Bifurcation Preservation as a Physics Diagnostic for Neural Phase-Field Surrogates

Alejandro Salinas-Medina [1 2 3] **Anisleidy Gonzalez-Mitjans** [2] **Xue Liu** [1 3 4]

Abstract

A common approach for evaluating neural surrogates of phase-field equations is aggregate field error against a reference solver, a measure that can overlook bifurcations: abrupt shifts between qualitatively distinct outcomes, e.g., whether a phase-field droplet dissolves or persists. We propose evaluating neural phase-field surrogates in terms of their capacity for bifurcation preservation. We demonstrate the diagnostic on the Cahn-Hilliard (CH) critical droplet boundary using a droplet-aware Fourier Neural Operator, which reaches a moderate held-out rollout error, with relative $L^2 = 0.153$, however, it shifts the threshold-stable $m = 8$ dissolve/persist boundary by 16.2%. This persists for different random seeds, training budgets, exact zero-mode conservation, a U-Net baseline, and an entire 2D (m, R_0) transition curve including out-of-training regimes. We then compared Allen-Cahn (AC), conservative AC, and CH models using the same epoch-checkpoint protocol. AC preserved the bifurcation once training had converged but fails when undertrained into the CH rollout-error band, and adding mass conservation to AC reduces preservation from $13/24$ checkpoints to $4/24$. Conservation is implicated as a contributing CH-specific feature, with operator order, free-energy curvature, and branch stability as remaining candidates.

[1]School of Computer Science, McGill University, Montreal, Quebec, Canada [2]Montreal Neurological Institute-Hospital, McGill University, Montreal, Quebec, Canada [3]Mila – Quebec AI Institute, Montreal, Quebec, Canada [4]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates. Correspondence to: Alejandro Salinas-Medina <alejandro.salinas@mail.mcgill.ca>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

1. Introduction

Neural surrogates for Partial Differential Equations (PDEs) are commonly evaluated using either MSE (mean-squared error) or relative L2 error compared to an existing solver benchmark, averaged over both space and time. This method is effective for smooth problems where small errors observed in the field are reflective of small errors occurring in the problem domain. However, this approach is problematic in the presence of bifurcations, where small changes in initial conditions result in the system exiting one regime and entering another, qualitatively different regime (Strogatz, 2018). An example of such problems is the Phase-field equations (Chen, 2002). The Cahn-Hilliard (CH) equation (Cahn & Hilliard, 1958) models conserved phase separation and coarsening, the Allen-Cahn (AC) equation (Allen & Cahn, 1979), depicting nonconserved order-parameter dynamics and interface motion driven by curvature, and the conservative Allen-Cahn (CAC) equation (Rubinstein & Sternberg, 1992) which is an extension of the Allen-Cahn equation with a conservation constraint for mass and the second order operator. Generally, in the sharp-interface region, the dynamics of CH is controlled by the curvature of the interface, which is considered a chemical potential, following the Gibbs-Thomson effect (Pego, 1989; Voorhees, 1985; Alikakos et al., 1994). In the droplet setting, this produces a critical-radius effect: a small droplet of one phase, surrounded by the other, will either dissolve or persist depending on whether its initial radius R_0 is below or above a critical value R_i (Dai & Pego, 2005). Predicting the wrong side of R_i is qualitatively wrong, even if the predicted field looks visually similar to the reference on average.

We ask a simple question: if a neural surrogate has low rollout error, does it also predict the right side of the bifurcation? In this paper we show that the answer can be no. Our main contributions are:

- We propose *bifurcation preservation* as a complementary diagnostic for neural phase-field surrogates, building on the publicly available SAV reference solver of Groves (Groves et al., 2025).
- We show empirically that a droplet-aware Fourier Neural Operator achieves moderate held-out rollout error

but still fails the threshold-stable $m = 8$ bifurcation diagnostic across seeds. A dense $m = 12$ evaluation reveals a second qualitative all-persist bracket failure at the 30-epoch training horizon.

- We extend the diagnostic to a 2D (m, R_0) parameter sweep across five interface scales, including two regimes outside the FNO-B training distribution, and show that the dissociation between rollout error and bifurcation preservation extends across the full transition curve $R_i^{(T)}(m)$ rather than localizing to a single regime.
- We test a set of robustness controls: seed variance, extended non-critical droplet budgets up to 50 trajectories, a longer 100-epoch training schedule, a trained zero-mode-projected FNO, and a U-Net baseline. These controls show that the failure is not removed by additional non-critical droplet exposure, longer training, hard zero-mode conservation, or changing to a convolutional architecture.
- We run a controlled mechanism sweep on three reference systems under a matched architecture, data budget, training protocol, and epoch-checkpoint schedule: Allen-Cahn (AC; second-order, non-conservative), conservative Allen-Cahn (CAC; second-order, mass-conserving), and CH (fourth-order, mass-conserving). AC preserves the bifurcation at convergence but fails when undertrained into the CH rollout-error band; CAC, which differs from AC only by adding conservation, preserves in only $4/24$ checkpoints versus $13/24$ for AC and $1/18$ for CH. Conservation is implicated as a contributing CH-specific feature.

