42, 4.56]$, all above $\tau_{\mathcal{G}}$; `nonlinear_latent_every_seed`), while TimesNet's latent shows no inflation even though its NRD reaches 0.950.

Table 13: Backbones on SWIRL per seed, frozen/NRD.

seed	Transformer-JEPA	TimesNet-style	PatchTST-JEPA
0	0.047/0.957	0.000/0.867	0.000/0.000
1	0.160/0.896	0.000/0.967	0.000/0.060
2	0.108/0.902	0.000/0.921	0.000/0.000
3	0.020/0.500	0.000/0.847	0.000/0.000
4	0.027/0.687	0.000/0.917	0.000/0.000
5	0.027/0.589	0.000/0.660	0.000/0.000
6	0.160/0.909	0.000/0.926	0.007/0.000

H.1 TRANSFORMER BUDGET STUDY

At the TCN's budget (150 episodes, 2000 steps) the transformer is unstable: NRD detection is 0.048-0.560 over 3 seeds, and the frozen detector gives 0.05-0.11 or no measurable operating point. A sweep on seed 0 with 400 episodes gives 0.827 at 3000 steps and width 64, 0.353 at width 128, and 0.957 at 4000 steps and width 64. The final configuration over 7 seeds gives NRD 0.777 (sd 0.169) against frozen 0.078 (Figure 9). The co-adaptation finding is backbone-invariant, and the transformer is budget-limited: its seed spread misses the pre-set 0.10 stability criterion. Its γ_{eff} stays at 0.90-0.97, so no deterministic backbone opens the variance channel on the swirl.

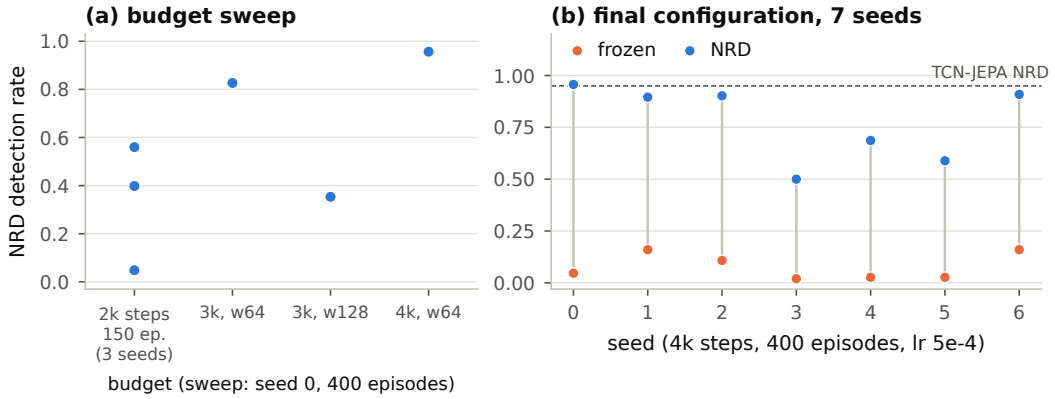

Figure 9: Transformer-JEPA on SWIRL: (a) budget sweep; (b) the final configuration over 7 seeds. The dashed line marks TCN-JEPA's NRD detection.

H.2 CO-ADAPTATION ACROSS ALL REPRESENTATIONS

Figure 10 places every representation on the same axis: the frozen linear null, a calm-fitted MLP null fitted post hoc, and the representation's own co-adapted predictor where one exists. Only the co-adapted predictor moves detection substantially.

