EVENT FIDELITY REORDERS NEURAL-OPERATOR SELECTION AT PDE TRANSITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Average field error can favor a forecast that misses the shock, phase change, or pattern onset defining a PDE transition. We introduce Critical-Point-Bench (CPB), which evaluates event fidelity before ranking models by field error within the transition band. CPB abstains when no candidate meets the declared event criteria. A conditional-risk bound explains why global error loses control when transition events are rare. On the final Allen–Cahn holdout, CPB selects the Fourier Neural Operator (FNO) over LinearNO and reduces critical normalized error from 0.00172 to 0.000414, a 75.9% decrease, while preserving phase recovery. On trajectory-held-out PDEBench, the model with the lowest aggregate error recovers only 21.0% of post-transition events, consistent with CPB's validation-stage abstention. Event fidelity therefore changes which model wins at a transition and reveals when no candidate has learned the defining physics.

1 INTRODUCTION

Aggregate field error cannot say where an operator fails or whether it realizes the event of interest. On one-dimensional viscous Burgers, for example, an FNO reaches aggregate normalized root mean square error (nRMSE) 0.002, yet its error near shock formation is $1.87\times$ its smooth-side error. The same scalar score says nothing about whether the predicted front appears in the right place.

Transitions demand two kinds of fidelity. The predicted field should remain accurate near the change, and the model should reproduce the event that defines the change. These requirements differ across front formation, phase selection, and pattern onset, but their order is common. A model that misses the event is not rescued by a small average error elsewhere.

Critical-Point-Bench (CPB) makes this order explicit. Each task defines a signed coordinate around its transition, a local field-error measure, and binary and continuous event criteria. Validation candidates must first reproduce the event. CPB then ranks the eligible models by transition-band error and abstains if the eligible set is empty. Discovery data shape the task definition, validation data choose the model, and a final holdout measures the resulting decision. Reference solutions are needed for validation, but the selected operator incurs no extra inference-time computation.

The experiments span three transition mechanisms and include matched interventions that improve one region while testing their cost elsewhere. On Allen–Cahn, event-aware ranking changes the selected architecture and reduces critical-region error. On PDEBench trajectories, the aggregate winner recovers only 0.210 of post-transition events, making abstention the scientifically meaningful result. CPB contributes a selection principle that gives event capability priority over local precision, while preserving aggregate accuracy as a complementary view of the full distribution.

2 RELATED WORK

PDE operator benchmarks. PDEBench (Takamoto et al., 2022), APEBench (Koehler et al., 2024), The Well (Ohana et al., 2024), PDEArena (Gupta & Brandstetter, 2023), and CFDBench/CFDONEval (Luo et al., 2023; Liu et al., 2025) provide broad coverage of equations, resolutions, geometries, and rollout horizons. Several contain shock-bearing or pattern-forming data. Their standard summaries

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

are nevertheless aggregated by task or test distribution. CPB is designed to be used alongside these suites: it conditions the same prediction error on a task-declared transition coordinate and checks the event associated with that coordinate.

Conditional evaluation, robustness, and abstention. Statistically, $E_s(B)$ is a conditional risk evaluated within physically meaningful slices. Robustness studies vary parameter shift, boundary conditions, resolution, rollout horizon, or input perturbations and show that in-distribution accuracy need not predict robustness (Shikhman, 2026). Error-localization and spectral-bias analyses further explain why high-frequency or spatially concentrated structures can be vulnerable (Rahaman et al., 2019). CPB turns this conditional view into a model-selection rule: event capability defines the candidate set, local error ranks eligible models, and an empty set prompts abstention. Here abstention is a statement about the candidate family rather than an uncertainty score for one example. Data-slicing and group-robust methods instead discover problematic validation subsets or optimize worst-group loss (Chung et al., 2020; Sagawa et al., 2020). Constrained optimization separates feasibility from an objective (Gelbart et al., 2014), while selective classification rejects individual examples to trade coverage for risk (Geifman & El-Yaniv, 2017); CPB instead applies one task-defined eligibility-and-ranking rule to a finite model set.

Methods for transition-bearing dynamics. RiemannONets and Cut-DeepONet target discontinuities and sharp fronts (Peyvan et al., 2024; Dang et al., 2026); phase-field operators target microstructure evolution (Oommen et al., 2022). Local neural-operator time-steppers have also been embedded in continuation and Krylov methods for bifurcation analysis (Fabiani et al., 2026), while learned ODEs and an application-specific operator transformer study bifurcation generalization or forecasting (Huh et al., 2025; van de Wetering & Zhu, 2026). Early-warning work instead asks whether a system is approaching a tipping point (Scheffer et al., 2009; Bury et al., 2021). These are method, system- analysis, or detection studies. CPB evaluates standard learned solution operators conditional on signed transition proximity and paired event recovery.

Operator architectures. The evaluated FNO, DeepONet, convolutional, and attention models represent distinct global-basis, spectral, local-convolutional, and token-mixing designs (Li et al., 2021; Lu et al., 2021; Kovachki et al., 2023; Brandstetter et al., 2022; Pfaff et al., 2021). We additionally include the recent LinearNO, which replaces quadratic point attention by independently normalized point-to-slice and slice-to-point maps (Hu et al., 2026). LinearNO therefore provides a current operator baseline; CPB supplies the shared measurement of how established and newer design choices behave near a qualitative change.

3 Transition-Aware Evaluation

3.1 Decision object

For input x, let $\mathcal{G}_s(x) = y$ and $\widehat{\mathcal{G}}_s(x) = \widehat{y}$ denote the reference solution and the learned prediction for system s. CPB requires a task definition $\Theta_s = (\delta_s, h_s, q_s, \ell_s^{\mathrm{evt}}, \mathcal{A}_s, r_s^\star, c_s^\star)$, fixed before final-holdout access. The scalar $\delta_s(x)$ is a transition coordinate computed from the input or declared task metadata; it is not assumed to be a Euclidean distance. Its zero defines the operational boundary $\mathcal{C}_s^\delta = \{x : \delta_s(x) = 0\}$, its sign labels the two declared sides, and h_s fixes the transition band. The paired event functional $q_s(\widehat{y}, y) \in \{0, 1\}$ gives a binary agreement indicator, while ℓ_s^{evt} and the task-specific aggregation $\mathcal{A}_s$ define a continuous event discrepancy. The thresholds $r_s^\star$ and $c_s^\star$ set the minimum recovery and maximum discrepancy required for eligibility. For a coordinate bin B, CPB reports

$$
\begin{aligned}
E_s(B) &= \mathbb{E}\left[\frac{\|\widehat{\mathcal{G}}_s(x) - \mathcal{G}_s(x)\|_2 / \sqrt{N_s}}{\sigma_s} \,\middle|\, \delta_s(x) \in B \right], \\
R_s(B) &= \Pr\Big[q_s(\widehat{\mathcal{G}}_s(x), \mathcal{G}_s(x)) = 1 \,\Big|\, \delta_s(x) \in B \Big], \\
C_s(B) &= \mathcal{A}_s\Big(\Big\{ \ell_s^{\mathrm{evt}}(\widehat{\mathcal{G}}_s(x), \mathcal{G}_s(x)) : \delta_s(x) \in B \Big\} \Big).
\end{aligned}
\tag{1}
$$

Here N_s is the number of spatial values and σ_s is the root mean square of the training targets. The fixed normalization remains finite when an individual target approaches zero. The current benchmark

sets $h_s = 0.15d_s$, with d_s a task scale chosen before testing, and retains the full signed profiles because one band ratio cannot distinguish a step, peak, and one-sided ramp. Mathematically, $E_s(B)$ is a conditional risk. In the complete model card $(E_-, E_0, E_+, E_{\mathrm{agg}}, R_+, C_+)$, the first three errors refer to the smooth, critical, and post-transition regions, and E_{agg} is the full-distribution error. CPB uses this card under Θ_s rather than conditional risk alone.

3.2 Why aggregate error cannot certify an event

The need for an explicit event check follows from a simple conditional-risk argument. Let A be the post-transition region with probability $\pi = \Pr(x \in A) > 0$, and let $d(\widehat{y}, y)$ denote normalized field error. Suppose the event admits a statistic z with margin $\gamma > 0$: disagreement $q_s(\widehat{y}, y) = 0$ requires $\|z(\widehat{y}) - z(y)\| \geq \gamma$. For some $L > 0$, assume z is L-Lipschitz with respect to d.

Proposition 1 (Aggregate-to-event bound)**.** *Under these conditions, the post-transition event error obeys*

$$1 - R_s(A) \leq \frac{L}{\gamma}\,\mathbb{E}[d(\widehat{y}, y) \mid x \in A] \leq \frac{L}{\pi\gamma}\,\mathbb{E}[d(\widehat{y}, y)]. \tag{2}$$

Proof. An event mismatch requires $\|z(\widehat{y}) - z(y)\| \geq \gamma$, and hence $d(\widehat{y}, y) \geq \gamma/L$. Markov's inequality gives the first bound; the risk decomposition over A and its complement gives the second. □

Thus a global score certifies event recovery only when it is small relative to both the frequency and margin of the event. Near a rare or weakly expressed transition, the guarantee becomes uninformative. The same decomposition allows two models with nearly identical aggregate error to exchange accuracy inside and outside the critical band. CPB therefore measures event recovery directly and ranks capable models by the conditional error that aggregate risk dilutes.

3.3 Systems and comparison matrix

Table 1 distinguishes three coordinate classes: an inviscid compression approximation for viscous Burgers, a symmetry-informed mean-bias coordinate for Allen–Cahn, and the analytic linear-instability coordinate for Swift–Hohenberg (Hopf, 1950; Cole, 1951; Allen & Cahn, 1979; Swift & Hohenberg, 1977). The initial 1D comparison uses 256 grid points and 800/200/600 training/validation/test samples; two descriptive extensions use 64×64 grids and 600/150/400 samples. Governing equations, data construction, event detectors, and solvers appear in the Appendix.

We compare a Fourier Neural Operator (FNO) (Li et al., 2021), DeepONet (Lu et al., 2021), a residual convolutional neural operator (CNO), a patch-attention Transformer, and LinearNO (Hu et al., 2026). A factorial panel of 125 trained models supplies the initial comparison; a separate study with identical sample counts and training schedules uses 800/400/1,200 training/validation/final-holdout examples per 1D system, makes all choices on validation data, and evaluates ten new seeds per reachable configuration. Across the initial comparison, selection, intervention, matched-control, and external experiments, we train 765 models; the Appendix gives the complete accounting. Eight timing-only cells, with three short runs each, define the compute-matched comparisons.

3.4 Validation-to-test selection

Both selectors read the same validation model cards and operate hierarchically. Within each architecture, the aggregate path minimizes E_{agg}; CPB retains candidates satisfying $R_+ \geq r_s^\star$ and $C_+ \leq c_s^\star$, then minimizes E_0. The resulting architecture-level cards are compared once more by the same respective path. Event thresholds, the 2% practical-equivalence tolerance at both stages, and tie orders are fixed before holdout access. These task tolerances encode the fidelity required by each application. An empty CPB set records that the tested configurations have not yet reached that fidelity. Algorithm 1 specifies the decision rule.

Table 1: Task definitions used by CPB. Coordinate status distinguishes an analytic onset from a physics-informed approximation or task coordinate. Event requirements determine which models are eligible for local-error ranking.

System	signed transition coordinate (status)	binary event q_s	continuous post-event discrepancy C_+	event requirements
Burgers	$(-\min_x \partial_x u_0 - 1/T)/(2\pi)$ (inviscid approximation; viscous finite-time task)	matching front presence and periodic location	post-event 95th percentile circular position error	$R_+ \geq 0.95$, $C_+ \leq 0.015625$
Allen–Cahn	mean bias m (symmetry-informed task coordinate)	matching phase sign and interface count within one	mean per-sample maximum of normalized mean error and half the interface-count error	$R_+ \geq 0.98$, $C_+ \leq 0.10$
Swift–Hohenberg	$r - 0$ (analytic linear-instability coordinate)	matching pattern presence and, when present, dominant wavenumber	mean per-sample maximum of stabilized amplitude and normalized-wavenumber errors	$R_+ \geq 0.95$, $C_+ \leq 0.20$

Table 2: Independent validation-to-holdout selection. Lower is better for E_{agg}, E_0, and C_+; higher is better for R_+. Internal results average ten training seeds and a shared 1,200-example holdout per seed; PDEBench uses eight seeds and trajectory-cluster inference. CPB abstains when no validation candidate meets both declared event requirements.