1.1. The Cahn-Hilliard droplet bifurcation

We work with the Cahn-Hilliard equation in two dimensions on a unit periodic square,

$$\frac{\partial \phi}{\partial t} = \nabla^2 \mu, \qquad \mu = \frac{dF}{d\phi} - \epsilon_m^2 \nabla^2 \phi, \tag{1}$$

with double-well free energy $F(\phi) = (\phi^2 - 1)^2$ and interfacial parameter ϵ_m. The parameter $\epsilon_m = mh/(2\sqrt{2}\,\mathrm{atanh}(0.9))$ is set through an integer mesh count m and the grid spacing h, following the Groves convention (Groves et al., 2025).

If we initialize ϕ as a circular droplet of radius R_0 (positive inside, negative outside) on a 128×128 grid, two things can happen as time advances. If R_0 is small, curvature drives the droplet to shrink and eventually dissolve. If R_0 is large, the droplet survives and slowly approaches a stable equilibrium radius. The boundary between these two outcomes is the critical initial radius R_i.

We are interested in the *finite-horizon* version of this boundary, $R_i^{(T)}(\epsilon_m)$, defined as the bracket midpoint between the largest dissolving R_0 and the smallest persisting R_0 when each trajectory is evolved to time $T = 1.0\, t_{\mathrm{char}}$, where t_{char} is the characteristic time for two-dimensional diffusion across the domain.

Reference solver. We generate reference trajectories with the scalar auxiliary variable (SAV) solver from (Groves et al., 2025), which is energy-stable and mass-conserving by construction. For each trajectory we extract a single scalar radius $R(t)$ from the $\phi = 0$ contour by computing its centroid and averaging the distances of contour points to the centroid. A trajectory is classified as *persist* if a $\phi = 0$ contour exists at $t = T$ and $R(T)/R(0) \geq 0.5$, and as *dissolve* otherwise.

Reference boundaries. We build reference $R_i^{(T)}$ values for three regimes, $m \in \{4, 8, 12\}$, by a coarse log-spaced sweep of R_0 followed by a refinement inside the resulting bracket. All three boundaries reach relative bracket uncertainty below 5% (Table 1). We also test sensitivity to the persist threshold across $\{0.3, 0.5, 0.6, 0.7\}$.

Table 1. Reference finite-horizon critical radii at $T = 1.0\, t_{\mathrm{char}}$ on a 128×128 periodic grid. Bracket uncertainty $\delta R_i / R_i$ is below 5% in all cases. The third column reports behavior under variation of the persist threshold from 0.3 to 0.7.

m	$R_i^{(T)}$	$\delta R_i / R_i$	threshold sensitivity
4	0.0875	2.86%	shifts at threshold 0.6 (stress-test)
8	0.1166	2.78%	**stable** from 0.3 to 0.7 (main gate)
12	0.1335	1.12%	stable through 0.6, shifts at 0.7

The $m = 8$ regime is exceptional: its boundary value does not move at all across the four thresholds we tested. We therefore use $m = 8$ as the *main* quantitative test condition. The denser $m = 12$ sweep provides a secondary qualitative test: its reference boundary is stable through threshold 0.6, but its predicted boundary may fail to bracket. The $m = 4$ regime remains a stress test because its reference boundary shifts under stricter persist thresholds.

2. Method

2.1. The diagnostic

Given a neural surrogate g_θ and a reference $R_i^{(T)}(\epsilon_m)$, we define the *boundary error* (BE) as

$$\mathrm{BE}(\epsilon_m) = \frac{\left|\widehat{R}_i^{(T)}(\epsilon_m) - R_i^{(T)}(\epsilon_m)\right|}{R_i^{(T)}(\epsilon_m)}, \tag{2}$$

where $\widehat{R}_i^{(T)}$ is computed by running g_θ autoregressively on the same R_0 sweep used for the reference and classifying

each predicted trajectory by the same dissolve/persist rule. We declare an effect "real" only when BE exceeds both a 5% floor and the bracket uncertainty $\delta R_i/R_i$.

2.2. Two FNO models

We compare two Fourier Neural Operator surrogates (Li et al., 2020; Kovachki et al., 2023) sharing the same architecture (4 Fourier layers, 32 modes per direction, width 64) and trained for 30 epochs with one-step MSE loss, while monitoring relative L^2 error as a validation metric:

- **FNO-A:** trained on spinodal trajectories only (25 per ϵ_m, three values of m).
- **FNO-B:** the same spinodal data plus 17 "non-critical" droplet trajectories in total across the three m values (5 at $m = 4$, 7 at $m = 8$, 5 at $m = 12$), where "non-critical" means R_0 values clearly far from $R_i^{(T)}$, either much smaller or much larger.

