DECISION-LOCALIZED CHARACTERISTIC DEFECTS FOR STOCHASTIC NEURAL OPERATOR SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Two stochastic neural operators can agree on global trajectory statistics yet predict different outcomes for the same path-dependent decision. With a small diagnostic-rollout budget, estimating these outcomes directly is noisy. Decision-localized characteristic defects (DLCD) select among pretrained operators by measuring exposure, drift, and covariance-sensitive increment errors in reference-defined decision regions. A fixed bank of physical observables, reusable reference calibration, and trusted local transitions provide the selection signal. At 512 diagnostic paths per candidate-condition, mean held-out regret across eight newly trained stochastic Allen–Cahn pools is 0.02510, compared with 0.02880 for regularized direct event estimation and 0.02693 for localized coefficient matching with matched information sources. The matched-information comparison yields 4.2% lower regret on a second Allen–Cahn cohort and 9.3% on six stochastic 2-D Kolmogorov pools. A history-conditioned decomposition explains the connection between local discrepancies and path decisions; ablations examine localization and aggregation. The gains are concentrated at smaller rollout budgets. Converting measured runtimes into additional direct-estimation paths produces an effective tie.

1 INTRODUCTION

Global trajectory statistics can conceal differences in the decisions made from stochastic model rollouts. In our Allen–Cahn example, a trajectory begins in the negative basin. The relevant event occurs when its spatial mean crosses $+0.65$ and stays above $+0.55$ for 0.15 time units. Its reference probability under nominal conditions is 0.2178. Operators with similar typical fields or energy spectra can give different estimates of this commitment event. The same difficulty arises for first-passage and other path-dependent observables (Zou et al., 2026).

The task is to choose one operator from a fixed pretrained library for a specified path decision. Selection may use trusted reference dynamics and B diagnostic trajectories per candidate and physical condition. Separate held-out candidate and reference trajectories determine the loss of the selected operator. Directly estimating the event on the diagnostic paths is a natural approach, although rare events and small differences between candidates make the resulting rankings noisy.

DLCD examines local dynamics along reference-defined routes toward the decision. For Allen–Cahn, the spatial-mean transition region receives full localization weight on $[-0.50, 0.30]$. Each candidate is scored using its occupancy of that region and the increments of a fixed bank of physical observables, compared with reference saved-time drift and covariance-sensitive targets. Reference trajectories calibrate the score, with short trusted transitions covering states that lack sufficient local reference support. Candidate event outcomes play no role in the DLCD score (Figure 1).

Our evaluation separates three questions: how DLCD compares with direct event estimates at fixed path budgets, what it adds beyond coefficient matching with the same information sources, and how its extra local-transition cost affects selection. Against the matched-access coefficient control, held-out regret falls by 6.8% across eight newly trained Allen–Cahn pools, 4.2% on an independent Allen–Cahn cohort, and 9.3% across six 2-D Kolmogorov pools. Individual pools and near-extrapolation conditions can reverse these gains; direct estimation becomes competitive under the timing-derived larger path budget. A history-conditioned decomposition connects local transition discrepancies to the decision, while the practical finite-witness score is evaluated empirically.

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.

2 Related Work

Stochastic operators and their evaluation. Fourier neural operators learn parametric solution maps (Li et al., 2021; Kovachki et al., 2023), and Neural SPDEs model random forcing in continuous spatiotemporal dynamics (Salvi et al., 2022). Stochastic alternatives include Wiener-chaos and neural-basis expansions, martingale operators, and operator flow matching (Shi et al., 2026a;b; Bahmani et al., 2025; Hidajat, 2026; Shi et al., 2025). SPDEBench, StocBench, and TRIE assess surrogate fidelity across statistical and dynamical criteria (Zhu et al., 2025; Pfister et al., 2026; Srikishan et al., 2026). DLCD addresses the subsequent deployment decision: selecting an already-trained operator for a particular path event.

Decision-focused selection. Ranking-and-selection methods allocate noisy simulation budgets, sometimes using common random numbers to reduce variance (Nelson and Matejcik, 1995; Lee and Nelson, 2016). Goal-oriented SDE learning uses bounds on path observables to train surrogate drifts (Zou et al., 2026); our candidates are already trained, and their reference decision outcomes are unavailable at selection. Shared physics responses can recover rankings for squared Hilbert-space losses (Liang and Liu, 2026). Here, the selection loss combines nonquadratic path functionals estimated from finite rollouts. Appendix L compares these settings.

Auditing local dynamics. Conditional-moment tests, Stein discrepancies, and martingale projections assess local properties of stochastic laws (Muandet et al., 2020; Li et al., 2026; Liu et al., 2016; Lorch et al., 2024; Bleile et al., 2026; Blanchet et al., 2026). Generator matching uses related operators to train generative models (Holderrieth et al., 2025). For PDE surrogates, residual correction, a posteriori estimates, semigroup consistency, and cross-PDE stress tests expose different forms of model error (Cao et al., 2023; Jha, 2024; Qiu et al., 2026; Shikhman, 2026a;b;c). DLCD concentrates such an audit on reference-defined decision regions. Goal-oriented and dual-weighted error analysis motivate this emphasis on the task (Endtmayer et al., 2020; Chakraborty et al., 2025; Toscano et al., 2026); the deployed score is an empirical selector rather than a certified error bound.

3 Decision-Localized Characteristic Defects

Each candidate supplies B diagnostic paths per condition. DEV-REF fixes localization, NULL-REF fixes null calibration, and trusted one-step transitions are charged separately. The calibrated score determines the selected candidate; held-out decision losses are reserved for evaluation. Appendix N specifies the information available at each stage and defines the reported metrics.

One selection instance. Consider a held-out Allen–Cahn pool under the nominal condition C0. Eight trained operators each generate 512 diagnostic trajectories. The task is the basin-commitment event introduced in Section 1; its reference probability, estimated on a separate ORACLE-REF population, is $7137/32768 = 0.217804$. DEV-REF trajectories determine the spatial-mean transition region, and NULL-REF fixes the score calibration before any held-out candidate is selected. For each candidate, DLCD measures occupancy of that region and the local increments of its physical witnesses, using trusted one-step transitions when reference support is sparse. The lowest calibrated score determines the selected operator. Its event-probability, commitment-time, and distributional errors are computed later on independent DEC-MODEL trajectories. DIRECT-D1 forms estimates of these same outcomes from its diagnostic trajectories; matched-access P2 instead fits local drift and covariance coefficients. This example fixes the roles of the data used by each selector; the full split and cost accounting appear in Appendix N.

3.1 Reference dynamics

For parameter λ, let

$$dU_t^\lambda = [A_\lambda U_t^\lambda + F_\lambda(U_t^\lambda)]\,dt + G_\lambda(U_t^\lambda)dW_t, \tag{1}$$

where W is a Q-Wiener process, with covariance $C_\lambda(u) = G_\lambda(u)QG_\lambda(u)^*$. Assume A_λ generates a strongly continuous semigroup, $e_i \in D(A_\lambda^*)$, and $G_\lambda(u)Q^{1/2}$ is Hilbert–Schmidt on the states considered. For smooth cylindrical $\phi(u) = \psi(\langle u, e_1\rangle, \ldots, \langle u, e_m\rangle)$ with bounded first and second

derivatives,

$$\mathcal{L}_\lambda \phi(u) = \sum_i \psi_i(z) \big[\langle u, A_\lambda^* e_i \rangle + \langle F_\lambda(u), e_i \rangle \big] \tag{2}$$

$$+ \frac{1}{2} \sum_{i,j} \psi_{ij}(z) \langle C_\lambda(u) e_i, e_j \rangle. \tag{3}$$

The numerical saving scheme determines the characteristic targets. For each witness, the saved-time reference drift is $b^*_{\phi,h}(u) = \mathbb{E}_{\mathrm{ref}}[\phi(U_{k+1}) - \phi(U_k) \mid U_k = u]/\Delta t$. We compute it analytically for scheme-matched smooth projections and use local reference regression or trusted one-step simulation for nonlinear physical summaries. Throughout, $\mathcal{L}_{h,\Delta t}\phi = b^*_{\phi,h}$ denotes this saved-time quantity; the continuous generator enters only where its discretization is matched.

3.2 Localized characteristic residuals

For each candidate, we measure occupancy of the decision region, conditional mean increments within it, and covariance-sensitive products of centered increments. At saved times t_k, define

$$\Delta M^\phi_{n,k} = \phi(U_{n,k+1}) - \phi(U_{n,k}) - \Delta t\, \mathcal{L}_{h,\Delta t}\phi(U_{n,k}). \tag{4}$$

For smooth witnesses, the continuous covariance rate is $\Gamma_{\phi,\chi}(u) = \langle G(u)^* D\phi(u), QG(u)^* D\chi(u) \rangle$. The saved-time target in Eq. (7) is instead $\Gamma^*_{h,\Delta t}(u) = \Delta t^{-1} \mathbb{E}_{\mathrm{ref}}[(\Delta\phi - \Delta t\, b^*_\phi(u))(\Delta\chi - \Delta t\, b^*_\chi(u)) \mid U_k = u]$, where $b^*_\phi(u) = \mathbb{E}_{\mathrm{ref}}[\Delta\phi \mid U_k = u]/\Delta t$; scheme targets, reference-local regression, and trusted bursts approximate this saved-time quantity, with state-compression and Monte Carlo errors assessed through null calibration.

The primary localization depends only on spatial mean m. Let $s(x) = 0$ for $x \le 0$, $s(x) = 3x^2 - 2x^3$ for $0 < x < 1$, and $s(x) = 1$ for $x \ge 1$. We use

$$a_{A1}(m) = s((m + 0.58)/0.08) s((0.38 - m)/0.08), \tag{5}$$

so unit weight lies on $[-0.50, 0.30]$ with smooth 0.08 edges. A1 was defined from transition geometry and DEV-reference support without candidate DEC/ORACLE outcomes.

For trajectory n, define normalized exposure $\bar{W}_n = T^{-1} \sum_k a_{n,k} \Delta t$, candidate exposure $p^\theta_a = N^{-1} \sum_n \bar{W}^\theta_n$, and $\Delta p_a = p^\theta_a - p^{\mathrm{ref}}_a$. NULL-REF centering and scaling give $T_{\mathrm{occ}} = |\Delta p_a - \mu_{\mathrm{null}}| / \max(\sigma_{\mathrm{null}}, 10^{-8})$. Localized drift and covariance-sensitive second-characteristic residuals are

$$\widehat{R}^\phi_{d|a} = \frac{\sum_{n,k} a_{n,k} \Delta M^\phi_{n,k}}{\sum_{n,k} a_{n,k} \Delta t + 10^{-8}}, \tag{6}$$

$$\widehat{R}^{\phi,\chi}_{q|a} = \frac{\sum_{n,k} a_{n,k} (\Delta M^\phi_{n,k} \Delta M^\chi_{n,k} - \Delta t \Gamma^*_{h,\Delta t}(U_{n,k}))}{\sum_{n,k} a_{n,k} \Delta t + 10^{-8}}. \tag{7}$$

The second group is covariance-sensitive at the saving interval. After normalization, drift mismatch contributes an additional $O(\Delta t)$ term (Appendix C). Candidate-condition exposure below 10^{-6} is flagged as unsupported rather than assigned a stable localized score. The drift bank has 13 witnesses: spatial mean; real/imaginary Fourier modes $k = 1{:}4$; free energy; positive-phase fraction; and low/high spectral energy. The second-characteristic group uses nine diagonal quadratic-variation channels plus six fixed cross-mode pairs. Positive-phase fraction is nonsmooth and therefore uses a matched scheme/trusted-reference characteristic rather than a classical cylindrical derivative.

Saved-time discretization and nonlinear state dependence can shift nominal martingale residuals away from zero. We therefore calibrate each channel against the same saved-time reference scheme and local state, dispatching to analytic/scheme targets when available, local NULL-REF calibration for the dominant supported cases, and trusted one-step bursts when required. In the reported Allen audit, 10.6% of targets are analytic/scheme-discrete, 84.7% use local reference calibration, and 4.7% use trusted bursts; no evaluated channel is unsupported.

For group residual vector z_g and NULL-REF covariance S_g, define $\Sigma_g = 0.9 S_g + 0.1\, \mathrm{diag}(S_g) + 10^{-6} I$ and

$$T_g = \sqrt{z_g^\top \Sigma_g^{-1} z_g / d_g}, \qquad T_{\mathrm{DLCD}} = \sqrt{(T^2_{\mathrm{occ}} + T^2_{\mathrm{drift}} + T^2_{\mathrm{cov}})/3}. \tag{8}$$

The deployable selector is $S_{\mathrm{DLCD}} = \sqrt{(T_{E0}^2 + T_{E1}^2 + T_{\mathrm{DLCD}}^2)/3}$, where $E0$ groups marginal/projected-field/CRPS summaries and $E1$ groups temporal/path summaries. We call T_{DLCD} the characteristic score and S_{DLCD} the DLCD selector. For non-analytic targets, the fixed Allen calibration conditions on $[m, F, v_+, \log E_{\mathrm{low}}, \log E_{\mathrm{high}}, t/T]$: NULL-REF supplies 12 equal-frequency mean bins and four time bins, and a ridge-linear local predictor ($\lambda = 10^{-4}$) is used when at least 200 reference increments are available and the local Mahalanobis radius is at most 3.5. Otherwise 64 trusted one-step bursts are drawn from the candidate saved state, with at most 128 triggered states (8,192 solver transitions) per candidate. The appendix and supplement give the exact dispatch and second-characteristic-target rules.