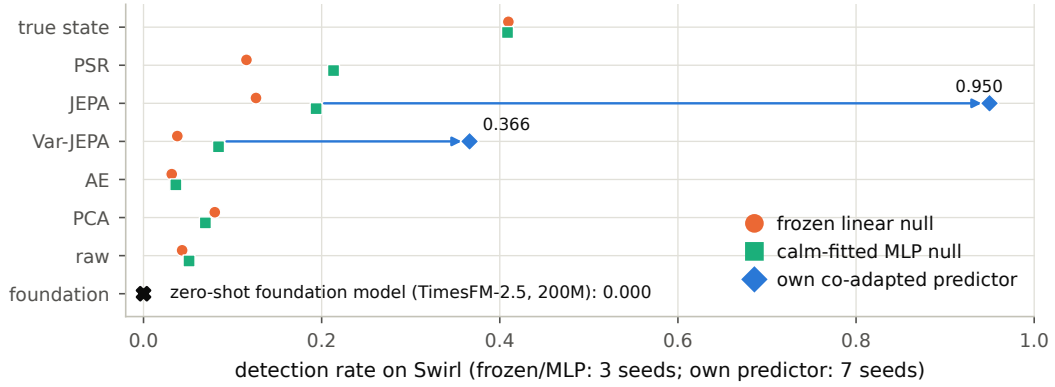

Figure 10: **Which null?** Frozen linear null (3-seed dump), calm-fitted MLP null (3-seed dump) and the representation's own co-adapted predictor (7 seeds) on SWIRL. Post-hoc nulls help little; the co-adapted predictor is a different object. A 200M-parameter zero-shot foundation model scored the same way detects nothing.

H.3 RECENT DETECTOR FAMILIES

Figure 11 reports every detector family of Section 5.3 on the same 3-seed latent dumps, with the 7-seed NRD result for reference.

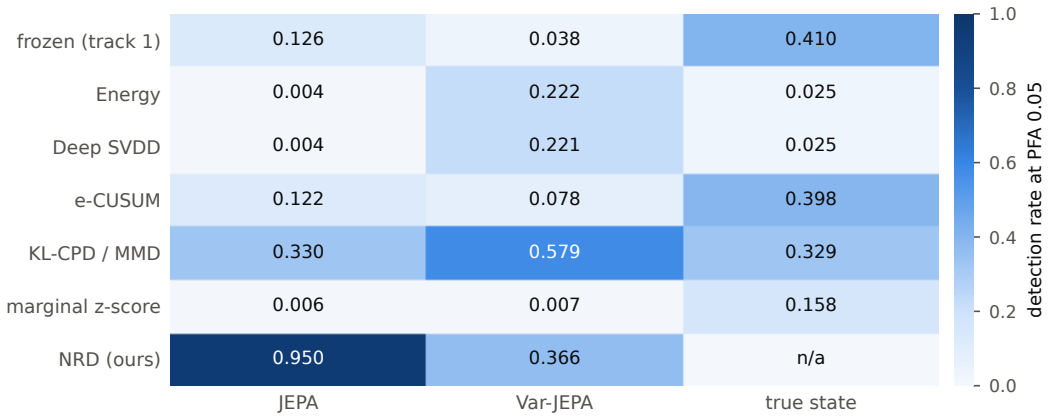

Figure 11: Recent detector families on the same latents (3-seed dumps), all at achieved PFA 0.05. The NRD row is the 7-seed result of Table 1. TRUE-STATE has no learned predictor, so it has no NRD entry.

H.4 AN INDEPENDENT ABLATION SWEEP: DELAY, CALM METRIC, GRID DENSITY, STANDARDIZATION

The ablations of this subsection were produced by one sweep that retrains under the published protocol and then answers every question by re-scoring the *same* trained model, so each comparison is within-run. On SWIRL with 7 seeds, the JEPA arm reproduces the published numbers exactly (frozen 0.231, NRD 0.950; Table 1), which is the strongest check on the pipeline we can run. The Var-JEPA arm does not: its objective draws latent samples on the accelerator, so a re-run is a different draw, and this sweep's Var-JEPA lands at frozen 0.074 and NRD 0.837 against the 0.090 and 0.366 of Table 1. Section H.7 discusses what that spread means for the claims that use it. All Var-JEPA rows below are read against this sweep's own baseline.

Detection delay. Table 14 reports the mean delay, the average first alarm time after ν over detected episodes, at the same achieved PFA as the detection rate. On JEPA latents NRD alarms $3.0\times$ sooner than the frozen detector (21.4 against 64.1 steps), so the gain is not a matter of eventually catching more episodes: the evidence arrives faster.

Table 14: Detection rate and mean delay (steps after ν) at achieved PFA 0.05, SWIRL, JEPA latent, per seed.

	frozen		NRD	
seed	detection	delay	detection	delay
0	0.191	58.0	0.937	22.4
1	0.277	70.2	0.966	19.8
2	0.233	66.5	0.960	17.1
3	0.307	60.5	0.920	20.8
4	0.247	58.4	0.930	23.3
5	0.127	79.0	0.953	24.7
6	0.233	55.9	0.983	21.8

The calm metric. NRD fits Σ_c on calm calibration residuals. Table 15 replaces it by a diagonal, a shrinkage estimator and a scaled identity. The gain is insensitive to the choice, which is what the theory expects: the evidence comes from the drift channel of a residual the calm predictor cannot explain, not from a finely estimated metric.

