FROM LEAST ACTION TO LATENT FORCING TESTS FOR ROOT-CAUSE ATTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Root-cause attribution in latent dynamical systems asks which hidden process was perturbed and when. We study sparse additive forcing in latent stochastic differential equations using the classical innovation likelihood-ratio statistic from model-based fault detection and isolation. On an Euler–Maruyama grid with diagonal diffusion, the diffusion-normalised action evaluated at the profile-maximising latent forcing is exactly the per-atom generalised log-likelihood ratio. The oracle detection laws follow from standard Gaussian testing. Our focus is the identification, reconstruction and reporting conditions needed to use this statistic in learned latent coordinates. We derive a deterministic condition under which representation, drift and diffusion errors preserve the oracle attribution ranking. On synthetic latent-target benchmarks, the resulting one-pass estimator achieves comparable atom accuracy to an iterative solver while being substantially cheaper to evaluate. Physical-system simulations and real-data experiments also expose the importance of observability and the limits of transferring latent forcing attribution to observed-variable diagnosis. These results support latent forcing tests as an efficient and theoretically grounded approach to root-cause attribution when latent coordinates are meaningful and sufficiently observable.

1 INTRODUCTION

A disturbance can enter one process and produce its largest measured response elsewhere. In a coupled tank system, extra inflow into upper Tank 3 changes the level of lower Tank 1 through nominal hydraulic dynamics. Identifying the most abnormal sensor therefore need not identify the entry point. This distinction also motivates incident diagnosis in service systems (Ikram et al., 2022, Han et al., 2025). We study the state and time at which an additive disturbance enters a partially observed dynamical system. Our running example uses calibrated flow proxies to reconstruct upper-tank levels, giving the attribution target a physical meaning (Johansson, 2000). It also makes the central difficulty explicit. A useful attribution needs both a meaningful state coordinate and enough measurement information to reconstruct its motion.

Classical fault detection and isolation provides a natural statistical foundation. Subtracting predicted motion gives an innovation, and covariance weighting measures its departure from nominal variability (Willsky & Jones, 1976, Basseville & Nikiforov, 1993). In a latent stochastic differential equation, this calculation becomes a forcing test in reconstructed coordinates. Predictive accuracy, however, does not identify those coordinates. Representation-learning results require additional structure or supervision to identify latent variables (Hyvärinen et al., 2019, Brehmer et al., 2022, von Kügelgen et al., 2023). Even an identified coordinate may be poorly observed, and reconstruction error can change which disturbance receives the highest score. These issues motivate our question of when an innovation test in a learned representation supports a named root-cause attribution. We assume that the modelled state and known inputs contain the relevant processes. An omitted common cause lies outside this scope.

This separation also guides evaluation. A synthetic generator exposes the true forcing and allows us to isolate score geometry from reconstruction. A physical simulation supplies independently defined state names and fault units. Observed benchmarks test whether the attribution survives projection back to measured variables. We use these settings to examine distinct requirements rather than pooling incompatible notions of root accuracy.

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.

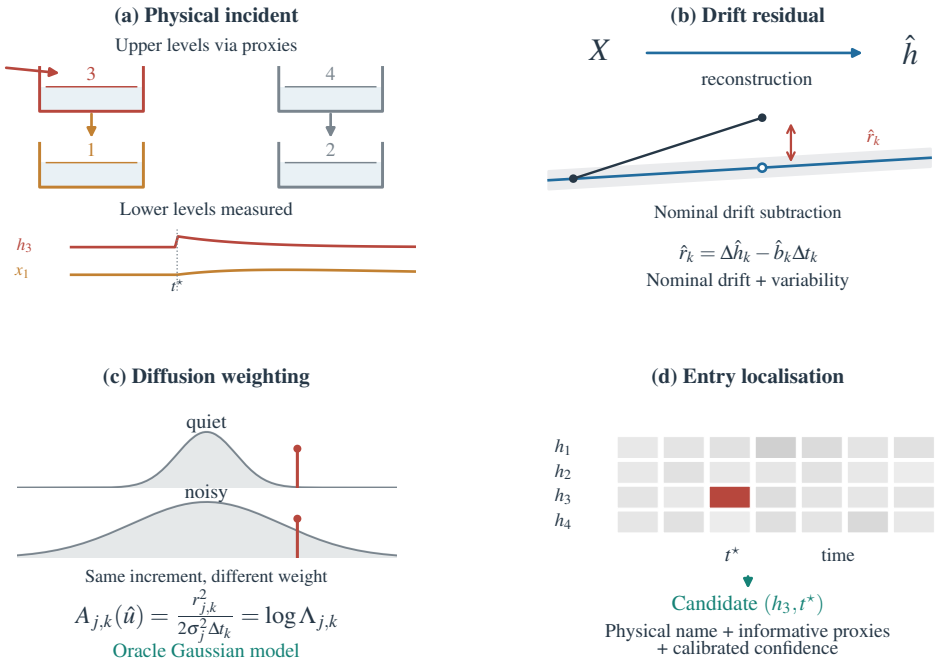

Figure 1: **Latent forcing attribution.** In this schematic tank incident, extra inflow enters indirectly observed Tank 3 and produces a downstream response in measured Tank 1. Reconstructing the state, subtracting nominal motion and weighting by stochastic variability produces a state–time action map. The exact action–likelihood identity concerns the oracle Gaussian model. The learned score and its physical interpretation require separate estimation and reliability checks. Traces and action cells are illustrative, not measured trajectories.

We introduce LARCA (Least-Action Root-Cause Analysis) to make these requirements explicit. For diagonal Gaussian transitions, diffusion-weighted action at the profile forcing equals the classical generalised log-likelihood ratio. We retain this established test and derive a deterministic condition under which errors in state reconstruction, drift and diffusion preserve its oracle ranking (Theorem 1). External anchors fix the coordinate meanings, while a procedure using disjoint data bounds errors among reported roots (Theorem 2). Experiments separately test noise geometry, observability and physical semantics. Our one-pass score matches the iterative solver's average atom accuracy at much lower scoring cost in the internal ablation. Weak proxies, a stronger filter-based comparator and losses to AERCA on observed-variable benchmarks delimit this result. Together, the analysis and experiments identify when latent forcing attribution is justified and where additional reconstruction or calibration is necessary.

2 Related Work

Innovation likelihood-ratio tests, covariance-weighted residuals and observer-based diagnosis are established tools of fault detection and isolation (Willsky & Jones, 1976, Basseville & Nikiforov, 1993, Gertler, 1998, Isermann, 2006). Our known-state singleton score belongs to this family. We study the additional identification, approximation and reporting conditions needed in learned latent coordinates. Causal representation learning establishes identifiability from auxiliary information or interventions (Hyvärinen et al., 2019, Brehmer et al., 2022, von Kügelgen et al., 2023). We assume external anchors and analyse the stability of incident rankings. AERCA and RCD instead diagnose interventions in observed-variable systems (Han et al., 2025, Ikram et al., 2022). This difference in targets motivates separate latent and observed-variable evaluations, including the latter's negative results. Appendix A develops these comparisons, the connection to path-space control, and their implications for evaluation.

3 LATENT DYNAMICS AND ATTRIBUTION

Latent dynamics. In the running example, the state contains all four tank levels. Lower levels are measured directly and upper levels are reconstructed from calibrated flow proxies. A state can therefore have a physical name while remaining unmeasured. The nominal model conditions on known operating inputs. Its state must contain the relevant dynamics. The analysis does not identify interventions in the presence of an additional, omitted common-cause process.

We observe $X_{t_k} \in \mathbb{R}^p$ at times $0 = t_0 < \cdots < t_K = T$, with $\Delta t_k = t_{k+1} - t_k$. A nominal latent SDE and decoder describe the observations,

$$dZ_t = b_\theta(Z_t, t)\, dt + \Sigma_\theta^{1/2}(Z_t, t)\, dW_t, \qquad X_{t_k} \sim p_\theta(x_{t_k} \mid Z_{t_k}).$$

The drift b specifies expected motion. The diffusion covariance Σ specifies the local uncertainty around that motion. An incident adds a progressively measurable forcing u to the drift, changing the nominal path law $\mathbb{P}^{\mathrm{nom}}$ to $\mathbb{P}^u$. We assume that the forcing is supported on a small set of coordinate–time *atoms*[1] $(j, k) \in [d] \times \{0, \ldots, K-1\}$. The target is this support. Forcing describes extra latent drift relative to the nominal model. Its interpretation as a physical intervention requires external state semantics. In the tank example, an atom is one tank during one sampling interval. A short inflow into Tank 3 occupies one such cell. The later movement of Tank 1 need not represent another intervention.

On the Euler–Maruyama grid, with coefficients evaluated at the start of each step,

$$Z_{k+1} = Z_k + (b_k + u_k)\Delta t_k + \Sigma^{1/2}\sqrt{\Delta t_k}\,\varepsilon_k, \quad \varepsilon_k \sim \mathcal{N}(0, I), \qquad r_k \coloneqq Z_{k+1} - Z_k - b_k \Delta t_k. \tag{1}$$

We suppress state and time indices on Σ and use constant diagonal diffusion in deployment. The *excess increment* r_k is the movement left unexplained by the drift. Under the null, it is conditionally Gaussian with covariance $\Sigma \Delta t_k$, making Σ^{-1} the metric for measuring departures. Throughout, σ_j^2 denotes the diffusion *variance* in coordinate j.

Propagation and forcing. Suppose an extra inflow enters Tank 3 during interval $k^\star$. The nominal drift already describes how water leaving Tank 3 changes Tank 1. Once the state has moved, that same drift predicts the ensuing downstream evolution from the current state. Under the specified transition model,

$$\mathbb{E}[r_{j,k} \mid Z_k, u_k] = u_{j,k} \Delta t_k.$$

The entry atom therefore has a non-zero mean excess increment, while an unforced downstream atom retains mean zero even when its level is far from the earlier operating point. Process noise still produces random excess increments at both locations. Attribution asks which departure is strongest relative to that noise. A reconstruction error or an omitted dynamical input can also leave unexplained motion, which is why the representation assumptions matter.

Forcing action. Under standard Girsanov conditions, the path-KL cost of adding drift is the expected diffusion-weighted forcing energy (Appendix D). Its discrete pathwise counterpart is $A(u) = \frac{1}{2}\sum_k \Delta t_k\, u_k^\top \Sigma^{-1} u_k$. A direction with high nominal noise carries less action for the same forcing magnitude. With diagonal diffusion this cost separates into atoms,

$$A_{j,k}(u) = \frac{\Delta t_k\, u_{j,k}^2}{2\sigma_j^2}, \qquad (j^\star, k^\star) = \arg\max_{j,k} A_{j,k}(u). \tag{2}$$

Section 4 derives the forcing estimate at which this action becomes statistical evidence. The total continuous-time action is invariant under smooth invertible reparameterisations, and the discrete quadratic under consistent linear changes of basis (Proposition 5). Assigning that action to named coordinates requires additional structure.

Coordinate identification. An encoder can mix the four levels and still predict the same sensor histories if the decoder compensates. A pulse on h_3 then spreads across the new coordinates. Only an external correspondence between a coordinate and the tank level makes the name meaningful. Figure 2 distinguishes this identification issue from weak measurement information.

[1] An atom is one candidate state coordinate during one sampling interval.

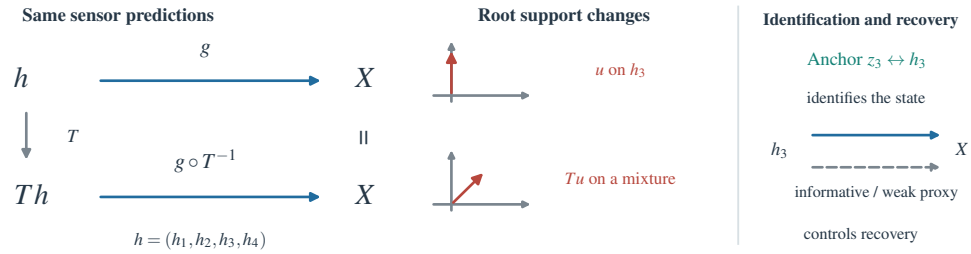

Figure 2: **Coordinate identification and observability.** Mixing the tank levels and compensating in the decoder preserves sensor predictions but changes forcing support. Calibration can identify a coordinate with Tank 3. The strength of its proxy then determines how well the measurements reveal that level. All four tanks remain inside the modelled state in either case.

An invertible change of latent basis, compensated in the decoder, preserves the observed law while changing the support of the forcing. Proposition 2 formalises this ambiguity in Appendix B.

Proposition 2 separates predictive adequacy from attribution semantics. Even a correctly specified predictive law cannot, by itself, attach a process name to a latent coordinate. LARCA therefore distinguishes an observed shock, a latent action set or equivalence class, an anchored coordinate, and a named root with action responsibility. External domain metadata supplies the anchoring status. Adequate observability and calibrated confidence are further requirements for the strongest claim. These are conditions on the model and measurements, not a test for omitted confounding. Raw atom scores are evaluated in a declared common basis. They alone do not justify a named attribution (Section 4.4).

4 Latent Forcing Tests

4.1 Local Forcing Tests

We hold the reconstructed path fixed and compare a local forcing alternative with nominal motion. Profiling chooses the amplitude that maximises the transition likelihood.[2] For a Gaussian transition, the maximiser explains the excess increment exactly.

Fix an atom (j,k). Conditional on the latent state and nominal coefficients, test $H_1 : u_{j,k} \neq 0$ against $H_0 : u_{j,k} = 0$, with all other atoms held at zero. The excess increment in (1) has mean $u_{j,k}\Delta t_k$ and variance $\sigma_j^2 \Delta t_k$. Profiling this local alternative connects the dynamical action to a statistical test.

Proposition 1 (The coordinate action is the profile log-likelihood ratio)**.** *Let Σ be diagonal with entries σ_j^2. For the single-atom test above, the profile maximiser is $\widehat{u}_{j,k} = r_{j,k}/\Delta t_k$, and*

$$\log \Lambda_{j,k} \coloneqq \log \frac{p(r_{j,k} \mid \widehat{u}_{j,k})}{p(r_{j,k} \mid 0)} = \frac{r_{j,k}^2}{2\sigma_j^2 \Delta t_k} = \frac{\Delta t_k\, \widehat{u}_{j,k}^2}{2\sigma_j^2} = A_{j,k}(\widehat{u}). \tag{3}$$

The log-density is $-(r_{j,k} - u_{j,k}\Delta t_k)^2/(2\sigma_j^2 \Delta t_k)$ up to a constant. Its quadratic vanishes at the profile maximiser, and subtracting the null log-density gives (3). The full proof is in Appendix B.

The equality holds at the likelihood-maximising forcing. Minimising $A(u)$ alone would instead return zero. A regularised path-space control objective also solves a different problem. Here the local alternative determines the forcing estimate, and its action measures the evidence against nominal motion. This gives a closed-form score, a noise metric fixed by the nominal covariance, and an oracle null law.

[2] A profile likelihood maximises over a nuisance parameter, here the forcing amplitude, separately for each candidate entry atom.

Correlated diffusion. For general positive-definite Σ, the singleton log-likelihood ratio is

$$\log \Lambda_{j,k} = \frac{(e_j^\top \Sigma^{-1} r_k)^2}{2\Delta t_k \, e_j^\top \Sigma^{-1} e_j}.$$

It reduces to (3) when Σ is diagonal. The full-vector identity continues to hold, but a coordinate test must constrain the alternative before profiling. The support-wise form is given in Appendix B. We deploy the diagonal case.

4.2 One-pass attribution

The incident provides sensor measurements, not the true state used in Proposition 1. In the tank example, reconstruction combines measured lower levels with calibrated upper-tank flow proxies to estimate all four levels. The nominal model then predicts the movement from each reconstructed state. Subtracting that prediction leaves an estimated forcing residual in the same physical coordinates used to name the root.

The model is trained on fault-free windows. Given an incident, the encoder $q_\phi(Z_{0:K} \mid X_{0:K})$ supplies a latent path $\widehat{Z}$, from which we compute

$$\begin{aligned} \widehat{r}_k &= \widehat{Z}_{k+1} - \widehat{Z}_k - \widehat{b}(\widehat{Z}_k)\Delta t_k, \\ \widehat{A}_{j,k} &= \frac{\widehat{r}_{j,k}^2}{2\widehat{\sigma}_j^2 \Delta t_k}, \qquad \widehat{m} = \arg\max_{j,k} \widehat{A}_{j,k}. \end{aligned} \tag{4}$$

Encoding, drift subtraction, diffusion normalisation and ranking form one pass over the window. Once the latent path and nominal drift have been evaluated, forming and scanning the action map costs $O(Kd)$ and requires zero optimiser steps. Encoder and drift evaluation add their own computational costs. Every atom is scored, with no temporal mask, path smoother or gradient evaluation (Appendix C). A score-selected pulse window is a separate ablation.

Because the path and nominal coefficients are estimated, $\widehat{A}$ is a *plug-in pseudo-GLRT*. Its ranking and its calibration need separate checks. A named-root report additionally requires identified coordinates and the selective procedure below.

4.3 Ranking stability

Suppose the true-state score ranks Tank 3 at the injection time above every other cell. Approximation errors move both that score and its nearest competitor. If each moves by at most ε_A, a gap larger than $2\varepsilon_A$ preserves their order. The following theorem gives the error budget in a common, anchored coordinate system on a uniform sampling grid.

Three approximations can move a score. Reconstruction changes the increments, drift error changes the movement subtracted from them, and diffusion error changes their statistical weight. A persistent reconstruction offset cancels from a consecutive difference, although it can still affect the drift evaluation. The theorem therefore tracks increment error separately from state error.

Theorem 1 (Plug-in ranking stability)**.** *Let r be the oracle excess increment and $\widehat{r} = r + e$ the deployed one, and let $\widehat{\sigma}_j^2 = \sigma_j^2(1+\delta_j)$. Assume*

- *(i) encoder error $\max_k \|\widehat{Z}_k - Z_k\|_\infty \le \varepsilon_Z$, with increment error $\max_k \|(\widehat{Z}_{k+1} - Z_{k+1}) - (\widehat{Z}_k - Z_k)\|_\infty \le \varepsilon_\Delta$ ($\le 2\varepsilon_Z$).*
- *(ii) a Lipschitz drift estimate, $\|\widehat{b}(\widehat{z}) - b(z)\|_\infty \le L_b \|\widehat{z} - z\|_\infty + \varepsilon_b$.*
- *(iii) diffusion error $\max_j |\delta_j| \le \rho < 1$, and*
- *(iv) bounded oracle excess increments $\max_{j,k} |r_{j,k}| \le R$.*

Then

$$\begin{aligned} \max_{j,k} |e_{j,k}| &\le \varepsilon_r := \varepsilon_\Delta + (L_b \varepsilon_Z + \varepsilon_b)\Delta t, \\ \max_{j,k} |\widehat{A}_{j,k} - A_{j,k}| &\le \varepsilon_A := \frac{2R\varepsilon_r + \varepsilon_r^2 + \rho R^2}{2\sigma_{\min}^2 (1-\rho)\Delta t}. \end{aligned} \tag{5}$$

If the oracle action gap satisfies $\Delta^\star := A_{m^\star} - \max_{m \neq m^\star} A_m > 2\varepsilon_A$, *then* $\arg\max_m \widehat{A}_m = \arg\max_m A_m = m^\star$.

The bound is sufficient and deterministic. Its constants must bound the scored path. Nominal quantiles are not uniform bounds on future incidents. Positive diagonal rescaling preserves each action when diffusion is transformed consistently. A dense rotation can change the attribution target and cannot align one side of the comparison alone. Agreement with the oracle also does not imply that the oracle identifies the planted root. Section 5.1 evaluates the bound as a finite-path diagnostic. It provides no exact null law for encoded scores.

Oracle calibration. Under the nominal Gaussian model, $2A_{j,k} \sim \chi^2_1$. A pulse with level increment a gives non-centrality $a^2/(\sigma_j^2 \Delta t)$, and a Bonferroni threshold controls the family-wise error across dK atoms. These laws concern true states and known coefficients. Detection, exact support recovery and top-1 ranking are distinct events. Appendix B gives the laws, Proposition 3 and the event definitions in Table 2.

4.4 Selective reporting

Returning to Tank 3, three questions remain after finding the largest action. Does the coordinate identify that physical tank? Do the measurements contain enough information to reconstruct its motion? Among incidents receiving such a report, how often is the named root wrong? Anchors address the first question, observability and approximation checks address the second, and held-out selective calibration addresses the third.

The observed-residual layer LARCA-Obs (LARCA in observed coordinates) compares calibrated one-step sensor residuals with the latent action map. For a linear decoder, $K^{r,\tau}_{i,k} = |C_i e_r|^2 \delta_{k\tau}$ maps an entry atom to contemporaneous residual evidence, with per-sensor scale and intercept fitted on calibration incidents. This compares residual evidence after drift subtraction, not the delayed response of raw sensor levels. Alignment determines eligibility for action-gap calibration rather than a per-incident veto. Definitions, calibration and ablations are in Appendix H.6.

For an eligible, anchored representation, a named-root claim is reported when a label-free test-time confidence κ exceeds a fixed threshold $\widehat{\tau}$. A labelled *development* split selects the confidence family, a disjoint *selection* split fixes the threshold for target coverage, and a disjoint i.i.d. *validation* split measures accepted errors. If m accepted validation incidents contain e errors, Theorem 2 (Appendix B) bounds the accepted-error rate by the Clopper–Pearson upper bound $U(e, m, \delta)$ with probability at least $1-\delta$. Below the threshold, the method withholds the specific root and reports the weaker supported result. This is selective prediction, with coverage denoting the fraction of incidents receiving a named attribution.[3] The confidence and threshold must be fixed before validation. The guarantee uses one split and the same incident distribution at deployment. It does not follow from the oracle χ^2 law or from pooling reused incidents across random splits.