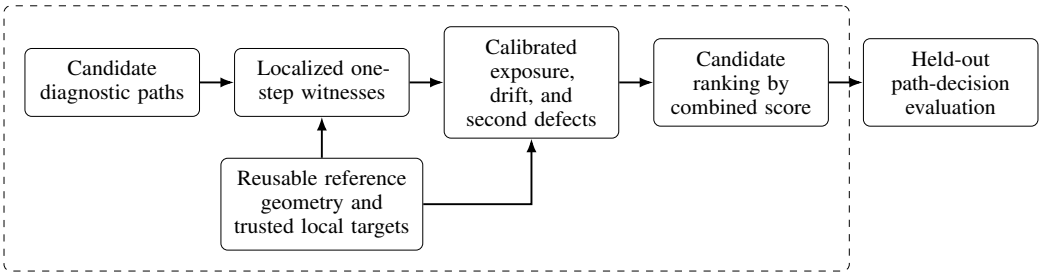


Figure 1: What each stage can access. Reference geometry fixes where candidates are examined; saved-time reference and trusted-transition targets define what is compared. DLCD uses candidate diagnostic paths but not their D1–D4 outcome summaries. DIRECT may estimate D1 from these same paths. All held-out DEC-MODEL and ORACLE-REF outcomes remain outside the selection boundary.

3.3 Decision functional and theory

For candidate θ, let τ be the commitment time for the primary decision D1 (specified in Section 4), $\bar{\tau} = \min(\tau, T)/T$, and $\mathrm{RMST} = \mathbb{E}[\bar{\tau}]$. The primary held-out loss is

$$\ell_{D1} = 0.60|P_\theta(\tau \le T) - P_{\mathrm{ref}}(\tau \le T)| + 0.25|\mathrm{RMST}_\theta - \mathrm{RMST}_{\mathrm{ref}}| + 0.15 W_1(\bar{\tau}_\theta, \bar{\tau}_{\mathrm{ref}}). \quad (9)$$

The fixed weighting prioritizes commitment probability while retaining timing/distribution information; component and weight sensitivities are reported separately. For selector S, $\hat{k}(S) = \arg\min_k S_k$ and regret is $R(S) = \ell_{\hat{k}(S)} - \min_j \ell_j$.

Let H_k denote the observed trajectory history through saved time k, including deterministic decision-automaton state. Let $P_k(\cdot|H_k)$ and $Q_k(\cdot|H_k)$ be reference and candidate one-step kernels, and define $v_K(H_K) = g(H_K)$ and $v_k(H_k) = \int v_{k+1}(H_k, x) P_k(dx|H_k)$.

Proposition 1 (History-conditioned decision decomposition)**.** *For bounded terminal path functional g, assume the reference kernels $P_k(\cdot|H_k)$ and hence the continuation recursion are defined for Q-almost-every candidate-reachable history. Then*

$$\mathbb{E}_Q g - \mathbb{E}_P g = (\mathbb{E}_{\mu_Q} v_0 - \mathbb{E}_{\mu_P} v_0) + \sum_{k=0}^{K-1} \mathbb{E}_Q[((Q_k - P_k) v_{k+1})(H_k)]\,. \quad (10)$$

If the initial law is shared, the first term vanishes.

This identity accommodates non-Markov candidate architectures because H_k may contain complete observed history. It also covers every component of Eq. (9): commitment probability and RMST are expectations of bounded path functionals, and because $\bar{\tau} \in [0, 1]$, $W_1(\bar{\tau}_\theta, \bar{\tau}_{\mathrm{ref}}) = \int_0^1 |P_\theta(\bar{\tau} \le s) - P_{\mathrm{ref}}(\bar{\tau} \le s)|\, ds$, so Proposition 1 applies pointwise to the bounded threshold indicators before integration. Under the shared condition-specific initial distribution used here, the initial-law term vanishes; Appendix H gives the resulting D1-specific upper bound by summed one-step continuation discrepancies. A history-level witness expansion bounds each localized term by one-step witness discrepancies plus approximation and leakage remainders (Appendix H). The deployed T_{DLCD} averages signed residuals over histories and times, allowing cancellation. A cellwise non-cancelling ablation in Section 5.4 tests how much this aggregation affects selection.

For normalized trajectory-cluster weights $\bar{W}_n = T^{-1}\sum_k a_{n,k}\Delta t$, $N_{\text{eff}} = (\sum_n \bar{W}_n)^2/\sum_n \bar{W}_n^2$. Let W denote the generic trajectory-level localization weight distributed as $\bar{W}_n$. Under a local alternative with contribution Y and $\delta_W = \mathbb{E}[WY]/\mathbb{E}W$:

Proposition 2 (First-order localization criterion)**.** *For i.i.d. trajectory clusters with* $0 \le W \le 1$, $\mathbb{E}W > 0$, $0 < \mathrm{Var}(Y) < \infty$, *and* $0 < \mathrm{Var}[W(Y-\delta_W)] < \infty$, *the first-order Gaussian/delta-method approximation gives localization a larger SNR if and only if*

$$\frac{\delta_W^2(\mathbb{E}W)^2}{\mathrm{Var}[W(Y-\delta_W)]} > \frac{(\mathbb{E}Y)^2}{\mathrm{Var}(Y)}. \tag{11}$$

Localization improves signal-to-noise ratio precisely when the gain in local defect signal outweighs the reduction in effective sample size.

4 Experimental Design

Primary system and decisions. Our primary system is stochastic Allen–Cahn on a periodic unit interval (Berglund, 2022), simulated to $T = 6$ on $N_x = 256$ points with $\Delta t = 2^{-11}$; the convergence run uses $N_x = 512$ and $\Delta t = 2^{-12}$. Low-mode-dominant noise has spectrum $q_k \propto [1 + (|k|/8)^2]^{-2}$ and reference-selected amplitude $\sigma^\star = 0.45$. The primary decision D1 is basin commitment: starting at spatial mean $m \le -0.65$, the mean crosses $+0.65$ and remains at least $+0.55$ for 0.15 time units. D2–D4 respectively test a DEV-selected free-energy event ($\bar{F} = 0.140$), sustained positive-phase volume, and local nucleation. The four $(\kappa, \sigma, h_{\text{ext}})$ conditions are C0 $(0.0050, 0.4500, 0)$, C1 $(0.00425, 0.4275, -0.010)$, C2 $(0.00575, 0.4860, +0.010)$, and near-extrapolation C3 $(0.00625, 0.5175, +0.020)$.

Candidate library and splits. The library contains noise-conditioned FNO, latent fixed-point Neural SPDE, an NCDE–FNO hybrid (Kidger et al., 2020), and Structured/WCE candidates. Development uses pools A–E. Eight held-out pools F–M contain 64 independently trained candidates (two seeds per family per pool), with separate training and validation seeds and trajectories. TRAIN, VAL, DEV-REF, NULL-REF, DIAG-MODEL, DEC-MODEL, ORACLE-REF, and CONV-REF have disjoint roles. Each independently trained candidate pool is one scientific replicate.

The budget B counts candidate DIAG-MODEL trajectories per condition; we evaluate $B \in \{128, 256, 512, 1024, 2048\}$ and use $B = 512$ for detailed comparisons. DIRECT may calculate D1 from these paths, whereas DLCD and P1–P3 do not use candidate D1–D4 summaries. No selector sees held-out DEC-MODEL/ORACLE outcomes before selection. Reference/null calibration and trusted bursts are accounted for separately from B. Budget comparisons share trajectory indices over 1,000 matched subsamples; Appendix A gives direct-event sampling precision.

Comparator information sets. $E0/E1$ combine conventional distributional, temporal, and path diagnostics; KCM/MMR, KMD, and SE-MPD use DEV-only tuning. We call the fixed $E0/E1$ + SE-MPD stack the BASE selector. A1-SE-MPD and A1-KMD receive the same fixed localization but no trusted generator. A0 uses the same characteristic construction globally, isolating localization inside the trusted-physics audit.

DIRECT-D1 estimates the commitment probability, normalized RMST, and censored-time W_1 from diagnostic paths and applies the same 0.60/0.25/0.15 decision loss. DIRECT-D1-SHRINK regularizes probability and RMST using development-frozen rules. Neither receives trusted local-transition targets. We also test adaptive allocation of a fixed total path budget and a meta-ranker trained on A–E development outcomes, frozen before F–M evaluation. The latter receives richer development information than DLCD; none of these controls sees held-out outcomes.

Physics controls P1–P3 receive the same candidate diagnostic trajectories, localization, trusted dynamics, reference calibration, saved-time grid, and physical state as DLCD. Candidate decision summaries are excluded from all four scores. P1 averages localized one-step residuals; P2 fits local drift and covariance coefficients before comparison with trusted targets; P3 matches a finite-difference generator on smooth projections. Each uses the shared outer score $\sqrt{(T_{E0}^2 + T_{E1}^2 + T_P^2)/3}$ and NULL-REF standardization, with method-specific trusted-transition query counts reported separately. P2 omits DLCD's standalone exposure group; Appendix F defines the exposure-augmented P2 follow-up and details the control estimators and costs.

Figure 1 shows the information boundary; Appendix C provides the complete stage-by-stage contract.

Transfer and implementation. To transfer DLCD, specify the path decision and use trusted reference dynamics to choose the progress coordinate, localization, and witnesses. Calibrate against the saved-time numerical scheme, then freeze the witness bank, support thresholds, covariance estimates, and score. At selection, compute localized increments from diagnostic paths and obtain reference targets analytically, by local regression, or through short trusted bursts. The supplement fixes the dispatch rules and information boundary.

Cost and complexity. Once candidate trajectories exist, characteristic aggregation is linear in saved transitions and the fixed witness channels; the group-covariance linear solves are fixed-dimensional. At $B = 512$, recorded median online wall clock is 97.9 s for BASE and 104.7 s for DLCD on the common A100-80GB stack. DIRECT-D1 takes 44.1 s at $B = 512$ because it avoids reference-calibration and trusted-burst work. A timing-derived sensitivity converts these measured anchors to a path-equivalent budget $B_{\mathrm{eq}} = \mathrm{round}[512(104.7/44.1)] = 1216$ without consulting held-out outcomes; at that budget DIRECT-D1 regret is 0.02508 versus the timing-study DLCD aggregate 0.02520. This is a path-equivalent sensitivity, not a separately timed equal-wall-clock run at $B = 1216$: the fixed-path comparison isolates diagnostic-rollout efficiency, whereas the extrapolated comparison indicates the compute boundary. DLCD's median triggered-burst load is about 2,300 solver transitions/candidate (p90 6,200; cap 8,192). Appendix J gives the full component breakdown and separates reusable DEV/NULL cost from online and evaluation-only resources.

5 RESULTS

Unless stated otherwise, Allen regret is averaged equally over C0–C3 within each candidate pool.

5.1 FRESH HELD-OUT SELECTION AND OUTCOME-AWARE CONTROLS

We first test whether the characteristic score carries information beyond conventional trajectory diagnostics. On the 40 C0 development candidates, Spearman association with D1 loss rises from 0.476 for BASE to 0.708 for the characteristic score; the association remains 0.531 after controlling for BASE rank and model family (Appendix D). Figure 2 then compares the frozen selectors pool by pool on fresh F–M candidates, including the unfavorable pools.

At $B = 512$, mean regret is 0.03040 for raw DIRECT-D1 and 0.02510 for DLCD (17.4% lower; 7/8 pools favorable). Regularized DIRECT-D1-SHRINK improves to 0.02880; DLCD achieves 12.8% lower regret with 6/8 favorable pools and a descriptive paired 95% interval of $[-0.00589, -0.00151]$ for the difference. The advantage varies between 8.8% and 12.8% for $B \leq 512$, then falls to 5.5% and 2.7% at $B = 1024$ and 2048. At the larger, timing-derived DIRECT-D1 budget of $B = 1216$, the regrets are effectively tied (0.02508 versus 0.02520). This path-equivalent comparison uses measured $B = 512$ runtime anchors; it is not a separately timed $B = 1216$ experiment.

Relative to BASE, DLCD reduces $B = 512$ regret from 0.03250 to 0.02510 (22.8%; 7/8 pools favorable), with descriptive paired 95% interval $[-0.0105, -0.0043]$ and bootstrap relative-reduction interval 14.6–27.8%; pool M reverses. A disjoint DEC/ORACLE reevaluation gives 0.03236 versus 0.02618 (19.1% lower) and candidate-loss rank Spearman 0.936. By condition, C0/C1/C2 improve by 31.0%, 19.0%, and 37.0%, while C3 is 1.9% worse. Appendix E gives the poolwise direct-outcome rows and development-to-held-out attenuation.