System	Selector	Choice	E_{agg}	E_0	R_+	C_+
Burgers	Both	FNO	0.0022	0.0016	0.998	0.000
Allen–Cahn	Aggregate	LinearNO	0.0011	0.0017	1.000	0.000
	CPB	FNO	0.0017	0.0004	0.999	0.002
Swift–Hohenberg	Both	FNO	0.0716	0.0055	0.980	0.031
PDEBench Burgers	Aggregate	FNO	0.0801	0.0511	0.210	0.500
	CPB	ABSTAIN	–	–	–	–

Algorithm 1 Hierarchical CPB validation-to-test selection

Require: Validation cards $\mathcal{V} = \bigcup_a \mathcal{V}_a$ grouped by architecture; event thresholds $(r_s^\star, c_s^\star)$; practical-equivalence tolerance $\eta = 0.02$

for all architectures a **do**
  $\mathcal{A}_a \leftarrow \{m \in \mathcal{V}_a : R_+(m) \geq r_s^\star \ \wedge \ C_+(m) \leq c_s^\star\}$
  if $\mathcal{A}_a \neq \varnothing$ **then**
    $\widetilde{m}_a \leftarrow \mathrm{Choose}_\eta(\mathcal{A}_a)$
  end if
end for
$\mathcal{M} \leftarrow \{\widetilde{m}_a : \mathcal{A}_a \neq \varnothing\}$
if $\mathcal{M} = \varnothing$ **then**
  return NO ELIGIBLE CANDIDATE: abstain for this task
end if
$m^\star \leftarrow \mathrm{Choose}_\eta(\mathcal{M})$
  Choose_η minimizes E_0; resolve η-ties by E_+, E_-, E_{agg}, parameters, training cost, then stable ID
Fix $m^\star$, its saved weights, the task definition, and this rule
Evaluate once on the final holdout; report the complete model card

Each tuning fit receives 3,900 updates and 300 validation evaluations. Every model and control is chosen without access to the 1,200-example holdout. Confirmatory uncertainty re-executes the cross-architecture stage over the architecture-level candidates selected by the initial tuning, while resampling paired final-training seeds and stratified validation/holdout examples. It then forms simultaneous 95% max-statistic intervals over the five estimable endpoints; the external endpoint yields abstention and therefore no numerical contrast. Intervention results are also compared with field and event behavior away from the transition.

4 RESULTS

4.1 INDEPENDENT SELECTION CHANGES THE ALLEN–CAHN DECISION

The selectors agree on Burgers and Swift–Hohenberg but differ on Allen–Cahn (Table 2): aggregate error selects LinearNO; CPB selects FNO from the same event-eligible set. On hold-out, critical nRMSE falls from 0.00172 to 0.000414 (contrast -0.001311, simultaneous 95% CI $[-0.001607, -0.001015]$): 76% relatively, but 0.131% of the training-target RMS absolutely. Phase recovery is noninferior (contrast -0.00098, CI $[-0.00376, 0.00180]$). FNO instead has 52.6% higher aggregate nRMSE and higher C_+ (0.002029 versus 0.000343); both meet the event requirements. CPB therefore makes a stable local-precision trade-off: both choices recover the phase, while FNO resolves the transition band more closely. Bootstrap re-selection chooses LinearNO by aggregate error in 19,986/20,000 draws and FNO by CPB in all 20,000 draws.

The external endpoint tests abstention rather than numerical superiority. On held-out PDEBench trajectories, aggregate validation selects FNO with aggregate nRMSE 0.0801, but its post-transition recovery is 0.210 and continuous discrepancy is 0.500. In a post-hoc eight-timepoint trajectory analysis, FNO captures 0.887 of first front-detector crossings but only 0.107 of crossings sustained through later queried times. None of the 16 validation candidates satisfies the event requirements chosen on validation, so CPB returns NO ELIGIBLE CANDIDATE before final-test access. The result calls for a stronger candidate set; distinguishing architecture, optimization, event definition, and data as its cause is a separate modeling question.

4.2 THE CROSS-PDE MATRIX REVEALS THREE TRANSITION SIGNATURES

Fourteen of 15 one-dimensional system–model cells have greater critical- than smooth-side mean error; Burgers CNO is flat ($1.0001\times$). Neighboring band widths retain 13–14 elevated cells, all three profile shapes, and the Allen–Cahn choice. That discovery panel already showed the same trade-off: aggregate nRMSE favored LinearNO, whereas FNO had 60% lower critical error with complete phase recovery (Appendix).

4.3 LOCAL PROFILES DISTINGUISH ACCURACY FROM EVENT CAPABILITY

The three signatures in Figure 1 imply different decisions. Burgers mean nRMSE rises $0.033 \to 0.085 \to 0.168$ across regimes. FNO preserves the front despite a $1.87\times$ critical-to-smooth ratio; DeepONet instead places a sharp front nearly half a period away (position error 0.488), a conclusion stable across the detector grid.

Allen–Cahn forms a peak: CNO, LinearNO, and DeepONet have ratios 1.34, 2.57, and 2.56, while phase recovery stays 0.989–1.000, making local field error discriminative. Swift–Hohenberg forms a ramp, with mean nRMSE $0.001 \to 0.049 \to 1.110$. FNO and CNO recover post-onset patterns at 0.985 and 0.493; the other families have negligible amplitude. Here the event criterion is decisive because a near-onset field can remain small.

Table 3: The 125-fit original-panel cross-PDE comparison. Each entry reports $E_{\mathrm{agg}}/E_0/R_+$: aggregate nRMSE, critical-band nRMSE, and post-transition event recovery, averaged over five independently trained models. The joint view separates local field accuracy from recovery of the changing event; complete regime tuples and fit-level uncertainty appear in the Appendix.

Dimension	system	FNO	DeepONet	CNO	Transformer	LinearNO
1D	Burgers	.002/.002/1.000	.493/.383/.000	.017/.008/.993	.018/.018/.983	.013/.013/.996
1D	Allen–Cahn	.002/.0009/.996	.113/.266/.992	.002/.002/.998	.002/.002/.999	.001/.002/1.000
1D	Swift–Hoh.	.060/.006/.985	.611/.060/.000	.510/.055/.493	.614/.062/.000	.612/.060/.000
2D	Burgers	.007/.004/.995	.461/.305/.010	.017/.014/.995	.078/.061/.988	.016/.014/.997
2D	Swift–Hoh.	.362/.014/1.000	.614/.035/.000	.459/.025/1.000	.615/.036/.000	.614/.035/.000

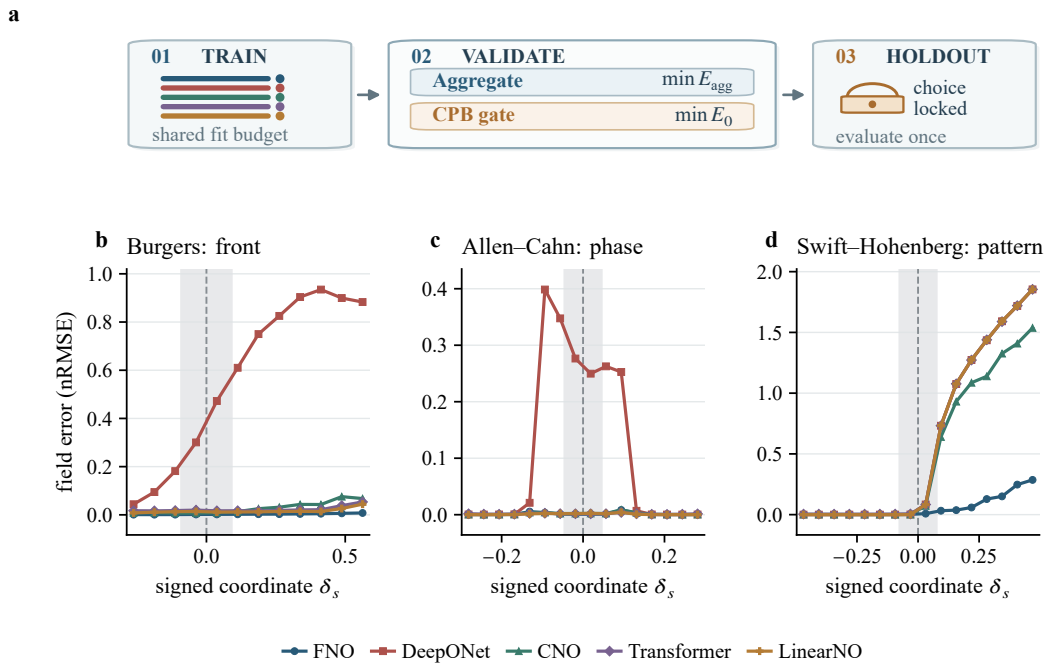

Figure 1: Transition-aware evaluation reveals distinct local-error profiles. (a) Equal-budget candidates follow aggregate or event-gated CPB validation rules within and then across model families; choices are locked before one final holdout evaluation. (b–d) Signed-coordinate profiles for Burgers, Allen–Cahn, and Swift–Hohenberg. Lines show fit means, bands show standard errors, and shading marks critical regions.

4.4 Transition profiles extend across dimension and distribution

The two 64×64 extensions show the same descriptive separation (Figure 2). FNO and CNO recover both events at rates at least 0.995; LinearNO and the Transformer retain the front but not the pattern. DeepONet's front label is threshold-sensitive (0.010–0.971), but its sharpness ratio 0.383 consistently indicates smoothing (Appendix).

The same concentration appears outside the controlled generators. On PDEBench 1D Burgers (Takamoto et al., 2022), an FNO trained on 3,000 trajectories has aggregate nRMSE 0.13, while the sharpest target quartile has $3.58\times$ the error of the smoothest quartile (0.215 versus 0.060). This split lacks a per-sample breaking coordinate, so it establishes distributional transfer of the sharpness-conditioned error gap rather than a second transition-coordinate experiment.

5 Testing Remedies for Transition Failures

5.1 Spectral capacity improves an event-capable operator

The local profiles motivate a spectral-capacity test for the front and pattern tasks. For FNO, per-sample nRMSE correlates with target energy above its 16-mode cutoff in Burgers (0.426) and Swift–Hohenberg (0.679). Across the transition, the corresponding high-mode fractions increase from 3.5×10^{-11} to 7.2×10^{-3} and from approximately zero to 3.5×10^{-6}. Allen–Cahn provides the contrasting mechanism: the correlation is -0.02, high-mode energy stays near 10^{-10}, and the architecture-dependent peak is instead associated with sensitivity to the mean bias.