Grid density. Table 16 runs both detectors under the published grid, a denser grid and a fine grid, as Proposition 1 and Section C.2 discuss. Adding alternatives near one lowers the threshold to 1.0484 and 1.0050, well below the measured inflation, and frozen detection still does not improve.

Table 15: NRD under four calm metrics (SWIRL, JEPA latent, 7 seeds).

calm metric Σ_c	detection	delay
full covariance (ours)	0.950	21.4
diagonal	0.947	29.0
shrinkage toward the identity	0.949	21.9
scaled identity	0.950	30.2

Table 16: Variance grids: threshold and detection (SWIRL, 7 seeds). Every grid keeps the pure-drift alternative $g = 1$.

			JEPA		
grid	points	$\tau_{\mathcal{G}}$	frozen	NRD	Var-JEPA frozen
published $\{1, 1.5, 2.5, 4\}$	4	1.2164	0.231	0.950	0.074
denser	8	1.0484	0.224	0.950	0.078
fine	16	1.0050	0.221	0.950	0.078

Table 17: The evidence ceiling of Proposition 1 evaluated per seed on SWIRL (Var-JEPA, `final_swirl7.json`). The rate is per step; the ceiling is the total over the 150-step post-change window, against $\log(1/\alpha) \approx 3.00$ and a restart price of 0.0513 per step.

seed	$\bar{g} = R_{\mathrm{lin}}$	$d_{\mathrm{IS}}(\bar{g})$	rate per step	ceiling (nats)
0	1.1249	0.00721	0.0288	4.32
1	1.0520	0.00131	0.0052	0.78
2	1.0781	0.00290	0.0116	1.74
3	1.1042	0.00508	0.0203	3.05
4	1.0961	0.00434	0.0174	2.60
5	1.0982	0.00452	0.0181	2.71
6	1.0683	0.00223	0.0089	1.34

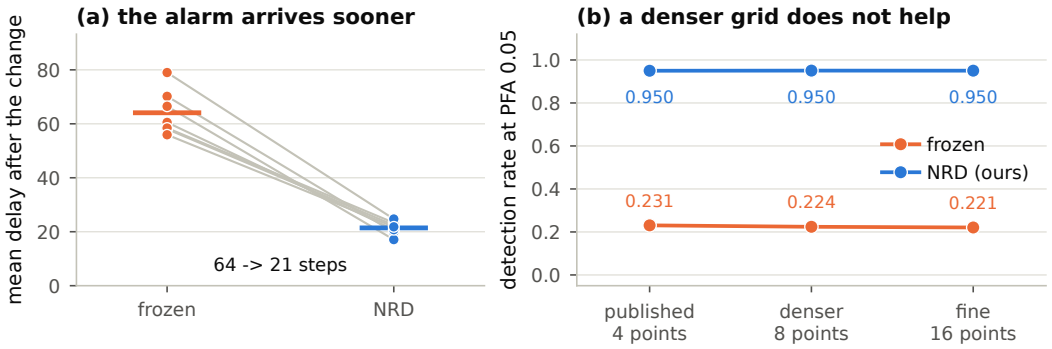

Figure 12: **Speed, and what a denser grid buys.** SWIRL, JEPA latent, 7 seeds, at achieved PFA 0.05. (a) Mean delay after the change, paired by seed. (b) Detection under three variance grids; the threshold falls from 1.2164 to 1.0050, and frozen detection does not rise.

The standardization family. Theorem 3 says that dividing the residual by the head's own scale anchors the calm energy at one. Table 18 separates the anchor from the head. Standardizing and then scoring against a fixed metric is the trap the design rule forbids (0.109). Standardizing and then re-fitting the calm metric of the standardized residual removes the anchor and recovers detection (0.878), and rescaling by the calm-average scale behaves like NRD itself (0.837 against 0.837), as it must, since a constant rescaling is absorbed by the calm metric. The practical rule is therefore sharper than "never standardize": never score an anchored residual against a metric that was not fitted on it.