The comparison between FNO-A and FNO-B addresses potential concerns regarding distribution shift. While FNO-A is restricted to spinodal data, FNO-B incorporates non-critical droplet trajectories. This reduces the possibility that failures observed near the critical boundary are due only to a lack of exposure to droplet-like structures during training, and instead tests whether the model handles the near-bifurcation regime correctly.

2.3. Training and evaluation protocol

For both models, we hold out the near-critical droplets (R_0 values within $\pm 30\%$ of the reference boundary) from training and use them only for evaluation. We report aggregate rollout error (relative L^2 and MSE) on three subsets: spinodal validation, all droplets, and the held-out near-boundary droplets. We then compute the predicted boundary $\widehat{R}_i^{(T)}$ and the boundary error per ϵ_m and per threshold.

2.4. Additional baselines

To test whether the failure is specific to vanilla FNO or to lack of mass conservation, we evaluate two additional baselines. First, we train a zero-mode-projected FNO, denoted FNO-B+P, using the same training data as FNO-B. The projection is applied during both training and autoregressive rollout, so the loss is computed on the mass-conserving prediction:

$$\phi^P_{t+\Delta t} = \phi^{\text{raw}}_{t+\Delta t} - \langle \phi^{\text{raw}}_{t+\Delta t} \rangle + \langle \phi_t \rangle.$$

Second, we train a U-Net baseline on the same training dataset. The U-Net uses an encoder-decoder convolutional architecture with skip connections, providing a non-spectral baseline for the same diagnostic.

3. Results

We organize the results in three stages. Stage 1 (Sec. 3.1-3.2) establishes that FNO-B has low rollout error yet fails the CH $m = 8$ bifurcation diagnostic. Stage 2 (Sec. 3.3-3.5) rules out simple explanations: training horizon, mass conservation, architecture, and parameter regime. Stage 3 (Sec. 3.6) asks whether the failure is intrinsic to FNO on phase-field bifurcations or specific to Cahn–Hilliard, and answers with a controlled mechanism sweep across Allen–Cahn, conservative Allen–Cahn, and Cahn–Hilliard.

Stage 1: Establish the failure

3.1. FNO-B has much lower rollout error than FNO-A

Table 2 reports relative-L^2 rollout error on droplet trajectories. FNO-A struggles on droplets, with relative-L^2 above 1.0 on every test subset, indicating poor transfer from spinodal-only training to droplet trajectories. FNO-B is much better: on non-critical droplets like the ones it was trained on, its relative L^2 is 0.025, and on held-out near-boundary droplets it is 0.153. That is a factor of about $9\times$ improvement on near-boundary droplets compared to FNO-A, and a factor of about $50\times$ improvement on the easier non-critical subset.

Table 2. Rollout error of FNO-A and FNO-B on droplet trajectories. FNO-B is much closer to the reference solver than FNO-A.

	FNO-A		FNO-B	
Subset	rel. L^2	MSE	rel. L^2	MSE
all droplets	1.32	1.62	0.088	0.041
held-out near-boundary	1.35	1.63	0.153	0.081
non-critical training-like	1.29	1.59	0.025	0.003

3.2. FNO-B fails the bifurcation diagnostic at $m = 8$

Despite the strong improvement in rollout error, FNO-B fails the bifurcation diagnostic at the main test condition. Table 3 compares predicted and reference boundaries.

Table 3. Boundary behavior at default threshold 0.5 for FNO-B. The $m = 8$ row gives the main quantitative boundary-error result. The dense $m = 12$ row gives a qualitative bracket failure because FNO-B classifies all dense $m = 12$ radii as persistent.

m	role	$R_i^{(T)}$ ref	$\widehat{R}_i^{(T)}$ FNO-B	boundary error
4	stress-test	0.0875	-	all dissolve
8	main gate	0.1166	0.0977	16.2%
12	dense support	0.1335	-	all persist