The retained-mode intervention confirms the spectral assignment. Across three independent fits at 4, 8, and 16 modes, Burgers critical nRMSE falls by 71% ($0.00688 \to 0.00198$), while Swift–Hohenberg falls by 90% ($0.05960 \to 0.00601$) and critical event recovery rises from 0.875 to 0.981. Allen–Cahn remains below 0.002 with complete event recovery. Additional 32- and 48-mode single-fit

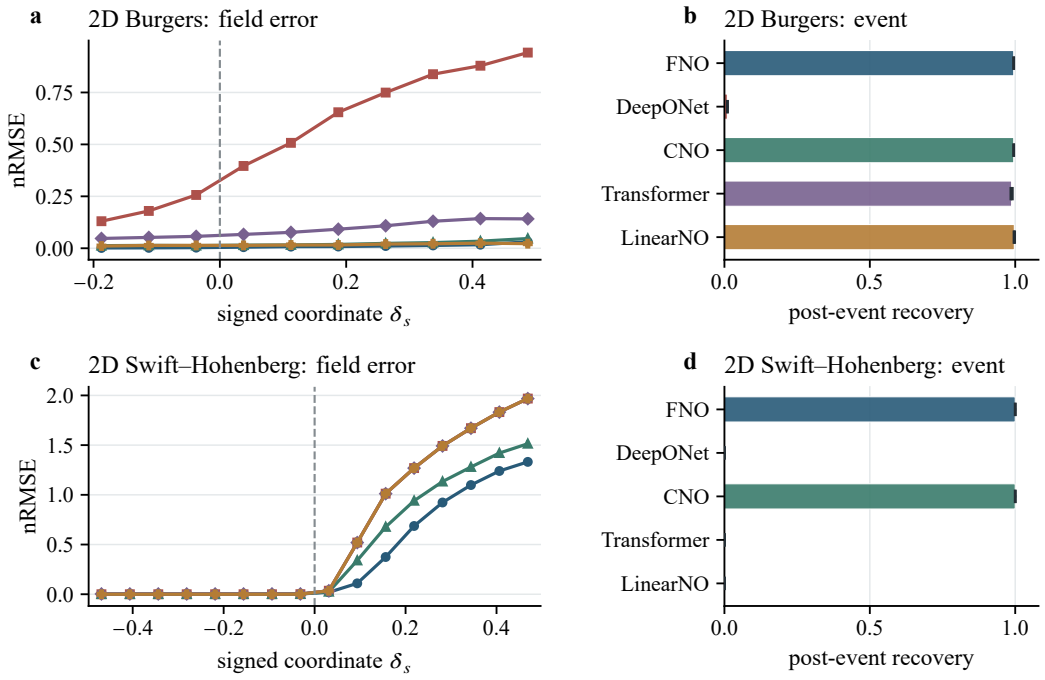

Figure 2: Two-dimensional extensions retain the front–pattern separation. (a,c) nRMSE versus signed coordinate. (b,d) Post-transition event recovery. Curves and horizontal bars show means over five fits with standard errors.

extensions do not show a consistent further gain. The contrast identifies when spectral capacity is a productive intervention rather than a generic scaling prescription (Appendix). For the validation-selected Swift–Hohenberg FNO, post-onset local-relative error increases with signed coordinate at slope 0.437 (simultaneous 95% CI $[0.288, 0.585]$). The ramp therefore remains after normalizing by target amplitude and reflects approximation quality rather than vanishing onset amplitude alone.

5.2 Matched controls preserve the critical-error ordering

The independent Allen–Cahn critical-error ordering survives both plausible scaling explanations (Figure 3). We compare FNO and LinearNO at four parameter-matched levels and four update-matched levels, with five paired seeds per cell. FNO has lower critical error in all eight comparisons. Every simultaneous 95% interval for LinearNO-minus-FNO lies above zero: parameter- matched lower bounds range from 3.38×10^{-5} to 1.42×10^{-3}, and compute-matched lower bounds from 6.63×10^{-4} to 1.60×10^{-3}. The matched comparisons isolate a persistent local advantage for FNO from parameter count and optimization access. Aggregate and downstream performance remain separate questions.

5.3 Transition-focused interventions improve local fidelity

Eight-seed interventions show that transition emphasis changes local accuracy. Relative to uniform FNO training, Burgers resampling changes critical nRMSE by -1.96×10^{-4} (simultaneous 95% CI $[-2.71, -1.20] \times 10^{-4}$), and $5\times$ weighting changes it by -2.45×10^{-4} (CI $[-3.60, -1.29] \times 10^{-4}$). Binary and continuous event structure are preserved in both cases. Increasing Swift–Hohenberg FNO capacity from 4 to 16 modes yields a critical-error contrast of -0.0678 (CI $[-0.0959, -0.0396]$), increases event recovery by 0.977, and reduces continuous discrepancy by 4.91. These paired contrasts measure the effect of each intervention in the tested training setup. They do not isolate which feature of the learned representation produces the gain.

The broader picture includes behavior away from the transition. Burgers emphasis transfers some error to smooth and aggregate regimes; the Swift smooth-side relative ratio is unstable because its

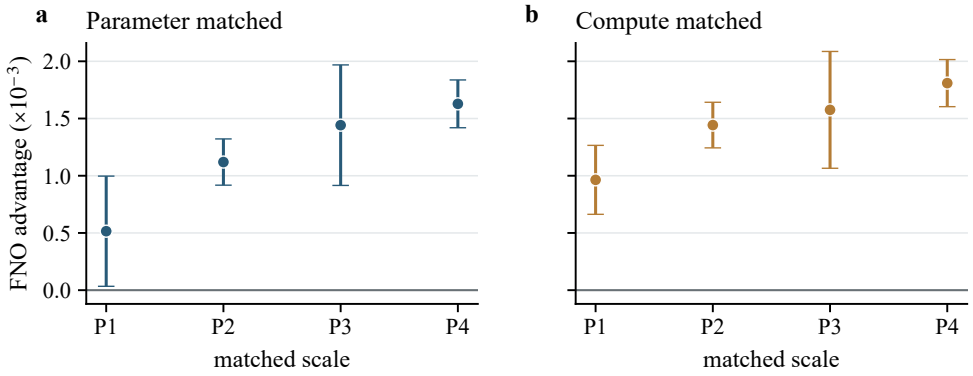

Figure 3: **FNO lowers critical nRMSE in all eight matched Allen–Cahn comparisons.** Points show LinearNO-minus-FNO error for parameter- and compute-matched levels. Bars are simultaneous 95% intervals over five paired seeds per cell.

baseline is nearly zero. The interventions therefore establish significant local gains, while their value for a full deployment also depends on global behavior.

5.4 Continuous structure rejects an apparent binary repair

In an additional exploratory control, a periodic-gradient loss chosen on validation raises DeepONet critical recovery from 0.001 to 0.784. The accompanying continuous measurements change the decision: critical nRMSE increases by 4.8%, predicted/reference sharpness falls from 1.20 to 0.19, circular position error remains 0.488, and post-transition recovery is 0.023. The loss smooths the misplaced front rather than locating it. Binary recovery is therefore accepted as an improvement only when transition-local field error and continuous event structure agree.

6 Sensitivity and Transfer Checks

The decision is stable to nearby definitions. Across critical-band widths $0.10d_s$, $0.15d_s$, and $0.20d_s$, 13–14 of 15 one-dimensional cells retain critical-to-smooth ratios above one, all three profile signatures remain, and aggregate-to-CPB selection still changes from LinearNO to FNO on Allen–Cahn. Across the full scale–offset–noise perturbation cube, selector stability is 1.000 for Burgers, 0.891 for Allen–Cahn, and 0.725 for Swift–Hohenberg; all three are stable at the chosen coordinate. These reruns measure sensitivity around that coordinate; whether it uniquely captures the intended physical construct remains a scientific question. In a post-hoc reference-only alignment analysis, detector-defined event boundaries remain inside every band but need not lie at zero (Swift–Hohenberg: 0.0513 versus band half-width 0.075). Detector grids and continuous outcomes preserve every decisive structural comparison. Reference refinement changes fields by at most 1.01×10^{-3} nRMSE and leaves all 25 event labels unchanged.

Two PDEBench checks extend the finding. A sharpness split exposes a $3.58\times$ extreme-quartile error gap. Independently tuned, trajectory-held-out selection abstains because every candidate fails an event criterion. Coordinate, detector, reference, and external checks appear in the Appendix.

7 Discussion

CPB changes model selection because event fidelity and field fidelity carry different information. Aggregate and local error rank the Allen–Cahn models in opposite orders; on PDEBench, the aggregate winner reproduces only 0.210 of post-transition events. Continuous measures expose failures hidden by a binary indicator, and the Burgers interventions show that a local improvement can raise aggregate error by 13.3% or 27.6%. These distinctions disappear in a whole-distribution average.

Table 4: Event geometry and persistence constrain the interpretation of local error. These measurements do not change the selected models. Intervals use bootstrap resampling over independent training fits; full definitions appear in the Appendix.

Analysis	result	supported interpretation
Allen–Cahn interfaces	LinearNO: 0.0496 cells $[0.0456, 0.0539]$; FNO: 0.0768 $[0.0728, 0.0805]$	both are sub-cell; lower FNO E_0 is not an interface-geometry gain
Swift–Hohenberg onset	reference boundary 0.05130 inside the ± 0.075 band; no detectable FNO shift; CNO delay 0.02589 $[0.02191, 0.02971]$	the linear zero and finite-time detector onset are distinct
PDEBench persistence	FNO sustained-onset capture 0.107 $[0.078, 0.142]$; prediction reversal 0.828	time-conditioned event inconsistency supports abstention, not a unique causal diagnosis

The transition coordinate must come from the physics of the task. Burgers uses an inviscid compression estimate, Allen–Cahn uses mean bias under a specified perturbation, and Swift–Hohenberg uses a linear onset threshold. Sensitivity analysis shows how each choice affects the decision, while the Allen–Cahn comparison demonstrates the value of the resulting local view: because both finalists recover the phase, the 0.001311 critical-error advantage can resolve the selection without becoming a universal ranking of the architectures.

Abstention as information. The hierarchical bootstrap repeats the validation-stage choice and carries its uncertainty into the holdout comparison. If no candidate meets the event requirements, abstention reveals a gap in the candidate set without pretending to distinguish among architecture, optimization, event definition, and data.

From evaluation to modeling. High event recovery with elevated local error motivates capacity or sampling near the transition; weak event recovery motivates a broader model family; and binary–continuous disagreement motivates a closer look at the detector. The present evidence covers periodic grids, three transition mechanisms, two descriptive 2D extensions, and PDEBench trajectories. Irregular, graph-based, interacting, and oscillatory systems are natural next tests.

8 Conclusion

Neural operators near a PDE transition should be selected for the event they must reproduce, then for the accuracy with which they resolve its neighborhood. CPB implements this ordering with a task-defined transition coordinate, event criteria, and transition-band error. On Allen–Cahn it selects the more accurate event-capable model; on PDEBench it asks for a better candidate family when none captures the post-transition dynamics. Event-aware selection turns that contrast into a useful scientific result rather than a single aggregate leaderboard.

The hierarchy gives each outcome a distinct scientific meaning. Failure of the event test points to a missing behavior in the candidate family. Event recovery with large transition-band error points instead to insufficient resolution of the local dynamics. Agreement of the event and local criteria supports selection, while disagreement between binary and continuous observables calls for revisiting the event definition. CPB therefore converts a model ranking into a set of testable hypotheses about representation, local accuracy, and the transition observable itself.

AI use statement

OpenAI Codex was used for coding assistance and language editing. The authors reviewed all AI-assisted content and take full responsibility for the manuscript, implementation, and reported results.

Reproducibility statement

The anonymous release specifies all transition coordinates, solvers, splits, models, statistical comparisons, machine-readable outputs, and figure/table regeneration settings.

REFERENCES

Samuel M. Allen and John W. Cahn. A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. *Acta Metallurgica*, 27(6):1085–1095, 1979. doi: 10.1016/0001-6160(79)90196-2.

Johannes Brandstetter, Daniel Worrall, and Max Welling. Message passing neural PDE solvers. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.03376.

Thomas M. Bury, R. I. Sujith, Induja Pavithran, Marten Scheffer, Timothy M. Lenton, Madhur Anand, and Chris T. Bauch. Deep learning for early warning signals of tipping points. *Proceedings of the National Academy of Sciences*, 118(39):e2106140118, 2021. doi: 10.1073/pnas.2106140118.

Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. Automated data slicing for model validation: A big data–AI integration approach. *IEEE Transactions on Knowledge and Data Engineering*, 32(12):2284–2296, 2020. doi: 10.1109/TKDE.2019.2916074. URL https://doi.org/10.1109/TKDE.2019.2916074.

Julian D. Cole. On a quasi-linear parabolic equation occurring in aerodynamics. *Quarterly of Applied Mathematics*, 9(3):225–236, 1951. doi: 10.1090/qam/42889.

Ha Dang, Sebastian Schmidt, and Juergen Hesser. Smooth piecewise cutting for neural operator to handle discontinuities and sharp transitions. *arXiv preprint arXiv:2605.19823*, 2026. URL https://arxiv.org/abs/2605.19823.

Gianluca Fabiani, Hannes Vandecasteele, Somdatta Goswami, Constantinos Siettos, and Ioannis G. Kevrekidis. Enabling local neural operators to perform equation-free system-level analysis. *Nature Machine Intelligence*, 8:1127–1141, 2026. doi: 10.1038/s42256-026-01265-1. URL https://www.nature.com/articles/s42256-026-01265-1.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Michael A. Gelbart, Jasper Snoek, and Ryan P. Adams. Bayesian optimization with unknown constraints. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pp. 250–259. AUAI Press, 2014. URL https://auai.org/uai2014/proceedings/individuals/107.pdf.

Jayesh K. Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized PDE modeling. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=dPSTDbGtBY.