Table 1: Fresh F–M held-out regret across candidate-path budgets. Each pool averages C0–C3 equally; entries then average the eight pools. Lower is better.

B	BASE	DIRECT-D1	DIRECT-SHRINK	DLCD	vs. SHRINK
128	.0430	.0467	.0445	.0406	8.8%
256	.0383	.0394	.0369	.0336	8.9%
512	.0325	.0304	.0288	.0251	12.8%
1024	.0287	.0261	.0254	.0240	5.5%
2048	.0249	.0230	.0225	.0219	2.7%

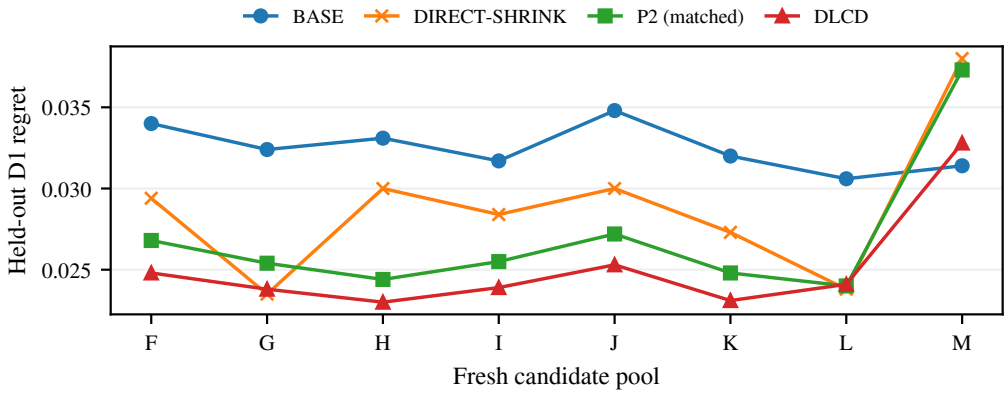


Figure 2: Fresh F–M held-out selection at $B = 512$: poolwise regret of DLCD, conventional BASE, development-regularized direct target estimation, and matched-access localized coefficient matching (P2). Lower is better. The natural development-candidate score associations appear separately in Appendix D.

5.2 Matched-access attribution, development-outcome supervision, and replication

At $B = 512$, P2 reaches 0.02693 versus 0.02510 for DLCD (6.8% lower; 7/8 favorable pools). The paired mean difference is -0.00183 (descriptive 95% interval $[-0.00288, -0.00077]$; exact two-sided pool-sign sensitivity $p = 0.0703$). Both methods use the same diagnostic and trusted-physics information; P2 omits the separate exposure group. The comparison therefore measures the full DLCD score, and Appendix F specifies an exposure-augmented P2 follow-up to refine the attribution. Localization-matched A1-SE-MPD and A1-KMD reach 0.02815 and 0.02870; global A0, using the same characteristic construction, reaches 0.02795. The corresponding DLCD reductions are 10.8%, 12.5%, and 10.2%. At matched wall clock, P2 uses $B = 633$ in 104.6 s, while DLCD uses $B = 512$ in 104.7 s; regret is 0.02641 versus 0.02520 (4.6%). A ridge meta-ranker trained with A–E candidate outcomes reaches 0.02540 on F–M, close to DLCD's 0.02510, with a 4/4 pool split. Its development labels provide supervision unavailable to DLCD; both selectors were frozen before F–M evaluation.

A second held-out Allen cohort N–U repeats the BASE and P2 comparisons. BASE reaches 0.03215 versus 0.02926 for DLCD (9.0%), while the same frozen P2 specification reaches 0.03055 versus 0.02926 (4.2%; 6/8 pools favorable). The N–U candidate-training seeds are disjoint from F–M and no P2 hyperparameter is retuned; the eight-pool exact two-sided sign sensitivity is $p = 0.2891$. Across the two Allen cohorts jointly, 13/16 pools favor DLCD over P2; Appendix K reports this explicitly secondary pooled sensitivity, while the cohort-separated analyses remain primary.

5.3 Stochastic 2-D Kolmogorov transfer

We next test a separately constructed stochastic 2-D Kolmogorov library, using the physical setting of StocBench (Pfister et al., 2026) but a different model-selection task. The decision is an extreme-enstrophy event: cross the reference-development 90th percentile and stay above the 85th percentile for $0.05T$. We evaluate nominal K0, K1 (1.10× viscosity), K2 (0.90× forcing), and near-extrapolation K3 (0.85× viscosity and 1.10× forcing). Each of six pools contains eight candidates: two seeds each of noise-conditioned FNO, Neural SPDE, an order-2 Wiener-chaos operator (Shi et al., 2026a;b), and Operator Flow Matching. This Wiener-chaos family differs from learned-basis Neural Chaos (Bahmani et al., 2025).

The audit localizes on enstrophy Z, ramping from q_{65} to q_{75}, retaining full weight through q_{85}, and tapering to zero at q_{95}. Four complex Fourier modes and five physical summaries form the 13 drift channels; the second-characteristic group uses the same saved-time calibration principle as Allen–Cahn. Appendix G and the supplementary JSON give all channels, support rules, and calibration settings.

Table 2: Decision-relevant fairness comparisons. Allen rows use equal-weight C0–C3 pool means; path-budget rows use $B = 512$ unless shown. Timing rows use DLCD 0.02520; other F–M rows use the path-budget aggregate 0.02510. Here Δ = DLCD − control; negative Δ and positive reduction favor DLCD.

Comparison	Control	DLCD	Δ	Reduction
DIRECT-D1 raw	0.03040	0.02510	-0.00530	17.4%
DIRECT-D1-SHRINK	0.02880	0.02510	-0.00370	12.8%
DIRECT-D1 timing-derived ($B = 1216$)	0.02508	0.02520	+0.00012	-0.5%
DEV-supervised meta-ranker	0.02540	0.02510	-0.00030	1.2%
P2 localized coefficient match	0.02693	0.02510	-0.00183	6.8%
P2 wall-clock matched	0.02641	0.02520	-0.00121	4.6%
BASE	0.03250	0.02510	-0.00740	22.8%
N–U P2 frozen	0.03055	0.02926	-0.00129	4.2%
N–U BASE	0.03215	0.02926	-0.00289	9.0%

Against P2 localized coefficient matching, mean $B = 512$ regret is 0.04313 versus 0.03913 for DLCD (9.3% lower), with 5/6 pools favorable. Mean paired contrast is -0.00400, with a descriptive 95% t interval of $[-0.00687, -0.00113]$. Secondary reductions are 5.2% at $B = 256$ and 7.9% at $B = 1024$. The six-pool exact two-sided sign sensitivity is $p = 0.2188$. These values come from the frozen six-pool processed records. For independent event-level replication and reanalysis, we provide one fully specified $T = 10$ 2-D event-reconstruction protocol and executable tests; the original source-run horizon is not encoded in those processed records (Appendices G and M).

Table 3: 2-D Kolmogorov $B = 512$ transfer against matched-access P2.

Pool	P2	DLCD	Reduction
KF1	.0432	.0375	13.2%
KF2	.0415	.0382	8.0%
KF3	.0447	.0390	12.8%
KF4	.0408	.0420	-2.9%
KF5	.0461	.0402	12.8%
KF6	.0425	.0379	10.8%
Mean	.04313	.03913	9.3%

5.4 Mechanism and operating boundaries

The fixed bank approximates the reference D1 continuation value on 2,048 candidate-reachable histories with $R^2 = 0.58$ overall and 0.66 inside A1, versus 0.32 outside reference support. Its development-reference fit has $R^2 = 0.66$ overall. A cellwise non-cancelling aggregation yields regret 0.02542, close to the signed selector's 0.02510, while history-absolute and tail-max alternatives reach 0.02618 and 0.02705. Eight independent DEV/NULL reference redraws produce regret 0.02496–0.02602; a reference-only automatic bank reaches 0.02582. Together these fits and ablations assess the proposed mechanism; decision-regret guarantees require stronger control of witness approximation and residual cancellation (Appendix H).

Local calibration improves the reported null size/bias over a global null (5.7%/0.13 versus 9.8%/0.58 SD); independent NULL-TEST size is 5.2%. On development pools, localization improves regret relative to an otherwise global characteristic audit (0.02298 versus 0.02874), but C3 reverses. Increasing saved snapshots from 129 to 257 leaves 19/20 selections unchanged. Pool, not resampled trajectory or reference redraw, is the scientific unit. Appendices C–K contain the calibration, ablation, secondary-decision, and multiplicity details; adverse correction, Burgers, and fractional-system cases are retained in Appendices I–J.

6 Discussion

The results identify a finite-rollout regime for selecting stochastic operators with reusable reference dynamics. Relative to localized coefficient matching (P2), regret falls by 6.8% in the first Allen–Cahn cohort, 4.2% in the independent cohort, and 9.3% in the six-pool Kolmogorov study. Allen near-extrapolation condition C3 and Kolmogorov pool KF4 reverse the trend. Because P2 omits a separate exposure group, the matched-access comparison establishes the contribution of the combined score. Exposure-augmented P2 and a directly timed equal-clock DIRECT experiment are the remaining attribution tests (Appendix N).

The computational comparison narrows this regime. DLCD uses trusted local transitions in addition to candidate trajectories; converting the measured $B = 512$ runtimes into extra DIRECT trajectories yields an effective tie. A separately timed $B = 1216$ experiment would establish the end-to-end crossover. The history-conditioned identity relates local kernel discrepancies to path decisions, while the signed finite-witness score remains an empirical proxy. Deployment depends on amortizing reference calibration, maintaining local support, and balancing short trusted transitions against additional full trajectories.

Reproducibility Statement

Appendices A–N document the study populations, proofs, event reconstruction, and comparator access rules. The anonymous supplement contains processed pool-level data, analysis scripts, event-replay implementations, statistical-unit checks, and resource accounting. Original checkpoints, complete execution logs, continuation-anchor arrays, and raw six-pool 2-D event histories fall outside this processed-data release. Appendix G specifies a $T = 10$ protocol for independent 2-D replication and event reanalysis; the published six-pool aggregates remain unchanged.

AI Use Statement

Generative AI was used for limited manuscript editing and to assist with drafting small implementation fragments for plotting and file handling. All substantive system logic, study procedures, analyses, and reported results were developed and verified independently by the author.

References

B. Bahmani, I. G. Kevrekidis, and M. D. Shields. Neural chaos: A spectral stochastic neural operator. *Journal of Computational Physics*, 539:114233, 2025.

N. Berglund. *An Introduction to Singular Stochastic PDEs: Allen–Cahn Equations, Metastability, and Regularity Structures*. EMS Press, 2022.

J. Blanchet, J. Wiesel, E. Zhang, and Z. Zhang. Empirical martingale projections via the adapted Wasserstein distance. *Annals of Applied Probability*, 36(1):547–606, 2026.

F. Bleile, S. Lumpp, and M. Drton. Efficient learning of stationary diffusions with Stein-type discrepancies. In *AISTATS*, PMLR 300:1234–1242, 2026.

L. Cao, T. O'Leary-Roseberry, P. K. Jha, J. T. Oden, and O. Ghattas. Residual-based error correction for neural operator accelerated infinite-dimensional Bayesian inverse problems. *Journal of Computational Physics*, 486:112104, 2023.

A. Chakraborty, T. Wick, T. Rabczuk, and X. Zhuang. Multigoal-oriented dual-weighted-residual error estimation using PINNs. *Machine Learning for Computational Science and Engineering*, 1:13, 2025.

B. Endtmayer, U. Langer, and T. Wick. Two-Side a Posteriori Error Estimates for the Dual-Weighted Residual Method. *SIAM Journal on Scientific Computing*, 42(1):A371–A394, 2020.

K. Hidajat. Martingale neural operators: Learning stochastic marginals via Doob–Meyer factorization. arXiv:2605.15806, 2026.

P. Holderrieth et al. Generator matching: Generative modeling with arbitrary Markov processes. In *ICLR*, 2025.

P. K. Jha. Residual-based error corrector operator to enhance accuracy and reliability of neural operator surrogates of nonlinear variational boundary-value problems. *Computer Methods in Applied Mechanics and Engineering*, 419:116595, 2024.

P. Kidger, J. Morrill, J. Foster, and T. Lyons. Neural controlled differential equations for irregular time series. In *NeurIPS*, 33:6696–6707, 2020.

N. Kovachki et al. Neural operator: Learning maps between function spaces with applications to PDEs. *JMLR*, 24(89):1–97, 2023.

S. Lee and B. L. Nelson. General-purpose ranking and selection for computer simulation. *IIE Transactions*, 48(6):555–564, 2016. doi:10.1080/0740817X.2015.1125043.

Y. Li, H. Lu, and X. Song. Kernel minimum distance estimation and testing with conditional moment restrictions: A unified framework. arXiv:2607.16605, 2026.

Z. Li et al. Fourier neural operator for parametric partial differential equations. In *ICLR*, 2021.

H. Liang and F. Liu. Shared physics responses recover hidden rankings in neural operator libraries. arXiv:2608.20441, 2026.