This distinction matters for deployment. An action score ranks candidate entries within a model, whereas the reporting rule decides whether that ranking supports a specific claim at the chosen coverage. We validate the latter decision using labelled incidents, including its errors, after fixing both the confidence family and threshold. A representation may therefore supply useful rankings yet fail to support reliable selective reports. Conversely, low error at very low coverage says little about the incidents for which the method abstains. We report both quantities and retain unsuccessful calibration results in the evaluation.

5 Experiments

We test three consequences of the proposed structure. Diffusion normalisation should matter when coordinate noise levels differ. Learned attribution should deteriorate when observations conceal an intervention even if the oracle remains accurate. Physical anchors should let the reported coordinate name a process defined outside the learned model. Figure 3 presents these tests.

[3] Selective risk is the error rate conditional on issuing a report. Reducing coverage can improve this rate, but does not guarantee that a confidence score separates correct and incorrect roots.

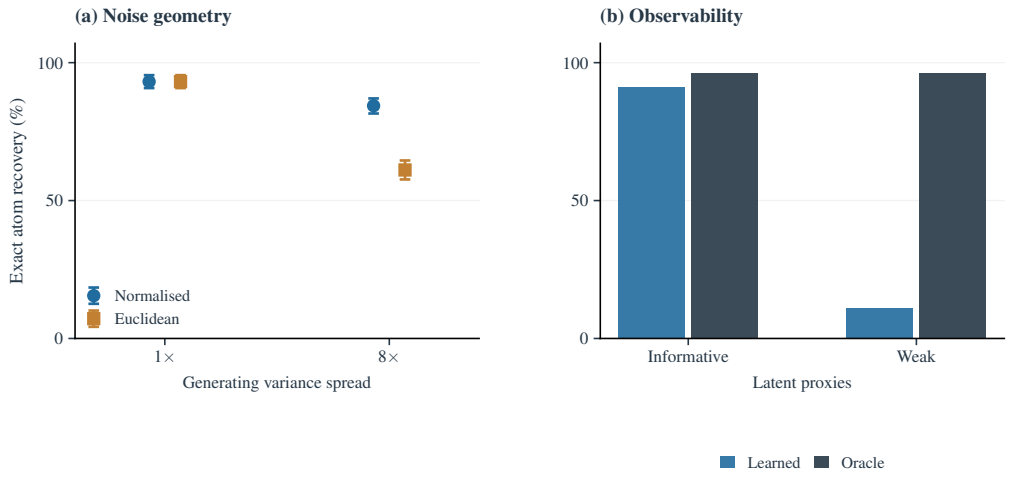

Figure 3: **Noise geometry and observability.** (a) Matched one-pass scores tie at unit generating variance spread and separate at eightfold spread. Points show seed means with 95% seed-bootstrap intervals at the common reference amplitude $a_0 = 0.728$. The conditions are distinct regimes. (b) The synthetic oracle remains strong when a decoder column is weakened, while learned atom recovery collapses. These are regime means from Table 9. The physical fault-scale study is shown separately in Fig. 4.

The latent-target suite uses ten regimes and ten seeds, with 32 incidents per regime–seed cell. Matched arms share the nominal model and observations. The oracle GLRT reads the generating path. AC*@1 measures exact coordinate–time recovery in a declared basis. Time-Acc measures onset recovery. Named-coordinate accuracy additionally requires anchored labels. Missing metric entries determine the common regimes used for each aggregate and are documented in Appendix H.1. A separate matched Gaussian-transition study tests threshold events. Regimes, tuning budgets and scan protocols are in Appendix H.1, Appendix H.8 and Appendix J.

5.1 Noise geometry and observability

At unit generating variance spread, the matched normalised and Euclidean one-pass scores tie at 93.1% atom accuracy. At eightfold spread, normalisation improves accuracy by +23.3 percentage points (84.4% against 61.1%). Figure 3a therefore tests the proposed metric where it can affect the ranking, alongside a homogeneous control. The separate solver-based spread study is retained in the archival supplement.

At $d = 6$, $K = 64$, $\Delta t = 1/64$ and $\sigma^2 = 1.3$, the predicted 90% detection amplitudes are 0.728 for an atom and 0.798 for onset. Gaussian simulations attain the specified scan error and power, while top-1 ranking and exact support have different success rates. Pooling heterogeneous variances miscalibrates null coordinates and reduces detection power. Full operating characteristics and learned-path comparisons are in Appendix B.2.

Numerical likelihood maximisation and oracle null checks agree with the derived identity and laws. With correlated diffusion, coordinate attribution must use the constrained test (Appendix G).

In the weak-observability synthetic regime, shrinking a decoder column conceals a latent intervention from the observations. The oracle still attains 96.2% exact-atom accuracy, while no learned arm exceeds 10.9% (Fig. 3b). External coordinate names alone do not repair missing observational information.

We report finite-path diagnostics and encoded-score null checks in the appendix. These checks do not establish a deployable error bound or an exact χ^2 law for learned scores (Appendix J,Appendix G.1).

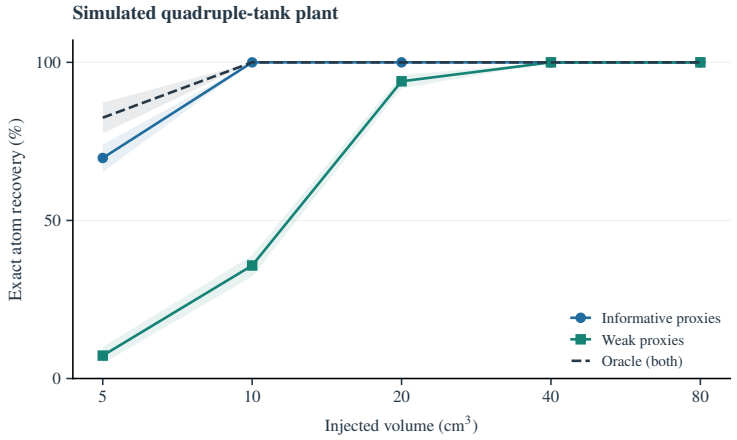

Figure 4: **Physical fault scale in the simulated plant.** Roots name upper tanks and fault size is injected volume. The learned score approaches the true-state oracle with informative local flow proxies. Weakened proxies require larger faults. Shading shows 95% seed-bootstrap intervals. The oracle curves coincide because only the observation map changes. This task and its physical units differ from the synthetic tests in Fig. 3.

Table 1: **Internal solver ablation on the latent target.** Ten regimes × ten seeds, 3200 incidents per arm. Atom and onset accuracies are proportions, averaged over nine and ten common regimes respectively. Optimiser steps are those actually executed. Time is median per-incident scoring time with a fitted model (Appendix H.5). The oracle reads true latents and is a reference rather than a feasible method. This comparison measures the cost of the authors' iterative objective, not an efficiency advantage over filter-based FDI. Bold marks the best feasible value in each performance column, including ties. We exclude the oracle from this comparison. Full ablations are in Table 6.

Method	Steps	Atom @1	Onset	Time (ms)
One-pass forcing test	0	**0.7948**	0.8119	**0.43**
Prepass	50	0.7924	**0.8478**	74.97
Prepass and iterative solver	550	0.7920	0.8475	983.76
Oracle dynamics and latents	0	0.9271	0.8931	–

5.2 Physical attribution

The quadruple-tank simulation uses published mass balances and dimensions (Johansson, 2000). Lower tank levels are measured. Upper levels are reconstructed from calibrated flow proxies. With informative proxies, the deployed score reaches the oracle ceiling by 10 cm^3. Weak proxies reduce recovery to 36% while the oracle remains at 100% (Fig. 4). Across configurations, the prespecified 20 cm^3 endpoint is 97–100%. The Euclidean score is similar, so this simulation supports observability and physical semantics, not a diffusion- weighting advantage. A held-out pump shift increases the deployed false-alarm rate from 6.1% to 10.9%. The matched-calibration guarantee does not cover the shift. Full protocols and results are in Appendix J.

A separate exploratory comparison adds a known-physics EKF innovation GLR and a fixed-duration forcing scan (Appendix L). For an eight-step, 20 cm^3 fault with informative proxies, the pointwise duration scan improves exact onset-atom recovery from 3.3% to 72.5%. The EKF scan reaches 79.2%. With weak proxies the EKF remains stronger. Removing upper proxies leaves exact timing poor even with filtering and smoothing. These results support duration modelling while leaving indirect reconstruction open.

5.3 SOLVER ABLATION

The deployed score reaches 0.7948 atom accuracy in 0.43 ms, compared with 0.7920 in 983.8 ms for the 550-step solver (Table 1). The 50-step prepass improves onset accuracy by $+0.0359$ over the one-pass score, so it remains useful when timing is the primary objective.

We average paired differences within regime–seed cells and bootstrap seed blocks jointly across regimes. The mean one-pass minus solver difference is 0.28 percentage points, with a 90% interval of $[0.07, 0.49]$, within a ± 1-point equivalence margin. The analysis uses 2880 paired incidents and 10 seed clusters. The margin was selected after the aggregate results were known, so this analysis of existing runs is not a comparison with a margin fixed in advance. It supports comparable average atom accuracy at these operating points, with finite-cluster uncertainty (Appendix J).

5.4 OBSERVED TARGETS AND SELECTIVE REPORTING

SWaT and MSDS ask which observed variable or event is implicated over an incident. On SWaT, deployed-head AC@1 is 3.7%, below AERCA's 20.2% and the bridge-residual ablation's 13.0%. Chance is 2.2%. The residual ablation leads AERCA only at looser retrieval budgets, where chance also rises. The two paired intervals excluding zero were selected after inspecting seven operating points and are uncorrected. Across seeds, the tight-budget deficit is unstable. Curves and full operating-point results are in Appendix H.2.

MSDS Recall@3 is 15.2% against AERCA's 41.2%, scored on the 40 of 50 trials where AERCA returned output. AERCA diverged on the other 10 trials. Our method returned output on all trials. Tables 11 and 14 report the common-model head ablations and failure accounting. The pulse mask is disabled for these sustained incidents. These results limit transfer from local latent forcing tests to observed-variable diagnosis. Further contemporary baselines need a protocol-compatible evaluation (Appendix H.2).

Our five-seed matched MSDS pilot does not support duration as a sufficient remedy. Under observation-availability alignment, singleton Recall@3 is 16.0% versus 5.1% for the fixed-family duration scan (Appendix L). The representation and observed-variable projection remain substantial limitations.

For the synthetic result, the confidence family is selected on disjoint development data. Synthetic atom accuracy rises from 53.9% at full coverage to 97.7% at 40% coverage. On the predeclared three-way split, Theorem 2 gives a named-root error upper bound of 10.9% at 47% realised coverage. The analogous SWaT result is loose and provides validation evidence only. Its confidence-family choice used validation labels, so it does not satisfy Theorem 2. On MSDS, none of the evaluated confidence scores usefully separates reliable roots (Appendix H.7). A high forcing score therefore does not by itself supply a reliable named-root claim.

6 DISCUSSION AND LIMITATIONS

We connect classical innovation testing to root-cause attribution in learned latent dynamics. The action identity supplies a one-pass score, while coordinate identification, ranking stability and selective calibration determine what we can claim from it. Our experiments support the roles of noise geometry and observability and show comparable average atom accuracy to our iterative solver at lower scoring cost. The equivalence margin was chosen after inspecting the existing aggregate results, so an independent comparison with a margin fixed in advance remains necessary.

Our physical evidence comes from simulation with calibrated local proxies. The known-physics EKF performs better with weak proxies, and exact timing remains poor without upper-tank proxies. We also lose to AERCA on observed-variable benchmarks. Duration modelling alone does not close the MSDS deficit. Omitted common causes fall outside our model, and operating-point shifts can invalidate calibration. The executed scale study reaches 4096 states and 16,777,216 candidate atoms, but increases fault strength to target fixed oracle power (Appendix K). It therefore establishes computational reach under that schedule, not fixed-fault robustness. Better reconstruction from indirect measurements, validation under distribution shift and field evaluation are the next requirements for reliable deployment.

Ethics statement

A named root can encourage action on a coordinate whose meaning or reliability the data do not support. We therefore require external anchoring, distinguish weaker claims and report held-out selective error bounds. The weak-action-gap stress case attains only 68.8% named-root precision despite satisfying every reporting condition (Appendix H.3), illustrating the remaining risk. The SWaT evaluation uses released testbed data and reports attribution quality. No human subjects, personal data or demographic attributes are involved.

Reproducibility

The source release includes experiment runners, configurations, seeds, per-incident metrics and table generators. Reported values and figures derive from stored artefacts, with source hashes and automated numerical checks. Appendix I documents provenance and claim rules, and Appendix H.2 records preprocessing and failures. SWaT and MSDS are subject to their providers' terms and are not redistributed.

AI Use Statement

Generative AI assisted methodological discussion, mathematical checks, experiment design, coding, diagnostic interpretation, literature searches and manuscript preparation. Synthetic data use documented generators, and measurements derive from stored experiment artefacts. Automated checks do not constitute independent human verification. The authors remain responsible for reviewing the AI-assisted work and for the final claims, code and conclusions.

References

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

Logan M. Bhamidipaty, Tommy Bruzzese, Caryn Tran, Rami Ratl Mrad, and Maxinder S. Kanwal. DynaDojo: An extensible platform for benchmarking scaling in dynamical system identification. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/32093649cbbcff773d9a991d8c30a7fe-Abstract-Datasets_and_Benchmarks.html.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S. Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Kailash Budhathoki, Lenon Minorics, Patrick Blöbaum, and Dominik Janzing. Causal structure-based root cause analysis of outliers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2357–2369. PMLR, 2022. URL https://proceedings.mlr.press/v162/budhathoki22a.html.

Wei Chen, Shigui Li, Jiacheng Li, Junmei Yang, John Paisley, and Delu Zeng. Dequantified diffusion-schrödinger bridge for density ratio estimation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8427–8452. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25aj.html.

János Gertler. *Fault Detection and Diagnosis in Engineering Systems*. Marcel Dekker, 1998.

Nikita Gushchin, Sergei Kholkin, Evgeny Burnaev, and Alexander Korotin. Light and optimal schrödinger bridge matching. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17100–17122. PMLR, 2024a. URL https://icml.cc/virtual/2024/poster/34581.

Nikita Gushchin, Daniil Selikhanovych, Sergei Kholkin, Evgeny Burnaev, and Alexander Korotin. Adversarial schrödinger bridge matching. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. doi: 10.52202/079017-2845. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a30d82c09902ced8aa3f49c5ee4a3d9b-Abstract-Conference.html.

Xiao Han, Saima Absar, Lu Zhang, and Shuhan Yuan. Root cause analysis of anomalies in multivariate time series through granger causal discovery. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=k38Th3x4d9.

Benjamin Herdeanu, Juan Nathaniel, Carla Roesch, Jatan Buch, Gregor Ramien, Johannes Haux, and Pierre Gentine. CausalDynamics: A large-scale benchmark for structural discovery of dynamical causal models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/file/39a3f7dc00e4723a5ca7808b04fc4d4f-Paper-Datasets_and_Benchmarks_Track.pdf.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, 2019.

Azam Ikram, Sarthak Chakraborty, Subrata Mitra, Shiv Kumar Saini, Saurabh Bagchi, and Murat Kocaoglu. Root cause analysis of failures in microservices through causal discovery. In *Advances in Neural Information Processing Systems*, 2022. URL https://papers.nips.cc/paper_files/paper/2022/file/c9fcd02e6445c7dfbad6986abee53d0d-Paper-Conference.pdf.

Rolf Isermann. *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*. Springer, 2006. doi: 10.1007/3-540-30368-5.

Jahoon Jeong and Hyunsoo Yoon. Pgrf-net: A prototype-guided relational fusion network for diagnostic multivariate time-series anomaly detection. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3hS7EtL4bV.

Karl Henrik Johansson. The quadruple-tank process. a multivariable laboratory process with an adjustable zero. *IEEE Transactions on Control Systems Technology*, 8(3):456–465, 2000. doi: 10.1109/87.845876.

Mingjie Li, Zeyan Li, Kanglin Yin, Xiaohui Nie, Wenchi Zhang, Kaixin Sui, and Dan Pei. Causal inference-based root cause analysis for online service systems with intervention recognition. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3230–3240. ACM, 2022. doi: 10.1145/3534678.3539041. URL https://doi.org/10.1145/3534678.3539041.

Lokesh Nagalapatti, Ashutosh Srivastava, Sunita Sarawagi, and Amit Sharma. Robust root cause diagnosis using in-distribution interventions. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/403ba123fbb0b967ed166025ceb3c9ad-Abstract-Conference.html.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, 1954. doi: 10.1093/biomet/41.1-2.100.

Charalampos Shimillas, Kleanthis Malialis, Konstantinos Fokianos, and Marios Polycarpou. Low rank transformer for multivariate time series anomaly detection and localization. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZtPIBpVojC.

Gideon Stein, Maha Shadaydeh, Jan Blunk, Niklas Penzel, and Joachim Denzler. CausalRivers: Scaling up benchmarking of causal discovery for real-world time-series. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a205fda871b0f6c1e18a7ad7325eb6cf-Abstract-Conference.html.

Anna Varbella, Kenza Amara, Blazhe Gjorgiev, Mennatallah El-Assady, and Giovanni Sansavini. PowerGraph: A power grid benchmark dataset for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.nips.cc/paper_files/paper/2024/hash/c7caf017cbbca1f4b368ffdc7bb8f319-Abstract-Datasets_and_Benchmarks_Track.html.

Julius von Kügelgen, Michel Besserve, Wendong Liang, Luigi Gresele, Armin Kekić, Elias Bareinboim, David M. Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Alan S. Willsky and Harold L. Jones. A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic Control*, 21(1):108–112, 1976. URL https://willsky.lids.mit.edu/publ_pdfs/14_pub_IEEE.pdf.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

APPENDIX CONTENTS

A Extended related work and discussion **14**

B Proofs **15**

C The deployed estimator is one pass **18**

D The path-KL origin of the action **18**

E Sharper bounds for plug-in ranking stability **19**

F Supporting theory **19**

G Numerical verification protocol **20**

H Additional experimental results **21**

I Implementation, reproducibility and claim rules **36**

J Matched validation protocols **37**

K Stress test for rapid attribution at scale **45**

L Exploratory innovation-scan and filter comparisons **56**

A Extended related work and discussion

A.1 Model-based fault detection and isolation.

Willsky and Jones' innovation GLR detects and estimates jumps in linear systems (Willsky & Jones, 1976). Change-detection theory develops likelihood scans and CUSUM (Basseville & Nikiforov, 1993), while structured residuals and observer-based fault diagnosis separate faults from predicted dynamics (Gertler, 1998, Isermann, 2006). Our singleton score is the same Gaussian mean-shift test when the state is known. Covariance normalisation, noncentral χ^2 power and cancellation of model-explained downstream motion are established foundations. A filter-based GLR instead propagates each candidate fault's signature through observation innovations and their predictive covariance. It can use temporal information when a state has no local proxy. Our pointwise encoder does not provide that capability. Appendix L evaluates a known-physics EKF comparator and a nominal smoother diagnostic. The contribution studied here is the identification, plug-in ranking stability and selective reporting needed when applying innovation isolation to learned latent coordinates.

A.2 Representation identifiability and incident attribution.

Auxiliary-variable and weakly supervised methods identify latent structure under additional information or interventions (Hyvärinen et al., 2019, Brehmer et al., 2022, von Kügelgen et al., 2023, Zhang et al., 2023). These works motivate our distinction between predictive adequacy and named coordinates. Our ranking bound separately controls approximation error on an incident.

A.3 Local tests and path-space inference.

Schrödinger bridges optimise path laws under distributional constraints (Gushchin et al., 2024a,b, Chen et al., 2025). We condition on a reconstructed path and profile a local drift alternative, giving an explicit Gaussian test without iterative forcing optimisation. Unlike CUSUM, which accumulates evidence for persistent change, our singleton test localises a transient coordinate–time intervention (Page, 1954).

A.4 Observed-variable diagnosis.

AERCA, CIRCA, RCD and causal outlier attribution analyse observed variables or mechanisms (Han et al., 2025, Li et al., 2022, Ikram et al., 2022, Budhathoki et al., 2022). Our target is the latent state and entry time. On observed targets AERCA outperforms our deployed forcing head (Section 5.4).

A.5 Assumptions and evaluation targets

An identifiability theorem characterises which latent models the available data can distinguish. A ranking bound instead asks whether estimation error changes a decision within a declared coordinate system. Our plug-in result addresses the second question and assumes that anchors have already resolved the first. For example, paired observations before and after interventions provide weak supervision for causal representation learning (Brehmer et al., 2022), while unknown interventions can identify nonparametric representations under explicit conditions (von Kügelgen et al., 2023). We do not infer those conditions from a low reconstruction loss. In the tank experiment, calibration supplies the correspondence between coordinates and physical levels. Weakening a flow proxy then tests reconstruction with that correspondence held fixed.

Likewise, recovering an observed variable over an incident and recovering an exact latent coordinate–time atom require different evidence. AERCA learns a Granger-causal model with exogenous variables and reports anomalous variables and time steps (Han et al., 2025). RCD treats a failure as an intervention and localises causal discovery around the affected system (Ikram et al., 2022). We evaluate our observed-variable projection separately because success on the latent target cannot establish superiority on these tasks. Our SWaT and MSDS results expose this distinction empirically. They also motivate stronger temporal representations and projection operators rather than a broader accuracy claim.

A.6 Temporal evidence and deployment

Our local alternative holds the reconstructed path fixed and profiles one forcing amplitude. It yields a direct score once encoding and drift evaluation are complete. A filter-based innovation GLR can instead propagate candidate fault signatures through successive measurements. That distinction matters when a state lacks a local proxy or a fault lasts several intervals. Our known-physics EKF comparison and fixed-duration scan examine these cases in Appendix L. The EKF's advantage with weak proxies shows that the reconstruction method can matter more than the choice of a local score.