Eberhard Hopf. The partial differential equation $u_t + uu_x = \mu u_{xx}$. *Communications on Pure and Applied Mathematics*, 3(3):201–230, 1950. doi: 10.1002/cpa.3160030302.

Wenjie Hu, Sidun Liu, Peng Qiao, Zhenglun Sun, and Yong Dou. Transolver is a linear transformer: Revisiting physics-attention through the lens of linear attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 408–416, 2026. doi: 10.1609/aaai.v40i1.37003. URL https://ojs.aaai.org/index.php/AAAI/article/view/37003.

In Huh, Changwook Jeong, and Muhammad Alam. Context-informed neural ODEs unexpectedly identify broken symmetries: Insights from the poincaré–hopf theorem. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 26311–26337, 2025. URL https://proceedings.mlr.press/v267/huh25a.html.

Aly-Khan Kassam and Lloyd N. Trefethen. Fourth-order time-stepping for stiff PDEs. *SIAM Journal on Scientific Computing*, 26(4):1214–1233, 2005. doi: 10.1137/S1064827502410633.

Felix Koehler, Simon Niedermayr, Rüdiger Westermann, and Nils Thuerey. APEBench: A benchmark for autoregressive neural emulators of PDEs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3822. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d9875ebcf74bccdc5076acab0dbee62c-Abstract-Datasets_and_Benchmarks_Track.html.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023. URL https://jmlr.org/papers/v24/21-1524.html.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Menghan Liu, Jianhuan Cen, Ziyang Zhou, Haolong Fan, Hongji Li, Ping Wei, Guohang Peng, Changye He, Yuzhe Qin, Yutong Lu, and Qingsong Zou. CFDONEval: A comprehensive evaluation of operator-learning neural network models for computational fluid dynamics. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 5752–5760, 2025. doi: 10.24963/ijcai.2025/640. URL https://www.ijcai.org/proceedings/2025/640.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3:218–229, 2021. doi: 10.1038/s42256-021-00302-5.

Yining Luo, Yingfa Chen, and Zhen Zhang. CFDBench: A large-scale benchmark for machine learning methods in fluid dynamics. *arXiv preprint arXiv:2310.05963*, 2023. URL https://arxiv.org/abs/2310.05963.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond B. Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich R. Kerswell, Suryanarayana Maddu, Jonah Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam H. Parker, Miles Cranmer, and Shirley Ho. The well: a large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1430. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4f9a5acd91ac76569f2fe291b1f4772b-Abstract-Datasets_and_Benchmarks_Track.html.

Vivek Oommen, Khemraj Shukla, Somdatta Goswami, Rémi Dingreville, and George Em Karniadakis. Learning two-phase microstructure evolution using neural operators and autoencoder architectures. *npj Computational Materials*, 8:190, 2022. doi: 10.1038/s41524-022-00876-7.

Ahmad Peyvan, Vivek Oommen, Ameya D. Jagtap, and George Em Karniadakis. RiemannONets: Interpretable neural operators for riemann problems. *Computer Methods in Applied Mechanics and Engineering*, 426:116996, 2024. doi: 10.1016/j.cma.2024.116996.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W. Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.03409.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 2019. URL https://proceedings.mlr.press/v97/rahaman19a.html.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization.

In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Marten Scheffer, Jordi Bascompte, William A. Brock, Victor Brovkin, Stephen R. Carpenter, Vasilis Dakos, Hermann Held, Egbert H. van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260):53–59, 2009. doi: 10.1038/nature08227.

Lennon Shikhman. Diagnosing failure modes of neural operators across diverse PDE families. *Transactions on Machine Learning Research*, 2026. URL https://arxiv.org/abs/2601.11428.

Jack Swift and Pierre C. Hohenberg. Hydrodynamic fluctuations at the convective instability. *Physical Review A*, 15(1):319–328, 1977. doi: 10.1103/PhysRevA.15.319.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2210.07182.

Johannes J. van de Wetering and Ben Zhu. Neural operator transformers capture bifurcating drift wave turbulence in fusion plasma simulations. *arXiv preprint arXiv:2603.05730*, 2026. URL https://arxiv.org/abs/2603.05730.

A Experimental protocol and confirmatory evidence

This appendix develops the complete CPB decision chain. It defines transition coordinates and event functionals, reports every one- and two-dimensional system–model cell, and documents the separate validation-to-holdout experiment. It then connects spectral capacity, matched scaling, resampling, weighting, and a periodic-gradient objective to the transition signatures. Detector and coordinate sensitivity, continuous structure, refined references, and trajectory-held-out PDEBench data test each link independently.

The scoped core ledger contains 765 outcome-bearing separately trained model fits. The 125-fit cross-PDE matrix supplies the discovery panel; 240 equal-budget development fits and 170 new final-choice fits test internal selection; 48 registered intervention fits and 80 matched scaling fits test remedies and architecture; and 48 external fits support validation-only tuning and final evaluation on PDEBench. Fifty-four targeted controls complete these defined core blocks. Compute matching separately uses eight timing-only cells, each with three separately initialized short training repeats (24 total); these resource-only runs are outside the fit count. Ancillary single-fit range extensions, Sobolev candidate and final-control runs, and the legacy PDEBench discovery fit are also reported outside this scoped count. Evaluation-only replays, detector sweeps, and 36,000 coordinate re-selections are not training fits.

Throughout, the primary evidence is normalized field error and event recovery as functions of a signed transition coordinate. Aggregate error remains a reference quantity. This organization keeps the decision question visible: which models satisfy the declared event criteria, and which eligible model has the lowest transition-band error?

Validation-to-holdout protocol The 125-fit discovery matrix motivated the selection question but does not serve as its confirmatory test: its original test panel had already been inspected while the benchmark narrative was developed. We therefore retain that matrix as a discovery comparison and construct a separate validation-to-holdout experiment. Development, validation, and final-holdout samples are generated with distinct seeds and contain 800/400/1,200 examples per 1D system. The final panel is fixed before new training. No final-holdout prediction is produced until every selected checkpoint and all internal intervention and scaling controls are complete.

For Burgers, Allen–Cahn, and Swift–Hohenberg, each of the five operator families receives eight architecture-specific candidates and the same two development seeds. Every fit uses batch size 64, exactly 3,900 optimizer updates, 300 validation evaluations, and the same training and validation samples. The candidate grids vary learning rate, regularization, width, depth, and the family-specific

Table 5: Core outcome-bearing experiment ledger: 765 separately trained model fits. Holdout evaluation, timing-only calibration (8 cells; 24 short runs), and other ancillary exploratory controls fall outside this scoped total.

Evidence block	Scope	Comparison	Fits	Decision role
Cross-PDE factorial panel	3 one-dimensional + 2 two-dimensional PDEs	5 operator families x 5 seeds	125	profile signatures and cross-system coverage
Targeted mechanism controls	capacity, modes, and transition emphasis	18 endpoints x 3 fits	54	mechanism and remedy checks
Equal-budget development	3 PDEs	5 families x 8 candidates x 2 seeds	240	validation-only selector tuning
Independent final choices	3 PDEs	17 reachable configurations x 10 seeds	170	held-out selector comparison
Registered interventions	Burgers + Swift–Hohenberg	6 settings x 8 seeds	48	resampling, weighting, and spectral capacity
Parameter-matched scaling	Allen–Cahn	4 levels x 2 families x 5 seeds	40	parameter-count control
Compute-matched scaling	Allen–Cahn	4 levels x 2 families x 5 seeds	40	optimizer-budget control
External development	PDEBench Burgers	2 families x 8 candidates x 2 seeds	32	external validation-only tuning
External final choices	PDEBench Burgers	2 reachable configurations x 8 seeds	16	trajectory-held-out selector comparison
		Total outcome-bearing model fits	**765**	

representation parameter; candidate count and optimization access do not vary by architecture. A failed or non-finite fit remains in the record and makes the corresponding architecture incomplete rather than triggering a replacement run.

The aggregate and CPB paths read the same validation model cards. Selection is hierarchical: each path first chooses one candidate within every architecture, then applies that same path to the resulting architecture-level cards. The aggregate path minimizes validation E_{agg} at both stages. CPB retains candidates only when both the task-specific binary-recovery condition and continuous-structure condition are satisfied, then minimizes E_0 at both stages; an architecture with no eligible candidate contributes no CPB card.

Values within 2% of the best primary metric are treated as practically equivalent at each stage. Aggregate ties are ordered by E_0, E_+, E_-, parameter count, training time, and stable identifier. CPB ties are ordered by E_+, E_-, E_{agg}, parameter count, training time, and stable identifier. If the cross-architecture CPB set is empty, the selector returns NO ELIGIBLE CANDIDATE; it does not substitute the least inaccurate ineligible candidate. This outcome is reported as abstention when the capability conditions are unmet. The selected configurations, task schema, tie rules, and ten final training seeds are committed before holdout access.

Table 6: New-experiment design. Counts denote separately trained fits; coordinate rows denote deterministic selector re-executions rather than new training fits. Reachable final configurations were chosen after validation selection and before holdout access.

Block	Fixed comparison	Scale
Internal equal tuning	3 systems $\times$ 5 families $\times$ 8 candidates $\times$ 2 seeds	240 fits
Internal final selection	17 configurations reachable by either selector	170 fits
Primary interventions	Burgers uniform/resample/weight; Swift modes 4/8/16	48 fits
Parameter matching	4 FNO–LinearNO levels, 5 paired seeds	40 fits
Compute matching	4 FNO–LinearNO levels, 5 paired seeds	40 fits
Coordinate robustness	3 scales $\times$ 5 offsets $\times$ 4 noise levels $\times$ 200 repeats, 3 systems	36,000 reruns
External equal tuning	2 families $\times$ 8 candidates $\times$ 2 seeds	32 fits
External final selection	2 aggregate-reachable configurations	16 fits

The confirmatory family comprises six contrasts fixed before holdout access: two Allen–Cahn selection contrasts, two Burgers sampling interventions, the Swift–Hohenberg 16-minus-4-mode contrast, and the external CPB selection contrast. A synchronized hierarchical bootstrap resamples paired training seeds outside and paired test samples inside. The aggregate and CPB cross-architecture stages are re-executed in every bootstrap draw over the architecture-level candidates chosen by the initial two-seed tuning. The resulting intervals include second-stage selection instability conditional

on those within-architecture choices; they do not represent uncertainty over the full 40-candidate tuning grid. Five endpoints yield numerical contrasts and receive simultaneous 95% max-statistic intervals over 20,000 draws. The external endpoint does not meet the capability conditions chosen before holdout access, so no CPB model is selected; it remains in the family as a non-rejection rather than being replaced by a post hoc numeric comparison. A missing seed, duplicate record, unmatched sample identifier, non-finite value, update mismatch, validation-cadence mismatch, or inconsistent dataset identity terminates confirmatory analysis. Fixed holdout profile bins with zero coordinate support are retained with count zero and an undefined per-bin estimate; bin edges are never adapted to the holdout. These display-only gaps do not alter any confirmatory scalar estimand.

Interventions have an additional joint acceptance rule. Critical error must improve with its simultaneous upper bound below zero; binary recovery may fall by at most 0.01; continuous discrepancy may increase by at most 0.02; and aggregate and smooth-side error may each worsen by at most 5%. This rule separates a local precision gain from a detector-only change or a transfer of error to another regime.

Confirmatory selection and controls The confirmatory comparison appears in Table 2 of the main text.

The two selectors agree on Burgers and Swift–Hohenberg and differ on the Allen–Cahn and external decisions. For Allen–Cahn, aggregate validation selects LinearNO, whereas CPB selects FNO. The holdout model cards show a local-precision trade-off: FNO lowers critical nRMSE from 0.0017248 to 0.0004135 (absolute contrast -0.0013113; 76% relative), while its aggregate nRMSE is 52.6% higher and both post-transition phase-recovery values remain essentially one. In 20,000 hierarchical bootstrap draws, aggregate selection chooses LinearNO in 19,986 draws and CPB chooses FNO in all 20,000. This concentration demonstrates that the local-error ordering is attached to the selector, rather than to one favorable fitted pair; it does not establish downstream phase improvement.