Q. Liu, J. D. Lee, and M. I. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *ICML*, PMLR 48:276–284, 2016.

L. Lorch, A. Krause, and B. Schölkopf. Causal modeling with stationary diffusions. In *AISTATS*, PMLR 238:1927–1935, 2024.

K. Muandet, W. Jitkrittum, and J. Kübler. Kernel conditional moment test via maximum moment restriction. In *UAI*, PMLR 124:41–50, 2020.

B. L. Nelson and F. J. Matejcik. Using Common Random Numbers for Indifference-Zone Selection and Multiple Comparisons in Simulation. *Management Science*, 41(12):1935–1945, 1995. doi:10.1287/mnsc.41.12.1935.

S. Pfister, B. Holzschuh, and N. Thuerey. StocBench: A benchmark for generative modeling of stochastic dynamics. arXiv:2608.22309, 2026.

Y. Qiu, W. Dahmen, and P. Chen. Variationally correct operator learning: Reduced basis neural operator with a posteriori error estimation. *Computer Methods in Applied Mechanics and Engineering*, 461(B):119228, 2026. doi:10.1016/j.cma.2026.119228.

C. Salvi, M. Lemercier, and A. Gerasimovics. Neural stochastic PDEs: Resolution-invariant learning of continuous spatiotemporal dynamics. In *NeurIPS*, 35:1333–1344, 2022.

D. Shi, L. Lin, A. Han, L. Thompson, J. M. Hernández-Lobato, Z. Wang, and J. Gao. Expanding the chaos: Neural operator for stochastic (partial) differential equations. arXiv:2601.01021, 2026a.

D. Shi, L. Thompson, A. Han, P. Hu, J. Gao, and J. M. Hernández-Lobato. Wiener Chaos Expansion based Neural Operator for Singular Stochastic Partial Differential Equations. arXiv:2603.08219, 2026b.

Y. Shi, Z. Ross, D. Asimaki, and K. Azizzadenesheli. Stochastic process learning via Operator Flow Matching. In *NeurIPS*, 38:42788–42828, 2025.

L. J. Shikhman. Diagnosing failure modes of neural operators across diverse PDE families. *TMLR*, 2026a.

L. J. Shikhman. Semigroup consistency as a diagnostic for learned physics simulators. In *ICML 2026 Workshop on AI for Physics*, 2026b.

L. J. Shikhman. A diagnostic software suite for auditing learned PDE simulators. arXiv:2606.18200, 2026c.

B. Srikishan, J. E. Santos, N. Muralidhar, and C. D. Young. TRIE: An evaluation framework for stochastic PDE surrogates. arXiv:2607.00196, 2026.

J. D. Toscano, D. T. Chen, V. Ooomen, J. Darbon, and G. E. Karniadakis. A variational framework for residual-based adaptivity in neural PDE solvers and operator learning. *npj Artificial Intelligence*, 2:32, 2026. doi:10.1038/s44387-026-00084-4.

Y. Zhu, Z. Li, D. Shi, L. Thompson, O. Nash, J. M. Lara Rangel, S. Li, B. Chen, R. Zhu, Q. Meng, and H. Ni. SPDEBench: An Extensive Benchmark for Learning Stochastic PDEs. arXiv:2505.18511v3, 2025; revised 2026.

J. Zou, H. C. Lie, and Y. Marzouk. Goal-oriented learning of stochastic differential equations using error bounds on path-space observables. arXiv:2603.20467v2, 2026.

A ALLEN–CAHN BENCHMARK, DECISIONS, AND REFERENCE REGIME

The primary benchmark is stochastic Allen–Cahn on $[0, 1]$ with periodic boundary conditions and $T = 6$,

$$dU = [\kappa U_{xx} + U - U^3 + h_{\text{ext}}]\, dt + \sigma Q^{1/2} dB.$$

Here B denotes cylindrical Wiener noise, so $Q^{1/2}dB$ has the Q-Wiener covariance used in the reference-dynamics formulation. The reference solver is semi-implicit spectral Euler (implicit Laplacian, explicit cubic, $2/3$ de-aliasing) with $N_x = 256$ and $\Delta t = 2^{-11}$; convergence uses $N_x = 512$ and $\Delta t = 2^{-12}$. The audit stores 129 snapshots ($\Delta t_{\text{audit}} = 6/128$). The noise spectrum is $q_k \propto [1 + (|k|/8)^2]^{-2}$ with Hermitian symmetry and trace normalization. Initial conditions are

$$u_0(x) = -0.85 + \sum_{k=1}^{4} [a_k \cos(2\pi kx) + b_k \sin(2\pi kx)], \qquad a_k, b_k \sim \mathcal{N}(0, 0.03^2),$$

clipped to $[-1.10, -0.60]$. Training draws $\kappa \sim U[0.004, 0.006]$, $\sigma \sim U[0.90\sigma^\star, 1.10\sigma^\star]$, and $h_{\text{ext}} \sim U[-0.015, 0.015]$. A reference-only sweep over $\sigma \in \{0.25, 0.35, 0.45, 0.55, 0.65\}$ gives C0 D1 probabilities $\{0.032, 0.086, 0.218, 0.341, 0.474\}$; $\sigma^\star = 0.45$ is selected in the 10–30% regime.

D1 is the first piecewise-linearly interpolated time when the spatial mean crosses $+0.65$ and remains at least $+0.55$ for 0.15, from an initial mean at most -0.65; incomplete dwell by T is censored. The D1 loss uses weights $0.60/0.25/0.15$, making commitment probability the majority component while retaining timing and distribution information. Six alternative weightings in the supplement preserve the qualitative development conclusion, with relative reductions from 22.9% to 34.2%. D2 uses the free energy

$$F(u) = \int_0^1 \left[\frac{\kappa}{2} |u_x|^2 + \frac{1}{4}(u^2 - 1)^2 - h_{\text{ext}} u \right] dx$$

with fixed threshold $\bar{F} = 0.140$. D3 is a sustained positive-phase-fraction event; D4 is a sustained local physical-box nucleation event. The primary D1 loss is the weighted probability/RMST/Wasserstein loss in the main text.

Table A.1: Reference event regime from 32,768 ORACLE-REF trajectories per condition.

Cond.	D1 count / P [95% CI]	D2 count / P	D3 count / P	D4 count / P	refine ΔP	null FP
C0	7137 / 0.217804 [0.2134,0.2223]	8087 / 0.246796	6294 / 0.192078	3637 / 0.110992	+0.004	5.4%
C1	4647 / 0.141815 [0.1381,0.1456]	5598 / 0.170837	3796 / 0.115845	2261 / 0.069000	+0.006	4.6%
C2	9958 / 0.303894 [0.2989,0.3089]	11013 / 0.336090	8977 / 0.273956	5145 / 0.157013	-0.003	5.8%
C3	12644 / 0.385864 [0.3806,0.3911]	13794 / 0.420959	11504 / 0.351074	7012 / 0.213989	+0.008	6.1%

Direct-outcome selection and sampling resolution. At the C0 reference rate $p = 0.217804$, an empirical commitment probability based on B independent rollouts has standard error $\sqrt{p(1-p)/B}$: 0.0365, 0.0258, 0.0182, 0.0129, and 0.0091 at $B = 128, 256, 512, 1024, 2048$. At this same rate, a 4,096-path reference probability estimate has binomial standard error 0.00645 (0.00545, 0.00719, and 0.00761 at the C1–C3 reference rates), so the reusable reference probability target is substantially more precise than a $B = 512$ candidate probability estimate but remains finite. For two independent candidate estimates near the C0 rate, the difference standard error at $B = 512$ is 0.0258. DIRECT-D1 therefore computes the full Eq. (9) loss directly on each candidate's DIAG-MODEL paths using a DEV-REF target. DIRECT-D1-SHRINK is selected using A–E/DEV only and applied unchanged to F–M: if x of B paths commit, the probability estimate is the Jeffreys posterior mean $(x + 1/2)/(B + 1)$; normalized RMST is shrunk toward the DEV-REF target as $\tilde{r} = (Br + 32r_{\text{ref}})/(B + 32)$, where the pseudo-count 32 is selected by leave-one-development-pool-out error; and the censored-time W_1 term remains empirical because the development smoothing grid did not improve leave-one-pool-out error. No fresh F–M outcome enters these choices. Neither selector accesses DEC-MODEL/ORACLE candidate outcomes. The supplement releases the complete A–E leave-one-pool-out trace over the evaluated RMST pseudo-count and W_1 smoothing grids. Table E.2 reports the resulting finite-budget comparison.

B CANDIDATE POPULATION AND MODEL CONFIGURATIONS

Pools A–E use seed pairs $\{1,2\},\ldots,\{9,10\}$ per family. Fresh pools F–M use seed pairs $\{11,12\},\ldots,\{25,26\}$, adding 64 independent fits. Each pool uses 10,000 TRAIN and 2,000 VAL trajectories with non-overlapping data/noise seeds. Shared evaluation references are held out. Other split sizes are DEV-REF 4,096, NULL-REF 16,384, DIAG-MODEL 8,192/candidate/condition, DEC-MODEL 8,192/candidate/condition, ORACLE-REF 32,768/condition, and CONV-REF 2,048 high-resolution trajectories.

Table B.1: Candidate-family configurations. No D1–D4 loss enters candidate fitting or architecture selection.

Family	Params	Batch	Epochs	LR	Objective	Time/seed
Noise-conditioned FNO	1,367,681	16	250	0.001	relative path-L2 + 0.10 * mean relative L2 over first 24 Fourier coefficients across saved times	2.8 h/seed
Neural SPDE	2,941,824	8	220	0.0003	mean relative pathwise L2 over all saved space-time snapshots	5.6 h/seed
NCDE-FNO	2,183,105	8	240	0.0005	relative path-L2 + 0.05 * mean normalized one-step derivative discrepancy on saved intervals	4.9 h/seed
Structured/WCE	1,624,772	16	200	0.0008	0.65 * normalized chaos-coefficient loss + 0.35 * trajectory reconstruction relative L2	3.4 h/seed

Noise-conditioned FNO uses four width-64 Fourier blocks with 24 modes. Neural SPDE uses latent dimension 64 and fixed-point evaluation with 24 spatial modes. NCDE–FNO is a custom CDE/Fourier hybrid with hidden dimension 64 and four width-48 Fourier blocks. Structured/WCE uses Wiener-chaos order 2, 16 KL coordinates, 153 Hermite-chaos terms, and four width-48 Fourier blocks.

C COMPARATOR INFORMATION SETS AND CALIBRATION

Table C.1: Information contracts. DLCD and matched-physics controls exclude candidate D1–D4 summaries; outcome-aware controls may estimate D1 only from finite DIAG-MODEL paths. All held-out DEC/ORACLE outcomes remain evaluation-only.

Selector class	Uses at selection	Excludes
DLCD / P1–P3	Candidate diagnostic trajectories; trusted local physics; DEV/NULL calibration; reference-only task geometry	Candidate diagnostic D1–D4 summaries; all DEC/ORACLE outcomes
DIRECT-D1 / SHRINK / adaptive	Candidate diagnostic trajectories; DEV-REF D1 target; development-fixed regularization/allocation rules	All DEC/ORACLE candidate outcomes; fresh-outcome tuning
DEV-supervised meta-ranker	A–E development outcome labels plus fixed diagnostic features	F–M DEC/ORACLE outcomes and any fresh tuning
Evaluation	DEC-MODEL and ORACLE-REF	Feedback into any frozen selector

E0 uses four-time scalar Wasserstein summaries, 32 projected-field sliced-Wasserstein projections, mean/covariance summaries, and CRPS; same-noise path error is auxiliary. E1 uses autocorrelation, lagged cross-covariances, joint-time MMD, signature/conditional-signature MMD, and a DEV-early-stopped GRU discriminator. KCM/MMR and KMD use $X = [m, F, \Re\hat{u}_1, \Im\hat{u}_1, \kappa, \sigma, h, t/T]$ with DEV/NULL-only tuning. SE-MPD evaluates 2-D, 4-D, and drift-compensated 4-D variants using DEV data and carries the DEV-selected variant into evaluation.

The DLCD calibration dispatch is fixed before candidate evaluation. Within each Allen condition, the calibration state is $z = [m, F, v_+, \log(E_{\text{low}} + 10^{-8}), \log(E_{\text{high}} + 10^{-8}), t/T]$. NULL-REF

means/standard deviations standardize the local regression coordinates; spatial mean and t/T remain on their raw scales for bin assignment. The spatial mean is divided into 12 equal-frequency NULL-REF bins and time into $[0, .25), [.25, .50), [.50, .75), [.75, 1]$. When an analytic or matched scheme-discrete target is unavailable, the target is predicted inside the corresponding mean–time cell by ridge regression ($\lambda = 10^{-4}$) on standardized $(F, v_+, \log E_{\mathrm{low}}, \log E_{\mathrm{high}})$. Local regression is used only with at least 200 reference increments and Mahalanobis radius at most 3.5. Otherwise 64 trusted one-step reference bursts are launched from the candidate saved state; at most 128 states are triggered per candidate, giving the 8,192-transition cap. Drift targets are conditional mean one-step increments divided by Δt; the second-characteristic target in Eq. (7) is the reference-centered increment-product rate at the same saved-time state. For 64 trusted bursts the corresponding sample cross-covariance uses the unbiased $n - 1$ denominator. P1 uses a separately matched raw-second-moment target for its raw-increment statistic; P2 matches centered local covariance coefficients. The deployed audit uses 10.6% analytic/scheme targets, 84.7% local calibration, 4.7% trusted bursts, and 0.0% unsupported channels. The shrinkage covariance is $0.9S + 0.1\,\mathrm{diag}(S) + 10^{-6}I$. The machine-readable reference contract is supplied in `calibration_dispatch_config.json`; its historical as-released version is retained for provenance comparison.