The action identity therefore does not establish a universal computational advantage over fault diagnosis. Our runtime comparison isolates the cost of an iterative forcing objective on a common learned model. Nor does oracle Gaussian calibration automatically transfer to encoded scores. We require disjoint data to select and validate the reporting rule, and its guarantee assumes the same incident distribution at deployment. The operating-point shift demonstrates why this assumption needs scrutiny. Appendix K further separates growing scan size, fault strength and measurement quality. Its target-power schedule increases fault amplitude with scan size, so high accuracy under that schedule cannot establish robustness for a fixed physical fault.

B Proofs

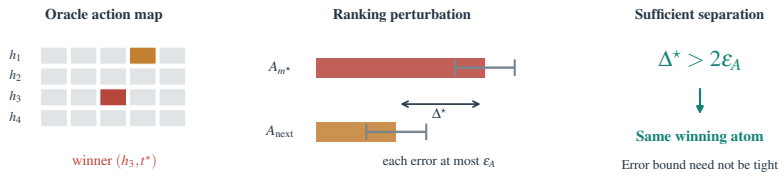

Figure 5: The action gap protects the winning atom. If each score changes by at most ε_A, the gap can shrink by at most $2\varepsilon_A$. Error bars are schematic and do not represent confidence intervals. Preserving the oracle winner does not ensure that the oracle identifies the true root.

Proposition 2 (Predictive equivalence does not identify a named root)**.** *Let $X_t = g(Z_t)$ and let a controlled latent path Z^u with control u induce the observed incident law. For any invertible linear $T \in \mathbb{R}^{d\times d}$, set $\widetilde{Z} = TZ$, $\widetilde{g} = g \circ T^{-1}$ and $\widetilde{u}_t = Tu_t$. The observed law of X is unchanged, but the named coordinate carrying the control can change. Coordinate-wise latent root attribution is therefore not identifiable without anchors, interventions, proxies, grouping or structural constraints that fix the latent coordinate system. For a nonlinear diffeomorphism the same holds with $\widetilde{u}_t(\widetilde{z}) = J_T(z)u_t(z)$ and Itô corrections.*

Proposition 2 (latent-coordinate non-identifiability). Let $T \in \mathbb{R}^{d\times d}$ be an invertible linear latent transformation preserving the permitted proxy model. Define $\widetilde{Z} = TZ$ and $\widetilde{g}(\widetilde{z}) = g(T^{-1}\widetilde{z})$. Then $\widetilde{g}(\widetilde{Z}_t) = g(Z_t) = X_t$ for every path. Because T is linear and constant, $d\widetilde{Z}_t = T\,dZ_t$, so the transformed control is exactly $\widetilde{u}_t = Tu_t$ with no correction term. The observed law is unchanged while action can move between transformed coordinates. Knowing d removes uncertainty about the *size* of the representation but not about its *names*. Exclusive anchors or an equivalent fixed proxy map remove the remaining freedom. For a nonlinear diffeomorphism T, Itô's formula gives

$$d\widetilde{Z}_t = \Big(J_T(Z_t)b(Z_t,t) + \widetilde{u}_t + \tfrac{1}{2}\mathrm{tr}\big(\nabla^2 T \cdot a(Z_t,t)\big)\Big)dt + J_T(Z_t)\sigma(Z_t,t)\,dW_t,$$

with $\widetilde{u}_t = J_T(Z_t)u_t(Z_t)$, so control, drift and diffusion transform together. The non-identifiability conclusion still holds, but the coordinatewise action decomposition (2) is stated only for the linear case used by the attribution model. □

Proposition 1 (the identity). By (1) the transition log-density in coordinate j is $-(r_{j,k} - u_{j,k}\Delta t_k)^2/(2\sigma_j^2\Delta t_k)$ up to a constant free of u. It is maximised at $u_{j,k} = r_{j,k}/\Delta t_k$, where the quadratic vanishes. Subtracting the null value $-r_{j,k}^2/(2\sigma_j^2\Delta t_k)$ gives the first equality of (3), and substituting $\widehat{u}$ into $(\Delta t_k/2)u^2/\sigma_j^2$ gives the second. □

Proposition 1 (the identity), non-diagonal case. The main-text proof covers diagonal Σ. For general symmetric positive-definite Σ and a candidate support $S \subseteq [d]$ at step k, write P_S for the coordinate selection matrix of S. Profiling the transition log-density over controls supported on S gives

$$\widehat{u}_S = \tfrac{1}{\Delta t_k}\, P_S^\top \big(P_S \Sigma^{-1} P_S^\top\big)^{-1} P_S \Sigma^{-1} r_k,$$
$$\log \Lambda_S = \tfrac{1}{2\Delta t_k}\, r_k^\top \Sigma^{-1} P_S^\top \big(P_S \Sigma^{-1} P_S^\top\big)^{-1} P_S \Sigma^{-1} r_k.$$

For $S = [d]$ this collapses to $r_k^\top \Sigma^{-1} r_k/(2\Delta t_k)$, the full-vector identity, which we verify to machine precision. For a singleton $S = \{j\}$ it equals $(e_j^\top \Sigma^{-1} r_k)^2/(2\Delta t_k\, e_j^\top \Sigma^{-1} e_j)$, which coincides with the coordinate action $r_{j,k}^2/(2\sigma_j^2 \Delta t_k)$ only when Σ is diagonal. The top-1 agreement of 58.2% and Kendall τ of 0.270 reported in Section G measure the resulting disagreement over an SPD ensemble $\Sigma = AA^\top + 0.1dI$ with $A_{ij} \sim \mathcal{N}(0,1)$. The number is ensemble-dependent, which is why we state the ensemble. □

Girsanov action identity. Let $\mathbb{P}^{\mathrm{nom}}$ be induced by $dZ_t = b\, dt + \sigma\, dW_t$ and $\mathbb{P}^u$ by $dZ_t = (b + u_t)\, dt + \sigma\, dW_t$. Under Novikov and absolute-continuity conditions, Girsanov's theorem writes the controlled-to-nominal log-likelihood ratio as a stochastic integral plus one half of the diffusion-normalised forcing energy. The stochastic integral has zero expectation under $\mathbb{P}^u$, so taking that expectation yields the action identity of Section 3 with $a = \sigma\sigma^\top$. □

Theorem 1 (plug-in ranking stability). Since $e_k = (\widehat{Z}_{k+1} - Z_{k+1}) - (\widehat{Z}_k - Z_k) - [\widehat{b}(\widehat{Z}_k) - b(Z_k)]\Delta t$, the encoder enters only through the *increment* of its error. A persistent offset cancels exactly, so assumptions (i)–(ii) and the triangle inequality give $\max_{j,k} |e_{j,k}| \le \varepsilon_\Delta + (L_b \varepsilon_Z + \varepsilon_b)\Delta t = \varepsilon_r$. For the action,

$$\widehat{A} - A = \frac{(r+e)^2 - r^2(1+\delta)}{2\sigma^2(1+\delta)\Delta t} = \frac{2re + e^2 - \delta r^2}{2\sigma^2(1+\delta)\Delta t},$$

and bounding each term with (iii)–(iv) and $1+\delta \ge 1-\rho$ gives $|\widehat{A} - A| \le \varepsilon_A$ as in (5). Finally, on the uniform-error event every atom satisfies $|\widehat{A}_m - A_m| \le \varepsilon_A$, so $\widehat{A}_{m^\star} \ge A_{m^\star} - \varepsilon_A > \max_{m \ne m^\star} A_m + \varepsilon_A \ge \max_{m \ne m^\star} \widehat{A}_m$. □

B.1 Oracle detection laws

A larger pulse is easier to distinguish from ordinary stochastic motion, but the threshold must account for every state and time searched. In the tank example, a volume pulse V into a tank of area S_j corresponds to a level increment $a = V/S_j$. The formula below describes that increment in the ideal Gaussian model. Its application to reconstructed levels needs the separate estimation and calibration checks. For true latent increments and known diagonal diffusion, standardisation gives

$$\begin{aligned} 2A_{j,k} &\sim \chi_1^2, \qquad & 2A_k &= \textstyle\sum_j 2A_{j,k} \sim \chi_d^2, \\ 2A_{j,k} \mid u^\star &\sim \chi_1^2(\lambda), & \lambda &= \frac{\Delta t_k (u^\star)^2}{\sigma_j^2} = 2A^\star. \end{aligned} \tag{6}$$

Thus the true forcing action determines the non-centrality and detection power. Scanning $N = dK$ atoms, a Bonferroni threshold $q_\alpha = F_{\chi_1^2}^{-1}(1 - \alpha/N)$ controls family-wise error at most α.

From a physical pulse to a detection probability. For a pulse with level increment a, the non-centrality is $\lambda = a^2/(\sigma_j^2 \Delta t)$. A larger pulse separates the forced transition from nominal noise. Searching more cells raises the threshold it must cross. The following result expresses this trade-off as a required increment.

Proposition 3 (Target-power atom detectability)**.** *Let*

$$\lambda_{\alpha,\beta} = \inf\{\lambda \ge 0 \mid \Pr[\chi_1^2(\lambda) > q_\alpha] \ge 1 - \beta\}. \tag{7}$$

If the generator plants $u^\star = a/\Delta t$ at one step, with $a \ge 0$, the true atom clears the scan threshold with probability at least $1 - \beta$ once

$$a \ge a_{\mathrm{detect}}(\alpha, \beta) = \sigma \sqrt{\Delta t\, \lambda_{\alpha,\beta}}. \tag{8}$$

For Tank 3, multiplying the required level increment by its area converts this condition into an injected-volume scale. The prediction belongs to the ideal Gaussian transition model. Fine-grid hydraulic integration and reconstruction from noisy proxies introduce discrepancies that must be measured separately. Under heterogeneous diffusion, each root's target-power amplitude uses its own variance. The common-amplitude experiment in Table 28 instead holds the physical increment fixed across roots, so their powers differ.

Detection, support and ranking ask different questions. A threshold can admit the true entry along with a false positive, or it can reject all cells even though the true entry ranks first. Table 2 separates these outcomes. At $\alpha = .05$ and $\beta = .1$, Proposition 3 gives a 90% atom-detection target. Requiring every null atom to remain below threshold gives an 85% lower bound for exact support by a union bound. Neither probability predicts top-1 accuracy. Appendix B.2 gives the support calculation and the onset-scan analogue. Section 5.1 measures the events separately.

Table 2: **Three outcomes of the same action map.** For one planted entry $m^\star$, thresholding and ranking answer different attribution questions.

Outcome	Question	Success event
Detection	Does the entry clear the nominal scan threshold?	$2A_{m^\star} > q_\alpha$
Exact support	Does thresholding select only the planted entry?	$\{m \mid 2A_m > q_\alpha\} = \{m^\star\}$
Top-1 ranking	Does the entry have the largest action?	$\arg\max_m A_m = m^\star$

B.2 Target-power detection and support recovery

With the oracle transition law in (6), scan $N = dK$ atoms at $q_\alpha = F^{-1}_{\chi^2_1}(1 - \alpha/N)$.

Proof of Proposition 3 and the support bound. This follows from $\lambda = a^2/(\Delta t\, \sigma^2)$ and monotonicity of the non-central tail. Exact thresholded support also requires every null atom to stay below threshold. A union bound gives

$$\Pr(\{m \mid 2A_m > q_\alpha\} = \{m^\star\}) \geq 1 - \alpha - \beta. \tag{9}$$

For an onset scan, use $S_k = \sum_j 2A_{j,k}$ and $q^{\text{time}}_\alpha = F^{-1}_{\chi^2_d}(1 - \alpha/K)$. The planted onset follows $\chi^2_d(\lambda)$, so its target-power amplitude uses d degrees of freedom and K comparisons. Atom threshold crossing, exact support and top-1 ranking are distinct events. At $\alpha = .05$ and $\beta = .1$, the atom detection target is 90% and the support lower bound is 85%. Neither is a predicted top-1 accuracy. Under heterogeneous diffusion, each amplitude uses the planted coordinate's variance. Section 5.1 tests these oracle predictions separately from performance on learned paths.

B.3 Split-calibrated selective prediction

Theorem 2 (Split-calibrated selective prediction, restated)**.** *Let each incident carry a confidence κ and a correctness indicator $C \in \{0, 1\}$, with $C = 1$ iff the named root is correct. Split a pool of incidents into disjoint* selection *and* validation *sets, and assume the validation incidents are i.i.d. draws from the deployment distribution. Fix a target coverage c and set $\widehat{\tau}$ to the $(1 - c)$-quantile of κ on the selection set only. Let m be the number of validation incidents with $\kappa \geq \widehat{\tau}$ and e the number of those with $C = 0$. For $\delta \in (0, 1)$, let $U(e, m, \delta)$ be the Clopper–Pearson upper $(1 - \delta)$-confidence bound. Then, with probability at least $1 - \delta$ over the validation draw, $\Pr(C = 0 \mid \kappa \geq \widehat{\tau}) \leq U(e, m, \delta)$.*

The threshold $\widehat{\tau}$ is a fixed, deterministic function of the selection set alone, and the validation set never enters its construction. Conditional on $\widehat{\tau}$, the accepted validation incidents $\{i : \kappa_i \geq \widehat{\tau}\}$ are m independent Bernoulli trials of the event $C = 0$ with common probability $p = \Pr(C = 0 \mid \kappa \geq \widehat{\tau})$, so the accepted error count is $\text{Binomial}(m, p)$. Observing e errors in m trials, the Clopper–Pearson construction inverts the binomial tail. The upper bound $U(e, m, \delta)$ solves $\Pr[\text{Binomial}(m, u) \leq e] = \delta$, and by that inversion $\Pr[p > U(e, m, \delta)] \leq \delta$ for every true p. Substituting p gives the stated bound. □

Two conditions are load-bearing, because violating either silently invalidates the interval. First, the binomial step needs conditional *independence* of the accepted validation incidents given $\widehat{\tau}$, which is stronger than the exchangeability a conformal rank argument would need. We therefore assume i.i.d. validation draws rather than claiming a distribution-free guarantee. Second, the bound applies to *one* selection/validation split. Repeating the split over the same pool and pooling all accepted predictions into a single binomial count is invalid because the same incidents recur across splits so the pooled predictions are dependent, each split induces a different $\widehat{\tau}$ and hence a different conditional error probability, and the pooled count has a far smaller effective sample size than its nominal m, producing a spuriously tight interval. We consequently report one predeclared split as the result covered by the theorem and summarise the spread of per-split bounds only as a stability diagnostic, never pooled. Subject to these conditions the bound holds for *any* confidence signal. It says nothing about whether κ is informative, only that the reported upper bound on the false-claim rate is valid.

C THE DEPLOYED ESTIMATOR IS ONE PASS

The action–evidence correspondence yields a closed-form attribution operator. This appendix verifies that its deployed implementation evaluates the operator directly and records how it differs from the iterative comparator.

What the deployed estimator executes. `control_backend: innovation` dispatches to `sb_rexcss/rca/innovation.py`. The implementation encodes the window, differences consecutive latents and subtracts $\widehat{b}\,\Delta t$. It then squares the excess increments, divides by $2\widehat{\sigma}^2\Delta t$ and ranks all atoms. Only the separate pulse-window ablation zeroes steps outside a score-selected window. Its centre is a single reduction over the score array. There is no loop over iterations, no gradient, and no call into the solver module. Reported per-incident diagnostics include `iteration_count`$= 0$ and `gradient_evaluations`$= 0$.

Iterative comparator. The optimised arm performs a 50-step coarse prepass followed by 500 solver steps. The prepass-only arm therefore executes 50, not zero, iterations. Table 1 separates these from the deployed score. Gaussian smoothing is a separate ablation. These are internal comparisons of different objectives.

How the claim is enforced. `tests/test_innovation_one_pass.py` monkey-patches both `_adjoint_control_gradient` and `_proximal_step` with functions that raise, then runs the deployed path. The test fails if either is ever reached. Further tests check that the returned score equals the closed-form GLRT to machine precision, that the window mask is the single $\arg\max$ and nothing else, that the smoother is off by default, and that the diffusion normalisation reorders atoms under heteroscedasticity, and that it provably cannot under a scalar σ.

Cost. Forming and scanning the deployed action map costs $O(Kd)$ after reconstruction and drift evaluation. The internal 550-step solver takes 983.8 ms versus 0.43 ms for the one-pass score (Table 21). This does not measure relative cost against classical FDI.

D THE PATH-KL ORIGIN OF THE ACTION

The main text states the least-action motivation in one paragraph. This is the derivation, kept here because the paper's conclusion is that the variational framing supplies the *metric* and not the algorithm.

Let $\mathbb{P}^{\mathrm{nom}}$ be induced by $dZ_t = b\,dt + \sigma\,dW_t$ and $\mathbb{P}^u$ by $dZ_t = (b + u_t)\,dt + \sigma\,dW_t$. The idealised bridge problem is

$$\mathbb{P}^\star = \arg\min\left\{\mathrm{KL}(\mathbb{P}\|\mathbb{P}^{\mathrm{nom}}) : \mathbb{P}_0 = \nu_0,\ \mathbb{P}_T = \nu_T^{\mathrm{inc}}\right\},$$

and under Novikov and absolute-continuity conditions Girsanov's theorem writes the controlled-to-nominal log-likelihood ratio as a stochastic integral plus one half of the diffusion-normalised forcing energy. The stochastic integral has zero expectation under $\mathbb{P}^u$, so

$$\mathrm{KL}(\mathbb{P}^u\|\mathbb{P}^{\mathrm{nom}}) = \tfrac{1}{2}\,\mathbb{E}_{\mathbb{P}^u}\int_0^T \|u_t(Z_t)\|^2_{a^{-1}}\,dt, \qquad a = \sigma\sigma^\top. \tag{10}$$

The path-KL supplies the diffusion metric. By Proposition 5, the energy is invariant under $u \mapsto Tu$ and $a \mapsto TaT^\top$. This identity assumes the same initial distribution. Otherwise an initial-distribution KL term must be added. The endpoint-constrained bridge remains a global optimisation over path laws. Proposition 1 instead conditions on the realised path and profiles one Gaussian transition. It solves that local likelihood problem, not the displayed bridge.

E SHARPER BOUNDS FOR PLUG-IN RANKING STABILITY

Theorem 1 uses one global error budget ε_A and one global excess increment bound R. Both can be localised, and doing so matters because the runner-up atom in a typical incident has a far smaller excess increment than the root.

Proposition 4 (Per-atom condition)**.** *Under the assumptions of Theorem 1, define the atomwise budget*

$$\varepsilon_A(m) = \frac{2|r_m|\varepsilon_r + \varepsilon_r^2 + \rho\, r_m^2}{2\sigma_{\min}^2(1-\rho)\,\Delta t}.$$

If $A_{m^\star} - \varepsilon_A(m^\star) > \max_{m \neq m^\star} \left[A_m + \varepsilon_A(m)\right]$ *then* $\arg\max_m \widehat{A}_m = m^\star$.

Proof. Identical to the last step of Theorem 1, replacing the uniform ε_A by its atomwise value in each of the two inequalities. □

Proposition 4 is never weaker than Theorem 1 and is strictly stronger whenever the excess increments are unequal, which they are by construction under an intervention. Off-root atoms have $|r_m| \approx \sigma\sqrt{\Delta t}$ while the root carries the planted amplitude in addition. We state the uniform version in the main text because it is the one that reads as a clean sufficient condition, and use the atomwise version when reporting fractions that satisfy the bound.

Why the encoder enters through its increment. Assumption (i) of Theorem 1 bounds $\varepsilon_\Delta = \max_k \|(\widehat{Z}_{k+1} - Z_{k+1}) - (\widehat{Z}_k - Z_k)\|_\infty$, not ε_Z alone. This is not a cosmetic refinement. An excess increment is formed by differencing consecutive states, so a *persistent* encoder offset, the dominant failure mode of a decoder pseudo-inverse, cancels exactly and contributes nothing. Only the part of the encoder error that changes between steps survives. We always have $\varepsilon_\Delta \leq 2\varepsilon_Z$, and The archival supplement records the earlier nominal-quantile diagnostics. Those quantiles are not uniform error bounds.

What the bound does not give. It is deterministic and worst-case, so it is sufficient and not necessary, and it says nothing about the *law* of $\widehat{A}$. A small uniform perturbation can protect an argmax without preserving an exact χ^2 law. It still constrains distributional distances, so it does not permit arbitrary changes of the score distribution. The encoded null in the stored experiment has a small empirical KS discrepancy. Nominal calibration remains necessary whenever the oracle assumptions are not established.

F SUPPORTING THEORY

Proposition 5 (Reparameterisation invariance of the action)**.** *Let T be invertible with $\widetilde{u} = Tu$ and $\widetilde{a} = TaT^\top$. Then $\widetilde{u}^\top \widetilde{a}^{-1} \widetilde{u} = u^\top a^{-1} u$, while the Euclidean energy $\|\widetilde{u}\|^2$ is not invariant unless T is orthogonal.*

Proof. $\widetilde{u}^\top \widetilde{a}^{-1} \widetilde{u} = u^\top T^\top T^{-\top} a^{-1} T^{-1} T u = u^\top a^{-1} u$. For the Euclidean form take $u \neq 0$ and $T = \mathrm{diag}(2, 1, \ldots, 1)$. □

Numerically, the Girsanov form is invariant to relative error 3.4×10^{-10} across random transformations while the Euclidean energy changes by factors 2.1–22.6. Rescaling units preserves the population action when forcing and diffusion transform consistently. Redundant sensors and additional proxies do not change that population quantity under the same representation change, although they can change its learned estimate through representation error. Anchors give coordinates physical

names. Total action alone does not identify coordinate support. Proposition 5 is the population-level counterpart of the finite-sample dose–response in the archived solver dose–response study. The metric matters exactly when coordinates differ, and is vacuous when they do not.

Theorem 3 (Sparse action-gap recovery)**.** *Assume the latent coordinate system is fixed by anchors, interventions, proxies or structural constraints. Let $S^\star$ be the true sparse forcing support and let $\widehat{A}$ satisfy* $\max_{r,k} |\widehat{A}_{r,k} - A^\star_{r,k}| \le \epsilon_n$*. If*

$$\min_{(r,k)\in S^\star} A^\star_{r,k} - \max_{(r,k)\notin S^\star} A^\star_{r,k} > 2\epsilon_n,$$

then thresholding $\widehat{A}$ between the two sides recovers $S^\star$.