Burgers and Swift–Hohenberg provide complementary stability cases. Both selectors choose FNO on validation. Their final model cards combine low critical error with post-transition recovery 0.998 and 0.980, respectively. Selector agreement is informative here: transition-aware evaluation confirms that the aggregate winner also recovers the declared event, rather than forcing a change on every system.

Table 7: Preregistered primary contrasts. Estimates are variant minus baseline; negative values favor the CPB choice or intervention for error endpoints. Intervals are simultaneous 95% max-statistic intervals over all five estimable numerical endpoints.

Contrast	estimate	lower	upper
Allen critical: CPB − aggregate	-0.001311	-0.001607	-0.001015
Allen recovery: CPB − aggregate	-0.000978	-0.003760	0.001804
Burgers resampling − uniform	-0.000196	-0.000271	-0.000120
Burgers weighting − uniform	-0.000245	-0.000360	-0.000129
Swift modes 16 − modes 4	-0.067779	-0.095908	-0.039650

The Allen critical interval is strictly favorable, and the recovery interval lies above the preregistered -0.01 noninferiority margin. Both Burgers interventions and the Swift retained-mode comparison also have upper bounds below zero. The external family member is nonestimable by design because the validation capability set is empty; it is retained as a non-rejection in the familywise accounting.

All eight matched contrasts favor FNO. For parameter matching, estimates rise from 5.15×10^{-4} to 1.63×10^{-3} and every simultaneous lower bound is positive; the smallest is 3.38×10^{-5}. Under matched optimizer access, estimates range from 9.64×10^{-4} to 1.81×10^{-3}, with the smallest lower bound 6.63×10^{-4}. The critical-error ordering therefore cannot be explained by parameter count or by additional training updates within these comparisons; this control does not establish overall architectural superiority.

The two Burgers interventions meet the critical-error, binary, and continuous criteria but exceed the 5% aggregate and smooth-side transfer limits. The Swift intervention improves aggregate error, recovery, and continuous structure; its only unmet criterion is a smooth-side relative ratio whose baseline is nearly zero. These outcomes estimate controlled effects of the planned interventions within this paired training setup while preserving the broader performance requirement as a separate

Table 8: Planned interventions separate transition-local effects from performance across the full trajectory. A check mark denotes satisfaction of the chosen criterion. All interventions significantly improve critical error and preserve binary and continuous event structure.

Intervention	ΔE_0	ΔR_+	ΔC_+	ΔE_{agg}	transition/event	all criteria
Burgers resampling	-1.96×10^{-4}	$+0.0007$	0.000	$+13.3\%$	✓	–
Burgers weighting	-2.45×10^{-4}	$+0.0005$	0.000	$+27.6\%$	✓	–
Swift modes 16	-0.0678	$+0.977$	-4.906	-88.2%	✓	–

full-trajectory test. Preregistration prevents outcome-driven intervention choice but does not identify a unique causal mechanism.

External capability check External tuning compares FNO and LinearNO with eight candidates and two development seeds per family. Aggregate validation selects FNO. None of the 16 validation candidates satisfies both $R_+ \geq 0.95$ and $C_+ \leq 0.015625$, so CPB returns NO ELIGIBLE CANDIDATE before final-test access. The aggregate-reachable FNO and LinearNO configurations are then trained with eight new seeds and evaluated on held-out trajectories. The FNO model card is $E_{\mathrm{agg}} = 0.0801$, $E_0 = 0.0511$, $R_+ = 0.210$, and $C_+ = 0.500$. The low event recovery agrees with the validation result on an independent split. CPB therefore withholds endorsement among the tested configurations. The empty eligible set does not identify representation as its unique cause.

Post-hoc observables After the confirmatory analysis, we examined further observables in the saved sample-level predictions. It performs no training, inference, threshold selection, or model selection; all results here are post hoc and exploratory. Intervals use 20,000 percentile-bootstrap draws with training fit as the independent unit.

First, we estimate the finite-horizon event boundary as the coordinate cut that minimizes balanced error for the reference detector state. The Allen–Cahn, Burgers, and Swift–Hohenberg cuts are 0.00113, 0.05391, and 0.05130, respectively, all inside the chosen bands (half-widths 0.045, 0.090, and 0.075). Thus the bands cover the detector-defined transition neighborhood, but the analytic or proxy coordinate zero need not equal finite-time event onset. At the held-out coordinate resolution, Swift–Hohenberg FNO has no detectable boundary shift; CNO is delayed by 0.02589 (95% CI $[0.02191, 0.02971]$), while the other three families never cross the pattern threshold and have no estimable boundary.

Second, an Allen–Cahn zero-level-set analysis reports position only when reference and prediction contain the same nonzero number of interfaces. FNO and LinearNO have sub-cell conditional position errors of 0.0768 and 0.0496 cells, with exact-count rates 0.9985 and 0.9998. This reinforces the bounded main-text interpretation: FNO's lower E_0 is not an interface-geometry improvement over LinearNO.

Third, we regroup PDEBench predictions into 1,500 trajectories at eight queried times. FNO captures 0.887 (95% CI $[0.876, 0.899]$) of reference first front-detector crossings, but 0.828 of its trajectories later fall below the threshold; it captures only 0.107 (CI $[0.078, 0.142]$) of reference crossings that remain sustained. These are inconsistencies across separately queried time-conditioned predictions, not autoregressive rollout failures. Across all 15 detector settings, neither FNO nor LinearNO meets the selection condition; even FNO's highest recovery is 0.347, with $C_+ = 0.500$. Case-level summaries for these analyses are included in the anonymous release.

B BENCHMARK SPECIFICATION AND COMPLETE RESULTS

Metrics, systems, and event detectors For a physical system s, let $\mathcal{G}_s : x \mapsto y$ be the reference solution operator and let $\widehat{\mathcal{G}}_s$ be a learned operator. A signed transition coordinate $\delta_s(x)$ defines the operational boundary $\mathcal{C}_s^\delta = \{x : \delta_s(x) = 0\}$. Negative values denote the smooth or pre-transition side, while positive values denote the post-transition side. The coordinate need not be a Euclidean distance or an exact samplewise basin boundary. For a coordinate bin B, the reported field error is

$$E_s(B) = \mathbb{E}\left[\frac{\|\widehat{\mathcal{G}}_s(x) - \mathcal{G}_s(x)\|_2/\sqrt{N_s}}{\sigma_s} \,\middle|\, \delta_s(x) \in B \right], \tag{3}$$

where N_s is the number of spatial grid values and σ_s is the root mean square of all training targets for system s. Thus, the normalization is fixed across test samples and remains finite when an individual target is nearly zero.

Each system also defines a paired event functional $q_s(\widehat{y}, y) \in \{0, 1\}$ that tests task-specific agreement between a prediction and reference. Its conditional recovery rate is

$$R_s(B) = \Pr\Big[q_s(\widehat{\mathcal{G}}_s(x), \mathcal{G}_s(x)) = 1 \;\Big|\; \delta_s(x) \in B\Big]. \tag{4}$$

Each task also defines a per-sample continuous event discrepancy ℓ_s^{evt} and aggregation $\mathcal{A}_s$:

$$C_s(B) = \mathcal{A}_s\Big(\Big\{\ell_s^{\mathrm{evt}}(\widehat{\mathcal{G}}_s(x), \mathcal{G}_s(x)) : \delta_s(x) \in B\Big\}\Big). \tag{5}$$

The critical band is $|\delta_s| \le 0.15 d_s$, where d_s is the sampled coordinate scale. The remaining negative and positive samples define the smooth and post-transition regimes. We use

$$\rho_s = E_s(\text{critical})/E_s(\text{smooth}) \tag{6}$$

as a compact summary, but retain the full signed-coordinate curves because a transition may produce a peak, step, or one-sided ramp.

All systems use periodic domains. The one-dimensional grid has 256 points and the two-dimensional grid has 64×64 points. The initial 1D split contains 800/200/600 training/validation/test samples; the descriptive 2D split contains 600/150/400. Test inputs densely span the relevant control interval. The separate confirmatory split is specified in the validation-to-holdout protocol above.

Table 9: Systems, transition coordinates, and event functionals. The coordinate is positive on the post-transition side. For Burgers, T is the prediction time and u_0 is the realized initial field.

System	transition	signed coordinate δ_s	event functional
1D Burgers	front formation	$(-\min_x \partial_x u_0 - 1/T)/(2\pi)$	matching front presence and circular position
Allen–Cahn	phase selection	$m - 0$	matching phase sign and interface-count tolerance
1D Swift–Hoh.	pattern onset	$r - 0$	matching pattern presence and dominant wavenumber
2D Burgers	front formation	$(-\min(\partial_x u_0 + \partial_y u_0) - 1/T)/(2\pi)$	front sharpness
2D Swift–Hoh.	pattern onset	$r - 0$	pattern presence and radial wavenumber

The coordinates have different evidential status. Burgers uses an input-dependent inviscid compression proxy for front formation in a viscous finite-time task. Allen–Cahn uses the initial mean bias as a symmetry-informed task coordinate. Swift–Hohenberg uses the analytic linear-instability coordinate $r = 0$, while finite-horizon pattern detection remains a separate measurement. These labels prevent selector robustness from being interpreted as proof of a universal physical distance.

The 1D Burgers equation is $u_t + uu_x = \nu u_{xx}$ on $[0, 1]$, with $\nu = 2 \times 10^{-3}$ and $T = 0.5$. Inputs have the form $u_0(x) = A\sin(2\pi x) + \eta(x)$, where η is a smooth low-mode perturbation. Since the inviscid breaking time is $t^\star = -1/\min_x \partial_x u_0$, the realized-input coordinate in Table 9 reduces to $A - A_{\mathrm{crit}}$ only when $\eta = 0$, with $A_{\mathrm{crit}} = 1/(2\pi T) \simeq 0.318$. The reference uses the Cole–Hopf solution (Hopf, 1950; Cole, 1951). The 2D extension uses the analogous compression margin based on $\partial_x u_0 + \partial_y u_0$.

Allen–Cahn solves $u_t = \varepsilon^2 u_{xx} + u - u^3$ with $\varepsilon = 0.04$ (Allen & Cahn, 1979). The initial mean m indexes the dominant-phase selection neighborhood. Although the equation has $u \mapsto -u$ symmetry, the benchmark adds one fixed nonzero perturbation field across m; we therefore do not claim that $m = 0$ is an exact samplewise basin-boundary distance. Swift–Hohenberg solves $u_t = ru - (1 + \partial_{xx})^2 u - u^3$ in 1D, with the corresponding bi-Laplacian operator in 2D (Swift & Hohenberg, 1977). Linear stability gives pattern onset at $r = 0$. Because r is an equation parameter rather than a recoverable property of the small initial perturbation, it is supplied as a second input channel. Both gradient flows use spectral exponential time-differencing RK4 references (Kassam & Trefethen, 2005).

The nominal 1D Burgers detector requires sufficient maximum gradient and, when the reference contains a front, circular position agreement within about two grid cells. Allen–Cahn requires the dominant mean phase and interface count to match. Swift–Hohenberg requires agreement in pattern presence and, when a pattern is present, its dominant wavenumber. The 2D versions use front sharpness or radial spectral wavenumber.

Binary recovery is always paired with a continuous outcome. Burgers reports the ratio of predicted to reference front sharpness and circular position error. Swift–Hohenberg reports the predicted/reference pattern-amplitude ratio and exact dominant-wavenumber recovery. Allen–Cahn reports phase agreement and absolute interface-count error. These quantities prevent a threshold crossing from being mistaken for restoration of the physical structure.

Architectures and optimization All models map the complete input field to the complete output field at the same resolution. The implementations are self-contained and share the same training and evaluation code. The Fourier Neural Operator (FNO) uses four spectral blocks with low Fourier modes and pointwise residual branches (Li et al., 2021). DeepONet uses a branch network on the discretized field and a coordinate trunk whose feature inner product forms the output (Lu et al., 2021). The Convolutional Neural Operator (CNO) is a residual convolutional encoder–decoder. The patch Transformer uses three pre-norm self-attention blocks. LinearNO uses four slice-factorized linear-attention blocks, four heads, and 64 latent slices (Hu et al., 2026). Its query map is normalized over slices and its key map over points, so it never constructs a point-by-point attention matrix.

Table 10: Default model settings and parameter counts for a one-channel Burgers input. Swift–Hohenberg has one additional scalar input channel and therefore a small increase in the lifting layer.