The 16.2% shift at $m = 8$ is well above both the 5% effect floor and the 2.78% bracket uncertainty. The boundary shifts downward, in the over-persistence direction: $\widehat{R}_i^{(T)} = 0.0977 < R_i^{(T)} = 0.1166$. By FNO-B's reckoning, droplets

in the band $0.0977 < R_0 < 0.1166$ are predicted to persist when the SAV reference says they should dissolve. Raw FNO-B over-stabilizes subcritical near-boundary droplets through two distinct failure modes (smooth delayed dissolution and contour re-emergence with non-physical large radii) documented in Appendix B. The post-hoc correction experiment below shows that enforcing mass alone can instead over-dissolve the persisting branch. The initial coarse $m = 12$ sweep appeared resolution-limited. We therefore ran a denser R_0 evaluation around the $m = 12$ transition. The dense SAV reference gives a stable finite-horizon boundary $R_i^{(T)} = 0.1335$ for thresholds 0.3-0.6, shifting only at threshold 0.7. Under this denser evaluation, all three original FNO-B seed-sweep models (30 epochs) classify every tested $m = 12$ radius as persistent and therefore fail to form a valid dissolve/persist bracket. The $m = 12$ over-persistence failure is partially recovered by longer training: at 100 epochs (Sec. 3.3), two of three seeds form a valid dense $m = 12$ bracket. We therefore treat $m = 12$ as a training-sensitive qualitative failure mode, while $m = 8$ is the training-robust main quantitative gate.

The $m = 4$ row shows FNO-B classifying every held-out near-boundary droplet as dissolving, which is the opposite direction from $m = 8$. Because the $m = 4$ reference itself is threshold-sensitive (Table 1), we report this as informative but not decisive.

Seed variance. To check whether the $m = 8$ boundary shift is specific to a single initialization, we trained three additional FNO-B models with seeds 124-126. All three reproduced a large downward shift of the held-out $m = 8$ boundary at threshold 0.5. The predicted critical radius was $\widehat{R}_i^{(T)} = 0.0875$ for seeds 124 and 125, and $\widehat{R}_i^{(T)} = 0.0977$ for seed 126, compared with the reference $R_i^{(T)} = 0.1166$. The resulting boundary errors were 24.9%, 24.9%, and 16.2%, with mean $22.0\% \pm 5.0\%$. Thus the failure direction is stable across seeds, although the exact boundary location and autoregressive rollout error vary with initialization. The failure is also robust to non-critical droplet budget: increasing from 17 to 30 to 50 droplets does not reduce boundary error and instead shifts the failure mode from over-persistence to bracket collapse (Appendix C).

Table 4. Seed variance for FNO-B on the held-out $m = 8$ boundary at threshold 0.5. The downward boundary shift appears in all three additional seeds.

seed	$\widehat{R}_i^{(T)}$	boundary error	held-out rel. L^2
124	0.0875	24.9%	0.524
125	0.0875	24.9%	0.687
126	0.0977	16.2%	0.356

Table 5. Dense $m = 12$ evaluation for the original FNO-B seed-sweep models at threshold 0.5. The SAV reference forms a valid bracket at $R_i^{(T)} = 0.1335$, but all FNO-B seeds classify every dense $m = 12$ radius as persistent.

Seed	SAV $R_i^{(T)}$	FNO-B predicted bracket	Failure direction
124	0.1335	none	all persist
125	0.1335	none	all persist
126	0.1335	none	all persist

Stage 2: Rule out simple explanations

3.3. Training-horizon control

To test whether the bifurcation failure is an artifact of the 30-epoch schedule, we retrained FNO-B for 100 epochs on seeds 124-126 (Table 6; full per-seed rollout analysis in Appendix G). Longer training reduced held-out rollout error in all three seeds but did not consistently improve bifurcation behavior: the seed with the lowest rollout error of any FNO-B model in this paper produced no valid $m = 8$ bracket, while another seed with worse rollout error matched the reference at threshold 0.5. The dissociation between rollout error and bifurcation outcome is therefore sharper at convergence than at 30 epochs, and aggregate rollout error remains a poor predictor of the diagnostic on the standard one-step training protocol.

Table 6. 100-epoch FNO-B on the held-out $m = 8$ boundary at threshold 0.5. Compared to the 30-epoch results in Table 4, held-out rollout error decreases substantially in all three seeds, but bifurcation behavior does not improve uniformly: the seed with the lowest rollout error fails to form a valid bracket, while another with worse rollout error matches the reference.

seed	$\widehat{R}_i^{(T)}$	boundary error	held-out rel. L^2
124	none (all dissolve)	-	0.097
125	0.0925	20.6%	0.332
126	0.1166	0.0%	0.137

3.4. Mass-conservation and architecture controls

Two further controls test whether the m=8 failure can be removed by enforcing structure or by swapping the architecture (full protocols, per-trajectory classifications, and the FNO-B+P / U-Net seed sweep are reported in Appendices E and F).