Finite-step interpretation of the second-characteristic group. At a fixed saved state u, let $b^{\theta}_{j,h} = \mathbb{E}_\theta[\Delta\phi_j \mid u]/\Delta t$ and $q^{\theta}_{j\ell,h} = \mathrm{Cov}_\theta(\Delta\phi_j, \Delta\phi_\ell \mid u)/\Delta t$, and analogously for the reference. If the same reference-centered increment is used in both terms of Eq. (7), then the following is an *exact* finite-step identity:

$$\frac{\mathbb{E}_\theta[(\Delta\phi_j - \Delta t\, b^*_{j,h})(\Delta\phi_\ell - \Delta t\, b^*_{\ell,h}) \mid u]}{\Delta t} - q^*_{j\ell,h} = (q^{\theta}_{j\ell,h} - q^*_{j\ell,h}) + \Delta t(b^{\theta}_{j,h} - b^*_{j,h})(b^{\theta}_{\ell,h} - b^*_{\ell,h}).$$

The right-hand drift-mismatch term is $O(\Delta t)$ when finite-step drift rates are bounded. This conditional identity concerns the true saved-state reference targets; local-regression and finite-burst targets introduce additional approximation and sampling error. The q-group is covariance-sensitive, not exactly drift-orthogonal at the saved-time resolution. The 129-to-257-snapshot sensitivity in Section 5.4 leaves 19/20 selections unchanged ($\rho = 0.982$); P2 separately estimates local drift/covariance coefficients.

Table C.2: Numerical calibration support.

Calibration	empirical size	median state bias
Global null	9.8%	0.58 SD
Local reference	5.7%	0.13 SD
Triggered bursts	4.6%	0.09 SD
Scheme-discrete	5.9%	0.08 SD

D NATURAL CANDIDATE ASSOCIATION AND ARCHITECTURE CONTROLS

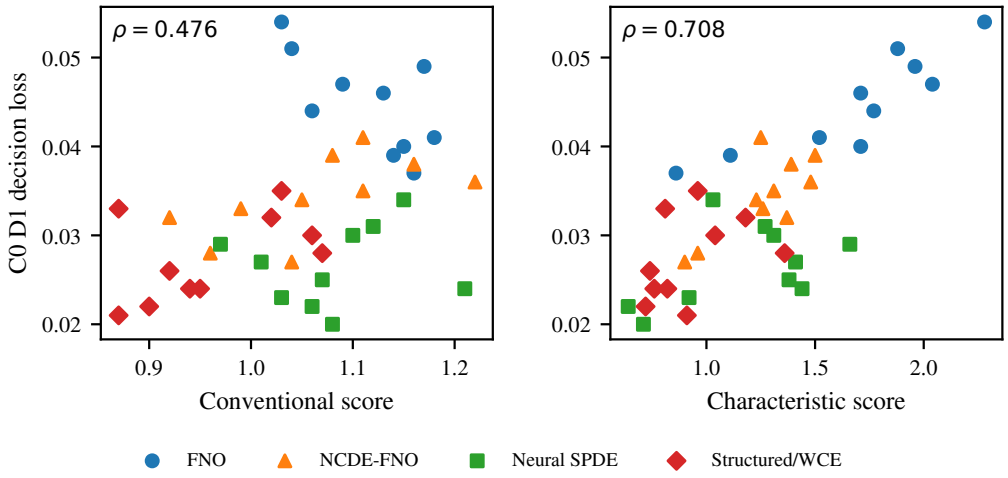


Figure D.1: Development-only natural C0 candidates: conventional and characteristic-only scores against held-out D1 loss, grouped by candidate family. This diagnostic association is distinct from the fresh F–M selection comparison in Figure 2.

Table D.1: Natural C0 association by family and leave-one-family-out sensitivity.

Analysis	family/exclusion	n	ρ(BASE,D1)	ρ(DLCD,D1)
within	FNO	10	-0.588	0.912
within	Neural SPDE	10	0.273	0.394
within	NCDE-FNO	10	0.748	0.624
within	Structured/WCE	10	0.439	0.456
LOFO	FNO	30	0.407	0.522
LOFO	Neural SPDE	30	0.658	0.842
LOFO	NCDE-FNO	30	0.422	0.739
LOFO	Structured/WCE	30	0.299	0.687

Across 40 C0 candidates, pooled Spearman correlations are 0.476 for BASE and 0.708 for DLCD. The within-family pooled values are 0.217 and 0.596. Partial Spearman association is 0.632 after residualizing ranked T_{DLCD} and D1 loss against ranked S_{BASE} alone, and 0.531 after additionally controlling family indicators. These answer different adjustment questions rather than representing conflicting estimates. At $B = 512$, family-restricted selection gains remain positive for all four candidate families; the machine-readable supplement supplies the exact family rows and conventional-quality-threshold sensitivities.

E HELD-OUT SELECTION AND D1 SENSITIVITY

Table E.1: Fresh F–M $B = 512$ poolwise outcome-aware comparison. Lower is better.

Pool	DIRECT-D1	DIRECT-SHRINK	DLCD	DLCD−SHRINK
F	.0306	.0294	.0248	-.0046
G	.0289	.0235	.0238	+.0003
H	.0315	.0300	.0230	-.0070
I	.0292	.0284	.0239	-.0045
J	.0308	.0300	.0253	-.0047
K	.0277	.0273	.0231	-.0042
L	.0236	.0238	.0241	+.0003
M	.0409	.0380	.0328	-.0052
Mean	.03040	.02880	.02510	-.00370

Table E.2: Fresh F–M outcome-aware budget curve.

B	DIRECT-D1	DIRECT-SHRINK	DLCD	DLCD reduction vs SHRINK
128	.0467	.0445	.0406	8.8%
256	.0394	.0369	.0336	8.9%
512	.0304	.0288	.0251	12.8%
1024	.0261	.0254	.0240	5.5%
2048	.0230	.0225	.0219	2.7%

The raw DIRECT-D1 $B = 512$ contrast is favorable for DLCD in 7/8 pools, with descriptive paired 95% interval $[-0.00759, -0.00301]$. Against DIRECT-D1-SHRINK, 6/8 pools favor DLCD and the corresponding interval is $[-0.00589, -0.00151]$. The two shrinkage reversals are G and L; the larger-budget convergence in Table E.2 is consistent with the finite-sampling interpretation rather than universal dominance of indirect diagnostics.

Table E.3: Fresh F–M BASE/DLCD and condition-level confirmation at $B = 512$.

Pool	BASE	DLCD	Δ
F	0.0340	0.0248	-0.0092
G	0.0324	0.0238	-0.0086
H	0.0331	0.0230	-0.0101
I	0.0317	0.0239	-0.0078
J	0.0348	0.0253	-0.0095
K	0.0320	0.0231	-0.0089
L	0.0306	0.0241	-0.0065
M	0.0314	0.0328	+0.0014
Mean	.0325	.0251	-.0074

Table E.4: Fresh condition-level $B = 512$ regret.

Condition	BASE	DLCD	Reduction
C0	0.03400	0.02346	+31.0%
C1	0.03100	0.02511	+19.0%
C2	0.03700	0.02331	+37.0%
C3	0.02800	0.02852	-1.9%

F MATCHED-ACCESS CONTROLS AND SUBSEQUENT HELD-OUT ALLEN COHORT

All P1/P2/P3 controls use the same F–M candidate trajectories, A1 localization, access to the trusted evolution/local transition reference, DEV/NULL calibration, saved-time grid, and projected physical state as DLCD. Each control excludes candidate D1–D4 summaries and all DEC/ORACLE outcomes during selection. "Matched access" means that the same information sources are available; trusted-transition query counts can differ and are reported separately.

Exact comparator estimators. Index localized saved transitions by $i = (n, k)$ with weight a_i, calibration state $z_i = [m, F, v_+, \log E_{\mathrm{low}}, \log E_{\mathrm{high}}, t/T]$, and witness increment $\Delta\phi_{ij} = \phi_j(U_{i,+}) - \phi_j(U_i)$. Let $b_j^*(z_i)$ be the trusted saved-time conditional drift, $q_{j\ell}^{*,\mathrm{raw}}(z_i) = \mathbb{E}_{\mathrm{ref}}[\Delta\phi_j \Delta\phi_\ell \mid z_i]/\Delta t$ the raw second-moment rate used by P1, and $q_{j\ell}^{*,\mathrm{cov}}(z_i) = \mathrm{Cov}_{\mathrm{ref}}(\Delta\phi_j, \Delta\phi_\ell \mid z_i)/\Delta t$ the centered rate used by P2. For any residual vector $r \in \mathbb{R}^d$, write

$$\mathcal{N}(r;\Sigma) = \sqrt{r^\top [0.90\Sigma + 0.10\,\mathrm{diag}(\Sigma) + 10^{-6} I]^{-1} r/d},$$

with Σ estimated on NULL-REF for the corresponding channel. P1 uses direct localized first-/second-characteristic residuals

$$r_{b,j}^{(1)} = \frac{\sum_i a_i(\Delta\phi_{ij} - \Delta t\, b_j^*(z_i))}{\sum_i a_i \Delta t + 10^{-8}}, \qquad r_{q,j\ell}^{(1)} = \frac{\sum_i a_i(\Delta\phi_{ij}\Delta\phi_{i\ell} - \Delta t\, q_{j\ell}^{*,\mathrm{raw}}(z_i))}{\sum_i a_i \Delta t + 10^{-8}}.$$

P2 first fits candidate-local coefficient functions $\hat{b}_j^\theta(z)$ and $\hat{q}_{j\ell}^\theta(z)$ from DIAG-MODEL transitions using the fixed six-dimensional feature map, the same 12 mean bins/four time bins, and ridge 10^{-4}; it then averages coefficient mismatch,

$$r_{b,j}^{(2)} = \frac{\sum_i a_i[\hat{b}_j^\theta(z_i) - b_j^*(z_i)]}{\sum_i a_i + 10^{-8}}, \qquad r_{q,j\ell}^{(2)} = \frac{\sum_i a_i[\hat{q}_{j\ell}^\theta(z_i) - q_{j\ell}^{*,\mathrm{cov}}(z_i)]}{\sum_i a_i + 10^{-8}}.$$

Thus $T_{P1} = \sqrt{[\mathcal{N}(r_b^{(1)};\Sigma_b)^2 + \mathcal{N}(r_q^{(1)};\Sigma_q)^2]/2}$ and T_{P2} is defined identically with $r^{(2)}$. In the reported comparator, P2 does not append a standalone T_{occ} group. A feature-controlled extension $T_{P2+\mathrm{occ}} = \sqrt{(T_{\mathrm{occ}}^2 + 2T_{P2}^2)/3}$ is specified as a follow-up diagnostic; no unmeasured regret is attributed to it. P3 is restricted to the fixed smooth projected observables and fits the candidate one-step finite-difference generator $\hat{G}_h^\theta \phi_j(z)$ with the same local regression; with trusted saved-time generator $G_h^* \phi_j(z)$,

$$r_{G,j}^{(3)} = \frac{\sum_i a_i[\hat{G}_h^\theta \phi_j(z_i) - G_h^* \phi_j(z_i)]}{\sum_i a_i + 10^{-8}}, \qquad T_{P3} = \mathcal{N}(r_G^{(3)};\Sigma_G).$$

Each control uses $S_{Pq} = \sqrt{(T_{E0}^2 + T_{E1}^2 + T_{Pq}^2)/3}$. The supplement supplies the machine-readable estimator contract and tested deterministic score-composition reference code.

Wall-clock match. The normalized compute cost is wall-clock relative to the deployed DLCD $B = 512$ stack,

$$C_{\mathrm{norm}}(B) = t_{P2}(B)/t_{\mathrm{DLCD}}(512).$$

P2 takes 84.7 s at $B = 512$ versus 104.7 s for DLCD, so the reported wall-clock path-count conversion is $B_{P2} = \mathrm{round}[512(104.7/84.7)] = 633$; held-out D1–D4 outcomes are not used in this conversion. The direct matched run is 104.6 s, hence $C_{\mathrm{norm}} = 104.6/104.7 = 0.999$. This comparison equalizes wall clock, not trusted-transition usage. DLCD's triggered reference bursts have a median of about 2,300 additional solver transitions per candidate (p90 about 6,200; cap 8,192); P2's target-evaluation proxy and DLCD burst transitions are reported as distinct resource axes in the supplement rather than treated as commensurate calls.

Table F.1: Matched-access controls at $B = 512$. Lower is better.