Model	principal settings	depth	spatial representation	1D params	2D params
FNO	width 64/32, modes 16/12	4	truncated Fourier basis	287,425	1,186,177
DeepONet	width 128, latent basis 64	4	branch–trunk basis	99,201	590,849
CNO	width 32	3	residual convolutional U-Net	59,073	163,713
Transformer	width 64, 4 heads, patch 4	3	quadratic patch attention	101,124	102,672
LinearNO	width 96, 4 heads, 64 slices	4	linear slice attention	430,849	431,041

Training minimizes the normalized root mean square error (nRMSE) corresponding to Equation (3) using Adam, initial learning rate 10^{-3}, and cosine decay. The 1D runs use 300 epochs and batch size 64. The 2D runs use 200 epochs; LinearNO uses batch size 8 and the other baselines batch size 64. Model selection within a run uses the lowest validation objective. Five independent parameter initializations and data-order randomizations are used for every system–model cell. Reported uncertainty is computed over these five fits, not over individual test samples.

For a scalar metric z_i from fit i, the main tables report $\bar{z} = 5^{-1}\sum_{i=1}^{5} z_i$. Confidence intervals use 10,000 percentile bootstrap resamples of the five fit-level values with fixed random state 12345. The test of greater critical than smooth error enumerates sign flips of the five paired differences $E_{0,i} - E_{-,i}$. Therefore, the smallest attainable exact one-sided p-value is $1/32 = 0.03125$.

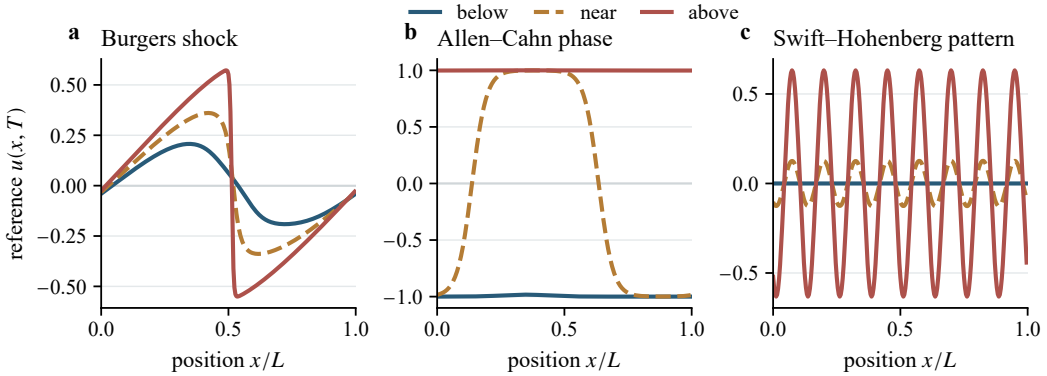

Figure 4: Representative one-dimensional reference solutions at the registered observation times, computed with the benchmark solvers and a fixed initial perturbation within each system. Low, near, and high controls are $A = (0.158, 0.318, 0.558)$ for Burgers, $m = (-0.12, 0, 0.12)$ for Allen–Cahn, and $r = (-0.25, 0.05, 0.30)$ for Swift–Hohenberg. These fields illustrate the front, phase, and pattern morphologies; they are not model predictions.

Table 11: Complete one-dimensional results. Entries are means over five independent fits. E_{agg} is aggregate nRMSE; E_-, E_0, and E_+ denote smooth, critical, and post-transition nRMSE; $\rho = E_0/E_-$; R_0 and R_+ are critical and post-transition event recovery. Very large Swift–Hohenberg ratios reflect a nearly vanishing smooth-side denominator and should be read with the full error triplet.

System	Model	E_{agg}	E_-	E_0	E_+	ρ	R_0	R_+
1D Burgers	FNO	0.002	9.4×10^{-4}	0.002	0.003	1.87	0.995	1.000
	CNO	0.017	0.008	0.008	0.029	1.00	0.996	0.993
	LinearNO	0.013	0.012	0.013	0.013	1.11	0.995	0.996
	Transformer	0.018	0.017	0.018	0.018	1.06	0.974	0.983
	DeepONet	0.493	0.128	0.383	0.774	3.00	0.001	0.000
Allen–Cahn	FNO	0.002	0.002	8.6×10^{-4}	0.002	0.54	1.000	0.996
	CNO	0.002	0.002	0.002	0.002	1.34	1.000	0.998
	LinearNO	0.001	8.4×10^{-4}	0.002	0.001	2.57	0.998	1.000
	Transformer	0.002	0.002	0.002	0.002	1.07	1.000	0.999
	DeepONet	0.113	0.104	0.266	0.070	2.56	0.989	0.992
1D Swift–Hoh.	FNO	0.060	9.5×10^{-4}	0.006	0.138	8.48	0.991	0.985
	CNO	0.510	0.001	0.055	1.176	56.92	0.886	0.493
	LinearNO	0.612	3.3×10^{-5}	0.060	1.413	2.14×10^{3}	0.875	0.000
	Transformer	0.614	0.004	0.062	1.413	23.15	0.875	0.000
	DeepONet	0.611	7.3×10^{-5}	0.060	1.412	2.04×10^{3}	0.875	0.000

Complete one-dimensional results The full table makes three distinctions explicit. First, small aggregate error does not imply a flat local profile. FNO has Burgers aggregate nRMSE 0.002 but $\rho = 1.87$. Second, event recovery and field error can disagree. DeepONet's Burgers critical error is high and its front is almost half a period out of position, whereas FNO, CNO, LinearNO, and the Transformer largely recover the event. Third, small pre-onset Swift–Hohenberg error does not imply post-onset pattern recovery. DeepONet, LinearNO, and the Transformer all have $R_+ = 0$.

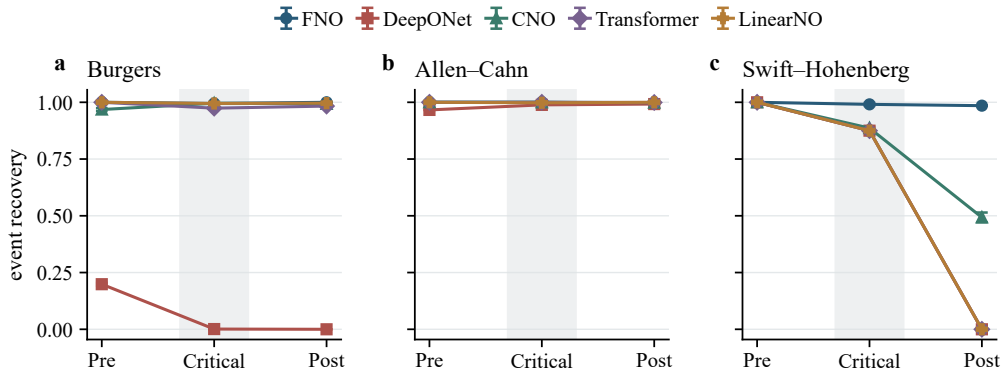

Figure 5: Event recovery by regime for all five operators and three 1D systems. Points show five-fit means and standard errors, with marker shape and color identifying the operator. The three columns correspond to front formation, phase selection, and pattern onset. This view exposes architecture-specific recovery that a single aggregate cannot resolve.

Fourteen of 15 system–model cells have mean $\rho > 1$ at the nominal band. The Allen–Cahn FNO is the exception; Burgers CNO exceeds one by just 0.013% and is statistically and numerically flat. The resulting pattern is systematic with an informative architecture-specific exception. The exact paired tests also prevent the large test-sample count from being mistaken for the number of independent training replicates.

Complete two-dimensional results The two-dimensional extensions retain the main architecture separation. On 2D Burgers, FNO, CNO, LinearNO, and the Transformer have post-front recovery near one, while DeepONet's nominal recovery is 0.010. The latter is threshold-sensitive and is therefore interpreted jointly with its continuous sharpness ratio in the detector-sensitivity analysis below. On 2D Swift–Hohenberg, FNO and CNO recover all post-onset patterns, whereas DeepONet,

Table 12: Five-fit uncertainty and paired tests for the 1D study. Brackets are 95% percentile bootstrap intervals over fits. The final column is the exact one-sided paired permutation p-value for $E_0 > E_-$. With five pairs, the smallest attainable value is 0.03125.

System	Model	ρ mean [95% CI]	E_0 mean [95% CI]	p
1D Burgers	FNO	$1.87\,[1.76, 1.97]$	$0.002\,[0.002, 0.002]$	0.0312
	CNO	$1.00\,[0.97, 1.03]$	$0.008\,[0.008, 0.009]$	0.4688
	LinearNO	$1.11\,[1.08, 1.13]$	$0.013\,[0.012, 0.014]$	0.0312
	Transformer	$1.06\,[1.02, 1.10]$	$0.018\,[0.016, 0.021]$	0.0312
	DeepONet	$3.00\,[3.00, 3.00]$	$0.383\,[0.382, 0.385]$	0.0312
Allen–Cahn	FNO	$0.54\,[0.46, 0.63]$	$8.6 \times 10^{-4}\,[7.1 \times 10^{-4}, 9.7 \times 10^{-4}]$	1.0000
	CNO	$1.34\,[1.13, 1.60]$	$0.002\,[0.002, 0.003]$	0.0312
	LinearNO	$2.57\,[2.19, 3.06]$	$0.002\,[0.002, 0.002]$	0.0312
	Transformer	$1.07\,[0.99, 1.12]$	$0.002\,[0.002, 0.002]$	0.0625
	DeepONet	$2.56\,[2.55, 2.57]$	$0.266\,[0.265, 0.268]$	0.0312
1D Swift–Hoh.	FNO	$8.48\,[5.58, 12.28]$	$0.006\,[0.005, 0.008]$	0.0312
	CNO	$56.92\,[39.00, 81.32]$	$0.055\,[0.054, 0.056]$	0.0312
	LinearNO	$2.14 \times 10^3\,[1.40 \times 10^3, 3.00 \times 10^3]$	$0.060\,[0.060, 0.060]$	0.0312
	Transformer	$23.15\,[11.56, 43.41]$	$0.062\,[0.061, 0.063]$	0.0312
	DeepONet	$2.04 \times 10^3\,[594.02, 4.60 \times 10^3]$	$0.060\,[0.060, 0.060]$	0.0312

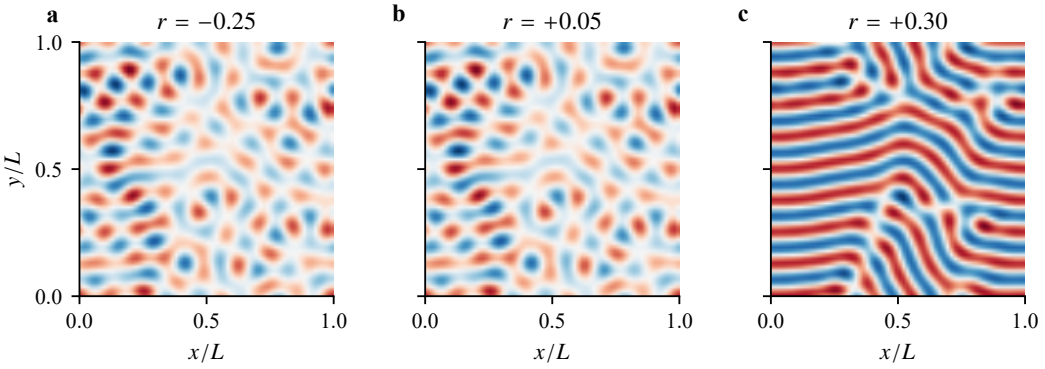

Figure 6: Two-dimensional Swift–Hohenberg reference solutions at $T = 60$ from the same initial perturbation. The controls are $r = -0.25, 0.05, 0.30$ from left to right; their RMS field amplitudes are 3.0×10^{-10}, 0.0196, and 0.423. Each panel uses its own symmetric color range to expose morphology, so color saturation should not be compared across panels. The nearly homogeneous first field is shown with amplified color solely for visibility.

Table 13: Complete two-dimensional results. Entries are means over five independent fits. E_{agg} is aggregate nRMSE; E_-, E_0, and E_+ denote smooth, critical, and post-transition nRMSE; $\rho = E_0/E_-$; R_0 and R_+ are critical and post-transition event recovery. Very large Swift–Hohenberg ratios reflect a nearly vanishing smooth-side denominator and should be read with the full error triplet.