A post-hoc zero-mode correction applied at rollout time eliminates the most dramatic false-persistence cases but over-dissolves the reference-persisting droplet $R_0 = 0.119793$, so zero-mode drift is a causal but not sufficient component of the failure. A trained zero-mode-projected FNO (FNO-B+P) preserves the spatial mean to $\sim 10^{-7}$ in all three seeds yet classifies every held-out m=8 radius as dissolving, trading the over-persistence failure for loss of

the persisting branch. A U-Net baseline trained on the same data exactly matches the reference boundary in one seed (BE $= 0.0\%$), shifts it by 20.6% in another, and forms no valid bracket in the third; rollout error and mean drift are informative but do not determine the diagnostic outcome (seed 126 has low drift and forms a bracket but still shifts the boundary by 20.6%). Together these controls rule out two candidate explanations: the failure is not removed by enforcing mass conservation during training, and it is not specific to the spectral architecture.

3.5. 2D parameter-sweep diagnostic

The bifurcation diagnostic generalizes naturally to multi-parameter families. To test whether FNO-B preserves the transition *curve* $R_i^{(T)}(m)$ rather than just the scalar boundary at $m = 8$, we extended the SAV reference and the FNO-B evaluation to five interface scales $m \in \{4, 6, 8, 10, 12\}$. The regimes $m = 6$ and $m = 10$ are *not* present in the FNO-B training set: they probe whether the surrogate has learned the transition curve as a function of the interface scale, or only memorized values at the three trained m regimes.

The SAV reference produces a smooth monotone transition curve with relative bracket uncertainty below 3% at every m: $R_i^{(T)}(m) = (0.0875,\ 0.1050,\ 0.1166,\ 0.1230,\ 0.1335)$ for $m \in (4, 6, 8, 10, 12)$. Each m-slice contains both dissolving and persisting trajectories, so the diagnostic is well-defined across the full parameter slice (Table 7, Figure 1). FNO-B fails to preserve this transition curve. At 30 epochs, all three seeds fail to form a valid bracket at every m tested except $m = 4$, where two of three seeds produce a bracket with boundary error (5.7%) marginally exceeding the 5% effect floor. At 100 epochs, some seeds recover isolated brackets (including an exact match (BE $= 0.0\%$) at $m = 8$ for seed 126) but no seed at any horizon recovers a valid bracket at more than two of the five m values. The out-of-training regime $m = 6$ yields zero valid brackets across all six FNO-B configurations tested; the other out-of-training regime $m = 10$ yields a valid bracket in only one configuration (100-epoch seed 124, BE $= 14.6\%$).

The 2D sweep therefore strengthens the central claim. **Bifurcation failure is not confined to a single radius threshold or a single interface regime: it appears as failure to preserve a transition curve across interface scales, and is most consistent at the out-of-training-distribution regime** $m = 6$. The diagnostic both localizes the surrogate's most reliable success (an exact $m = 8$ match in seed 126 at 100 epochs) and exposes its broadest failure (uniform bracket collapse at $m = 6$) within a single 2D slice. We do not claim that this generalizes across PDEs or geometries, only that the dissociation observed at $m = 8$ is a particular cross-section of a curve-level failure rather than a single-condition artifact.

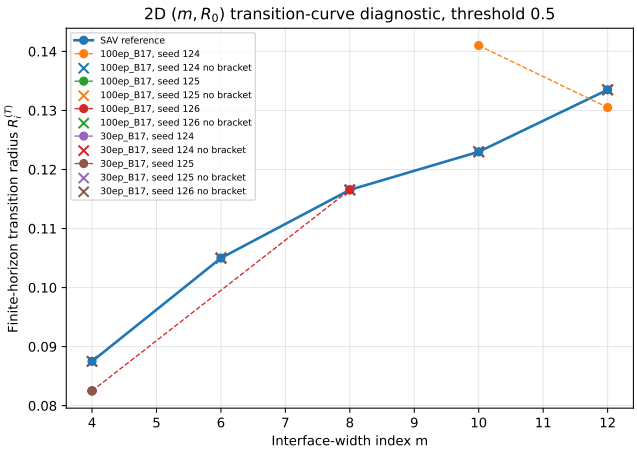

Figure 1. 2D (m, R_0) transition-curve diagnostic at threshold 0.5. The SAV reference (solid line, circles) defines a smooth monotone transition curve $R_i^{(T)}(m)$ at every tested interface scale, with relative bracket uncertainty below 3%. FNO-B configurations (seeds $\times$ training horizons) recover at most one bracket at four of the five m values; configurations that fail to form a valid dissolve/persist bracket are marked with red $\times$ at the SAV reference height. The out-of-training regime $m = 6$ yields no valid bracket across all six FNO-B configurations.

Stage 3: Localize the failure

Stage 1 and 2 establish that the CH $m = 8$ failure is robust to seed, training horizon, data budget, mass conservation, architecture, and parameter regime. We now ask whether the failure is intrinsic to FNO on phase-field bifurcations, or specific to Cahn-Hilliard.