Table 18: Scoring a Var-JEPA residual with the model's own predictive scale (SWIRL, 7 seeds, this sweep).

scoring	detection
NRD (calm-fitted Σ_c, no standardization)	0.837
per-state $\hat{\sigma}$, fixed metric (NRD-$\hat{\sigma}$)	0.109
per-state $\hat{\sigma}$, calm metric re-fitted	0.878
calm-average $\hat{\sigma}$	0.837

H.5 NONSTATIONARY CALM WINDOWS

The calm-fitted metric assumes the pre-change residuals are stationary. We perturb the evaluation stream of the same trained models with two effects a deployment really has, a within-episode cycle and a slow gain drift, and separate two deployments. In the *matched* deployment the calibration episodes carry the same nonstationarity, so thresholds see it; in the *mismatched* deployment thresholds come from clean calm episodes while the live stream drifts. Table 19 reports both. The matched columns say whether the advantage survives; the mismatched column says what an unmodelled cycle costs in false alarms, where the nominal level is 0.05. NRD is insensitive to both perturbations (0.954 when the stream is stationary, against 0.957 under a 10% cycle and 0.952 under a 10% gain drift), and stationary thresholds stay valid, with an achieved false-alarm rate of at most 0.060 on any seed and condition. The frozen detector moves more, in both directions (0.226 and 0.314 against 0.243), because the perturbation itself shifts the energy its variance channel measures.

Table 19: Nonstationary calm windows (SWIRL, JEPA latent, 3 seeds, models trained on stationary calm data).

	matched calibration		stationary thresholds	
calm window	frozen	NRD	achieved PFA	NRD detection
stationary	0.243	0.954	0.050	0.953
cycle, 5%	0.251	0.953	0.048	0.953
cycle, 10%	0.226	0.957	0.037	0.962
gain drift, 10%	0.314	0.952	0.002	0.991

H.6 FORECAST-THEN-DETECT BASELINES

The mainstream alternative to scoring a representation is to fit a forecaster on calm data and monitor its residuals. We fit LSTM one-step forecasters with the same budget, on the raw 56-dimensional observations and post hoc on the frozen JEPA latent, and score them with the same e-process, the same calm-fitted covariance and the same achieved-PFA calibration. Neither forecaster sees the change: both are fitted on calm training episodes only. Table 20 reports the result against the co-adapted predictor of the same encoder.

Table 20: Forecast-then-detect on SWIRL, with training-only calibration. The LSTM rows use 3 seeds, the frozen and NRD rows the 7 of Section H.4. Delays are in steps after the change.

detector	residual of	detection	delay
frozen (linear VAR null)	the frozen latent	0.231	64.1
LSTM, raw observations	the observations	0.047	73.3
LSTM, post hoc on the latent	the frozen latent	0.157	70.6
NRD (ours)	the frozen latent	0.950	21.4

H.7 VAR-JEPA IS A DIFFERENT DRAW ON EVERY RE-RUN

Var-JEPA's objective samples latent variables during training, and those draws come from the accelerator's generator. A re-run under identical seeds and settings therefore lands on a different model, while the deterministic JEPA arm reproduces bit for bit. Table 1 and the ablation sweep of Section H.4 agree exactly for JEPA (frozen 0.231, NRD 0.950) and disagree for Var-JEPA (NRD 0.366 against 0.837).

The disagreement is not noise in the measurement, and the three quantities separate cleanly. The frozen view is reproducible: detection 0.090 against 0.074, and an energy inflation of $R_{\text{lin}} = 1.089$ against 1.085, so Theorem 1 applies to both runs in the same way and the channel is closed in both. What moves is absorption: $R_{\text{raw}} = 0.959$ in the published run, below one on every seed, against 1.011 in the replication, below one on 3 of seven seeds. Theorem 2 says exactly this: absorption is the alignment $\mathbb{E}_Q\langle c, \delta\rangle$ between the regime shift and *that* model's correction, which is a property of the trained model and not of the architecture. A predictor that absorbs less leaves more in its residual, and the native detection rate follows: 0.837 where the published run had 0.366.