System	Model	E_{agg}	E_-	E_0	E_+	ρ	R_0	R_+
2D Burgers	FNO	0.007	0.002	0.004	0.010	2.32	1.000	0.995
	CNO	0.017	0.013	0.014	0.021	1.06	0.995	0.995
	LinearNO	0.016	0.012	0.014	0.018	1.19	0.991	0.997
	Transformer	0.078	0.050	0.061	0.102	1.23	0.954	0.988
	DeepONet	0.461	0.156	0.305	0.687	1.96	0.817	0.010
2D Swift–Hoh.	FNO	0.362	0.001	0.014	0.847	11.84	0.966	1.000
	CNO	0.459	8.5×10^{-4}	0.025	1.071	32.31	0.983	1.000
	LinearNO	0.614	7.6×10^{-6}	0.035	1.432	5.64×10^3	0.949	0.000
	Transformer	0.615	0.003	0.036	1.431	13.53	0.949	0.000
	DeepONet	0.614	2.7×10^{-5}	0.035	1.432	1.39×10^3	0.949	0.000

LinearNO, and the Transformer recover none. LinearNO's low critical nRMSE (0.035) does not prevent its post-transition pattern collapse, illustrating the need to report both E_s and R_s.

C STABILITY, CONTROLS, AND NUMERICAL CHECKS

Definition stability and replication The nominal critical band is $|\delta_s| \leq 0.15 d_s$. We compare the neighboring widths $0.10 d_s$ and $0.20 d_s$ on all 75 one-dimensional fits. Table 14 shows that at least 13 of 15 system–model cells retain $\rho_s > 1$, all three profile signatures persist, and aggregate versus critical model selection on Allen–Cahn remains LinearNO versus FNO.

Table 14: The profile signatures and Allen–Cahn selection reversal persist across three critical-band widths over all 75 one-dimensional fits. "Shapes" counts the Burgers step, Allen–Cahn peak, and Swift–Hohenberg ramp. The last column is the fraction of Burgers test inputs assigned to different regimes by the realized and nominal coordinates.

Band width	$\rho_s > 1$ cells	shapes	Allen selection	relabeled
$0.10 d_s$	13/15	3/3	LinearNO $\rightarrow$ FNO	17.8%
$0.15 d_s$	14/15	3/3	LinearNO $\rightarrow$ FNO	19.7%
$0.20 d_s$	14/15	3/3	LinearNO $\rightarrow$ FNO	19.3%

The coordinate control is consequential rather than cosmetic. At the nominal band, realized-input Burgers ratios for LinearNO and the Transformer are 1.11 and 1.06; nominal $A - A_{\mathrm{crit}}$ reduces them to 0.88 and 0.91 by mixing samples whose perturbations move the breaking boundary. The input-dependent compression margin therefore prevents a structured label error.

The confirmatory coordinate analysis perturbs scale, offset, and additive noise over $3 \times 5 \times 4$ settings and 200 deterministic repeats per system, yielding 36,000 selector re-executions. Burgers preserves both the selector and the expected step signature in every rerun. Allen–Cahn retains the CPB FNO choice in 89.1% of reruns. Swift–Hohenberg retains its selector in 72.5% across the full perturbation cube and in all reruns at the exact registered coordinate. The selector conclusion is therefore strongest for Burgers, highly stable for Allen–Cahn, and explicitly tied to accurate onset localization for Swift–Hohenberg. This analysis measures decision sensitivity around the registered coordinates; it does not by itself validate those coordinates as unique physical constructs.

Two additional training fits are added to every decisive retained-mode, largest-capacity, and Burgers intervention comparison. The resulting additions comprise 18 retained-mode fits, ten largest-capacity fits, and eight intervention fits. The independent unit remains a training fit.

At the largest capacity, every Swift–Hohenberg fit remains above $\rho_s = 1.5$: the three-fit ranges are 9.59–37.58 (FNO), 2.31×10^4–2.45×10^4 (DeepONet), 12.01–44.34 (CNO), and 37.20–44.48 (Transformer). Because the smooth-side denominator is nearly zero, persistence and curve shape are more informative than the ratio magnitude. Burgers DeepONet is much more stable at 3.004–3.012. Thus every replicated endpoint supports the scoped claim that the gap persists at the largest tested size.

Mechanism and capacity controls For the FNO baseline, we measure the fraction of target spectral energy above mode 16 and correlate it with per-sample nRMSE. The correlation is 0.426 for Burgers and 0.679 for Swift–Hohenberg. Their high-mode fractions increase from 3.5×10^{-11} to 7.2×10^{-3} and from approximately zero to 3.5×10^{-6} across the transition. Allen–Cahn instead has correlation -0.02 and nearly constant high-mode energy of order 10^{-10}. These results are consistent with spectral-resolution pressure on the front and pattern tasks, whereas the phase-selection peak has no analogous high-mode association. Together with the retained-mode intervention below, they motivate a spectral-capacity account without identifying a unique error mechanism.

Across three independent fits at retained-mode counts $\{4, 8, 16\}$, Burgers critical nRMSE changes $0.00688 \rightarrow 0.00212 \rightarrow 0.00198$ and Swift–Hohenberg changes $0.05960 \rightarrow 0.00817 \rightarrow 0.00601$. Their four-to-16-mode reductions are 71% and 90%, respectively. Swift–Hohenberg critical event recovery rises from 0.875 to 0.981. Allen–Cahn changes $0.00192 \rightarrow 0.00132 \rightarrow 0.00110$ while retaining complete event recovery. The single-fit extension to $\{32, 48\}$ provides no consistent additional gain and is interpreted only as a wider-range comparison.

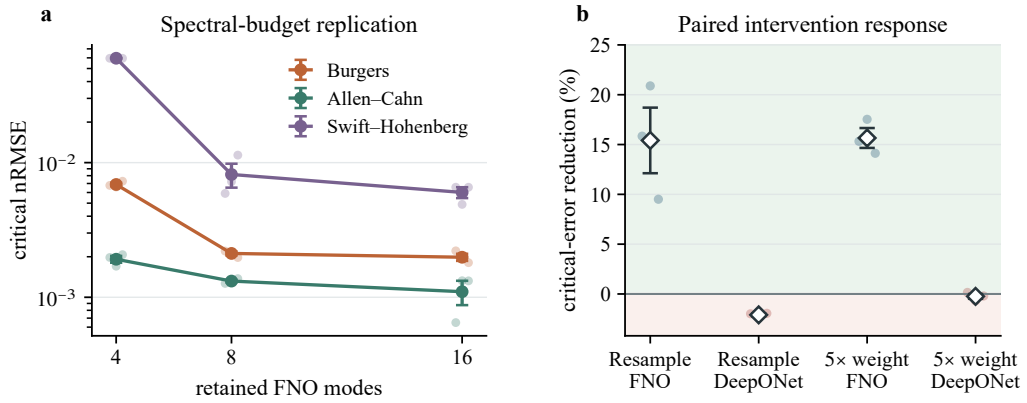

Figure 7: Spectral capacity and transition emphasis provide capability-matched improvements. (a) Critical nRMSE versus retained FNO modes; faint points are independent fits and error bars are standard errors. (b) Paired fit-level critical-nRMSE improvement from resampling or $5\times$ weighting, with diamonds and error bars marking the mean and standard error. Positive values denote lower error. Both interventions refine FNO without reducing event recovery, confirming that transition emphasis complements a represented event capability.

Across three architecture sizes, the largest tested model retains $\rho > 1.5$ on Swift–Hohenberg for all four original architectures and on Burgers for DeepONet in every replicated fit. The full size curves use one fit; the decisive largest endpoints use three independent fits. Capacity response is therefore mechanism-specific: increasing size within this range preserves the pattern and localization gaps.

Sampling, weighting, and loss controls The resampling control increases the realized fraction of critical Burgers training inputs to 56% while retaining the original validation and test sets. The weighting control instead multiplies critical-sample loss by five. Three paired training fits show that both controls improve FNO, which already locates the front, while leaving DeepONet localization unchanged.

Table 15: Paired three-fit Burgers interventions. Positive improvement denotes lower critical nRMSE. E_0 and E_1 are fit means, and the change is the mean paired percentage. Event columns give mean critical recovery.

Control	model	E_0	E_1	improvement	event $R_0 \to R_1$
Critical resampling	FNO	0.00171	0.00145	$+15.4\%$	$0.995 \to 0.995$
	DeepONet	0.38297	0.39100	-2.1%	$0.002 \to 0.003$
$5\times$ critical weight	FNO	0.00171	0.00144	$+15.7\%$	$0.995 \to 0.995$
	DeepONet	0.38297	0.38389	-0.24%	$0.002 \to 0.003$

The FNO gains begin from a small baseline, but they occur in every fit without event degradation. Their value is therefore precision refinement, not recovery of a missing event. DeepONet shows the complementary boundary: emphasizing the same samples did not create the localization behavior needed to place the front in these runs. This architecture-dependent response makes the controls a conditional remedy test rather than a blanket verdict on resampling or weighting.

This ancillary control followed an exploratory 43-fit, validation-only screen of structure-aware objectives. Within that screen, we compared four periodic-gradient weights under the criterion reported below and froze the selected value before training the final five-fit DeepONet and FNO panels. No final-holdout result entered either development step, but the control is post-hoc and outside both the confirmatory family and the 765-fit scoped core ledger.

The detector-free Sobolev control augments the field objective by

$$\mathcal{L} = \mathcal{L}_{\mathrm{nRMSE}} + \lambda_\nabla \mathcal{L}_\nabla, \qquad \mathcal{L}_\nabla = \frac{\left(\frac{1}{d}\sum_{j=1}^{d} \|D_j \widehat{y} - D_j y\|_2^2 / N_s\right)^{1/2}}{g_s}, \tag{7}$$

where D_j is the periodic forward difference scaled by the grid size in spatial direction j, d is spatial dimension, and g_s is the corresponding root-mean-square derivative scale computed from training targets. No critical labels or event detector outputs enter Equation (7).

Table 16: Exploratory validation-only comparison of four Burgers DeepONet gradient-loss coefficients following a broader structure-loss screen. Eligibility requires no more than a 5% increase in critical nRMSE relative to the same-initialization baseline. Among eligible candidates, critical event recovery is maximized.

λ_∇	critical nRMSE	change	critical recovery	eligible
0.2 (selected)	0.3780	+4.87%	0.905	yes
0.5	0.3856	+6.96%	0.921	no
1.0	0.3898	+8.15%	0.921	no
2.0	0.3932	+9.08%	0.921	no

Only $\lambda_\nabla = 0.2$ meets the validation criterion. We evaluate this selected coefficient on five separately trained final-test fits (seeds 0–4) for each of DeepONet and FNO. DeepONet's critical recovery increases from 0.001 to 0.784, but its critical nRMSE increases from 0.3835 to 0.4020. More decisively, post-transition sharpness falls from 1.202 to 0.188, circular position error remains 0.488, and post recovery reaches only 0.023.

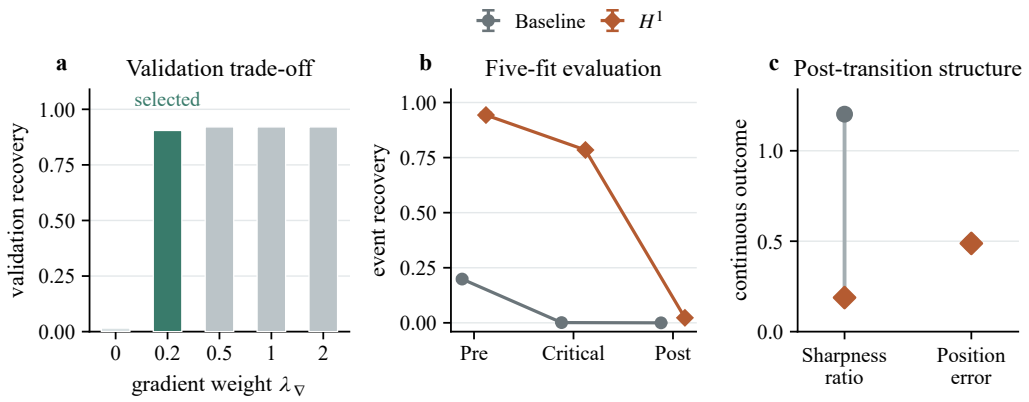

Figure 8: Continuous structure overturns an apparent binary improvement under the exploratory Sobolev control. (a) A validation-only comparison selects the gradient-loss coefficient under an at-most-5% critical-error constraint. (b) Across five test fits, recovery increases in the smooth and critical regimes but remains near zero after the transition; markers are five-fit means and error bars are standard errors. (c) The post-transition front is strongly smoothed and remains nearly half a period out of position.