3.6. Mechanism controls: AC and conservative AC

We test which structural feature(s) of CH drive the bifurcation failure by running matched-protocol experiments on two reference systems differing from CH along controlled axes. Allen-Cahn (AC) is second-order and non-conservative (Allen & Cahn, 1979):

$$\frac{\partial \phi}{\partial t} = -\big[\, 4\phi(\phi^2 - 1) - \epsilon_m^2 \nabla^2 \phi \,\big]. \tag{3}$$

Conservative Allen-Cahn (CAC) is second-order *and* mass-conserving:

$$\frac{\partial \phi}{\partial t} = \epsilon_m^2 \nabla^2 \phi - 4\phi(\phi^2 - 1) + \big\langle 4\phi(\phi^2 - 1) \big\rangle. \tag{4}$$

CAC differs from AC only by mass conservation and from CH only by operator order. All three systems use the same FNO-B architecture, data budget (25 spinodal + 17 non-critical droplet trajectories), one-step MSE loss, and seeds $\{123, 124, 125\}$. For AC and CAC we train each seed for 100 epochs with checkpoints at $\{1, 3, 5, 10, 20, 30, 50, 100\}$, generating 24 (seed, checkpoint) configurations per sys-

Table 7. 2D parameter-sweep summary at threshold 0.5. For each m, the table reports the SAV reference $R_i^{(T)}$, the relative bracket uncertainty, and the number of FNO-B configurations (out of 6: three seeds at each of two training horizons, 30 and 100 epochs) that produce a valid dissolve/persist bracket. Trained-support m values are $\{4, 8, 12\}$; $m \in \{6, 10\}$ are out of training distribution.

m	in training	SAV $R_i^{(T)}$	$\delta R_i/R_i$	valid / 6	best BE (configuration)
4	yes	0.0875	2.86%	3	$\approx 5.7\%$ (2 of 3 seeds, 100 ep)
6	**no**	0.1050	2.86%	0	–
8	yes	0.1166	2.78%	1	0.0% (seed 126, 100 ep)
10	**no**	0.1230	2.44%	1	14.6% (seed 124, 100 ep)
12	yes	0.1335	1.12%	1	2.25% (seed 124, 100 ep)

tem that map bifurcation outcome across the rollout-error range. For CH we aggregate the 18 FNO-B configurations from Sec. 3.1-3.3 and Appendix C. Bifurcation outcome is evaluated on the six-radius held-out grid $[0.88, 0.94, 0.97, 1.03, 1.06, 1.12]$ $R_i^{(T)}$ at threshold 0.5. Bifurcation regimes are comparable: $R_i^{(T)}/\epsilon_m$ is 7.3 for AC at $T = 20\,t_{\text{char}}$, 5.1 for CAC at $T = 20\,t_{\text{char}}$ (reference $R_i^{(T)} = 0.0769$, $\delta R_i/R_i = 1.19\%$), and 7.8 for CH $m = 8$. Per-seed and per-checkpoint detail in Appendix H.

AC preserves at convergence but fails when under-trained. At 100-epoch convergence, FNO-B-AC preserves the AC dissolve/persist boundary across three seeds (held-out rel. $L^2 \in [0.003, 0.015]$, boundary error $< 3\%$ at threshold 0.5). However, when AC is undertrained into the CH rollout-error band (rel. $L^2 \in [0.10, 0.20]$, epochs ≤ 3 plus one outlier at epoch 20), all five in-band AC checkpoints fail to form a valid bracket, classifying every held-out radius as persistent. The AC-CH dissociation is therefore not independent of rollout error: AC preserves only at low rollout error, and at CH-comparable rollout error AC also fails. The unresolved $T = 1$ AC regime ($R_i^{(T)}/\epsilon_m \approx 1.7$) is contour-resolution-limited and is reported as a separate diagnostic in Appendix D.

Adding mass conservation to AC substantially increases failure rate. Under the same protocol, CAC preserves the bifurcation in only 4 of 24 checkpoints, compared to 13 of 24 for AC – a 3.25× increase in failure rate from a single-axis change. CAC failures distribute across both shifted (8/24) and no-bracket (12/24) modes, mirroring the CH pattern (1/18 preserved, 8/18 shifted, 9/18 no-bracket). The matched-protocol comparison appears in Table 8 and Figure 2. CAC also produces shifted brackets at low rollout error (rel. $L^2 \approx$ 0.03-0.05) where AC is uniformly preserved, indicating that the diagnostic detects bifurcation failure at rollout errors where the aggregate metric reports near-perfect agreement.

Threshold sensitivity is physical. Unlike CH $m = 8$, the AC and CAC references at $T = 20$ are not threshold-invariant: $R_i^{(T)}$ shifts measurably across thresholds $\{0.3, 0.5, 0.7\}$. This reflects the absence of a stable persisting equilibrium in second-order phase-field dynamics: every droplet eventually shrinks under mean-curvature flow (Bronsard & Kohn, 1991), so what counts as "persisting" at fixed horizon depends on the threshold. The CH threshold-stability at $m = 8$ is therefore a feature of CH's stable equilibrium branch.