Method	Control regret	DLCD regret	Reduction
P1 direct drift/second-characteristic	0.02720	0.02510	+7.7%
P2 localized coefficient match	0.02693	0.02510	+6.8%
P3 finite-difference generator	0.02758	0.02510	+9.0%

Table F.2: Poolwise P2 path-matched comparison at $B = 512$. Lower is better.

Pool	P2	DLCD	Reduction
F	0.02680	0.02480	+7.5%
G	0.02540	0.02380	+6.3%
H	0.02440	0.02300	+5.7%
I	0.02550	0.02390	+6.3%
J	0.02720	0.02530	+7.0%
K	0.02480	0.02310	+6.9%
L	0.02400	0.02410	-0.4%
M	0.03730	0.03280	+12.1%

Table F.3: Outcome-aware and development-supervised fairness checks on F–M.

Comparator	Control regret	DLCD regret	Reduction	favorable pools
DIRECT-D1, $B = 512$	.03040	.02510	+17.4%	7/8
DIRECT-D1-SHRINK, $B = 512$	.02880	.02510	+12.8%	6/8
DIRECT-D1, time-normalized $B = 1216$	.02508	.02520	-0.5%	–
DEV-supervised meta-ranker	.02540	.02510	+1.2%	4/8

DIRECT-D1 takes 44.1 s at $B = 512$; the timing-derived conversion gives the path-equivalent $B = 1216$ for DLCD's 104.7 s online budget. This conversion equalizes the timing-derived path budget; $B = 1216$ was not separately timed end to end. The meta-ranker standardizes the fixed $(E0, E1, \mathrm{occ}, \mathrm{drift}, 2\mathrm{nd}, P2)$ feature vector using A–E only and fits ridge regression to development D1 loss. Leave-one-development-pool-out selection over $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ selects $\lambda = 10^{-2}$; the fitted ranker is then frozen for F–M. It therefore has strictly more development supervision than DLCD but no F–M outcome access.

Table F.4: Complementary fresh-Allen fairness and independent-confirmation checks. Lower regret is better.

Check	Control	DLCD	Result
A1-SE-MPD	.02815	.02510	10.8% lower
A1-KMD	.02870	.02510	12.5% lower
A0 global characteristic	.02795	.02510	10.2% lower
Disjoint DEC/ORACLE	.03236	.02618	19.1% lower; rank $\rho = .936$
Independent NULL-TEST	–	–	5.2% size; 0.11 null-SD median bias
Refined C0 / C3	–	–	rank $\rho = .969/.958$; 7/8 agreement each

Table F.5: Held-out N–U Allen cohort at $B = 512$ with the same P2 specification fixed from development.

Pool	BASE	P2	DLCD	DLCD vs P2
N	.0318	.0294	.0275	+6.5%
O	.0330	.0312	.0292	+6.4%
P	.0322	.0300	.0285	+5.0%
Q	.0311	.0310	.0318	-2.6%
R	.0341	.0311	.0288	+7.4%
S	.0307	.0296	.0276	+6.8%
T	.0328	.0303	.0284	+6.3%
U	.0315	.0318	.0323	-1.6%
Mean	.03215	.03055	.02926	+4.2%

The P1/P2/P3 aggregate regrets are 0.02720/0.026925/0.02758 versus 0.02510 for DLCD. P2 scientific-unit rows are released for both F–M and N–U. Under the P2 wall-clock match, P2 uses

$B = 633$ (104.6 s) and DLCD $B = 512$ (104.7 s), with regrets 0.02641 and 0.02520. The direct-outcome time-normalized sensitivity is less favorable to DLCD and is reported separately rather than folded into the measured P2 wall-clock comparison.

G STOCHASTIC 2-D KOLMOGOROV TRANSFER

The transfer uses six eight-candidate pools with noise-conditioned FNO, Neural SPDE, WCE-NO, and Operator Flow Matching. The physical setup follows the StocBench-aligned stochastic 2-D Kolmogorov regime (Pfister et al., 2026) on $[0, 2\pi]^2$, with vorticity dynamics $d\omega = [-(v \cdot \nabla)\omega + \nu\Delta\omega - \alpha\omega]dt + \epsilon\, d\eta$, $\nu = 10^{-3}$, $\alpha = 0.1$, nominal $\epsilon = 1.0$, stochastic forcing supported on Fourier modes $(6, 0), (7, 0), (5, 5), (8, 8)$, pseudo-spectral Euler–Maruyama reference integration on a 256×256 grid with solver step 10^{-4} and saved snapshots every 0.5, and a 64×64 model grid. The primary path decision is a reference-DEV-defined extreme-enstrophy persistence event: cross the reference 90th percentile and remain above the 85th percentile for $0.05T$. Conditions are K0 nominal, K1 $\nu \times 1.10$, K2 forcing amplitude $\times 0.90$, and K3 $\nu \times 0.85$ with forcing $\times 1.10$.

System-specific DLCD contract. Let Z denote enstrophy and q_r its DEV-REF percentile. The fixed K-A1 localization is

$$a_K(Z) = s\left(\frac{Z - q_{65}}{q_{75} - q_{65}}\right) s\left(\frac{q_{95} - Z}{q_{95} - q_{85}}\right),$$

so unit weight lies on $[q_{75}, q_{85}]$ with smooth reference-defined shoulders q_{65}–q_{75} and q_{85}–q_{95}. The 13 drift witnesses are the real/imaginary coefficients of Fourier modes $(6, 0), (7, 0), (5, 5), (8, 8)$; kinetic energy E; enstrophy Z; palinstrophy P; and low/high radial enstrophy energies ($|k| \le 8$ and $8 < |k| \le 16$). The second-characteristic group uses the corresponding eight Fourier real/imaginary diagonal channels plus E and Z, together with six fixed cross channels: Re/Im within each of the four modes, Re$(6, 0)$–Re$(7, 0)$, and Re$(5, 5)$–Re$(8, 8)$. The local calibration state is $[\log(Z + 10^{-8}), \log(P + 10^{-8}), \log(E_{low} + 10^{-8}), \log(E_{high} + 10^{-8})$, injection rate, $t/T]$, standardized on NULL-REF. Calibration uses 12 equal-frequency $\log Z$ bins, four time bins, ridge 10^{-4}, at least 200 reference increments, and Mahalanobis radius at most 3.5; unsupported local states dispatch to 64 trusted one-step bursts, with at most 128 triggered states (8,192 transitions) per candidate. Group shrinkage and outer score composition are identical to Allen–Cahn. The 2-D E0 stack uses four-time W1/CRPS summaries for $E, Z, P, \log(E_{high}/E_{low})$, 32 fixed field/Fourier projections, and forced-mode mean/covariance summaries; E1 uses autocorrelations, lagged cross-covariances, four-time joint MMD, level-4 signature/conditional-signature MMD, and the fixed DEV-early-stopped causal path discriminator. The machine-readable localization, witness, calibration, and score contract is released in `kolmogorov_dlcd_protocol.json`. The accompanying benchmark protocol records the public StocBench solver/configuration revision used to identify the physical constants, and the six released scientific-unit pool rows reconstruct the reported 2-D aggregates. The public benchmark training sequence length does not specify this study's evaluation horizon or its event interpolation rule.

Frozen event-reconstruction protocol for independent 2-D replication. For subsequent independently generated Kolmogorov cohorts, we fix the fully specified evaluation contract before candidate outcomes are observed: study horizon $T = 10$, the existing 0.5-unit audit interval (21 saved audit states), and an additional 0.05-unit scalar enstrophy event-monitor interval (201 observations including both endpoints). We retain the benchmark-aligned 256×256 reference solver, 10^{-4} integration step, 64×64 candidate grid, and K0–K3 physical conditions described above. A shared initial-state law comes from a separately seeded nominal-reference stationary bank after a 50-unit burn-in, with thinning checked on an independent pilot; reference and candidate roles have disjoint seeds. For each K condition, $4{,}096$ independent DEV-REF trajectories supply equal trajectory–time weighted enstrophy observations on the 201-point monitor grid. Empirical $q_{65}, q_{75}, q_{85}, q_{90}, q_{95}$ use linear sample-quantile interpolation and are frozen before DIAG/DEC candidate evaluation. The event is the first strict upward crossing of q_{90}, linearly interpolated between monitor observations, followed by continuous piecewise-linear occupation at or above q_{85} for $0.5 = 0.05T$ time units. Failed dwell attempts permit subsequent crossings; incomplete terminal dwell is censored at T; an initially above-q_{90} trajectory is not counted as an upward crossing. Normalize the censored crossing time by T and retain the same $0.60/0.25/0.15$ event-probability/RMST/W_1 loss. The independently frozen replication uses

eight new eight-candidate pools with $B = 512$ primary, $B = 256, 1024$ secondary, $1{,}000$ common-index resamples, $16{,}384$ NULL-REF and $8{,}192$ ORACLE-REF paths per condition, and a separately reported 256-path high-resolution convergence audit per condition. The source-independent reference and event replay is supplied as `kolmogorov_followup_event_protocol.json` and `kolmogorov_followup_event.py`. This frozen prospective protocol specifies how the transfer claim can be replicated; it does not retroactively assign an unrecorded horizon or quantile population to the six previously reported KF1–KF6 pool results.

Table G.1: Poolwise 2-D transfer against matched-access P2 at $B = 512$.

Pool	P2	DLCD	Reduction
KF1	0.0432	0.0375	+13.19%
KF2	0.0415	0.0382	+7.95%
KF3	0.0447	0.0390	+12.75%
KF4	0.0408	0.0420	-2.94%
KF5	0.0461	0.0402	+12.80%
KF6	0.0425	0.0379	+10.82%

Secondary-budget reductions are 5.2% at $B = 256$ and 7.9% at $B = 1024$. A separately fixed spectral-imbalance decision improves by 6.4%. On selected models, CRPS, one-step energy distance, enstrophy-spectrum discrepancy, invariant-measure discrepancy, and long-rollout spectral error improve by 3.8%, 5.1%, 6.7%, 2.6%, and 3.3%, respectively.

Reanalysis of the six released pools. The same $T = 10$ contract is also the fixed *recommended reconstruction convention* when original six-pool scalar event trajectories are recovered: 201 equispaced enstrophy monitor values on $[0, 10]$, separate conditionwise DEV-REF q_{85} and q_{90} from $4{,}096$ trajectories with NumPy linear interpolation, strict below-to-at/above q_{90} upcrossings, continuous piecewise-linear persistence above q_{85} for 0.5, retry after failed dwell, no event from an initially above-q_{90} path, and right censoring at 10. Reconstruct the original frozen candidate selections and evaluate the same $0.60/0.25/0.15$ loss from distinct DEC and ORACLE populations; record per-pool/per-condition event counts, quantile values, seeds, source hashes, and the match or mismatch with the released KF1–KF6 scientific-unit rows. The default is a concrete, canonical *reanalysis recommendation*; it is not evidence that the unbundled original six-pool runs used this horizon. If recovered execution metadata show a different setting, use that actual setting to authenticate the historical figures and retain $T = 10$ for the new independent cohort.

H Theory bridge, proofs, and localization mechanism

H.1 History-conditioned decomposition

For the reference recursion $v_k(H_k) = P_k v_{k+1}(H_k)$, under the candidate law,

$$\mathbb{E}_Q[v_{k+1}(H_{k+1}) - v_k(H_k)] = \mathbb{E}_Q[((Q_k - P_k)v_{k+1})(H_k)].$$

Summing over k telescopes to $\mathbb{E}_Q[g] - \mathbb{E}_{\mu_Q} v_0$. The reference recursion gives $\mathbb{E}_P[g] = \mathbb{E}_{\mu_P} v_0$, yielding Proposition 1. The support condition in Proposition 1 ensures that the reference continuation recursion is defined on histories reached under the candidate law.

Corollary 1 (D1 continuation-discrepancy bound)**.** *Assume the condition-specific initial law is shared. Let $g_e = \mathbf{1}\{\tau \le T\}$, $g_r = \bar{\tau}$, and $g_s = \mathbf{1}\{\bar{\tau} \le s\}$ for $s \in [0, 1]$, with reference continuations v^e, v^r, v^s. Define $d^a_k(H_k) = ((Q_k - P_k)v^a_{k+1})(H_k)$. Then*

$$\ell_{D1}(Q, P) \le \sum_{k=0}^{K-1} \mathbb{E}_Q \left[0.60|d^e_k| + 0.25|d^r_k| + 0.15 \int_0^1 |d^s_k| \, ds \right].$$

Indeed, Proposition 1 applies to g_e, g_r, and each g_s; the triangle inequality and Tonelli's theorem give the stated bound using $W_1(\bar{\tau}_Q, \bar{\tau}_P) = \int_0^1 |F_Q(s) - F_P(s)| \, ds$. The bound acts on task continuations,

not on the deployed signed characteristic compression, so it motivates a target-aligned one-step audit without turning T_{DLCD} into a certificate.