Proof. On the uniform action-error event, every active atom is at least its true value minus ϵ_n and every inactive atom is at most its true value plus ϵ_n. The assumed gap leaves a nonempty interval between these bounds, and any threshold inside it recovers the support. When the proxy map admits a latent transformation, the argument applies after quotienting coordinates by the permitted ambiguity relation and cannot identify a representative within a class. □

Theorem 3 is the deterministic counterpart of the distributional statement in Section B.1. Equation (6) says how large $A^\star$ is in expectation, (8) inverts the union-bound threshold into an amplitude, and Theorem 3 says what a perturbation budget of ϵ_n costs. The amplitude sweep of Appendix H.3 tests whether the quantity the theorem conditions on is measurable rather than degenerate.

G NUMERICAL VERIFICATION PROTOCOL

Table 3: Numerical verification of Proposition 1 and the operating characteristics of Section B.1, over 27 regime–seed cells and 82,944 null atoms. The last two rows are the non-diagonal diffusion study of Section 4.1. Generated from `exp/glrt_identity/`.

Quantity	Predicted	Measured	Agreement
Identity $A_{j,k} = \log \Lambda_{j,k}$ (3)	exact	3 independent routes	$2.66e-15$ abs.
Per-atom null $2A_{j,k}$ (6)	χ^2_1 (mean 1, var 2)	mean 0.9959, var 2.00	KS 0.0059
Per-step null $2A_k$ (6)	χ^2_6 (mean 6)	mean 5.9753	KS 0.0152
Non-centrality (6)	$1 + \lambda = 25.12$	25.13	$+0.0\%$
Non-diagonal Σ: full-vector identity	exact	n/a	0
Non-diagonal Σ: coordinate split	n/a	top-1 agreement 58.2%	Kendall τ 0.270

Three routes to the identity. Table 3, first row, compares (i) the closed-form action $r^2_{j,k}/(2\sigma^2_j \Delta t_k)$ as computed by the shipped scorer, (ii) the profile log-likelihood $\log p(Z_{k+1} \mid Z_k, \widehat{u}) - \log p(Z_{k+1} \mid Z_k, 0)$ evaluated at the analytic maximiser $\widehat{u} = r/\Delta t$, and (iii) a numerical maximisation of the transition log-density over $u_{j,k}$ by Brent search to tolerance 10^{-12}. Maximum absolute disagreement over 27 regime–seed cells is $2.66e-15$ ($5.26e-16$ after scaling by the statistic's magnitude). Agreement at double-precision round-off is what an algebraic identity should produce.

We are deliberate about how much this is worth. Route (iii) is genuinely independent. It never forms $\widehat{u}$ and shares no helper function with the scorer, so its agreement is evidence that the closed form is the maximiser. Route (ii) is *algebraically the same expression* once $\widehat{u} = r/\Delta t$ is substituted, so it agrees exactly by construction. We report it as a consistency check on the implementation, not as a third independent confirmation. A numerically verified algebraic identity is, in the end, a check that the code computes the algebra. The content of Proposition 1 is the algebra itself.

Null and non-centrality. Null atoms are drawn from fault-free windows of the same trained models used for attribution, $82{,}944$ in total. We report the KS distance to χ^2_1 rather than a p-value because at this n every test rejects. The relevant question is the size of the discrepancy, and 0.0059 is a descriptive marginal discrepancy, not evidence of independent encoded increments. Even independent pointwise reconstruction errors η_k contribute $\eta_{k+1} - \eta_k$ to the residual, with adjacent covariance $-\operatorname{Var}(\eta_k)$ before drift-error terms. Thus encoded scores need not have the oracle χ^2 law. Held-out whole-window calibration is used for the physical scan. The per-step check sums over $d = 6$ latent coordinates. The non-centrality check uses the benchmark operating point ($\Delta t = 1/64$, $\sigma^2 = 1.3$) and compares the mean of $2A$ at the true atom against $1 + \lambda$.

G.1 Encoded-score diagnostics

Nominal quantiles cannot replace the uniform pathwise bounds in Theorem 1. The finite-path check in Appendix J uses a common basis and actual maxima, and tests full atom agreement. The table below measures marginal score discrepancies, which are not calibration guarantees.

Table 4: **The plug-in gap.** Five arms sharing everything but the source of the latent path. The deployed one-pass estimator leaves the null essentially where the true-latent scores leave it in this diagnostic. The legacy pipeline shows a larger discrepancy. Panel 1 of 2.

Latent source	Time acc.	AC*@1	Null mean $2A$
oracle: true latents, true b, true Σ	0.8611	0.9815	0.9959
shipped one-pass: encoded latents	0.8025	0.9167	0.9974
+ Gaussian pre-smoother	0.7994	0.9167	0.9815
+ pulse window mask	0.8025	0.8179	0.1603
legacy solver pipeline (previously shipped)	0.8920	0.9198	0.2188

Table 5: Continuation of Table 4, panel 2 of 2.

Latent source	KS to χ^2_1 (all atoms)	(surviving)
oracle: true latents, true b, true Σ	0.0059	0.0059
shipped one-pass: encoded latents	0.0047	0.0047
+ Gaussian pre-smoother	0.0076	0.0076
+ pulse window mask	0.8671	0.0367
legacy solver pipeline (previously shipped)	0.9180	0.2663

Reading the plug-in gap. "Encoded" is the deployed maskless forcing score. "Windowed" is the score-selected pulse-window ablation. "Legacy" adds the solver-era prepass, smoothing and sharpening. These arms differ in post-processing as well as the path source. Small empirical KS discrepancies do not prove that encoded or selected-survivor scores have an exact χ^2_1 law. A score-selected mask may change the argmax atom even though its centre preserves the peak of the summed time score. It cannot be called an argmax-preserving perturbation without checking the relevant gap.

H Additional experimental results

Estimator provenance. The benchmark, likelihood checks, plug-in study, real-data comparisons and selective-reporting tables use the one-pass estimator. Historical sparsity, component-ladder and evidence-alignment results used the iterative solver and are compiled separately in `larca-archive.pdf`. They are not evidence for the deployed scorer. The full source-to-artifact map remains in the repository.

H.1 PER-REGIME BREAKDOWN OF THE LATENT-TARGET BENCHMARK

In the compact comparison tables, bold marks the best displayed feasible point estimate in each performance column, including ties. Oracle rows are separate references. Boldface does not indicate statistical significance.

The baseline labelled latent CUSUM operates on standardised innovations, not on latent states. With $z_{j,k} = \widehat{r}_{j,k}/(\widehat{\sigma}_j\sqrt{\Delta t_k})$, it computes $S^+_{j,k} = \max(0, S^+_{j,k-1} + z_{j,k} - 0.5)$ and $S^-_{j,k} = \max(0, S^-_{j,k-1} - z_{j,k} - 0.5)$, starting at zero, then scores $\max(S^+_{j,k}, S^-_{j,k})$. Its peak estimates the end of accumulated evidence, so exact-onset accuracy is not a full assessment of change detection. The state- deviation arm is a diagnostic ablation without drift subtraction. Neither the state-deviation comparison nor projecting observed-variable methods onto latent labels establishes superiority over model-based fault isolation.

Table 6: Full latent-target ablations behind Table 1. Rows share a nominal model and dataset, except the oracle which reads the generating path. MRR, atom accuracy and root-set recovery average nine common regimes. Named-coordinate accuracy averages eight and onset accuracy ten. The timing column is the median per-incident scoring cost. Panel 1 of 3.

Method	Steps	MRR	AC*@1
One-pass forcing test	0	**0.8434**	**0.7948**
+ pulse window	0	0.7612	0.7243
Pulse window, Euclidean	0	0.7520	0.7184
Pulse window and smoother	0	0.7626	0.7260
Prepass	50	0.8271	0.7924
Prepass and iterative solver	550	0.8270	0.7920
Iterative solver alone	500	0.8428	0.7934
Latent scan / CUSUM	0	0.0974	0.0358
Latent state deviation	0	0.0410	0.0174
Oracle dynamics and latents	0	0.9578	0.9271

Table 7: Continuation of Table 6, panel 2 of 3.

Method	AC@1	Time-Acc	Root-Set
One-pass forcing test	**0.8898**	0.8119	0.7767
+ pulse window	0.8242	0.8119	0.7177
Pulse window, Euclidean	0.8266	0.7759	0.7115
Pulse window and smoother	0.8262	0.8125	0.7194
Prepass	0.8875	**0.8478**	**0.7847**
Prepass and iterative solver	0.8871	0.8475	**0.7847**
Iterative solver alone	0.8887	0.8294	0.7747
Latent scan / CUSUM	0.3660	0.0609	0.0312
Latent state deviation	0.3141	0.0125	0.0146
Oracle dynamics and latents	0.9254	0.8931	0.9191

Table 8: Continuation of Table 6, panel 3 of 3.

Method	Time (ms)
One-pass forcing test	0.43
+ pulse window	0.65
Pulse window, Euclidean	0.46
Pulse window and smoother	3.26

Continued on next page

Table 8: Continued.

Method	Time (ms)
Prepass	74.97
Prepass and iterative solver	983.76
Iterative solver alone	902.61
Latent scan / CUSUM	1.28
Latent state deviation	**0.15**
Oracle dynamics and latents	–

Table 9: Per-regime breakdown of Table 1: the deployed one-pass estimator, AERCA, and the oracle GLRT on the generating latent path. Ten seeds, 32 incidents per cell. Oracle rows coincide when only the observation map changes and the latent process and interventions are held fixed. Generated from `exp/v2/`. A dash marks a metric that is undefined in that regime. Rotation ambiguity prevents comparison of coordinate labels. For historical weak-observability rows, a reporting condition suppressed coordinate accuracy while retaining raw atom accuracy. Those coordinate entries were not recorded. They are not mathematically undefined. Panel 1 of 2.

Regime	Method	AC@1	AC*@1
anchor upgrade	LARCA	91.6%	90.6%
anchor upgrade	AERCA	85.0%	82.8%
anchor upgrade	*oracle*	96.6%	96.2%
downstream symptom dominance	LARCA	92.8%	92.2%
downstream symptom dominance	AERCA	83.1%	80.6%
downstream symptom dominance	*oracle*	96.6%	96.2%
weak observability	LARCA	–	10.9%
weak observability	AERCA	–	6.9%
weak observability	*oracle*	–	96.2%
linear anchor	LARCA	92.2%	91.2%
linear anchor	AERCA	83.8%	81.2%
linear anchor	*oracle*	96.6%	96.2%
lorenz96	LARCA	92.8%	91.9%
lorenz96	AERCA	84.1%	81.6%
lorenz96	*oracle*	96.2%	95.9%
multiple roots	LARCA	99.7%	99.1%
multiple roots	AERCA	96.6%	96.2%
multiple roots	*oracle*	100.0%	100.0%
nonidentifiable rotation	LARCA	–	–
nonidentifiable rotation	AERCA	–	–
nonidentifiable rotation	*oracle*	–	96.2%
nonlinear decoder	LARCA	91.2%	89.7%
nonlinear decoder	AERCA	83.4%	80.9%

Continued on next page

Table 9: Continued.

Regime	Method	AC@1	AC*@1
nonlinear decoder	*oracle*	96.6%	96.2%
tep simulator	LARCA	91.6%	90.3%
tep simulator	AERCA	88.1%	85.9%
tep simulator	*oracle*	96.6%	96.2%
weak action gap	LARCA	60.0%	59.4%
weak action gap	AERCA	55.9%	54.7%
weak action gap	*oracle*	61.3%	60.9%

Table 10: Continuation of Table 9, panel 2 of 2.

Regime	Method	Root-Set	Time-Acc
anchor upgrade	LARCA	90.6%	78.8%
anchor upgrade	AERCA	82.8%	53.1%
anchor upgrade	*oracle*	96.2%	86.9%
downstream symptom dominance	LARCA	92.2%	81.6%
downstream symptom dominance	AERCA	80.6%	53.8%
downstream symptom dominance	*oracle*	96.2%	86.9%
weak observability	LARCA	10.9%	70.0%
weak observability	AERCA	6.9%	52.8%
weak observability	*oracle*	96.2%	86.9%
linear anchor	LARCA	91.2%	80.3%
linear anchor	AERCA	81.2%	53.8%
linear anchor	*oracle*	96.2%	86.9%
lorenz96	LARCA	91.9%	80.6%
lorenz96	AERCA	81.6%	55.0%
lorenz96	*oracle*	95.9%	86.6%
multiple roots	LARCA	82.8%	98.4%
multiple roots	AERCA	62.5%	77.5%
multiple roots	*oracle*	92.8%	98.4%
nonidentifiable rotation	LARCA	–	65.6%
nonidentifiable rotation	AERCA	–	48.1%
nonidentifiable rotation	*oracle*	96.2%	86.9%
nonlinear decoder	LARCA	89.7%	79.4%
nonlinear decoder	AERCA	80.9%	50.9%
nonlinear decoder	*oracle*	96.2%	86.9%
tep simulator	LARCA	90.3%	77.5%
tep simulator	AERCA	85.9%	54.7%
tep simulator	*oracle*	96.2%	86.9%

Continued on next page

Table 10: Continued.

Regime	Method	Root-Set	Time-Acc
weak action gap	LARCA	59.4%	99.7%
weak action gap	AERCA	54.7%	82.5%
weak action gap	*oracle*	60.9%	100.0%

Table 9 breaks the headline down regime by regime, with the oracle row giving a true-state reference for this latent target. The oracle reads the generating latent path, so its row is identical across every regime that varies only the observation map, and moves only where the latent drift or the injected intervention itself changes. This is a property of the design (the regimes are a sweep over *observability* at a fixed latent process) rather than a defect of the run, and it is what makes the oracle a common reference for measuring representation cost. Three readings matter.

First, the oracle row separates two error sources that the pooled number conflates. The gap between LARCA and the oracle is the cost of the learned representation, while the gap between the oracle and 100% is the intrinsic difficulty of the regime at this amplitude. Weak observability is the sharpest case. The oracle scores 96.2% while nothing built on a learned representation exceeds 10.9%. The regime hides a latent coordinate by shrinking its decoder column, so the control remains fully present in the latent path and is almost absent from the observations. The failure is therefore one of observability rather than of the estimand, which is the case Theorem 1 is about and the reason the claim hierarchy downgrades rather than guesses when anchors are missing.

Observed-variable methods projected onto a latent target are recorded only in the archival supplement. Their near-zero atom scores are not evidence of a competitive advantage. Native-target results remain in Tables 11 and 14.

The aggregation uses nine common regimes for raw atom accuracy, eight for coordinate accuracy after eligibility screening, and ten for timing. AC*@1 is a coordinate–time hit and implies the corresponding raw coordinate hit. The mismatch in available columns is a logging-policy defect, not a different mathematical target. Future runs record both raw metrics whenever the coordinate basis is shared, and report named-claim validity separately. The generator key `hidden_confounder` is retained for artefact compatibility but denotes weak observability, not an implemented latent common-cause process.

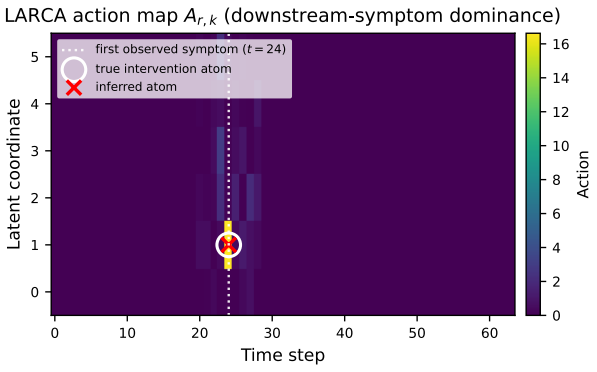

Figure 6: Recovered action map for one incident (`downstream_symptom_dominance`). The inferred atom (cross) coincides with the true one (circle), and that cell carries 40% of the action mass over 384 coordinate–time cells. Observed surprisal crosses its corrected shock threshold at the same step (dotted), matching the contemporaneous propagation operator of Section 4.4.

Figure 6 shows one recovered action map. The inferred atom coincides with the true one and carries most of the action mass, with the observed surprisal crossing its corrected threshold at the same step.

H.2 REAL-DATA PROTOCOL DETAILS

The bridge-residual ablation leads AERCA only at looser SWaT retrieval budgets, where chance also rises. Two paired t intervals on per-seed differences exclude zero, but were selected after inspecting seven operating points and are uncorrected. The tight-budget deficit is unstable across seeds. MSDS excludes the 10 trials where AERCA diverged from the paired recall comparison, while reporting failures separately. The pulse mask is disabled for these sustained incidents.

Tables 11 and 14 carry the real-data results summarised in Section 5.4, and this section records the protocol details that belong with them.

Table 11: SWaT attack attribution over 5 seeds, AERCA's own metrics, window w=10, 25 scored incidents. The operating points of Figs. 7 and 8, from the same score dumps and the same metric. Brackets are 95% percentile-bootstrap intervals resampling the *seed*. They describe seed spread and are not the paired t intervals on per-seed differences discussed above. The chance row is a uniformly random ranker under the same $\min(k, \#\text{attacked})$ normaliser, and is included because AC*@k at loose k is substantially metric fill. All LARCA rows share one trained model per seed and one predeclared latent-to-observed projection, so they differ only in the latent statistic. Bridge residual is the score an earlier version reported, kept as a diagnostic ablation. Panel 1 of 3.

Method	AC@1	AC@5	AC@10
Uniformly random ranker (chance)	2.2%	10.0%	20.0%
LARCA innovation (deployed, 0 iterations)	3.7% [2.3,5.6]	16.8% [13.4,20.9]	32.8% [28.8,36.9]
LARCA innovation + pulse window (ablation)	2.6% [2.2,2.9]	13.5% [11.2,15.9]	28.0% [25.6,30.6]
LARCA optimised control (ablation)	10.4% [7.1,13.0]	36.8% [31.6,42.0]	58.2% [52.4,64.0]
LARCA bridge residual (diagnostic)	13.0% [6.5,20.0]	36.6% [26.6,46.6]	59.9% [56.0,64.6]
AERCA (official)	20.2% [16.6,24.4]	32.3% [26.0,38.6]	44.5% [38.5,50.6]

Table 12: Continuation of Table 11, panel 2 of 3.

Method	AC*@1	AC*@25
Uniformly random ranker (chance)	0.7%	1.7%
LARCA innovation (deployed, 0 iterations)	8.0% [8.0,8.0]	3.8% [3.3,4.2]
LARCA innovation + pulse window (ablation)	8.0% [8.0,8.0]	4.0% [3.4,4.4]
LARCA optimised control (ablation)	1.6% [0.0,3.2]	2.4% [1.5,3.1]
LARCA bridge residual (diagnostic)	12.0% [4.8,20.8]	20.9% [16.0,26.4]
AERCA (official)	14.4% [8.8,20.0]	22.8% [18.8,26.8]

Table 13: Continuation of Table 11, panel 3 of 3.

Method	AC*@50	AC*@100
Uniformly random ranker (chance)	3.3%	6.7%
LARCA innovation (deployed, 0 iterations)	5.4% [4.4,6.3]	12.4% [9.7,15.6]
LARCA innovation + pulse window (ablation)	6.1% [4.9,7.2]	14.8% [12.6,16.5]
LARCA optimised control (ablation)	4.0% [2.6,5.2]	7.5% [5.4,9.5]
LARCA bridge residual (diagnostic)	31.9% [26.9,36.6]	48.4% [44.4,52.3]
AERCA (official)	27.2% [24.2,30.4]	35.3% [30.0,41.2]

Table 14: MSDS event localisation on AERCA's official event-level protocol over 50 randomised trials, scored with the deployed forcing head. Both methods are scored on the 40 trials where unmodified AERCA produced output. Brackets are 95% percentile-bootstrap intervals resampling the *trial*. They do not overlap, so the deficit is not sampling noise. Generated from `exp/v2_msds/`. Panel 1 of 2.

Method	Event Recall@1	Event Recall@3
LARCA innovation (deployed, 0 iterations)	4.6% $[3.9, 5.3]$	15.2% $[14.1, 16.3]$
AERCA (official)	10.7% $[6.1, 16.3]$	41.2% $[30.7, 51.8]$

Table 15: Continuation of Table 14, panel 2 of 2.

Method	Event Recall@5	Failure rate
LARCA innovation (deployed, 0 iterations)	30.5% $[27.7, 33.4]$	**0% (0/50)**
AERCA (official)	54.3% $[43.5, 64.9]$	20% (10/50)

One projection for every head. All LARCA rows on both datasets use the same predeclared latent-to-observed map. A latent atom score is attributed to observed variable i in proportion to $|C_{ir}|$, the magnitude of that variable's decoder loading. Because every head is projected identically and scored on the same trained model, a row difference is the latent statistic and nothing else. Earlier versions of this work reported only the bridge-residual head on real data while describing the one-pass estimator in the method section. That inconsistency is removed, and the residual head is kept as a labelled diagnostic ablation.

Contemporary baselines. ALoRa-Loc (Shimillas et al., 2026) and PGRF-Net (Jeong & Yoon, 2026) have public implementations, at `https://github.com/CharisShimillas/ALoRa` and `https://github.com/jahoonjeong/PGRF-Net`. We have not completed a protocol-compatible evaluation. Their published scores cannot be inserted into our SWaT table because windows, channel mappings and aggregation rules may differ. IDI (Nagalapatti et al., 2025) additionally requires an observed causal DAG and a designated target node, neither supplied by our present SWaT protocol. We make no claim that the observed-target comparison is complete.