Table 8. Mechanism-control matched-protocol comparison at threshold 0.5. Counts across all (seed, checkpoint) configurations.

System	preserved	shifted	no bracket
AC ($n = 24$)	13	5	6
CAC ($n = 24$)	4	8	12
CH ($n = 18$)	1	8	9

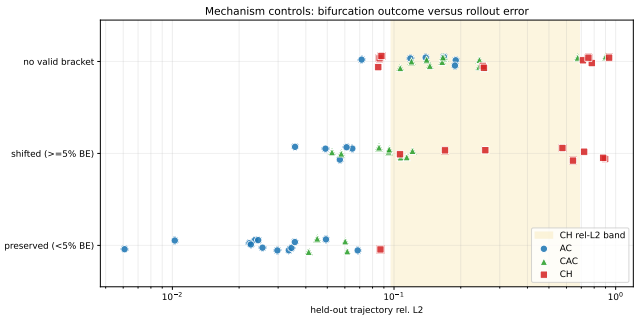

Figure 2. Mechanism-control scatter at threshold 0.5: held-out rollout error vs. bifurcation outcome for AC (blue), CAC (green), and CH (red). Shaded band marks the CH rollout-error range. CAC preservation rate falls below AC at matched protocol; CAC failures concentrate in the same rollout-error region as CH failures and also appear at low rollout error where AC is preserved.

Interpretation. Three findings emerge. (i) AC at convergence preserves the bifurcation, ruling out a universal FNO limitation. (ii) AC under-training reproduces the failure when rollout error enters the CH band, so rollout error is part of the explanation. (iii) Adding conservation to AC at fixed operator order moves CAC substantially toward CH-like failure rates, implicating conservation as a contributing CH-specific feature. CAC sits between AC and CH: same operator order as AC, same conservation as CH,

with failure rate intermediate. The residual gap to CH ($4/24$ preserved vs. $1/18$) is consistent with contributions from operator order, free-energy curvature, or branch stability, none of which are isolated by the present experiments. The diagnostic therefore functions not only as a pass/fail evaluation but as a mechanism-isolation tool: by varying the reference system along one structural axis at a time under a matched protocol, contributions of each axis to bifurcation failure can be quantified.

4. Discussion

Dissociation between rollout error and bifurcation outcome. Aggregate rollout error tells us how close the predicted field is to the reference, on average. Bifurcation preservation tells us whether the qualitative outcome (dissolves vs. persists) is the same. These can disagree, and the disagreement is sharper at convergence than at short training horizons: FNO-B substantially improves rollout error over FNO-A and still misclassifies the $m = 8$ bifurcation; the failure persists under seed, data-budget, and training-horizon controls; and at 100 epochs the seed with the lowest held-out rollout error of any FNO-B model in our study fails to form a valid $m = 8$ bracket, while another seed with worse rollout error matches the reference at the operational threshold. Dense $m = 12$ evaluation reveals a second qualitative all-persist bracket failure at the 30-epoch horizon, partially recovered by longer training. The 2D (m, R_0) sweep shows the dissociation across an entire transition curve: FNO-B forms a valid bracket in at most one of six configurations at four of five tested m values, and never forms a valid bracket at the out-of-training $m = 6$ regime, even though the SAV reference brackets cleanly there.Mechanistic field snapshots and spatial-mean traces underlying this dissociation are provided in Appendix A.

Scope of the empirical claims. We tested three phase-field reference systems (CH as the primary system, AC and conservative AC as mechanism controls) one bifurcation family per system (dissolve/persist for circular droplets), one spatial resolution (128×128), and finite horizons matched to each system's natural time scale ($T = 1.0\, t_{\mathrm{char}}$ for CH; $T = 1$ and $T = 20\, t_{\mathrm{char}}$ for AC; $T = 20\, t_{\mathrm{char}}$ for CAC). The main quantitative boundary-error signal comes from the threshold-stable $m = 8$ condition; the 2D parameter sweep (Sec. 3.5) extends this to a transition curve over five m values, including two out-of-training regimes, but the parameter slice is still 2D and is restricted to interface scale and droplet radius. The $m = 4$ case remains a threshold-sensitive stress test, while dense $m = 12$ evaluation provides a qualitative failure condition rather than a precise boundary-error estimate, and is partially recovered by longer training. We therefore do not claim universal generalization across PDEs, geometries, or bifurcation families. Instead, we present a controlled case study showing that rollout error, additional non-critical data, longer training on a one-step MSE loss, exact zero-mode conservation, and a simple architecture swap do not by themselves certify bifurcation preservation, and that the failure persists across a 2D parameter slice including out-of-training regimes.