If, on localized support, $v_{k+1}(H_k, x) = c_k(H_k) + \sum_j \alpha_{k,j}(H_k)\phi_j(x) + r_k(H_k, x)$ with $\|r_k(H_k, \cdot)\|_\infty \le \epsilon_k(H_k)$, constants cancel under $Q_k - P_k$ and

$$|a_k((Q_k - P_k)v_{k+1})(H_k)| \le a_k \sum_j |\alpha_{k,j}|\,|(Q_k - P_k)\phi_j| + 2a_k\epsilon_k.$$

The unlocalized contribution is bounded by $2(1 - a_k)\|v_{k+1}\|_\infty$. This motivates the fixed bank while retaining an explicit approximation and leakage remainder. This is a history-level inequality. Because the deployed score averages signed residuals before taking a group norm, history/time cancellation can occur when the coefficients $\alpha_{k,j}(H_k)$ vary. Accordingly, neither Proposition 1 nor this inequality implies that small T_{DLCD} certifies small decision error; the theory provides a mechanism-level rationale whose approximation quality is evaluated empirically below.

Proof of Proposition 2. For independent trajectory clusters, the unlocalized estimator $\bar{Y}_N = N^{-1}\sum_i Y_i$ has first-order variance $\mathrm{Var}(Y)/N$ and Gaussian squared SNR $N(\mathbb{E}Y)^2/\mathrm{Var}(Y)$. The localized self-normalized estimator $\widehat{\delta}_W = \sum_i W_i Y_i / \sum_i W_i$ has, by the delta method for the ratio of sample means,

$$\sqrt{N}(\widehat{\delta}_W - \delta_W) \Rightarrow \mathcal{N}\left(0, \frac{\mathrm{Var}[W(Y - \delta_W)]}{(\mathbb{E}W)^2}\right).$$

Its first-order squared SNR is therefore $N\delta_W^2(\mathbb{E}W)^2/\mathrm{Var}[W(Y - \delta_W)]$. Comparing the two squared SNRs and cancelling N gives Eq. (11).

H.2 Candidate-history bridge

Table H.1: Candidate-history theory-to-witness-bank bridge summary over 2,048 anchors.

Scope	anchors	cont./anchor	R^2	MAE	RMSE	cal. slope
all anchors	2048	1024	0.580	0.049	0.068	0.91
inside A1	1021	1024	0.660	0.043	0.061	0.94
outside A1	1027	1024	0.370	0.056	0.076	0.85
in reference support	1888	1024	0.610	0.047	0.065	0.93
out of reference support	160	1024	0.320	0.069	0.086	0.78
high-precision subset	512	4096	0.600	0.047	0.066	0.92

H.3 DEV stability and localization robustness

Table H.2: Five independent DEV replicas.

Rep.	overlap	N_{eff}/N	rank ρ	regret	agreement
1	0.82	0.126	0.95	0.0254	7/8
2	0.87	0.134	0.94	0.0249	8/8
3	0.79	0.121	0.91	0.0261	6/8
4	0.85	0.137	0.96	0.0250	7/8
5	0.81	0.129	0.93	0.0257	7/8

The 25-cell center/width robustness surface has cross-budget Spearman $\rho = 0.995$ between $B = 512$ and $B = 1024$. The deployed A1 point has $N_{eff}/N = 0.130$, $B = 512$ regret 0.0230, and $B = 1024$ regret 0.0207. The surface was evaluated after A1 was fixed from reference-only transition geometry and DEV-reference support; because the deployed point is also the best point on this coarse descriptive grid, the surface is not treated as independent confirmation. Severe center shifts or excessive narrowing degrade regret, exposing the contrast/sample-size tradeoff rather than a flat optimum.

Table H.3: Controlled localization-power summary across 576 cells.

Group	cells	local > global	local < global	mean diff.
ALL	576	345	209	+3.48 pp
delta over s 0.10	144	79	64	-1.30 pp
delta over s 0.20	144	92	52	+0.63 pp
delta over s 0.40	144	87	49	+4.72 pp
delta over s 0.80	144	87	44	+9.88 pp

H.4 Cancellation stress, automatic design, and full reference redraws

The fixed cellwise L_2 aggregation squares standardized residual means inside reference-defined calibration cells before cross-cell combination, preventing sign cancellation between cells. At F–M $B = 512$, deployed-signed/cellwise/history-absolute/tail-max regrets are 0.02510/0.02542/0.02618/0.02705. Rank correlations versus the deployed signed score are 1.000/0.962/0.931/0.887 and selected-candidate agreement is 8/8, 7/8, 6/8, and 5/8. Thus the closest theory-aligned non-cancelling variant remains within 1.3% relative regret of the deployed score and 5.6% below P2.

A reference-only automatic-design ablation replaces the Allen-specific witness/localization choice by the first four complex Fourier modes plus automatically available scalar invariants, with localization selected from reference transition-density/decision-gradient geometry. It reaches regret 0.02582, rank correlation 0.928 against the hand design, 6/8 selected-candidate agreement, and trusted-burst fraction 5.3% versus 4.7% for the deployed design.

Eight independent full DEV-REF/NULL-REF realizations give regrets 0.02510, 0.02548, 0.02496, 0.02582, 0.02536, 0.02602, 0.02519, and 0.02565. Their between-reference SD is 0.00037; rank correlations span 0.914–0.968 and selection agreement 6/8–8/8. These reference redraws are robustness units and are not counted as additional candidate-pool replications.

H.5 Bounded cluster-level concentration

Let (W_i, Y_i) be independent with $0 \le W_i \le 1$, $\mathbb{E}W_i = w_0 > 0$, $\delta_W = \mathbb{E}[W_iY_i]/w_0$, $|Y_i - \delta_W| \le M$, and $\mathrm{Var}(W_i(Y_i - \delta_W)) \le v$. If $N \ge 2w_0^{-2}\log(2/\delta)$, Bernstein plus Hoeffding and a union bound give, with probability at least $1-\delta$,

$$|\hat{\delta}_W - \delta_W| \le \frac{2}{w_0}\left[\sqrt{\frac{2v\log(4/\delta)}{N}} + \frac{2M\log(4/\delta)}{3N}\right].$$

This bound controls bounded cluster-level estimation error; downstream decision error additionally depends on witness approximation quality and reference support.

H.6 Technical assumptions and finite-case checks

The reference generator in Eqs. (2)–(3) is a smooth-cylinder identity: each basis vector belongs to $D(A_\lambda^*)$, the projected drift is integrable, and $G_\lambda(u)Q^{1/2}$ is Hilbert–Schmidt. It is *not* applied by classical differentiation to the nonsmooth positive-phase witness. The finite-step targets are conditional on the complete saved state for the reference Markov solver; the candidate can depend on the full observed history.

Proposition 1 and Corollary 1. Assume standard Borel history spaces, regular conditional reference and candidate one-step kernels, a bounded measurable terminal functional, and a version of the reference continuation recursion on Q-reachable histories. Let H_k include the decision automaton state. Conditional on H_k, $\mathbb{E}_Q[v_{k+1}(H_{k+1}) - v_k(H_k)] = [(Q_k - P_k)v_{k+1}](H_k)$. The finite sum telescopes without an independence or Markov assumption on the candidate: its endpoints are $\mathbb{E}_Q g - \mathbb{E}_{\mu_Q} v_0$. Under P, the corresponding recursion gives $\mathbb{E}_P g = \mathbb{E}_{\mu_P} v_0$. For D1, apply the identity separately to the event indicator, normalized censored time, and all threshold indicators of normalized censored time. The triangle inequality, $|\mathbb{E}Z| \le \mathbb{E}|Z|$, and Tonelli applied to nonnegative bounded

terms yield the stated $0.60/0.25/0.15$ bound. The assertion is about the continuation functions, not an upper bound on the aggregated signed DLCD statistic.

Finite-step drift and second characteristic. At fixed saved state u, let $m_\theta = \mathbb{E}_\theta[\Delta\phi \mid u] = \Delta t\, b_\theta$ and let $V_\theta = \mathrm{Cov}_\theta(\Delta\phi \mid u) = \Delta t\, q_\theta$. Writing $\Delta\phi - \Delta t\, b^* = (\Delta\phi - m_\theta) + \Delta t(b_\theta - b^*)$ and expanding the centered product gives $\Delta t^{-1}\mathbb{E}_\theta[(\Delta\phi - \Delta t b^*)(\Delta\chi - \Delta t c^*) \mid u] - q^* = q_\theta - q^* + \Delta t(b_\theta - b^*)(c_\theta - c^*)$ exactly. The cross terms vanish by conditional centering. This identity requires the reference target at the same saved state and step; empirical local regression and trusted-burst estimation add separate approximation errors.

Proposition 2. Because $0 \le W \le 1$ and Y has finite variance, the two required first and second moments exist. With $\delta_W = \mathbb{E}(WY)/\mathbb{E}W$ and $\mathbb{E}W > 0$, a joint central limit theorem for the numerator and denominator of $\sum_i W_i Y_i / \sum_i W_i$ and the delta method give the variance in Proposition 2. Both squared signal-to-noise ratios scale by N, which cancels; the criterion is an asymptotic first-order comparison, not a finite-N improvement guarantee. It applies to independently generated trajectory clusters, not to saved frames treated as independent replicates.

Cluster concentration. For $0 < \delta < 1$ and $N \ge 2w_0^{-2}\log(2/\delta)$, Hoeffding gives $\Pr(\bar{W} < w_0/2) \le \delta/2$. The centered independent summands $X_i = W_i(Y_i - \delta_W)$ satisfy $\mathbb{E}X_i = 0$, $|X_i| \le M$ and $\mathrm{Var}(X_i) \le v$. Two-sided Bernstein with $t = \log(4/\delta)$ bounds $|N^{-1}\sum_i X_i|$ by $\sqrt{2vt/N} + 2Mt/(3N)$ with probability at least $1 - \delta/2$. On their intersection, division by $\bar{W} \ge w_0/2$ yields the displayed nonasymptotic bound. All probabilistic assertions condition on the stated independent-cluster assumptions; correlated reference-quantile observations affect a separate calibration uncertainty.

The accompanying `test_theory_proofs.py` checks an exactly enumerable history-dependent process, both initial-law cases, a finite-step nonzero-drift counterexample, and the localization variance algebra. These executable tests complement, rather than replace, the proofs and do not authenticate historical simulation trajectories.

I Correction and decision-construction robustness

DLCD-CORR uses $L = L_{sup} + 0.10L_{drift} + 0.03L_{cov} + 3L_{ret}$, where the historical notation L_{cov} denotes the covariance-sensitive second-characteristic correction term at the saved-time scheme. Training uses AdamW with learning rate 3×10^{-5}, batch size 16, at most 40 epochs, and DEV-objective checkpointing. Hyperparameter selection excludes held-out D2/D3/C3 outcomes.

Table I.1: Six-checkpoint DLCD-CORR results. Values are shown at the recorded precision available for these endpoints.

Checkpoint	D2 pre	D2 post	D3 pre	D3 post	C3-D1 pre	C3-D1 post	D4 pre	D4 post
Neural SPDE s1	0.043	0.032	0.040	0.031	0.048	0.041	0.036	0.032
Neural SPDE s2	0.039	0.030	0.037	0.029	0.045	0.039	0.034	0.030
Neural SPDE s3	0.044	0.033	0.041	0.032	0.050	0.040	0.039	0.034
Structured/WCE s1	0.041	0.031	0.039	0.030	0.046	0.040	0.035	0.032
Structured/WCE s2	0.038	0.039	0.036	0.036	0.043	0.046	0.032	0.034
Structured/WCE s3	0.042	0.033	0.040	0.031	0.049	0.040	0.037	0.032

Mean relative improvements are 19.5% D2, 18.6% D3, 12.1% C3-D1, and 8.6% D4; Structured/WCE s2 is adverse on D2, C3-D1, and D4. Across D2 thresholds 0.130, 0.135, 0.140, 0.145, 0.150, mean improvement ranges from 16.9% to 19.5%.

Table I.2: D1-independent D2-alt transfer (mean improvement 9.4%). Pre-correction values were retained to three decimals in the original records; post-correction values use the available stored precision.

Checkpoint	Pre	Post	Improvement
Neural SPDE s1	0.046	0.03763	+18.2%
Neural SPDE s2	0.042	0.03507	+16.5%
Neural SPDE s3	0.047	0.04023	+14.4%
Structured/WCE s1	0.044	0.04396	+0.1%
Structured/WCE s2	0.040	0.04136	-3.4%
Structured/WCE s3	0.045	0.04023	+10.6%

J COST, DEPLOYMENT, AND SECONDARY TRANSFERS

Table J.1: Reference and candidate resource accounting. Shared reference resources are reusable across candidates; ORACLE/CONV resources are evaluation-only. Original generation time for reusable reference pools is excluded from the reported online selector wall clock.

Resource	Size	Scope	Role	Accounting
DEV-REF	4,096	shared	localization/design support	outside B; offline
NULL-REF	16,384	shared	null calibration	outside B; amortized lookup/work
DIAG-MODEL	B	candidate–condition	diagnostic paths	counted in B and online time
Trusted bursts	med. 2,300; p90 6,200; cap 8,192	candidate	fallback local targets	outside B; counted online
DEC-MODEL	8,192	candidate–condition	held-out decision evaluation	evaluation only
ORACLE-REF	32,768	condition	reference decision target	evaluation only
CONV-REF	2,048	shared high-res.	numerical refinement audit	evaluation only

The online timing comparison therefore measures marginal cost once reusable reference calibration data exist. For a new system, DEV-REF/NULL-REF solver-generation cost should be added when those pools must be created from scratch; the released artifact records their sizes but not the original generation wall clock.