SWaT label mapping, and a correction to an earlier incident count. An earlier version of this paper reported 30 scored incidents and described two unresolvable attack points, `MV504` and `DIT301`. We can no longer reproduce that count. The attack sheet we hold,

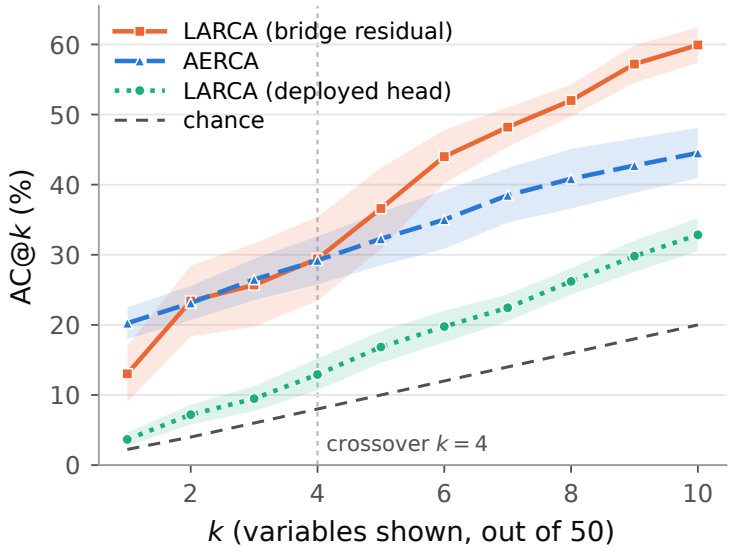

Figure 7: SWaT AC@k over the 50 observed variables. AERCA leads at $k \leq 3$. The bridge-residual score leads from $k = 4$. Bands show one standard error over 5 seeds. Chance uses the same $\min(k, \#\text{attacked})$ normaliser.

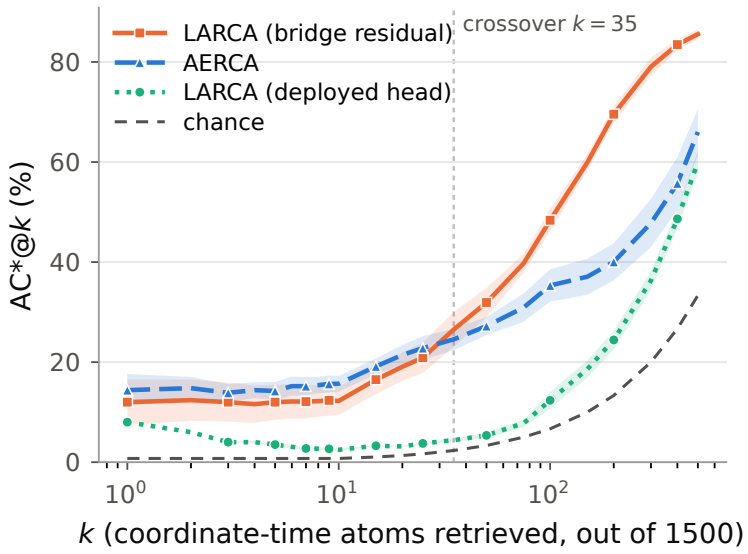

Figure 8: SWaT AC*@k over 1500 coordinate–time atoms. Loose-budget accuracy increasingly reflects metric fill: chance reaches 6.7% at $k = 100$ and 33.3% at $k = 500$.

`List_of_attacks_Final.xlsx`, contains neither string, and an audit of it finds no unresolvable points. It lists 25 of SWaT's 41 published attacks, and its own trailing note refers to an attack number beyond its last row. Independently, only 15,404 of the 54,621 attack-state samples in the abnormal recording fall inside any listed window, leaving 72% of attack time with no label row at all, in 13 episodes of at least one minute including one of 9.5 hours. The recording is complete and the label sheet is truncated, so the larger incident count came from a fuller attack list that we no longer have. Recovering it is a data-acquisition task, not a processing choice, and until it is recovered the honest count is the one we can verify.

The sheet also carries a year typo: five attacks are dated 2015-01-02 while the campaign ran from 2015-12-28 to 2016-01-02, so their windows selected zero samples and were previously dropped without a record. Our loader repairs a date only when the repair is self-verifying, accepting a corrected year solely if the resulting window lands *wholly* on attack-flagged samples. All five do, with 375/375, 281/281, 401/401, 297/297 and 1675/1675 samples, which a wrong year could not produce. A flag restores the unrepaired behaviour, and windows that still select no samples are now reported rather than skipped silently. AERCA's official loader carries the same year bug, so its published SWaT numbers are also on the unrepaired set. Our harness feeds both methods the same incidents, so the comparison is unaffected either way.

This leaves 25 scored incidents with a median of 10 attacked atoms each over a 30-step window and 50 sensors, and it raises the chance level of AC@1 from 2.1% to 2.2%. One caveat travels with the set: in a single incident, 9 of the 20 pre-onset samples that the protocol treats as nominal context are themselves unlabelled attack samples. The other 24 incidents have clean pre-onset context.

MSDS failure accounting. Recall is scored only on the trials where unmodified upstream AERCA produced output, with both methods scored on that same subset, because a worst-case fill for the diverging trials would flatter us by construction. The failure-rate column reports divergences separately over all trials.

H.3 Action-gap and downgrade stress

Table 16: Intervention-amplitude sweep (`linear_anchor`, 3 seeds, 44 evaluation incidents per amplitude). Exact-atom recovery and the rate of named-root reports generally increase with amplitude, alongside the *estimated* coordinate gap ($0.11 \rightarrow 0.25$), so the quantity Theorem 3 conditions on is measurable rather than degenerate. Panel 1 of 2.

Amplitude	Est. gap	Acc@1	Acc@5
0.1	0.11	0.0%	2.3%
0.2	0.11	2.3%	18.2%
0.3	0.12	15.9%	31.8%
0.4	0.12	27.3%	45.5%
0.5	0.13	47.7%	70.5%
0.6	0.14	65.9%	84.1%
0.7	0.15	86.4%	95.5%
0.8	0.17	93.2%	100.0%
0.9	0.21	100.0%	100.0%
1.0	0.25	100.0%	100.0%

Table 17: Continuation of Table 16, panel 2 of 2.

Amplitude	Root claim
0.1	0.0%
0.2	0.0%
0.3	0.0%
0.4	0.0%

Continued on next page

Table 17: Continued.

Amplitude	Root claim
0.5	0.0%
0.6	20.5%
0.7	88.6%
0.8	93.2%
0.9	90.9%
1.0	90.9%

Recovery degrades gracefully rather than collapsing abruptly (Table 16). AC*@1 falls from 100.0% at unit amplitude to 47.7% at amplitude 0.5 and 0.0% at 0.1, while the reporting rule issues no named root at all below amplitude 0.6. This is a ranking and claim-rate sweep, not a threshold-crossing power experiment. It therefore does not validate Proposition 3. The estimated gap rises from 0.11 to 0.25, and its association with recovery is a useful diagnostic. Testing Theorem 3 would additionally require an independently valid bound on the estimation error. A monotone estimated gap alone is insufficient.

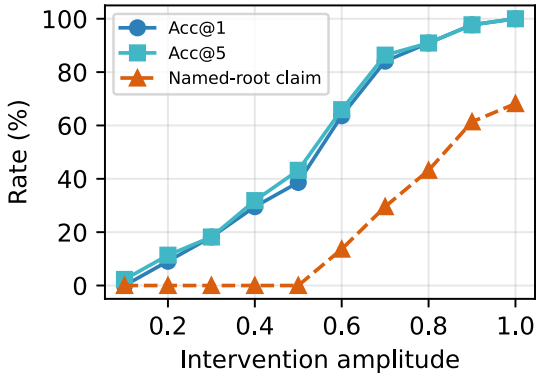

Figure 9: Intervention-amplitude sweep on `linear_anchor`. Exact-atom recovery and the rate of named-root reports generally increase with the intervention amplitude, and thus with the true action separation.

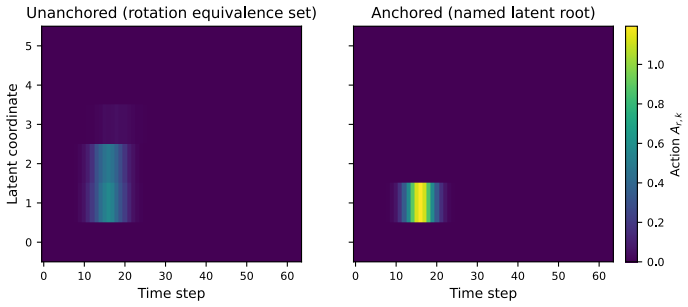

Figure 10: Unanchored and anchored action maps. Without coordinate anchors the correct report is a latent equivalence set. Anchors allow a named latent-root claim when the action and the observed evidence are aligned.

Figures 9 and 10 plot the amplitude sweep and the effect of anchoring on the action map. The latter is the visual form of the claim-level hierarchy, since without anchors the correct report is an equivalence set.

Table 18: Stress tests of the reporting conditions (9 regimes × 3 seeds, 324 incidents). "Named prec." is the fraction of issued named roots that are correct and is reported next to "Rule viol." because the two differ sharply. Obs-lat. R^2 is the median raw evidence-alignment diagnostic, reported but never used as a per-incident veto. It is not divided by an attainable-variance estimate. The separate operator study does not supply a universal ceiling for this table. Panel 1 of 3.

Regime	Level	Root claim	Equiv.
anchor upgrade	3.89	94.4%	5.6%
downstream symptom dominance	3.72	86.1%	13.9%
weak observability	2.00	0.0%	100.0%
linear anchor	3.72	86.1%	13.9%
multiple roots	3.44	72.2%	27.8%
nonidentifiable rotation	2.00	0.0%	100.0%
nonlinear decoder	3.89	94.4%	5.6%
tep simulator	3.83	91.7%	8.3%
weak action gap	3.78	88.9%	11.1%

Table 19: Continuation of Table 18, panel 2 of 3.

Regime	Named prec.	Set cov.
anchor upgrade	100.0%	–
downstream symptom dominance	100.0%	–
weak observability	–	–
linear anchor	100.0%	–
multiple roots	100.0%	–
nonidentifiable rotation	–	–
nonlinear decoder	100.0%	–
tep simulator	100.0%	–
weak action gap	68.8%	–

Table 20: Continuation of Table 18, panel 3 of 3.

Regime	Rule viol.	Obs-lat. R^2
anchor upgrade	0.0%	0.58
downstream symptom dominance	0.0%	0.57
weak observability	0.0%	0.44
linear anchor	0.0%	0.57
multiple roots	0.0%	0.57
nonidentifiable rotation	0.0%	0.46
nonlinear decoder	0.0%	0.58
tep simulator	0.0%	0.56
weak action gap	0.0%	0.54

Table 18 reports the stress suite under the structured operator Section 4.4 with alignment recorded as a diagnostic. Without coordinate anchors (rotation ambiguity and weak observability) the point-root-claim rate is exactly 0.0% and the method returns a latent equivalence set on 100% of incidents. The representation condition alone enforces this, including when the alignment condition is removed. Under anchored settings the method reports a specific root on 72.2–94.4% of incidents at 100% named-root precision, with a 0.0% rule-violation rate throughout.

We report named-root precision next to the rule-violation rate deliberately, because `weak_action_gap` separates them. The method reports a root on 88.9% of incidents with no rule violations, yet only 68.8% of the named roots are correct. This is an honest negative result

about how the action-gap threshold is calibrated. The gap statistic itself separates the regimes. Its median here is 0.122 against 0.21 in cleanly anchored ones, but the threshold is calibrated per regime as a lower quantile of the same held-out gap distribution it screens. Because every incident here has a weak gap, the threshold adapts down to 0.011 (against 0.04 in anchored regimes) and stops excluding anything. A self-referential calibration cannot flag a regime that is uniformly hard. Screening it requires an absolute floor on the gap, or a threshold calibrated on nominal windows where no intervention is present, which we have not implemented. A zero rule-violation rate therefore records compliance with the implemented rules. It does not establish that the reported roots are correct.

H.4 Paired solver comparison

Paired one-pass minus solver differences are averaged within regime–seed cells and bootstrapped in seed blocks jointly across regimes. The analysis uses 2880 paired incidents and 10 seed clusters, giving a mean atom-accuracy difference of 0.28 percentage points and a 90% interval of $[0.07, 0.49]$. The interval lies within a ± 1-point equivalence margin chosen after the aggregate results were known. This supports comparable average accuracy at these operating points with finite-cluster uncertainty. It is not a comparison with a margin fixed in advance. The independent replication in Appendix K remains specified but unrun.

H.5 Runtime of the internal solver ablation

Table 21: **What the iteration count costs.** Per-incident scoring time by arm, timed sequentially against one shared trained model on identical incidents, with a warm-up pass discarded. Model fitting is excluded because it is shared by every arm and would swamp the quantity under test. Panel 1 of 2.

Arm	Iter.	Median (ms)	p90 (ms)
innovation: encode, difference, normalise (shipped)	0	0.43	0.81
+ pulse window mask (pre-registered, superseded)	0	0.65	0.91
pulse window, no diffusion metric	0	0.46	0.78
pulse window + Gaussian smoother	0	3.26	3.65
latent scan / CUSUM	0	1.28	1.74
latent deviation (no drift subtraction)	0	0.15	0.18
coarse prepass only (50 steps)	50	74.97	88.09
500 solver steps, no prepass	500	902.61	1064.25
coarse prepass + 500 solver steps	550	983.76	1151.60

Table 22: Continuation of Table 21, panel 2 of 2.

Arm	× one-pass
innovation: encode, difference, normalise (shipped)	1.0×
+ pulse window mask (pre-registered, superseded)	1.5×
pulse window, no diffusion metric	1.1×
pulse window + Gaussian smoother	7.5×
latent scan / CUSUM	3.0×

Continued on next page

Table 22: Continued.

Arm	× one-pass
latent deviation (no drift subtraction)	0.34×
coarse prepass only (50 steps)	174×
500 solver steps, no prepass	2090×
coarse prepass + 500 solver steps	2278×

Timings come from a separate single-process run with a shared fitted model and identical incidents, rather than concurrent benchmark-arm durations. They compare the implemented objectives, not computational optimality against FDI.

H.6 Residual-evidence alignment

The observed-residual layer LARCA-Obs (LARCA in observed coordinates) calibrates one-step sensor residuals on a chronologically held-out fault-free tail to obtain conformal surprisal $E^{\mathrm{obs}}_{i,k}$. Here C_i denotes the ith row of the linear decoder, e_r selects latent coordinate r, and $\delta_{k\tau}$ restricts the comparison to the same transition. Before per-sensor calibration, predicted evidence is $\widehat{E}_{i,k} = \sum_{r,\tau} K^{r,\tau}_{i,k} \widehat{A}_{r,\tau}$.

For a linear decoder, the diagnostic kernel is

$$K^{r,\tau}_{i,k} = |C_i e_r|^2 \delta_{k\tau}. \tag{11}$$

For a nonlinear decoder, replace C_i by row i of its Jacobian evaluated on the reconstructed path. Each sensor's predicted evidence is a nonnegative scale of $\sum_{r,\tau} K^{r,\tau}_{i,k} \widehat{A}_{r,\tau}$ plus a nonnegative intercept, fitted on calibration incidents. The kernel shape is specified by the nominal model. It describes residual evidence, not raw physical propagation: one-step prediction removes motion explained by the nominal dynamics. Squared sensitivity matches the quadratic action's units. It is not established as empirically superior to linear sensitivity. Alignment is reported as an R^2 diagnostic and determines eligibility for gap calibration, rather than acting as a per-incident veto. The earlier solver-based operator ablations are in the archival supplement.

H.7 Selective prediction with a three-way split

Table 23: Named-root precision as a function of coverage, ranked by the confidence signal (synthetic: exact-atom AC*@1, SWaT: coordinate AC@1). This is the in-sample risk–coverage curve, with threshold chosen and evaluated on the same pool. The held-out counterpart is Table 24.

Coverage	Synthetic Acc@1	SWaT AC@1
100% (full coverage)	53.9%	17.0%
80%	65.9%	20.0%
60%	83.7%	22.5%
40%	97.7%	30.0%
30%	97.0%	36.7%
20%	100.0%	45.0%

Table 24: Selective reporting on **one predeclared** split (seed 0, never tuned), δ=0.05. The threshold is the target-coverage quantile of the confidence on the selection portion, applied to a disjoint validation portion. Realised coverage differs from target because a selection-portion quantile is applied to a finite validation portion. The synthetic three-way result satisfies Theorem 2. SWaT remains validation evidence because its confidence-family choice used validation labels. Panel 1 of 2.

Target cov.	Three-way realised cov.	Three-way error bound
40%	47%	10.9% (3/69)
30%	35%	11.6% (2/52)
25%	30%	10.3% (1/44)
20%	23%	8.4% (0/34)

Table 25: Continuation of Table 24, panel 2 of 2.

Target cov.	Two-way (superseded) realised cov.	Two-way (superseded) error bound
40%	37%	9.3% (3/81)
30%	24%	5.5% (0/53)
25%	22%	5.9% (0/49)
20%	17%	7.6% (0/38)

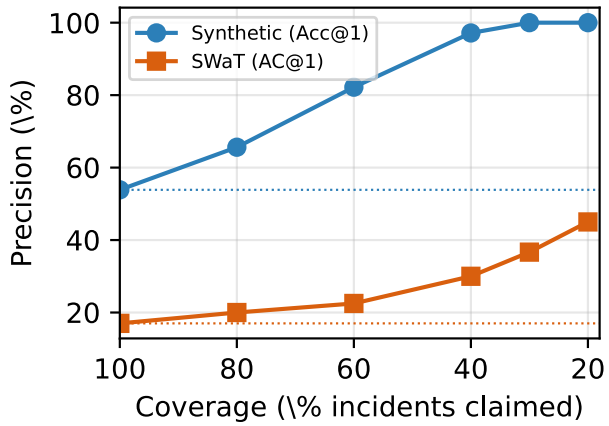

Figure 11: Selective-prediction risk–coverage. Reporting roots only above a domain-appropriate confidence threshold (standardised action magnitude on synthetic, ranking decisiveness on SWaT) raises named-root precision on both benchmarks. Dotted lines are the full-coverage baselines. These curves are in-sample, tabulated in Table 23. Table 24 gives the held-out comparison. Only the synthetic three-way result satisfies Theorem 2.

Figure 11 plots the in-sample error–coverage curves. Their held-out counterparts appear in Table 24, with the SWaT qualification stated above.

Development data choose the confidence family from five candidates. Selection data fix the threshold. Validation data count accepted errors. The synthetic family is standardised action magnitude (development AUC 0.955). Only the three-way synthetic rows in Table 24 satisfy Theorem 2. The SWaT family used validation labels, so its bounds remain descriptive. MSDS confidence scores do not usefully separate reliable roots. The earlier split construction and its correction are documented in the archival supplement.

H.8 MATCHED TUNING BUDGET FOR THE LEARNED BASELINE

A comparison against a baseline held at one fixed configuration is confounded by that choice, and the confound runs in our favour by default. LARCA carries five hyperparameters set in its configuration file, while the baseline was run at a single setting. That setting also matches none of the upstream author defaults exactly, using the same window and depth for every regime and a learning rate ten times the released one.

	upstream default	ours
learning rate η	10^{-4} (all releases)	10^{-3}
hidden layers L	1 linear, 8 nonlinear	2
window w	1 linear, 5 nonlinear	1

We therefore gave the baseline an explicit tuning budget under a protocol fixed before any evaluation data was examined. Six configurations were searched, spanning $\eta \in \{10^{-4}, 10^{-3}\}$ and $w \in \{1, 5\}$ at L=2 and including both upstream defaults verbatim. Selection used a development split of two seeds disjoint from every reported seed, on three regimes that discriminate (`linear_anchor`, `nonlinear_decoder`, `lorenz96`). The rank-deficient and weak-observability regimes are at chance for all methods and cannot separate configurations. The primary endpoint was declared in advance as pooled AC*@1. Every configuration received an identical budget of 250 epochs under the baseline's own patience-20 early stopping, and we record whether each arm stopped on its own or exhausted that budget.

Table 26: Baseline tuning search: 6 configurations $\times$ 3 regimes $\times$ 2 development seeds, 32 incidents per cell. AC*@1 subscripts are standard errors clustered on (regime, seed), since incidents within a cell share one trained model. Δ is the paired difference against the shipped setting over the six cells. Panel 1 of 2.

AERCA configuration	AC*@1	Time-Acc
w=1, L=2, η=10^{-3} **(shipped)**	$77.6_{\pm 0.5}$	45.3
w=1, L=2, η=10^{-4}	$77.6_{\pm 0.5}$	40.6
w=1, L=1, η=10^{-4} (upstream linear)	$77.6_{\pm 1.0}$	40.6
w=5, L=2, η=10^{-3}	$77.1_{\pm 1.0}$	45.3
w=5, L=8, η=10^{-4} (upstream nonlinear)	$75.5_{\pm 0.5}$	47.4
w=5, L=2, η=10^{-4}	$69.3_{\pm 3.2}$	40.6

Table 27: Continuation of Table 26, panel 2 of 2.

AERCA configuration	ΔAC*@1	Budget
w=1, L=2, η=10^{-3} **(shipped)**	–	converged
w=1, L=2, η=10^{-4}	$+0.00 \pm 0.81$	converged
w=1, L=1, η=10^{-4} (upstream linear)	$+0.00 \pm 0.81$	exhausted
w=5, L=2, η=10^{-3}	-0.52 ± 1.49	converged
w=5, L=8, η=10^{-4} (upstream nonlinear)	-2.08 ± 1.04	converged
w=5, L=2, η=10^{-4}	-8.33 ± 3.49	converged

No configuration beat the shipped setting on the primary endpoint. Three arms tie at 77.6% and the remainder are worse, two of them significantly so. The window-5 arms trade the two metrics against each other rather than dominating. The w=5, L=8 arm attains the best Time-Acc (47.4% against 45.3%) but its advantage there is not significant (+2.1 pp, 95% CI $[-1.3, +5.5]$) while its

deficit on the primary endpoint is (-2.1 pp, CI $[-4.1, -0.04]$). Lowering the learning rate to the released 10^{-4} either ties, at w=1, or costs 8.3 points, at w=5, so our one deliberate deviation from the upstream defaults is supported rather than merely inherited.