We therefore read Var-JEPA's detection rate as a distribution over training runs rather than as a constant. The claims that matter do not depend on where in that distribution a run lands: the frozen detector stays closed because $R_{\text{lin}} < \tau_{\mathcal{G}}$ in both runs, NRD-$\hat{\sigma}$ stays at zero because the anchor of Theorem 3 does not depend on the draw, and the deterministic JEPA arm, which carries the near-oracle result, reproduces exactly.

I ABLATIONS AND CONTROLS

1. **NRD-$\hat{\sigma}$ gives** 0.000 wherever it matters. The $\hat{\sigma}$-standardized residual scored against a fixed metric is the trap that the design rule forbids (Theorem 3); re-fitting the calm metric of that same standardized residual lifts detection to 0.878 (Section H.4).

2. **Post-hoc nulls are not enough.** The calm-fitted MLP null gives 0.194 on the JEPA latent, while the co-adapted head gives 0.950.
3. **Selectivity.** On SWIRL-LIN NRD detection is 0.002 while frozen is 0.963. The NRD advantage exists exactly where the change is *not* expressible by the calm predictor, and it vanishes where it is.
4. **PatchTST counterexample.** Mean-pooled patch codes destroy one-step temporal structure, and NRD detection falls to 0.009. This fixes the recipe's precondition.
5. **Transformer budget.** At 150 episodes and 2000 steps NRD detection is unstable (0.048-0.560). At 400 episodes, 4000 steps and learning rate 5×10^{-4} it reaches 0.777 (sd 0.169): backbone-invariant but budget-limited.
6. **Grid threshold.** g^* never leaves 1 on the swirl, where $R_{\text{lin}} = 1.089 < \tau_{\mathcal{G}}$. On ROTOR the channel opens with $\gamma_{\text{eff}} \in [2.42, 4.56]$ and g^* moves to 1.5-4.

J EXTENDED LIMITATIONS

Implementation fidelity and baselines. TimesNet and PatchTST are our in-context implementations under a shared JEPA objective, not the published stacks. The zero-shot foundation model is scored like every other detector, on the learned latent; running it as a forecaster on raw observations and adding further pretrained families (Ansari et al., 2024; Verdenius et al., 2024) is future work. **Theory.** Theorems 2 and 3 are exact under their hypotheses, and we do not characterize moderate violations of (M) or (O) beyond the one case where they happen: SWIRL-LIN violates marginal matching, so its deflation is reported as an empirical fact. Theorem 1 and Proposition 1 govern evidence growth, not the power of a particular calibrated threshold, and the diagnostic grid argmax is a standardized contrast consistent with but not implied by Theorem 1 (Section C.1). All claims are alignment claims under the data-processing inequality. **Provenance.** Three quantities (NRD-$\hat{\sigma}$, the Var-JEPA variants and the TimesFM run) come from experiment logs rather than machine-dumped files.

K EXTENDED DISCUSSION

Research questions and answers. *Q1 (bottleneck):* the frozen detector. The same latent that fails under a frozen linear VAR reaches the exact-likelihood ceiling under its own co-adapted predictor with a calm-fitted covariance. *Q2 (variance head):* Var-JEPA opens the variance channel against a linear null, the first learned representation to do so, and closes it against its own view (Theorem 2 and Theorem 3). *Q3 (when NRD reaches the oracle):* when the predictor is the one the representation was trained against, the change is not expressible by that predictor, and the encoder preserves one-step temporal order. It is blind otherwise.

Why recent anomaly-detection pipelines are compared at the detector level. Every representation must satisfy one contract: map a window to an 8-dimensional latent trained on calm data only, and be scored by the same frozen detector. The contract separates representation quality from detector choice. Recent time-series anomaly detectors are pipelines rather than representations. The Anomaly Transformer outputs an association-discrepancy score (Xu et al., 2022). Deep SVDD learns a hypersphere boundary (Ruff et al., 2018). VACE, RoCA and ProtoX-AD output geometric, contrastive one-class or prototype-based normality scores (Cencillo et al., 2026; Mou et al., 2025; Sánchez-Ferrera et al., 2026). Stripping their heads to keep an encoder would mean re-implementing each encoder, and any gap would be attributed to the re-implementation; keeping the heads breaks the contract. We therefore compare at the right level. Detector families (energy, one-class, kernel change point, e-CUSUM, marginal) run on the same latents. Backbones (causal transformer, PatchTST, TimesNet-style) run under the same objective. A zero-shot foundation model and real data complete the picture.