Table J.2: DLCD $B = 512$ component wall-clock medians.

Component	median (s)	IQR	p90
model generation	31.8	1.9	33.7
physical operator	14.9	0.8	15.9
calibration	6.2	0.4	6.7
aggregation	5.0	0.3	5.4
triggered bursts	26.7	3.8	30.9
shared null amortized	10.1	0.6	10.9
I/O storage	4.7	0.5	5.3

Component medians are taken marginally across runs and therefore need not sum to the median end-to-end runtime.

Table J.3: Fixed-total library budget deployment (eight F–M pools, aggregate regret).

Total B	Uniform DLCD	Halving DLCD	Adaptive DIRECT-D1	Uncert.-overlap DLCD
1024	.0365	.0357	.0361	.0349
2048	.0328	.0314	.0314	.0307
4096	.0292	.0278	.0278	.0271
8192	.0257	.0249	.0242	.0245

At total budget 4096, the outcome-only adaptive allocator reaches 0.0278 versus 0.0271 for DLCD uncertainty-overlap allocation; by 8192 the direct policy is slightly lower (0.0242 versus 0.0245), again showing convergence/crossover as direct outcome information becomes sufficiently precise. Fresh stochastic Burgers gives mean regret 0.0382 versus 0.03237 (15.3%) over pools F–H, with pool H adverse. Fractional Allen–Cahn gives 15.2% reduction at $s = 0.75$ and a 1.3% near tie at $s = 0.90$.

K Statistical inference and reproducibility details

Candidate pool is the scientific replication unit. Fresh F–M $B = 512$ paired contrasts (mean; descriptive 95% t interval; favorable pools) are BASE -0.0074 $[-0.0105, -0.0043]$, 7/8; raw DIRECT-D1 -0.00530 $[-0.00759, -0.00301]$, 7/8; DIRECT-D1-SHRINK -0.00370 $[-0.00589, -0.00151]$, 6/8; and path-matched P2 -0.001825 $[-0.00288, -0.00077]$, 7/8. BASE and P2 each have exact two-sided sign sensitivity $p = 0.0703$. Frozen P2 on N–U gives -0.0012875 $[-0.00231, -0.00027]$, 6/8 favorable, with exact two-sided sign sensitivity $p = 0.2891$; the six-pool Kolmogorov P2 comparison has 5/6 favorable and $p = 0.2188$. As a secondary pooled sensitivity only, combining the 16 disjoint F–M and N–U candidate-training pools gives mean DLCD-minus-P2 contrast -0.00155625, descriptive 95% t interval $[-0.002212, -0.000900]$, 13/16 favorable pools, and exact two-sided sign $p = 0.02127$; the two cohorts remain the primary reported analyses. The development-supervised meta-ranker is near tied (-0.00030; interval spans zero; 4/4 split). Intervals are descriptive and conditional on shared held-out reference resources.

As a descriptive leave-one-pool-out sensitivity, the relative DLCD reduction remains positive under every single-pool omission: 11.3–14.5% versus DIRECT-D1-SHRINK on F–M, 5.7–7.7% versus P2 on F–M, 3.8–5.2% versus P2 on N–U, and 8.5–11.6% versus P2 on the six 2-D pools. These are robustness summaries of the same scientific units, not additional replications.

Development and confirmatory cohorts remain separate. At $B = 512$, A–E development shows a 31.5% aggregate BASE reduction with 5/5 favorable pools versus 22.8% and 7/8 on F–M. Budget, threshold, localization, lesion, fractional, reference-realization, and aggregation grids are robustness analyses rather than additional independent replications.

The anonymous data package includes scientific-unit rows for F–M BASE/DLCD, DIRECT-D1 and DIRECT-D1-SHRINK, path-matched P2, N–U BASE/P2/DLCD, the development-supervised comparator, 2-D transfer, and correction checkpoints. It also releases the non-cancelling aggregation table, automatic-design pool rows, the full 8×8 reference-realization-by-pool matrix, direct time-normalized sensitivity, adaptive-outcome allocation summaries, and development-only comparator-selection traces. Calibration/configuration files, SHA-256 manifests, deterministic checks, reference implementations, and scripts for the principal tables and Figure 2 are included.

Numerical reconstruction. Scientific-unit rows are aggregated directly where available; otherwise the supplement labels the result as aggregate-only and provides its source summary, formula, and configuration. Rounded manuscript display values are not reused as inputs to downstream calculations, and development results remain separate from held-out cohorts when their metric, comparator, or resource contract differs.

L Closely related deployment-ranking applicability boundary

The comparison with Liang and Liu (2026) is important because both works address deployment-time ranking without fresh high-fidelity candidate labels, but their mathematical objects and information

contracts differ. Table L.1 states the boundary explicitly rather than treating either method as a drop-in baseline for the other.

Table L.1: Scope and assumption comparison with a particularly close recent deployment-ranking formulation.

Dimension	Liang–Liu shared-response ranking	DLCD
Candidate object	Finite library of deterministic operator responses/predictions	Finite library of stochastic candidate laws / trajectory generators
Primary loss geometry	Exact shared-response geometry for squared Hilbert loss; exact Bregman extensions; local tangent geometry plus remainder for general smooth losses	Nonquadratic path-functional regret built from event probability, normalized RMST, and censored-time W_1
Deployment signal	Shared physical response used to recover hidden loss differences/rankings	Localized exposure, drift, and covariance-sensitive second-characteristic defects calibrated on saved-time reference dynamics
Stochastic finite-B ranking noise	Not the central object of the exact squared-loss ranking result	Central operating variable; direct target estimates are explicitly compared across candidate-path budgets
History/path dependence	Loss is evaluated through the finite-library response geometry	Decision functional may depend on complete observed history and automaton state
Guarantee type	Exact ranking/certification results under stated loss/operator assumptions	History-conditioned decomposition and concentration rationale; deployed signed score is diagnostic rather than a decision-error certificate
Trusted/reference information	Governing-equation response around a shared anchor/target geometry	Reusable DEV/NULL calibration plus trusted local dynamics/bursts, accounted separately from candidate paths

A direct adaptation would therefore require changing either the loss geometry or the stochastic deployment object. For DLCD, the central empirical fairness question is instead whether the local-characteristic audit adds value beyond direct finite-budget target estimation and beyond a matched-access coefficient control; Sections 5.1–5.3 report those comparisons.

M SAVED-OBSERVABLE EVENT RECONSTRUCTION AND SOURCE-RUN LINKAGE

A persistent threshold event comprises an upward crossing of h followed by continuous occupation above $l < h$ for a duration d. The supplementary `event_reconstruction.py` implements a precisely specified saved-observable convention. For two consecutive observations $(t_k, y_k), (t_{k+1}, y_{k+1})$ with $y_k < h \le y_{k+1}$, the crossing time is $t^* = t_k + (h - y_k)(t_{k+1} - t_k)/(y_{k+1} - y_k)$. It qualifies if piecewise-linear interpolation remains above l throughout $[t^*, t^* + d]$ and the dwell ends by T. Failed attempts do not preclude a later qualifying crossing; the event time is the first qualifying crossing, not dwell completion. A trajectory without a completed dwell is right-censored at T. This convention cannot recover threshold excursions between saved observations. Its implementation tests interpolation, repeated crossings, terminal censoring, and initial-state policies; it also reconstructs event probability, normalized RMST, and empirical censored-time W_1 from supplied event arrays.

For Allen–Cahn, the specified study horizon and parameters are $T = 6$ and $(h, l, d) = (0.65, 0.55, 0.15)$, with an initial spatial mean at most -0.65. For the independent 2-D study, the released decision definition is $(h, l, d) = (q_{90}, q_{85}, 0.05T)$ using reference-development percentiles. The processed 2-D pool records do not include the study-specific horizon, quantile-estimation population, or original interpolation convention; the public StocBench training horizon does not determine these inputs. Consequently, we provide the 2-D cohort-level reported aggregates and a complete system-specific score specification, but not a claim of raw-event-level reconstruction. The replay implementation requires the independently recorded study T and condition-specific reference quantiles rather than substituting benchmark defaults. The fully specified independent replication protocol in Appendix G fixes $T = 10,201$ monitor observations, the condition-specific DEV-REF quantile population, and event replay for newly generated runs.

Poolwise comparisons are reproducible from released tables. Source-run replication additionally needs the executed code, role-specific seeds, candidate checkpoints, DIAG/DEC files, selected-model IDs, reference observations, and continuation-anchor arrays, as specified in the supplementary evidence-scope register.

N METRIC VALIDITY AND COMPARATOR INFORMATION CONTRACT

This appendix collects the definitions and access restrictions needed to interpret the central comparisons. The numerical values are the same results reported in the main text; the additional P2+exposure and truly timed equal-clock DIRECT controls are fully specified but not included as observed results.

A concrete selection instance. For the nominal Allen–Cahn condition, an input consists of eight independently trained stochastic operators, a predeclared $B = 512$ diagnostic-trajectory budget per candidate, the fixed D1 commitment rule, and reusable reference trajectories and trusted local dynamics. The selector returns one of the eight model indices. Its diagnostic paths cannot include DEC-MODEL trajectories or the ORACLE-REF outcomes used to score the chosen model. The reference C0 commitment probability is $7137/32768 = 0.217804$; this value fixes a reference task target, not a candidate's predicted outcome. The candidate's three D1 loss components are the absolute probability error, absolute normalized-RMST error, and W_1 distance between the candidate and reference censored event-time distributions. The minimum true loss within that candidate pool defines the regret baseline.

Table N.1: Access and attribution at selection time. All rows receive B candidate diagnostic paths; reference and candidate evaluation sets remain disjoint. DEV tuning is fixed before F–M testing.

Selector	D1 on DIAG	Trusted target	Exposure group	DEV outcomes	Status
DIRECT-D1 / SHRINK	yes	no	no	shrink only	observed
BASE	no	no	no	tuning only	observed
A1-localized controls	no	no	as specified	tuning only	observed
P2 coefficient match	no	yes	no	tuning only	observed
DLCD	no	yes	yes	tuning only	observed
DEV meta-ranker	no	feature-dependent	feature-dependent	yes	observed
P2 + exposure	no	yes	yes	tuning only	proposed
True-clock DIRECT-D1	yes	no	no	allocation only	proposed

P2 and DLCD share localization and information sources, but the number of trusted-transition queries can differ. DIRECT estimates D1 on DIAG-MODEL paths; neither it nor any physics selector sees held-out DEC-MODEL/ORACLE-REF losses. "Proposed" rows have no measured outcomes in the public results.

Scoring and statistical unit. For candidate j, ℓ_j is its held-out decision loss (Eq. (9)); selection regret is $\ell_{\hat{k}} - \min_j \ell_j$. Pool means average conditions equally; cohort means average the independent candidate-training pools equally. F–M and N–U are separate independently trained Allen–Cahn cohorts, not duplicate diagnostic resamplings. The six KF pools are a separate physical system. Repeated path subsamples, reference redraws, solver steps, and candidate conditions are not additional independent trained libraries. The exact pool-level two-sided sign sensitivities are 0.0703125 for 7/8, 0.2890625 for 6/8, and 0.21875 for 5/6 favorable outcomes. The paired t and bootstrap intervals in the paper are descriptive, conditional on the shared reference resources.

Where the method does not lead. The unfavorable pool M remains in the primary Allen comparison; C3 worsens by 1.9% against BASE, while KF4 has regret 0.0420 versus P2's 0.0408. The timing-derived $B = 1216$ DIRECT loss, 0.02508, is slightly below the timing-study DLCD loss, 0.02520. These findings constrain the fixed-path claim and avoid equating consistency of the estimator with accuracy on every decision or physical regime.

Specified attribution experiments. The exposure-augmented P2 score is defined, with no held-out retuning, by $T_{P2+\mathrm{occ}} = \sqrt{(T_{\mathrm{occ}}^2 + 2T_{P2}^2)/3}$ and the same outer $E0/E1$ combination as DLCD. It receives exactly the same per-candidate diagnostic paths and trusted-target provenance. The genuinely equal-clock DIRECT test measures rather than infers how many D1 paths can be generated and evaluated in DLCD's observed end-to-end wall-clock budget, with synchronized CPU/GPU timing, warmup, repeated runs, and all calibration/burst resources charged or separately amortized. Both experiments have complete author-side numerical planning profiles and cannot be interpreted as previously measured results.

Event-protocol scope. The primary Allen event definition is fully specified in Appendix A. The released six-pool Kolmogorov score has its recorded decision quantiles, physical conditions, and pool-level results, but the original source-run time horizon and raw event histories are not contained in this processed archive. The entirely new $T = 10$ replication protocol in Appendix G is a separate prospective experiment, not a post-hoc setting for those six pools.