Two limitations belong with this result. First, one arm, the upstream linear default, exhausted its epoch budget in all six cells and never triggered early stopping, so its numbers are a lower bound and its ceiling is unmeasured. It ties on the primary endpoint and is significantly worse on Time-Acc (-4.7 pp, CI $[-7.3, -2.1]$) even so. Second, the search covers the learning rate, window and depth, not the loss weights or the reconstruction thresholds. The claim this supports is that no searched configuration *beat* the shipped setting, not that the shipped setting is optimal.

H.9 Declared hyperparameters

All results use a single configuration per method, fixed across every regime, seed and incident, with no per-regime or per-incident adaptation. The deployed LARCA estimator uses diffusion normalisation, no temporal mask, no smoothing, and zero optimiser steps (Appendix C). The separate pulse-window ablation uses a half-width of four steps. The solver arms of Table 1, reported for comparison only, additionally use action weight λ_A=0.05, sparsity weight λ_1=0.001, learning rate 10^{-2} and warm initialisation. The latent nominal model is trained for 250 epochs at learning rate 10^{-3} with hidden width 128 and lag window 3. The reporting rule layer uses an action-gap threshold of 0.1, anchor stability 0.9 and anchor condition bound 10^4. The baseline uses window 1, two hidden layers of width 50, learning rate 10^{-3} and an epoch cap of 5000 under its own early stopping, which is the configuration Appendix H.8 evaluates against five alternatives. Data generation uses σ_{scale}=1.3 and incident amplitude 0.7 for the fixed-difficulty tables, with $\Delta t = 1/64$ and $\sigma^2 = 1.3$ at the benchmark operating point used by (8). These values were inherited from development on earlier pilot regimes rather than selected by a search on the reported benchmark. The budgeted search of Appendix H.8 is the only hyperparameter search whose protocol was fixed in advance, and it was applied to the baseline.

I Implementation, reproducibility and claim rules

Table and figure provenance. The repository table map records upstream generators and stored artefacts for both the active paper and archival supplement, identifying each estimator. The compact main comparison and the two separated appendix tables are generated by `iclr/scripts/make_presentation.py` from the full headline and runtime tables. The same script builds Fig. 3 from the matched scan seed means, per-regime table and physical summary. Its `figures-presentation/provenance.json` records source hashes and plotted values. The matched validation tables are generated by `make_review2_tables.py`. Their protocols are in Appendix J.

Every run writes its resolved configuration, manifest, aggregate and per-incident metrics, runtime, reporting decisions, raw action and evidence arrays, and standard output and error. The manifest records the commit, working-tree state, command, seed, data source, hardware, software versions and timestamps. Each workflow is driven by a versioned configuration file and a single entry point, and paper assets are derived only from raw outputs. A missing asset compiles as a labelled placeholder, which prevents an old or hand-entered value from being mistaken for a current one. The archived operator ablation is NumPy-only and regenerates its tables and figures from a self-contained script.

Claim rules. The source of d is stored for every dataset. Synthetic OU systems use a predeclared construction, simulator studies use selected physical state or subsystem dimensions, and real testbeds use documented process groupings and never tune d with incident labels. Latent-root accuracy is reported only where simulator states, injected latent inputs or equivalent external labels identify the target. A direct sensor fault is a negative control. MSDS and ordinary operational traces support observed event, delay, stability or documented subsystem metrics, and do not support hidden coordinate accuracy without additional anchors. The reporting conditions proceed from observed calibration and latent recovery to coordinate identification and stability, action separation, and counterfactual coverage. A later score cannot compensate for a failed earlier condition. Evidence alignment determines eligibility for held-out action-gap calibration. It remains a diagnostic

for each incident rather than a veto, for the power reasons in Section 4.4. These operational checks alone do not supply a statistical error guarantee, as the weak-gap stress case demonstrates.

J MATCHED VALIDATION PROTOCOLS

Reported runs and fixed decisions. The protocols below describe the matched scan study, finite-path check and physical benchmark reported in the paper. Full-run artefacts supply their measurements. Smoke checks establish implementation behaviour only. Failures and parameter changes remain in the run record. The solver-equivalence analysis uses a one-percentage-point margin, with half-point and two-point sensitivity analyses. That margin was chosen after the existing aggregate results were known. The independent comparison proposed in Appendix K would fix the margin before generating new test data.

Separate detection and recovery events. At $d = 6$, $K = 64$, $\Delta t = 1/64$ and $\sigma^2 = 1.3$, predicted 90% detection amplitudes are 0.728 for an atom and 0.798 for onset. Gaussian simulations attain the specified scan error and power, while top-1 ranking and exact support have different success rates. Pooling heterogeneous variances miscalibrates null coordinates and reduces power. The separate solver-based spread study remains in the archival supplement.

The Gaussian-transition check samples a single intervention and records $T_{m^\star} > q_\alpha$, any null threshold exceedance, exact thresholded support, and top-1 atom accuracy separately. The onset check uses $S_k = \sum_j T_{j,k}$ and q_α^{time}, recording onset threshold crossing and onset top-1 separately. Ten independent seeds and $10\,000$ windows per seed assess target power and null family-wise error. Under heterogeneous variance, each root uses its own variance in the amplitude prediction. Common noise draws pair the amplitude conditions.

The learned-representation check uses the existing anchored OU generator, its nonlinear observation variant, and heterogeneous diffusion. Each of ten seeds has 256 nominal training windows, 128 disjoint nominal evaluation windows, and 64 incidents per amplitude. The model is fitted once per seed and regime. The amplitude grid includes the homogeneous reference predictions, avoiding interpolation. In Table 28, every root is evaluated at the same amplitude $a_0 = 0.728$, computed using the reference variance 1.3. For a heterogeneous root j, its predicted power is instead

$$p_j(a_0) = \Pr\left[\chi^2_1\left(\frac{a_0^2}{\sigma_j^2 \Delta t}\right) > q_\alpha\right].$$

The table averages these probabilities over the planted roots. Its heterogeneous oracle power is consequently 0.748, rather than 0.900. Using a root-specific amplitude $a_{\mathrm{detect},j} = \sigma_j \sqrt{\Delta t \lambda_{\alpha,\beta}}$ would give the 90% target for each root, but that is not the common-amplitude comparison tabulated here. Oracle scores test the corresponding predicted law. Encoded scores measure the approximation discrepancy. Multiple roots and planted distractors are excluded from this single-atom theorem check.

Table 28: Matched scan events at the common homogeneous-reference amplitude $a_0 = 0.728$. Heterogeneous roots use this same amplitude, with oracle power calculated from each root's variance and then averaged. The target is 0.900 for the homogeneous rows and 0.748 for the heterogeneous mixture. Plug-in and ranking rows are separate diagnostics. Panel 1 of 2.

Regime	Score	Event	Oracle power
linear anchor	oracle	atom detection	0.900
linear anchor	oracle	exact support	–
linear anchor	oracle	atom top1	–
linear anchor	innovation	atom detection	0.900
linear anchor	innovation	exact support	–
linear anchor	innovation	atom top1	–
nonlinear decoder	oracle	atom detection	0.900
nonlinear decoder	oracle	exact support	–

Continued on next page

Table 28: Continued.

Regime	Score	Event	Oracle power
nonlinear decoder	oracle	atom top1	–
nonlinear decoder	innovation	atom detection	0.900
nonlinear decoder	innovation	exact support	–
nonlinear decoder	innovation	atom top1	–
heteroscedastic noise	oracle	atom detection	0.748
heteroscedastic noise	oracle	exact support	–
heteroscedastic noise	oracle	atom top1	–
heteroscedastic noise	innovation	atom detection	0.748
heteroscedastic noise	innovation	exact support	–
heteroscedastic noise	innovation	atom top1	–

Table 29: Continuation of Table 28, panel 2 of 2.

Regime	Score	Measured (95% CI)
linear anchor	oracle	0.902 [0.880, 0.922]
linear anchor	oracle	0.872 [0.852, 0.892]
linear anchor	oracle	0.970 [0.961, 0.980]
linear anchor	innovation	0.905 [0.884, 0.922]
linear anchor	innovation	0.791 [0.762, 0.817]
linear anchor	innovation	0.931 [0.908, 0.955]
nonlinear decoder	oracle	0.902 [0.880, 0.922]
nonlinear decoder	oracle	0.872 [0.852, 0.892]
nonlinear decoder	oracle	0.970 [0.961, 0.980]
nonlinear decoder	innovation	0.881 [0.859, 0.902]
nonlinear decoder	innovation	0.744 [0.714, 0.770]
nonlinear decoder	innovation	0.948 [0.931, 0.963]
heteroscedastic noise	oracle	0.752 [0.722, 0.783]
heteroscedastic noise	oracle	0.728 [0.697, 0.759]
heteroscedastic noise	oracle	0.869 [0.847, 0.887]
heteroscedastic noise	innovation	0.769 [0.742, 0.795]

Continued on next page

Table 29: Continued.

Regime	Score	Measured (95% CI)
heteroscedastic noise	innovation	0.675 [0.645, 0.703]
heteroscedastic noise	innovation	0.844 [0.816, 0.870]

Finite-path plug-in check. The check uses a positive diagonal map fitted on nominal paths solely to put the oracle diagnostic in the deployed coordinate units. Both excess increments and diffusion are transformed consistently, preserving each deployed action. For every incident it computes the maximum excess increment error and the actual maximum relative diffusion error. It marks the relative-error theorem unavailable when $\rho \geq 1$, without clipping. A separate direct bound, valid for positive variances, is

$$|\widehat{A}_m - A_m| \leq \frac{2|r_m|\varepsilon_r + \varepsilon_r^2}{2\widehat{\sigma}_j^2 \Delta t_k} + \frac{r_m^2}{2\Delta t_k} \left| \frac{1}{\widehat{\sigma}_j^2} - \frac{1}{\sigma_j^2} \right|.$$

The check reports full atom agreement and verifies the implication whenever the measured gap exceeds the bound. Because it uses true incident paths, it is a finite-path verification and cannot be used as a deployment confidence score.

Across $6{,}720$ incidents, no cases satisfying the bound disagree with the oracle argmax. The relative-error condition $\rho < 1$ holds throughout this suite. At the largest amplitude the argmax agrees with the oracle on 100.0% of incidents, while only 78.6% satisfy the sufficient condition. This fraction plateaus rather than approaching one. The deterministic worst-case bound therefore leaves many correct rankings unexplained, as expected for a sufficient condition. A probabilistic refinement remains open.

Table 30: Finite-path verification over $6{,}720$ incidents, 3 regimes $\times$ ten seeds. The fraction satisfying the bound is an oracle diagnostic, distinct from the statistical selective-error guarantee. Panel 1 of 2.

Amplitude	n	$\rho < 1$ (%)	Bound holds (%)
0.4	960	100.0	0.0
0.7	960	100.0	6.1
1	960	100.0	45.3
1.5	960	100.0	69.3
2	960	100.0	73.2
3	960	100.0	76.6
4	960	100.0	78.6

Table 31: Continuation of Table 30, panel 2 of 2.

Amplitude	Atom agreement (%)	Violations
0.4	56.4	0
0.7	89.5	0
1	98.9	0
1.5	100.0	0
2	100.0	0
3	100.0	0
4	100.0	0

Replication across computing environments. The same code and seeds were run on a workstation GPU with CUDA `torch` and as cluster array tasks with CPU `torch`. Over $6{,}720$ paired

incidents, the two runs make the same argmax-agreement decision 98.4% of the time. The largest difference in any amplitude cell is 0.2 percentage points for agreement and 1.5 points for the fraction satisfying the bound. The matched scan also yields 98.9% identical decisions over $69{,}120$ paired rows, with exact agreement in oracle columns that do not use a fitted model. These results suggest that the sufficient-condition frequency is more sensitive to numerical fitting differences than the ranking agreement on this suite.

Physical state semantics from a published process. The hydraulic benchmark uses Johansson's quadruple-tank mass balances and published tank dimensions (Johansson, 2000). Let h_i be water level, A_i tank area, a_i outlet area, and v_1, v_2 measured pump voltages. For example,

$$\dot{h}_1 = -\frac{a_1}{A_1}\sqrt{2gh_1} + \frac{a_3}{A_1}\sqrt{2gh_3} + \frac{\gamma_1 k_1}{A_1}v_1, \qquad \dot{h}_3 = -\frac{a_3}{A_3}\sqrt{2gh_3} + \frac{(1-\gamma_2)k_2}{A_3}v_2.$$

The other two balances exchange the paired indices and pump splits. Thus a root names a physical tank, independently of a learned representation. The stochastic extension, outlet-flow proxy sensors and fault protocol are introduced for this study. The experiment is a simulation derived from published physics, not a new laboratory dataset or a reproduction of the original sensor arrangement.

Lower-tank levels are measured with noise. Upper-tank states are inferred from noisy outlet-flow proxies using declared sensor calibration. A flow fault injects a known volume into one tank at an externally chosen time. We compare one-step pulses and eight-step disturbances at equal total injected volume. The volume grid is $5, 10, 20, 40, 80$ cm^3. Minimum-phase and nonminimum-phase valve settings, weakened upper proxies, and a held-out pump-voltage shift test different failure modes. The process is integrated at 0.1 s and scored at 1 s, so the score does not receive the exact simulator increments on its own grid.

Each seed has 128 nominal training episodes, 199 calibration episodes and 80 test episodes per condition, balanced over the four roots. All feasible methods receive the same observations, pump inputs and sensor calibration. A drift regressor fits square-root state features from nominal observations only. Comparisons include the normalised forcing score, Euclidean forcing score, two-sided CUSUM, nominal state deviation, a residual using the known hydraulic drift, and a labelled true-state oracle. Learned and known-physics arms estimate effective excess increment variance from nominal observations. They do not receive latent training labels or incident labels.

The primary endpoint is exact coordinate–onset accuracy at a pulse volume of 20 cm^3, over all four tank roots in each configuration. Table 32 reports that endpoint for every configuration, with an equal-weight summary across configurations explicitly marked as descriptive. Full curves, upper versus lower roots, coordinate accuracy, support hits, onset error, nominal false alarms and runtime are also reported. For sustained faults, a support hit and an onset hit are different events. Max-score thresholds use the finite-sample nominal order statistic for family-wise level $.05$. Its exchangeability interpretation applies to the matched operating condition, not the shifted one. Intervals resample seeds, and method contrasts preserve incident pairing.

Table 32: **Prespecified hydraulic primary endpoint.** Exact tank–time recovery at $20\,\text{cm}^3$ over all four roots, in percent. Each configuration uses ten seeds and 80 incidents per seed and arm. Configuration columns report the original endpoints. The final column is a descriptive equal-weight mean with a 95% seed-block bootstrap interval, preserving each seed across configurations. Intervals describe variation across fitted-model seeds. A degenerate interval at 100% does not imply perfect population recovery. Panel 1 of 2.

Method	Min. phase	Nonmin.	Weak proxies	Shift
CUSUM	46.0	49.6	26.4	44.2
Euclidean forcing	**100.0**	**100.0**	**97.5**	**100.0**
One-pass forcing test	**100.0**	**100.0**	97.0	**100.0**
Known-physics residual	**100.0**	**100.0**	97.0	**100.0**
State deviation	25.0	10.1	13.0	21.9
True-state oracle	100.0	100.0	100.0	100.0

Table 33: Continuation of Table 32, panel 2 of 2.

Method	Mean [95% CI]
CUSUM	41.56 [38.97, 44.53]
Euclidean forcing	**99.38** [99.19, 99.56]
One-pass forcing test	99.25 [98.97, 99.50]
Known-physics residual	99.25 [99.03, 99.50]
State deviation	17.50 [16.19, 18.81]
True-state oracle	100.00 [100.00, 100.00]

Table 34: Quadruple-tank dose–response, all 4 plant configurations, upper tanks. Figure 3c visualises the two proxy settings shown in Table 37. Panel 1 of 3.

Scenario	Arm	5	10
minimum phase	cusum	24	43
minimum phase	euclidean	66	100
minimum phase	innovation	70	100
minimum phase	oracle physics	82	100
minimum phase	physics residual	70	100
minimum phase	state deviation	4	14
nonminimum phase	cusum	24	48
nonminimum phase	euclidean	42	99
nonminimum phase	innovation	45	98
nonminimum phase	oracle physics	82	100
nonminimum phase	physics residual	44	99
nonminimum phase	state deviation	2	3
operating shift	cusum	23	43
operating shift	euclidean	61	100
operating shift	innovation	66	100
operating shift	oracle physics	82	100
operating shift	physics residual	67	100
operating shift	state deviation	3	8
weak upper proxies	cusum	0	3
weak upper proxies	euclidean	8	38
weak upper proxies	innovation	7	36
weak upper proxies	oracle physics	82	100
weak upper proxies	physics residual	5	34
weak upper proxies	state deviation	2	8

Table 35: Continuation of Table 34, panel 2 of 3.

Scenario	Arm	20	40
minimum phase	cusum	36	27
minimum phase	euclidean	100	100
minimum phase	innovation	100	100
minimum phase	oracle physics	100	100
minimum phase	physics residual	100	100
minimum phase	state deviation	46	72
nonminimum phase	cusum	43	34
nonminimum phase	euclidean	100	100
nonminimum phase	innovation	100	100
nonminimum phase	oracle physics	100	100
nonminimum phase	physics residual	100	100
nonminimum phase	state deviation	14	43
operating shift	cusum	38	28
operating shift	euclidean	100	100
operating shift	innovation	100	100
operating shift	oracle physics	100	100
operating shift	physics residual	100	100
operating shift	state deviation	30	67
weak upper proxies	cusum	1	0
weak upper proxies	euclidean	95	100
weak upper proxies	innovation	94	100
weak upper proxies	oracle physics	100	100
weak upper proxies	physics residual	94	100
weak upper proxies	state deviation	24	41

Table 36: Continuation of Table 34, panel 3 of 3.

Scenario	Arm	80
minimum phase	cusum	10
minimum phase	euclidean	100
minimum phase	innovation	100
minimum phase	oracle physics	100
minimum phase	physics residual	100
minimum phase	state deviation	88
nonminimum phase	cusum	18
nonminimum phase	euclidean	100
nonminimum phase	innovation	100
nonminimum phase	oracle physics	100
nonminimum phase	physics residual	100
nonminimum phase	state deviation	68
operating shift	cusum	12
operating shift	euclidean	100
operating shift	innovation	100
operating shift	oracle physics	100
operating shift	physics residual	100
operating shift	state deviation	84

Continued on next page

Table 36: Continued.

Scenario	Arm	80
weak upper proxies	cusum	0
weak upper proxies	euclidean	100
weak upper proxies	innovation	100
weak upper proxies	oracle physics	100
weak upper proxies	physics residual	100
weak upper proxies	state deviation	50

Table 37: Quadruple-tank pulse dose–response for upper-tank roots. Columns give injected volume in cm^3 and entries give exact tank–time accuracy in percent. The oracle receives true states. The two scenarios vary proxy informativeness. Panel 1 of 3.

Scenario	Arm	5	10
minimum phase	cusum	24	43
minimum phase	euclidean	66	100
minimum phase	innovation	70	100
minimum phase	oracle physics	82	100
minimum phase	physics residual	70	100
minimum phase	state deviation	4	14
weak upper proxies	cusum	0	3
weak upper proxies	euclidean	8	38
weak upper proxies	innovation	7	36
weak upper proxies	oracle physics	82	100
weak upper proxies	physics residual	5	34
weak upper proxies	state deviation	2	8

Table 38: Continuation of Table 37, panel 2 of 3.

Scenario	Arm	20	40
minimum phase	cusum	36	27
minimum phase	euclidean	100	100
minimum phase	innovation	100	100
minimum phase	oracle physics	100	100
minimum phase	physics residual	100	100
minimum phase	state deviation	46	72
weak upper proxies	cusum	1	0
weak upper proxies	euclidean	95	100
weak upper proxies	innovation	94	100
weak upper proxies	oracle physics	100	100
weak upper proxies	physics residual	94	100
weak upper proxies	state deviation	24	41

Table 39: Continuation of Table 37, panel 3 of 3.

Scenario	Arm	80
minimum phase	cusum	10
minimum phase	euclidean	100
minimum phase	innovation	100

Continued on next page

Table 39: Continued.

Scenario	Arm	80
minimum phase	oracle physics	100
minimum phase	physics residual	100
minimum phase	state deviation	88
weak upper proxies	cusum	0
weak upper proxies	euclidean	100
weak upper proxies	innovation	100
weak upper proxies	oracle physics	100
weak upper proxies	physics residual	100
weak upper proxies	state deviation	50

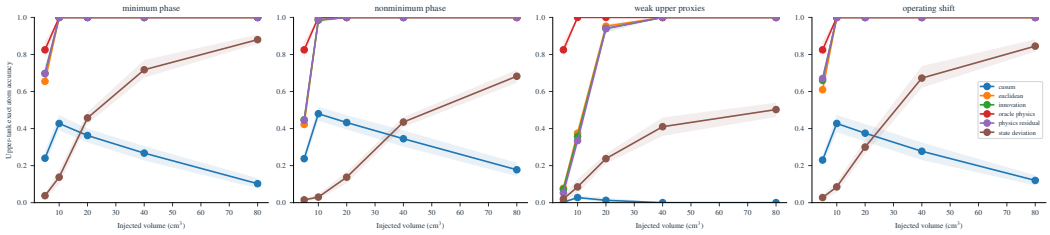

Figure 12: Exact atom accuracy for upper-tank faults across injected volumes. Upper states are inferred through outlet-flow proxies.

Table 40: Nominal alarm rates. The operating-shift row tests distribution shift and is not covered by the matched-calibration exchangeability condition.