For FNO, the same loss leaves the already correct event behavior essentially unchanged: critical recovery is 0.998 and post recovery is 1.000. The DeepONet result demonstrates that a thresholded critical-band number can improve even as physical front fidelity degrades. This is why CPB treats continuous structure measures and regime-separated recovery as necessary companions to a binary detector.

Detector and reference checks We apply the same detector grid to every fit. The 1D Burgers grid crosses sharpness thresholds $\{0.06, 0.08, 0.10, 0.12, 0.15\}$ with position tolerances $\{1, 2, 4\}$ cells. Allen–Cahn crosses interface levels $\{-0.05, 0, 0.05\}$ with count tolerances $\{0, 1, 2\}$. The 1D/2D Swift–Hohenberg grid crosses amplitude thresholds $\{0.025, 0.04, 0.05, 0.06, 0.075, 0.10\}$ with wavenumber tolerances $\{0, 1\}$. The 2D Burgers sharpness thresholds are $\{0.20, 0.25, 0.30, 0.35, 0.40\}$.

Allen–Cahn critical recovery remains 0.989–1.000 for every architecture and detector configuration. The 1D DeepONet front remains absent under every binary configuration because its position error is

Table 17: Post-transition conclusions remain stable across detector choices when supported by continuous structure. Results are averaged over five fits; the range column spans the full detector grid. The last column is continuous or threshold-free: a is pattern-amplitude ratio, k exact-wavenumber recovery, s shock-sharpness ratio, and e_x circular position error.

System	Model	nominal	range	continuous post outcome
1D Burgers	DeepONet	0.000	[0.000, 0.000]	$s = 1.202,\ e_x = 0.488$
	DeepONet+H^1	0.023	[0.000, 0.122]	$s = 0.188,\ e_x = 0.488$
	LinearNO	0.996	[0.914, 1.000]	$s = 1.002,\ e_x = 0.0003$
1D Swift–Hoh.	CNO	0.493	[0.469, 0.930]	$a = 0.521,\ k = 0.494$
	DeepONet	0.000	[0.000, 0.000]	$a = 0.022,\ k = 0.000$
	FNO	0.985	[0.985, 1.000]	$a = 0.989,\ k = 0.985$
	Transformer	0.000	[0.000, 0.000]	$a = 0.017$
	LinearNO	0.000	[0.000, 0.000]	$a = 0.00014,\ k = 0.002$
2D Burgers	DeepONet	0.010	[0.010, 0.971]	$s = 0.383$
	LinearNO	0.997	[0.984, 1.000]	$s = 1.001$
2D Swift–Hoh.	CNO	1.000	[0.984, 1.000]	$a = 0.704,\ k = 1.000$
	DeepONet	0.000	[0.000, 0.018]	$a = 0.0046,\ k = 0.000$
	FNO	1.000	[0.993, 1.000]	$a = 0.815,\ k = 1.000$
	Transformer	0.000	[0.000, 0.018]	$a = 0.0036,\ k = 0.000$
	LinearNO	0.000	[0.000, 0.018]	$a = 0.000039,\ k = 0.000$

close to one half-period. LinearNO's 1D Burgers post recovery remains high across the full grid. The 2D DeepONet front is different: its binary label spans 0.010–0.971, and the continuous sharpness ratio 0.383 supports only the narrower statement that the predicted front is substantially smoother than the reference.

Reference targets are deterministic functions of the stored operator inputs. For Swift–Hohenberg, random fields are drawn on the operator grid and then periodically interpolated to the solver grid. For 1D Burgers, a direct periodic Gaussian heat-kernel implementation of the Cole–Hopf solution avoids loss of precision near steep fronts. Allen–Cahn and the two-dimensional systems use the spectral time integrators described in Section B.

To assess numerical refinement, we select five realized inputs per system from the smooth side through the transition band and into the post-transition regime. Fine-grid initial conditions are reconstructed from the stored operator input by periodic Fourier interpolation. Each target is then recomputed at higher spatial resolution and, for time-stepped solvers, at half the time step.

Table 18: Reference refinement preserves all event labels while changing fields by at most 1.01×10^{-3} nRMSE. The middle column reports the maximum discrepancy over five realized inputs per system.

System	max production–refined nRMSE	unchanged event labels
1D Burgers	1.83×10^{-6}	5/5
Allen–Cahn	4.38×10^{-12}	5/5
1D Swift–Hohenberg	2.20×10^{-6}	5/5
2D Burgers	1.00×10^{-3}	5/5
2D Swift–Hohenberg	4.72×10^{-6}	5/5

Production/refinement pairs are 2048/4096 points for 1D Burgers, 512/1024 points with half time step for the 1D gradient flows, $192^2/256^2$ with half time step for 2D Burgers, and $64^2/128^2$ with half time step for 2D Swift–Hohenberg. All 25 event labels remain unchanged under refinement.

D Additional Theory and Held-Out Event-Margin Analysis

The next results clarify why an aggregate ranking can conflict with a transition ranking and when a binary event comparison is sensitive to its threshold. They concern any real-valued event score and therefore do not assume a particular neural-operator architecture.

Proposition 2 (When aggregate and critical rankings reverse)**.** *Let L be a nonnegative per-case loss on a population partitioned into a critical region C of probability $\pi \in (0, 1)$ and its complement. For two candidates A, B, write $d_C = \mathbb{E}[L_A - L_B \mid C] > 0$ and $d_{\bar{C}} = \mathbb{E}[L_A - L_B \mid \bar{C}] < 0$. Then A is*

worse on the critical region but better in aggregate exactly when

$$\pi < \frac{-d_{\bar{C}}}{d_C - d_{\bar{C}}}. \tag{8}$$

Proof. The law of total expectation gives $\mathbb{E}[L_A - L_B] = \pi d_C + (1-\pi)d_{\bar{C}}$. Solving the strict inequality $\mathbb{E}[L_A - L_B] < 0$ for π yields Equation (8). The denominator is positive under the stated signs. □

This identity applies directly to average per-case loss, including squared normalized field loss with a common normalization. Because nRMSE includes a square root and may use stratum-specific normalizers, the identity is a mechanism statement rather than an algebraic equality for every nRMSE column. It predicts that a small transition region can be outweighed by a modest advantage on common smooth cases.

Proposition 3 (Event disagreement requires score error)**.** *Let s be a reference score, $\hat{s}$ a predicted score, and τ the fixed event threshold. Set $Y = \mathbf{1}\{s \geq \tau\}$, $\hat{Y} = \mathbf{1}\{\hat{s} \geq \tau\}$, and $M = |s - \tau|$. If $Y \neq \hat{Y}$, then $|s - \hat{s}| \geq M$. Consequently, for any reference-defined subset A on which $M \geq m > 0$,*

$$\Pr(\hat{Y} \neq Y \mid A) \leq \frac{\mathbb{E}[|s - \hat{s}| \mid A]}{m}. \tag{9}$$

Proof. When s and $\hat{s}$ fall on opposite sides of τ, the interval between them contains τ; hence $|s - \hat{s}| \geq |s - \tau|$. On A, the mismatch indicator is at most $|s - \hat{s}|/m$. Taking its conditional expectation proves Equation (9). □

The bound does not imply that far-margin cases are always easier: a candidate may have much larger score error on those cases. In particular, a predictor that never emits a positive pattern event can appear accurate among near-margin negative references yet fail every far-margin positive reference.

Proposition 4 (Threshold perturbation certificate)**.** *Write $R(t) = \Pr[\mathbf{1}\{s \geq t\} = \mathbf{1}\{\hat{s} \geq t\}]$. For $|\delta| \leq h$,*

$$|R(\tau + \delta) - R(\tau)| \leq \Pr(|s - \tau| \leq h) + \Pr(|\hat{s} - \tau| \leq h). \tag{10}$$

Proof. Moving the threshold from τ to $\tau + \delta$ can change a reference or predicted binary label only if its score lies between the two thresholds. The agreement indicator can change only if at least one label changes. The union bound and $|\delta| \leq h$ give the result. □

Held-out event-margin comparison. We applied the registered thresholds to frozen final-holdout predictions, without retraining or checkpoint selection. The comparison covers 170 independently initialized fits (17 system–variant groups, ten fits each), all aligned to the same 1,200 holdout cases per system.

For each system we fixed the near and far strata from the reference-only quartiles of $|s - \tau|$; each contains 300 cases. We additionally report agreement inside the original critical coordinate band, $|\delta_s| \leq h_s$, with half-widths $h_s = 0.045$, 0.090, and 0.075 for Allen–Cahn, Burgers, and Swift–Hohenberg. Each table entry is the mean of ten fit-level agreements. The percentile interval for far minus near uses 10,000 bootstrap resamples of the ten independently trained fits, while the shared test cases remain fixed. It thus describes training-seed variation, not uncertainty from new test populations.

Table 19: Held-out event-state agreement from frozen 1D predictions. Rows average ten independent fits on the same 1,200 cases. Critical denotes $|\delta_s| \leq h_s$; near/far are the bottom/top quartiles of reference threshold distance ($n = 300$ each). The 95% interval for $\Delta = A_{\text{far}} - A_{\text{near}}$ resamples fits with cases fixed.

System	Model (variant)	All	Critical	Near	Far	Δ [95% CI]
Allen–Cahn	CNO (04)	1.000	1.000	1.000	1.000	$+0.000$ [+0.000, +0.000]
	CNO (06)	1.000	1.000	1.000	1.000	$+0.000$ [+0.000, +0.000]
	DeepONet (02)	0.999	0.995	0.997	1.000	$+0.003$ [+0.002, +0.004]
	FNO (08)	1.000	1.000	1.000	1.000	$+0.000$ [+0.000, +0.000]
	LinearNO (04)	1.000	1.000	1.000	1.000	$+0.000$ [+0.000, +0.000]
	Transformer (04)	1.000	1.000	1.000	1.000	$+0.000$ [+0.000, +0.000]
	Transformer (06)	1.000	1.000	1.000	1.000	$+0.000$ [+0.000, +0.000]
Burgers	CNO (08)	0.994	0.992	0.986	1.000	$+0.014$ [+0.013, +0.015]
	DeepONet (02)	0.554	0.176	0.423	1.000	$+0.577$ [+0.577, +0.577]
	FNO (04)	0.998	0.997	0.992	1.000	$+0.008$ [+0.006, +0.009]
	LinearNO (04)	0.994	0.984	0.977	1.000	$+0.023$ [+0.018, +0.028]
	Transformer (04)	0.982	0.975	0.952	1.000	$+0.048$ [+0.042, +0.054]
Swift–Hohenberg	CNO (02)	0.974	0.863	0.958	1.000	$+0.042$ [+0.041, +0.043]
	DeepONet (06)	0.552	0.849	0.957	0.000	-0.957 [-0.957, -0.957]
	FNO (04)	0.999	0.996	0.998	1.000	$+0.002$ [+0.001, +0.004]
	LinearNO (04)	0.552	0.849	0.957	0.000	-0.957 [-0.957, -0.957]
	Transformer (08)	0.552	0.849	0.957	0.000	-0.957 [-0.957, -0.957]

The reference event prevalence in the near/far strata is 0.510/0.517 for Allen–Cahn, 0.423/1.000 for Burgers, and 0.043/1.000 for Swift–Hohenberg. These class balances matter for interpreting agreement. In Burgers, DeepONet has only 0.176 critical-band agreement, although its far-margin agreement is 1.000. In Swift–Hohenberg, DeepONet, LinearNO, and Transformer agree on approximately 0.957 of near-margin cases but on none of the far-margin cases: the former stratum is mostly no-event references, while the latter consists entirely of events. FNO maintains at least 0.992 near-margin and 1.000 far-margin agreement in all three systems. Binary agreement alone does not measure event location or continuous structure, which remain separately reported in the benchmark. The comparison therefore supports the hierarchy of binary and continuous checks rather than replacing it with a new selector.