Scenario	Arm	Null alarm rate (%)
minimum phase	cusum	5.1 [3.4, 7.0]
minimum phase	euclidean	6.5 [4.6, 8.5]
minimum phase	innovation	5.9 [4.5, 7.1]
minimum phase	oracle physics	6.1 [4.5, 7.9]
minimum phase	physics residual	5.5 [4.1, 6.9]
minimum phase	state deviation	5.1 [4.0, 6.2]
nonminimum phase	cusum	4.4 [2.9, 5.8]
nonminimum phase	euclidean	5.1 [3.5, 6.7]
nonminimum phase	innovation	6.0 [4.6, 7.4]
nonminimum phase	oracle physics	5.9 [4.5, 7.4]
nonminimum phase	physics residual	6.2 [4.8, 7.6]
nonminimum phase	state deviation	4.8 [2.9, 6.6]

Continued on next page

Table 40: Continued.

Scenario	Arm	Null alarm rate (%)
operating shift	cusum	6.4 [4.6, 8.1]
operating shift	euclidean	9.6 [6.9, 12.5]
operating shift	innovation	10.9 [9.4, 12.5]
operating shift	oracle physics	6.2 [4.6, 8.0]
operating shift	physics residual	10.0 [8.6, 11.4]
operating shift	state deviation	100.0 [100.0, 100.0]
weak upper proxies	cusum	5.6 [4.2, 6.9]
weak upper proxies	euclidean	4.9 [3.8, 6.1]
weak upper proxies	innovation	6.5 [5.0, 8.1]
weak upper proxies	oracle physics	6.1 [4.5, 7.9]
weak upper proxies	physics residual	5.4 [4.0, 6.7]
weak upper proxies	state deviation	4.5 [2.9, 6.1]

K STRESS TEST FOR RAPID ATTRIBUTION AT SCALE

Status and purpose. This appendix reports an executed study, not a plan. It asks whether the one-pass statistic keeps useful latent-root accuracy as the number of coupled states and candidate times grows, and whether the computation stays small relative to the time available for diagnosis. DynaDojo motivates separating state dimension, sample size and target error in dynamical-system evaluation (Bhamidipaty et al., 2023). CausalDynamics builds difficulty ladders from thousands of coupled ordinary and stochastic differential equations with confounding and lags (Herdeanu et al., 2025). CausalRivers argues for scaling causal benchmarks on real networked infrastructure (Stein et al., 2025), and PowerGraph illustrates why propagated faults and system-level explanations matter in critical networks (Varbella et al., 2024). The study extends our hydraulic setting along those axes. It does not evaluate any of those benchmarks and establishes nothing about electrical-grid or river data.

The complete grid is 1189 shards, 2,359,840 scored incident–arm rows and 401,920 incidents, run on two identical A100 compute nodes. Generation, scoring, aggregation and the resource accounting are in `scripts/run_scale_stress.py`, `scripts/plan_scale_stress.py` and `scripts/aggregate_scale_stress.py`. The frozen protocol is `configs/scale_stress.yaml`. Physics, estimator and protocol checks are in `tests/test_scale_stress.py`.

A network with physical state names. We connect M quadruple-tank modules through declared, bounded-degree hydraulic transfer links. Every level remains a modelled state, giving $d = 4M$ coordinates. All nominal pump inputs are measured. The mass balance for tank i is

$$S_i\, dh_i = \left[q_i^{\mathrm{nom}}(h, v) + q_i^{\mathrm{fault}}\right] dt + S_i \sigma_i\, dW_i, \qquad u_i = q_i^{\mathrm{fault}}/S_i,$$

where S_i is tank area and the signed nominal inflow includes every declared transfer. Each link removes from its source exactly the volume added at its destination, so the routing matrix has unit column sums on every diverted tank and mass balance holds identically rather than approximately.

With one module and no links the drift and the equilibrium reduce exactly to Johansson's published plant, which is asserted in the test suite.

Link layouts, operating ranges and physical parameters are frozen before evaluation, from a layout seed disjoint from every data stream, so changing the size cannot re-roll the plant. Because a cascaded tank accumulates its parents' inflow, each drain is sized so that the steady level of every tank lies in a declared 8–14 cm band. The steady flows solve $(I - W)q =$ pump inflow exactly. A layout is admitted using only nominal quantities and that feasibility check. Integration uses four substeps per sampling interval, finer than the grid the estimator scores on, so numerical mismatch is present rather than assumed away. A boundary crossing raises and is recorded as a configuration failure. 0 occurred across the reported conditions. Solver feasibility is likewise deterministic per cell rather than seed-dependent: in every cell the budgeted arm either finished within the limit for all five seeds or for none of them.

As in the single-module plant, lower levels are directly measured and upper levels have calibrated flow proxies. An extra volume enters a declared upper tank during one sampling interval. The label is its physical tank and entry time. The estimator scans all dK state–time atoms and is not told that roots are confined to upper tanks. Downstream responses follow the dynamics without being relabelled as new roots.

Two scaling axes and controlled difficulty. We use $M \in \{1, 4, 16, 64, 256, 1024\}$ and $K \in \{64, 256, 1024, 4096\}$ at a fixed sampling interval, so the largest condition carries 4096 latent states and 16,777,216 state–time cells. The two axes move separately rather than only together. Incident counts and seed counts follow an allocation ladder keyed to dK and declared in the config before any result was inspected. Each cell reports the rule it used and its realised seed count, and every planned condition is listed in the run log whether or not it completed.

At each size we evaluate two fault schedules. A fixed physical-volume schedule exposes the cost of scanning more candidates at the same disturbance size. A second schedule uses the Gaussian reference

$$V_j = S_j \sigma_j \sqrt{\Delta t \, \lambda_{\alpha,\beta}(dK)},$$

where $\lambda_{\alpha,\beta}$ is the noncentrality reaching power $\beta = 0.9$ against the exact Bonferroni per-atom threshold for dK atoms at $\alpha = 0.05$. This separates multiplicity from computational scale. The simulator uses the nominal-fit diffusion of the same seed to choose the amplitude. Feasible estimators never receive true states or the generating diffusion, and the training stream that produces the fit is disjoint from calibration and incidents. It is a design reference for the oracle Gaussian model, not a promise of exact power for the finely integrated nonlinear plant or its learned representation.

That gap is the point of the schedule rather than a weakness of it. The reference volume is derived from Bonferroni multiplicity and a noncentral chi-square power calculation on the oracle model. Whether the nonlinear plant, the proxy inversion and the fitted drift actually deliver that power is an open question the experiment answers. A biased reconstruction, an inflated variance estimate or a drift that absorbs part of the disturbance would all show up as measured accuracy far from the target. Agreement is a confirmed prediction. It is not arranged by construction, and the fixed-volume schedule reports the same statistic with no adaptation at all.

Data separation and matched methods. Twenty seeds, 1000–1019, carry disjoint named random streams for nominal fitting, model selection, calibration and incident generation, so changing the test count cannot change training. Training windows are allocated by total sample count rather than by episode count, so a longer window does not silently buy a better reconstruction. Both counts are reported per cell. The ridge is chosen on a development split that no calibration or incident stream sees, from one pass of normal equations shared across the candidate path. Calibration uses 199 disjoint nominal windows per exact (d, K) condition, and the conformal maximum threshold at $\alpha = 0.05$ takes the rank-190 value of those 199 maxima.

Every feasible arm sees the same nominal data, the same reconstruction, the same incident windows and the same tuning budget. We compare the one-pass forcing test, Euclidean forcing, a two-sided CUSUM scan of standardised innovations, state deviation, and a least-action solver at two prespecified budgets, with a true-state oracle as a labelled reference only. The solver is given every advantage: it minimises the shooting objective with an exact adjoint built from the network's analytic

sparse Jacobian, a cached transposed operator and a backtracking line search, so the comparison measures the cost of iterating rather than an implementation penalty. Its gradient is verified against finite differences of the rolled-out objective in the test suite. One accepted iteration is one forward rollout plus one adjoint sweep. Actual iterations, actual forward passes and the achieved objective reduction are recorded. Because a solver incident costs roughly two budget-length path evaluations, the two solver arms are evaluated on a prespecified subgrid with 39 calibration windows, at which the conformal threshold is still exact at $\alpha = 0.05$. The one-pass arms run on the whole grid.

Accuracy, time and resources. The primary endpoints are exact physical tank–time recovery and median end-to-end scoring latency for a batch of one incident. We also report coordinate accuracy, onset error, set recovery, support hits, the nominal family-wise alarm rate, the 95th and 99th latency percentiles, peak memory and fitting cost, and we time reconstruction, nominal-drift evaluation and action-map construction separately. The action-map stage costs $O(Kd)$ once the path and drift are available. A dense drift or encoder can dominate total cost, which is why the fitted drift here is block sparse on the declared inflow neighbourhood and is asserted to be so. Latency is measured in a separate study, never from the parallel throughput runs: one host, one shard at a time, ten discarded warm-up evaluations and thirty timed repeats on a fixed incident subset. These nodes are shared, so rather than assert an idle machine we record the host's load average beside every timing shard and report it. The median during the pass was 64.7 on a 128-core host, so the recorded load stays well below the core count and the eight-thread timing job was not competing for the last core, and the competing work ran at lower scheduling priority. Contention inflates absolute times and leaves the fitted scaling exponent, which is a ratio across cells measured under the same conditions, comparatively robust. Absolute latencies should be read as upper bounds.

One declared deviation: the timing study uses one seed per cell rather than three. It was made on cost after the grid had run and before any latency was read, not on results. Latency at a fixed (d, K) is set by the dimensions and the declared sparsity, which the plant seed does not change, and the thirty within-shard repeats already measure run-to-run variation. Solver latency is separately covered by the ladder study's timed solver incidents across five seeds. Every other study keeps its original seed allocation. A per-incident limit of sixty seconds and the device's memory bound out every arm. Timeouts and infeasibility are recorded as outcomes rather than dropped, and large conditions where the solver cannot run are reported rather than omitted.

Accuracy against dimension and window. Table 41 and panel (a) of Figure 13 give the result. At the target-power reference the exact tank–time rate stays close to its design target across the whole grid. At the largest condition it is 91.9% (95% seed-block interval 88.1–95.0), at $d = 4096$ and $K = 4096$, that is 16,777,216 candidate cells, from 5 seeds. Coordinate accuracy there is 91.9%. At a fixed physical volume the same statistic degrades as the scan grows, which is the expected multiplicity cost and not an implementation limit: holding the disturbance fixed while the atom count rises raises the per-atom threshold. The measured nominal family-wise alarm rate is 4.8% against the 5% target, so the calibration that supports these rates is not being bought by over-alarming.

Table 41: Exact physical tank–time recovery for the deployed one-pass statistic across the size grid, with 95% seed-block bootstrap intervals. The upper block holds detection difficulty fixed through the target-power volume reference. The lower block holds the physical volume fixed at 20 cm^3 and therefore pays the multiplicity cost of a larger scan. Seeds are the realised count from the declared allocation ladder. Panel 1 of 2.

States d	Seeds	$K = 64$	$K = 256$
Target-power reference			
4	20	91.7	95.0
		[90.4, 92.9]	[94.2, 95.8]
16	20	92.3	94.6
		[91.2, 93.4]	[93.8, 95.5]
64	20	94.1	94.4
		[93.3, 94.9]	[93.6, 95.2]

Continued on next page

Table 41: Continued.

States d	Seeds	$K = 64$	$K = 256$
256	20	94.2 [93.4, 95.0]	94.1 [93.5, 94.8]
1024	20	94.5 [93.6, 95.3]	94.7 [93.8, 95.5]
4096	20	95.1 [94.0, 96.1]	94.5 [93.2, 95.8]
Fixed $20\,cm^3$			
4	20	51.6 [49.8, 53.4]	42.2 [40.5, 44.0]
16	20	46.2 [44.7, 47.8]	38.4 [36.2, 40.6]
64	20	55.4 [53.5, 57.1]	49.2 [47.0, 51.2]
256	20	46.5 [44.3, 48.7]	37.5 [35.9, 39.0]
1024	20	35.0 [32.7, 37.2]	28.0 [26.6, 29.3]
4096	20	30.2 [28.2, 32.1]	25.5 [24.0, 27.2]

Table 42: Continuation of Table 41, panel 2 of 2.

States d	Seeds	$K = 1024$	$K = 4096$
Target-power reference			
4	20	95.5 [94.6, 96.4]	94.8 [93.9, 95.7]
16	20	95.9 [95.1, 96.8]	93.8 [92.6, 94.9]
64	20	94.8 [94.1, 95.4]	94.3 [93.6, 95.1]
256	20	95.2 [94.4, 95.9]	94.5 [93.2, 95.5]
1024	20	92.3 [90.7, 93.7]	95.2 [94.0, 96.7]
4096	20	94.0 [92.1, 95.6]	91.9 [88.1, 95.0]
Fixed $20\,cm^3$			
4	20	30.8 [29.4, 32.3]	22.6 [21.2, 24.1]
16	20	32.4 [30.9, 34.0]	26.2 [24.6, 27.9]
64	20	39.6 [38.0, 41.4]	33.9 [32.0, 35.4]
256	20	30.7 [28.7, 32.7]	26.3 [24.6, 28.0]
1024	20	26.5 [24.9, 28.0]	22.7 [19.8, 25.8]
4096	20	18.5 [15.2, 22.1]	19.4 [15.6, 22.5]

Cost against scanned atoms. Table 43 and panel (b) of Figure 13 report latency. Over the measured range 256–16,777,216 atoms the median end-to-end cost of the deployed statistic grows with

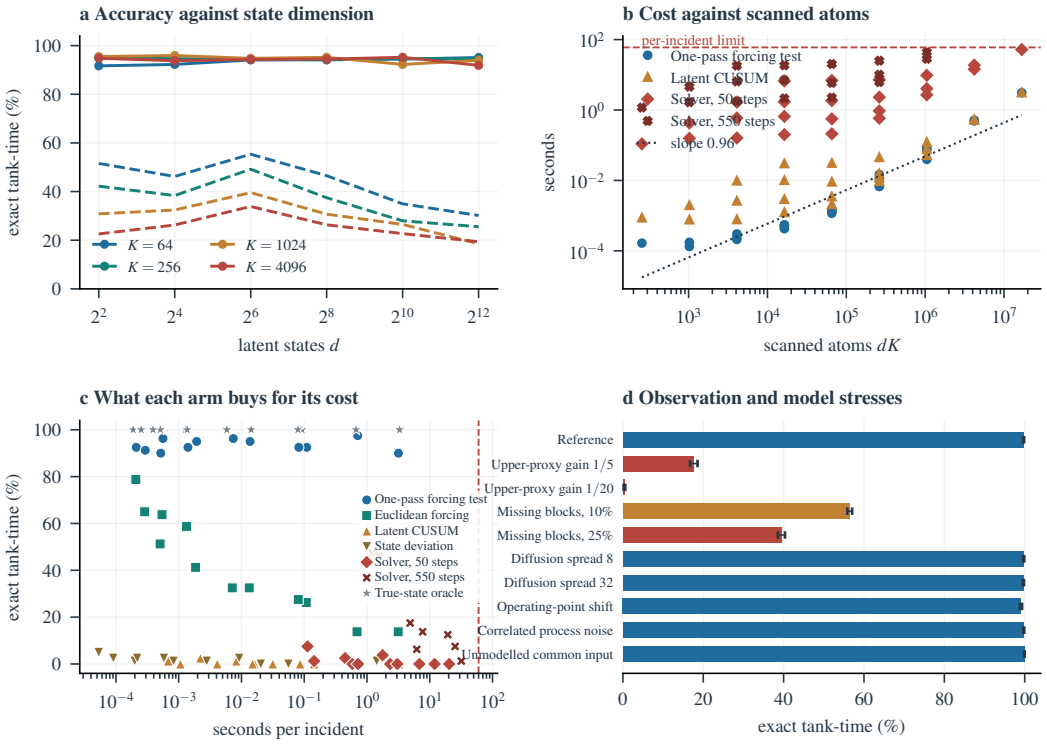

Figure 13: **Executed scale stress test.** (a) Exact tank–time recovery against state dimension, solid for the target-power reference and dashed for a fixed 20 cm^3 volume. Holding difficulty fixed separates multiplicity from scale. (b) Median scoring cost against scanned atoms dK from the sequential timing study on one host, with the fitted log–log slope and the prespecified sixty-second per-incident limit. (c) The operational frontier: accuracy against measured cost per incident, one point per arm and cell, with the same limit. (d) Exact tank–time recovery under each observation and model stress at 80 cm^3, one mechanism at a time, with seed-block intervals. Every panel is measured. None is illustrative.

a fitted log–log slope of 0.96, consistent with the $O(Kd)$ action-map term once reconstruction and drift are available. At the largest condition the median is 3,079 ms with a 99th percentile of 3,104 ms and peak resident memory 9.9 GiB. We report the observed range and do not extrapolate past the completed conditions.

The stage decomposition says where that time actually goes, and it is not where the $O(Kd)$ bookkeeping would suggest. At the largest condition the action map itself takes 2.4% of the total, against 43% for reconstruction and 55% for evaluating the nominal drift. Scanning sixteen million candidate cells is close to free. Paying for the representation the scan runs on is the cost. That is worth stating plainly because it locates the target for any future speedup, and because it is the reason the fitted drift is kept block sparse on the declared inflow neighbourhood: a dense drift would put the dominant term back to $O(Kd^2)$ and undo the scaling this appendix reports.

Table 43: Measured scoring latency for the deployed statistic, from the sequential timing study: ten discarded warm-up evaluations and thirty timed repeats per cell, one shard at a time on a single shared host whose load average is recorded beside every shard. Encode and drift columns are the share of the median spent on reconstruction and on nominal-drift evaluation. The remainder is the action map itself. Panel 1 of 2.

States d	Window K	Atoms dK	Median (ms)	p95 (ms)
4	64	256	0.17	0.18
4	256	1,024	0.13	0.15
16	64	1,024	0.17	0.19
4	1024	4,096	0.30	0.38
16	256	4,096	0.25	0.32
64	64	4,096	0.21	0.24
4	4096	16,384	0.42	0.47
16	1024	16,384	0.55	0.61
64	256	16,384	0.49	0.63
256	64	16,384	0.54	0.64
16	4096	65,536	1.16	1.25
64	1024	65,536	1.54	1.77
256	256	65,536	1.29	1.49
1024	64	65,536	1.44	1.57
64	4096	262,144	15	16
256	1024	262,144	6.76	7.69
1024	256	262,144	6.76	7.08
4096	64	262,144	11	12
256	4096	1,048,576	87	97
1024	1024	1,048,576	40	41
4096	256	1,048,576	72	78
1024	4096	4,194,304	488	491
4096	1024	4,194,304	513	520
4096	4096	16,777,216	3,079	3,092

Table 44: Continuation of Table 43, panel 2 of 2.

States d	Window K	p99 (ms)	Encode (%)	Drift (%)	Peak (GiB)
4	64	0.18	21	75	0.12
4	256	0.15	21	73	0.12
16	64	0.21	23	73	0.12
4	1024	0.45	27	66	0.12
16	256	0.42	26	66	0.14
64	64	0.37	28	67	0.16
4	4096	0.50	23	65	0.13
16	1024	0.61	27	64	0.14
64	256	0.69	29	63	0.21
256	64	0.84	36	58	0.33
16	4096	1.37	25	65	0.21
64	1024	2.99	23	71	0.21
256	256	1.50	38	53	0.55
1024	64	1.67	50	44	0.97
64	4096	18	8	89	0.50
256	1024	8.15	29	65	0.55
1024	256	7.26	44	50	1.79
4096	64	50	68	28	2.34
256	4096	97	10	86	1.78
1024	1024	43	31	63	1.79

Continued on next page

Table 44: Continued.

States d	Window K	p99 (ms)	Encode (%)	Drift (%)	Peak (GiB)
4096	256	107	68	28	3.07
1024	4096	499	12	84	3.06
4096	1024	552	65	31	3.06
4096	4096	3,104	43	55	9.91

The arm ladder, including what stops working. Table 45 and panel (c) of Figure 13 compare arms on the prespecified subgrid. The diffusion metric earns its place: dropping the inverse-variance weight costs exact tank–time accuracy at every size, because the effective innovation variance of a proxy-observed upper tank is orders of magnitude above that of a directly measured lower tank. The innovation CUSUM, with reference drift 0.5, behaves as its construction implies: at every size it names the right tank far more often than the right time, because accumulation is a reasonable coordinate statistic whose peak sits late in the run rather than at entry. Both of its rates fall as the dimension grows, and state deviation follows the same pattern. A CUSUM peak is an accumulation endpoint. Recovering onset requires a separate change-time estimate. These rows therefore do not exhaust classical FDI baselines. The two solver arms are dominated on both axes. At their prespecified budgets they reduce the shooting objective but do not reach the one-pass answer, and their cost is roughly two path evaluations per accepted iteration, a measured 7,094$\times$ the deployed statistic at matched cells, so at $d = 16$ and $K = 4096$ ($65{,}536$ atoms) the 550-step arm cannot complete one incident inside the sixty-second limit and is recorded infeasible rather than dropped.

Where that frontier falls is itself informative, and it is not where a reader might expect. The solver becomes infeasible at a *small* state dimension with a long window before it does at a large dimension with a short one, even though the second condition scans far more atoms. The reason is structural: a shooting solver rolls the path forward and sweeps the adjoint backward, both strictly sequential in time, so its cost carries a factor of K that no amount of parallelism removes. State dimension enters too, through the $O(d \cdot \text{degree})$ cost of each step, but only the K factor is sequential. The one-pass statistic has no such dependency, because every atom's innovation is computed from its own two adjacent states. That is a difference in kind rather than in constant factor, and it is the reason the gap widens rather than closes as the monitored horizon grows. This is the same conclusion the main comparison reaches at the benchmark operating point, now measured where the scan is large enough for the cost to matter operationally.

Table 45: Exact physical tank–time recovery by arm on the prespecified subgrid at the target-power reference, with 95% seed-block intervals. "$>$60 s" marks a cell where the arm could not finish one incident inside the prespecified per-incident limit. "n/e" marks a cell outside that arm's prespecified evaluation subset. The true-state oracle is a labelled reference, not a deployable method. Panel 1 of 2.

Arm	$d = 4096$, $K = 64$	$d = 4096$, $K = 256$
One-pass forcing test	95.0 [90.0, 100.0]	92.5 [88.8, 96.2]
Euclidean forcing	32.5 [21.2, 45.0]	27.5 [22.5, 31.2]
Latent CUSUM	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
State deviation	2.5 [0.0, 5.0]	0.0 [0.0, 0.0]
Solver, 50 steps	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
Solver, 550 steps	13.8 [6.2, 20.0]	1.2 [0.0, 3.8]
True-state oracle	100.0 [100.0, 100.0]	100.0 [100.0, 100.0]

Table 46: Continuation of Table 45, panel 2 of 2.

Arm	$d = 4096$, $K = 1024$	$d = 4096$, $K = 4096$
One-pass forcing test	97.5 [95.0, 100.0]	90.0 [83.8, 96.2]
Euclidean forcing	13.8 [6.2, 23.8]	13.8 [7.5, 20.0]
Latent CUSUM	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
State deviation	1.2 [0.0, 3.8]	1.2 [0.0, 3.8]
Solver, 50 steps	0.0 [0.0, 0.0]	$>$60 s
Solver, 550 steps	$>$60 s	$>$60 s
True-state oracle	100.0 [100.0, 100.0]	100.0 [100.0, 100.0]

Observation and model stresses. Table 47 and panel (d) of Figure 13 vary one mechanism at a time on the prespecified subset $d \in \{16, 256, 4096\}$, $K \in \{64, 1024\}$, reported at $80\,\mathrm{cm}^3$ where the unstressed reference reaches 99.7%. Three groups emerge, and the separation between them is the useful result.

Heterogeneous diffusion is absorbed. Raising the generating diffusion-variance spread from one to eight and to thirty-two leaves exact tank–time recovery at 99.7% averaged over both spreads, essentially unchanged from the reference, which is what the diffusion-normalised metric is for. Within-module correlated process noise is likewise survived by the diagonal score at 99.8%. A feasible block-whitened variant, fitted on the same nominal windows, is also reported so the diagonal choice is compared with the correctly constrained test rather than asserted.

Missing observations degrade the statistic gracefully and in a way the record explains. Blocks covering a tenth of the window cost little. A quarter brings exact tank–time recovery to 39.5%, and the exact locations and lengths of every block are recorded. Part of that loss is not a scoring failure at all: when the labelled atom itself falls inside a missing block it cannot be recovered by any method, and we report both the overall rate and the rate conditional on the labelled atom being observed (59.0%), together with the fraction of incidents in which the root was hidden (18.7%).

Weak observability breaks attribution, and we report that plainly. Reducing the upper-proxy gain to a fifth and to a twentieth of its calibrated reference drives exact tank–time recovery to 0.4% at the twentieth setting. Proxy inversion amplifies sensor noise as the inverse square of the gain, so the effective innovation variance of exactly the tanks that can host a root rises by that factor and the attribution target ceases to be identifiable at a fixed disturbance size. A fast implementation that loses the root as observability weakens demonstrates a computational benefit and not reliable physical attribution. This condition is what that sentence looks like when measured.

Two further conditions separate localisation from calibration, which matters more operationally than either alone. An operating-point shift applied at test time with unchanged calibration leaves localisation at 99.1% but invalidates the threshold: the nominal family-wise alarm rate rises to 81.5% against its 5% target, because the calibration windows never saw that operating point. The remedy is recalibration, not a different score, and the reported alarm rate is the diagnostic that says so. An unmodelled common input, a slow shared inflow on a contiguous region that the estimator's state does not contain, is a labelled scope failure. At the strength used here localisation survives at 99.9%, and we do not claim the local forcing interpretation or the nominal calibration guarantee holds for larger departures.

Table 47: Exact physical tank–time recovery for the deployed statistic under each observation and model stress, one mechanism at a time, at the larger fixed volume of $80\,\mathrm{cm}^3$, with 95% seed-block intervals. The smaller volume is in the stored rows. Averaging the two would mix an operating point near the floor with one near ceiling and make a degradation unreadable. The operating-shift row should be read together with its nominal alarm rate in the text: localisation survives while the frozen calibration does not. Panel 1 of 3.

Condition	$d = 16$, $K = 64$	$d = 256$, $K = 64$
Reference	99.8 [99.5, 100.0]	100.0 [100.0, 100.0]
Upper-proxy gain $1/5$	15.8 [12.5, 18.9]	18.0 [15.6, 20.3]
Upper-proxy gain $1/20$	1.4 [0.6, 2.3]	0.2 [0.0, 0.5]
Missing blocks, 10%	88.3 [86.1, 90.6]	89.2 [88.3, 90.2]
Missing blocks, 25%	70.5 [67.5, 73.0]	70.3 [67.7, 73.0]
Diffusion spread 8	99.8 [99.5, 100.0]	99.8 [99.5, 100.0]
Diffusion spread 32	99.7 [99.2, 100.0]	99.5 [99.1, 100.0]
Operating-point shift	99.8 [99.5, 100.0]	99.8 [99.5, 100.0]
Correlated process noise	99.8 [99.5, 100.0]	100.0 [100.0, 100.0]
Unmodelled common input	100.0 [100.0, 100.0]	100.0 [100.0, 100.0]

Table 48: Continuation of Table 47, panel 2 of 3.

Condition	$d = 4096$, $K = 64$	$d = 16$, $K = 1024$
Reference	99.2 [98.8, 99.7]	100.0 [100.0, 100.0]
Upper-proxy gain $1/5$	15.9 [13.1, 18.9]	20.8 [18.6, 23.1]
Upper-proxy gain $1/20$	0.2 [0.0, 0.5]	0.5 [0.0, 0.9]
Missing blocks, 10%	89.1 [86.2, 91.9]	42.5 [38.0, 46.6]
Missing blocks, 25%	66.4 [62.0, 70.2]	21.1 [19.2, 23.0]
Diffusion spread 8	99.5 [99.1, 100.0]	100.0 [100.0, 100.0]
Diffusion spread 32	99.2 [98.8, 99.7]	100.0 [100.0, 100.0]
Operating-point shift	99.2 [98.8, 99.7]	100.0 [100.0, 100.0]
Correlated process noise	99.4 [98.9, 99.8]	100.0 [100.0, 100.0]
Unmodelled common input	99.8 [99.5, 100.0]	100.0 [100.0, 100.0]

Table 49: Continuation of Table 47, panel 3 of 3.

Condition	$d = 256$, $K = 1024$	$d = 4096$, $K = 1024$
Reference	100.0 [100.0, 100.0]	99.4 [98.8, 100.0]
Upper-proxy gain $1/5$	20.2 [17.0, 23.0]	15.2 [13.0, 17.7]
Upper-proxy gain $1/20$	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
Missing blocks, 10%	25.5 [23.0, 28.0]	4.1 [2.3, 5.9]
Missing blocks, 25%	8.1 [5.6, 10.5]	0.5 [0.0, 0.9]
Diffusion spread 8	100.0 [100.0, 100.0]	99.5 [99.1, 100.0]
Diffusion spread 32	100.0 [100.0, 100.0]	99.5 [99.1, 100.0]
Operating-point shift	99.2 [98.3, 100.0]	96.4 [94.8, 98.0]
Correlated process noise	100.0 [100.0, 100.0]	99.5 [99.1, 100.0]
Unmodelled common input	100.0 [100.0, 100.0]	99.8 [99.5, 100.0]

Departures from the single-pulse model. Table 50 injects two or four roots at staggered times, and distributes one root's volume over eight steps. For these conditions we report set recovery and support hits alongside onset accuracy, because a top-one atom is the wrong endpoint for a multi-root incident. With four simultaneous roots at $80\,\mathrm{cm}^3$, set recovery is 99.8%: several roots acting at once do not confuse the scan, because each contributes its own atoms. Spreading one root's volume over eight steps is a different matter and the study reports it as a failure. Exact onset accuracy falls to 0.4% and support recovery only reaches 24.0%. The mechanism is the per-step amplitude: dividing the same volume over eight intervals divides the per-atom noncentrality by sixty-four, which puts the individual atoms below the scan threshold. A statistic that tests one atom at a time is the wrong instrument for a slow leak, and aggregating evidence over a window would be the right one.

Table 50: Multi-root and distributed incidents for the deployed statistic at $80\,\mathrm{cm}^3$. Set recovery is the fraction of true root coordinates appearing in the top-r atoms for r true roots. Support is the fraction of peaks falling on the true tank inside the true forcing interval. Panel 1 of 3.

Condition	$d = 16$ set	$d = 16$ support
Single pulse	100.0 [100.0, 100.0]	100.0
Two roots, staggered	100.0 [100.0, 100.0]	100.0
Four roots, staggered	100.0 [100.0, 100.0]	100.0
One root over eight steps	53.4 [51.1, 55.8]	31.1

Table 51: Continuation of Table 50, panel 2 of 3.

Condition	$d = 256$ set	$d = 256$ support
Single pulse	99.7 [99.2, 100.0]	99.7
Two roots, staggered	99.8 [99.6, 100.0]	100.0
Four roots, staggered	99.8 [99.7, 100.0]	100.0
One root over eight steps	33.1 [30.0, 36.4]	28.7

Table 52: Continuation of Table 50, panel 3 of 3.

Condition	$d = 4096$ set	$d = 4096$ support
Single pulse	99.4 [98.9, 99.8]	99.4
Two roots, staggered	99.8 [99.5, 100.0]	100.0
Four roots, staggered	99.6 [99.3, 99.8]	100.0
One root over eight steps	13.1 [10.6, 15.5]	12.2

Online scanning and time to attribution. The online analysis freezes a causal reconstruction and scores prefixes as observations arrive. One property of the statistic settles the multiplicity question that an online claim usually raises. Each arm here is causal and pointwise in time: the score at step k uses the encoded levels at k and $k + 1$, the fitted drift, and for CUSUM a running sum. A prefix ending at k therefore produces exactly the leading k rows of the whole-window map, so the running maximum over every repeated scan equals the whole-window maximum, and the level calibrated on dK atoms already covers the repetition. Rescanning after each observation costs no additional multiplicity. This is a property of the deployed statistic, not a general fact about scanning procedures, and it would not hold for a smoother, a non-causal encoder or a solver that refits the whole path.

Table 53 reports time from fault entry to the first correct reported root, decomposed into observation delay and computation. The median reported delay is 0 sampling intervals: the first scan taken after the disturbance enters already names the correct tank and the correct interval. That is a statement about this plant, not a general one. The forced tank is itself a modelled state whose level responds within the entry interval and is reconstructed directly from its own calibrated proxy, so no propagation delay is incurred before the evidence appears. A root observable only through its downstream response would pay that delay, and the single-module study's Tank 3 to Tank 1 example is exactly such a case.

The computation is not the binding constraint either way. The per-observation update is one encode of the new sample, one drift evaluation at the previous state and one normalisation: $O(d)$ work that does not grow with the window. At $d = 4096$ it measures 0.80 ms, about 1,254$\times$ inside one sampling interval, which is the operationally relevant margin for a plant that has to be diagnosed while it is still running.

Table 53: Online prefix scanning with a frozen causal reconstruction. Delay is the number of sampling intervals between fault entry and the first scan whose running peak is the labelled atom and clears the calibrated threshold. Per-scan cost is the measured incremental update for one new observation. Panel 1 of 2.

States d	Volume (cm^3)	Reported (%)	Median delay (s)
16	20	22.2	0
16	80	100.0	0
256	20	20.6	0
256	80	99.8	0
4096	20	11.2	0
4096	80	99.1	0

Table 54: Continuation of Table 53, panel 2 of 2.

States d	Volume (cm^3)	p90 delay (s)	Per-scan cost (ms)
16	20	0	0.03
16	80	0	0.03
256	20	0	0.08
256	80	0	0.08
4096	20	0	0.80
4096	80	0	0.80

What this supports, and what it does not. Within the completed conditions, the deployed statistic identifies the physical tank and the entry interval of a latent disturbance across four orders of magnitude of scan size, at a cost that grows about linearly in dK and stays far inside a diagnostic time budget, while a fairly implemented least-action solver at a fixed budget is both less accurate and, past a measured size, infeasible. Difficulty was controlled two ways, and the two schedules answer different questions: the target-power reference shows that multiplicity alone is handled by the calibrated threshold, and the fixed-volume schedule shows what a fixed disturbance costs as the search grows.

The limits are equally measured. Attribution fails when the proxies that observe the candidate tanks are weakened enough, and the failure is a loss of identifiability rather than of speed. A frozen threshold does not survive an operating-point shift even where localisation does. Set-valued endpoints, not a top atom, are the right target once several roots act at once. Nothing here transfers to electrical-grid, river or process data without its own study. The plant is a simulated hydraulic network whose parameters, topology and noise we chose, and the reference schedule is derived from a Gaussian oracle model rather than from the nonlinear plant that generates the data.

An independent solver comparison. A separate matched-margin replication of the original ten-regime synthetic suite, with the equivalence margin fixed at one percentage point before new incidents are generated, remains proposed and has not been executed. The scaling evidence above does not substitute for it: it compares arms on this hydraulic network, not on that suite.

L Exploratory innovation-scan and filter comparisons

These pilots address comparator coverage and fault duration. They were designed after the original results and use frozen candidate families within each run. They are exploratory rather than independent confirmation of a selected model. The raw outputs are `exp/opus55/filter_pilot` and `exp/opus55/msds_scan`. The runners record source and input hashes and refuse to overwrite a nonempty output directory. Tables are generated by `scripts.make_opus55_tables`. `tables-fdi/provenance.json` records their input hashes and aggregation. No test labels select the windows.

Constant-forcing interval alternative. For coordinate j and a candidate interval I, profile a single signed drift u_j shared across the interval. With constant diagonal variance rate,

$$\log \Lambda_{j,I} = \frac{(\sum_{k\in I} r_{j,k})^2}{2\sigma_j^2 \sum_{k\in I} \Delta t_k}.$$

This is the classical Gaussian interval GLR. A one-step interval recovers Proposition 1. It is distinct from masking around a selected singleton peak. We scan all coordinates and starts over lengths $\{1, 2, 4, 8, 16\}$ in the tank pilot, adding length 32 for MSDS. Tank scores are retrospective and indexed by candidate onset. MSDS scores use only completed windows ending at the current transition. Scanning durations adds multiplicity: the tank threshold is calibrated on the maximum over the entire coordinate, time and duration family, separately for each arm. A singleton χ^2 threshold is not reused for these scans.

Known-physics observation-innovation GLR. The EKF uses the published hydraulic equations, known process and sensor noise, known pump inputs, and a prior whose mean and variance match the nominal initial state distribution. It never receives the incident's true state. Let e_k, S_k, F_k, H_k, and G_k denote its observation innovation, predictive covariance, state and measurement Jacobians, and Kalman gain. A unit candidate state forcing has prediction-error signature $D_k^- = F_k D_{k-1}^+ + e_j \mathbf{1}\{k \in I\}$, measurement signature $g_k = H_k D_k^-$, and updated signature $D_k^+ = (I - G_k H_k) D_k^-$. Its profile score is

$$\log \Lambda_{j,I}^{\text{filter}} = \frac{(\sum_k g_k^\top S_k^{-1} e_k)^2}{2 \sum_k g_k^\top S_k^{-1} g_k}.$$

Zero-information candidates receive zero score. Later observations can contribute to an earlier candidate through its propagated signature. This is exact for a linear Gaussian model with fixed coefficients and approximate for the EKF linearisations. A test independently constructs the full batch observation covariance and verifies both impulse and interval scores against that likelihood. The EKF has more model information than the fitted-drift pointwise arm, so this comparison measures the gap to a physics-informed FDI comparator.

Tank protocol and findings. Three independent seeds each supply 32 nominal training episodes, 99 disjoint calibration episodes and 40 test episodes per condition. All faults enter upper tanks, while every method searches all four tank coordinates and all 96 transitions. We hold total volume fixed at 10 or 20 cm^3 across one-step and eight-step faults, using the original fine-grid simulator. Informative and weak-proxy settings use gains 1 and 0.2. The downstream-only condition supplies only lower-tank levels. A test verifies that changing removed proxy channels cannot change the filter output. Nominal extended Rauch–Tung–Striebel (RTS) smoothing followed by plug-in scoring is a separate known-physics reconstruction diagnostic. Its effective residual variance is estimated from nominal training episodes. It is not a learned, fault-aware smoother.

Table 55: Exploratory tank comparison at 20 cm^3. Percentages are mean ± sample standard deviation over three seed means, each based on 40 upper-tank incidents. Root means correct tank. Onset atom additionally requires the exact entry transition. These are ranking endpoints, not detection power. The true-state scan is an inaccessible reference. Pointwise inversion is undefined without upper proxies and is omitted there. All doses, arms, detection rates and null trials are retained in the accompanying seed-level CSV. Panel 1 of 2.

Sensors	Score	One-step pulse Root	One-step pulse Onset atom
Informative	Pointwise singleton	100.0 ± 0.0	100.0 ± 0.0
Informative	Pointwise scan	100.0 ± 0.0	100.0 ± 0.0
Informative	EKF singleton GLR	100.0 ± 0.0	100.0 ± 0.0
Informative	EKF duration GLR	100.0 ± 0.0	100.0 ± 0.0
Informative	RTS plug-in scan	100.0 ± 0.0	100.0 ± 0.0
Informative	True-state scan	100.0 ± 0.0	100.0 ± 0.0

Continued on next page

Table 55: Continued.

Sensors	Score	One-step pulse Root	One-step pulse Onset atom
Weak proxy	Pointwise singleton	95.8 ± 5.2	93.3 ± 7.2
Weak proxy	Pointwise scan	97.5 ± 2.5	84.2 ± 10.4
Weak proxy	EKF singleton GLR	100.0 ± 0.0	100.0 ± 0.0
Weak proxy	EKF duration GLR	100.0 ± 0.0	94.2 ± 3.8
Weak proxy	RTS plug-in scan	99.2 ± 1.4	2.5 ± 2.5
Weak proxy	True-state scan	100.0 ± 0.0	100.0 ± 0.0
Downstream only	EKF singleton GLR	57.5 ± 5.0	10.8 ± 5.2
Downstream only	EKF duration GLR	49.2 ± 6.3	9.2 ± 6.3
Downstream only	RTS plug-in scan	85.8 ± 3.8	0.8 ± 1.4
Downstream only	True-state scan	100.0 ± 0.0	100.0 ± 0.0

Table 56: Continuation of Table 55, panel 2 of 2.

Sensors	Score	Eight-step forcing Root	Eight-step forcing Onset atom
Informative	Pointwise singleton	73.3 ± 5.8	3.3 ± 1.4
Informative	Pointwise scan	96.7 ± 1.4	72.5 ± 8.7
Informative	EKF singleton GLR	82.5 ± 8.7	4.2 ± 2.9
Informative	EKF duration GLR	97.5 ± 2.5	79.2 ± 1.4
Informative	RTS plug-in scan	86.7 ± 1.4	72.5 ± 5.0
Informative	True-state scan	97.5 ± 2.5	84.2 ± 2.9
Weak proxy	Pointwise singleton	61.7 ± 8.8	0.0 ± 0.0
Weak proxy	Pointwise scan	86.7 ± 1.4	14.2 ± 10.1
Weak proxy	EKF singleton GLR	88.3 ± 5.8	1.7 ± 2.9
Weak proxy	EKF duration GLR	92.5 ± 2.5	37.5 ± 8.7
Weak proxy	RTS plug-in scan	95.0 ± 2.5	35.0 ± 4.3
Weak proxy	True-state scan	97.5 ± 2.5	84.2 ± 2.9
Downstream only	EKF singleton GLR	54.2 ± 10.1	1.7 ± 2.9
Downstream only	EKF duration GLR	45.0 ± 6.6	5.8 ± 1.4
Downstream only	RTS plug-in scan	89.2 ± 5.2	5.0 ± 5.0
Downstream only	True-state scan	97.5 ± 2.5	84.2 ± 2.9

With informative proxies, scanning durations improves eight-step onset-atom accuracy from 3.3% to 72.5%. The EKF scan reaches 79.2%. With weak proxies, the corresponding scan results are 14.2% and 37.5%. Thus filtering remains a material comparator when proxy inversion is noisy. In the downstream-only condition, impulse EKF onset-atom accuracy is 10.8%. The RTS interval score names the correct tank on 85.8% of one-step incidents but identifies the exact onset atom on only 0.8%. Temporal reconstruction can aid coordinate attribution without solving exact timing. This pilot does not establish the motivating downstream-only capability for the deployed pointwise method.

Matched MSDS score heads. Five model seeds each train one nominal latent SDE for 250 epochs with four latent coordinates and the original chronological training-split and normalisation procedure. All heads share that model, diffusion and absolute-decoder projection. The score family is fixed before inspecting labels. The same 161 events are reused across seeds, so variability describes model fitting, not five independent real datasets. No AERCA model is retrained in this pilot. It is an internal head comparison rather than a new cross-method result.

Table 57: MSDS event Recall@3, percentage mean ± seed standard deviation. The legacy convention assigns transition $k \to k+1$ to row k. Availability alignment assigns it to row $k+1$, when both observations exist. All fixed windows are reported without selecting one from test outcomes. Ranking uses labelled event boundaries. This is not a detection or selective-calibration guarantee.

Score	Legacy transition row	Available observation row
Singleton	17.8 ± 3.0	16.0 ± 2.2
Innovation CUSUM	6.2 ± 2.4	5.5 ± 1.7
Duration scan (all fixed windows)	7.7 ± 1.3	5.1 ± 0.8
Fixed window, 2 steps	15.4 ± 3.5	12.7 ± 3.4
Fixed window, 4 steps	10.4 ± 1.6	8.9 ± 1.4
Fixed window, 8 steps	8.1 ± 1.4	6.5 ± 0.6
Fixed window, 16 steps	7.1 ± 1.5	6.3 ± 0.8
Fixed window, 32 steps	4.7 ± 1.1	5.3 ± 1.3

Under availability alignment, Recall@3 is 16.0% for the singleton and 5.1% for the duration scan. The sustained-forcing alternative therefore does not resolve the real-data deficit in this representation and projection. These results argue against treating fault duration as a sufficient explanation of the losses in Section 5.4.