Implications of the structure-preserving and architectural baselines. The trained projection (FNO-B+P) and U-Net baselines jointly show that bifurcation preservation is not equivalent to mass conservation, rollout error, mean-drift control, or architecture choice. The mechanism-control sweep (Sec. 3.6) sharpens this further. The same FNO-B that fails CH $m = 8$ preserves the AC boundary at convergence, but fails when undertrained into the CH rollout-error band; and the addition of mass conservation alone (CAC), holding operator order and protocol fixed, reduces preservation rate from $13/24$ to $4/24$ checkpoints. Conservation is therefore directly implicated as a CH-specific feature contributing to the failure, with operator order, free-energy curvature, and branch stability as remaining candidates. Bifurcation preservation is a separate physical diagnostic whose failure pattern is decomposable into structural contributions when measured under a matched cross-system protocol.

Relation to physics-informed and structure-preserving surrogates. This work is complementary to physics-informed and structure-preserving neural surrogate methods. Physics-informed neural networks incorporate physical laws or PDE residual constraints into the training objective (Raissi et al., 2019), while broader physics-informed machine learning methods use governing equations, symmetries, and constraints to regularize learning from data (Karniadakis et al., 2021). Other structure-preserving models encode conservation-inspired inductive biases more directly; for example, Hamiltonian neural networks are designed to respect energy-like conservation structure in dynamical systems (Greydanus et al., 2019).

More closely related are neural surrogate approaches that encode conservation or geometric structure into the operator itself, including conservative neural operators, equivariant operator architectures, and message-passing solvers designed to preserve integral quantities by construction (Brandstetter et al., 2022). These methods target the same broad failure mode identified here: unconstrained rollout can drift away from the physically admissible manifold. Recent work has also focused on improving long-rollout stability through iterative refinement (Lippe et al., 2023), which addresses error accumulation in field predictions but is orthogonal to the qualitative preservation question we study. Our results suggest that preserving a conserved quantity is not by itself equivalent to preserving a bifurcation. The zero-mode-projected FNO eliminates mass drift but still loses the

persisting branch, indicating that transition-level diagnostics remain necessary even for conservation-aware surrogates.

Our focus is different from these prior approaches: we ask whether a trained surrogate preserves a qualitative physical transition. The projected-FNO experiment connects these directions by showing that enforcing a conserved quantity can remove mass drift, but does not by itself guarantee preservation of the bifurcation boundary.

General form of the diagnostic. In this work the bifurcation parameter is the scalar droplet radius R_0, and the outcome is binary: dissolve or persist. More generally, the same idea applies whenever a physical system has a controllable parameter λ and a qualitative outcome map $q(u_T)$, where u_T is the state at the evaluation horizon. A reference solver defines a transition set

$$\Gamma = \{\lambda : q(u_T(\lambda)) \text{ changes class}\}.$$

A surrogate can then be evaluated by the displacement between its predicted transition set $\widehat{\Gamma}$ and the reference transition set Γ. In one-dimensional settings this reduces to the scalar boundary error used here; in higher-dimensional parameter spaces, the same idea can be measured with a signed-distance or Hausdorff-type error between transition surfaces. A cheaper screening variant evaluates two subcritical and two supercritical cases and tests dissolve/persist ordering before triggering full boundary-error estimation.

Limitations. The reference boundaries are finite-horizon. Asymptotic ($T \to \infty$) boundaries from (Groves et al., 2025) are systematically smaller, and we use the finite-horizon variant because it is the one a neural surrogate trained on bounded-time trajectories can actually be tested against. The $m = 4$ result remains uninformative: we know FNO-B classifies all held-out near-boundary droplets at $m = 4$ as dissolving, but the $m = 4$ reference itself is threshold-sensitive, so we cannot interpret the direction of disagreement cleanly. Finally, the post-hoc correction experiment clarifies the $R_0 = 0.113$ case, but also shows that mass conservation alone is insufficient: zero-mode correction removes subcritical false persistence while causing the reference-persisting $R_0 = 0.119793$ case to dissolve. The trained zero-mode-projected FNO addresses the simplest structure-preserving baseline, but it does not solve the bifurcation problem. The remaining limitation is the absence of a stronger structure-preserving surrogate that jointly preserves mass and accurately models nonzero-mode/interface dynamics near topology-changing events. Our projected-FNO result suggests that such a model must preserve not only the zero mode, but also the curvature-sensitive nonzero-mode dynamics that determine whether the near-critical interface disappears or survives.The mechanism-control sweep implicates conservation but does not uniquely identify the CH failure source. CAC sits between AC and CH in failure rate ($4/24$ preserved vs. $13/24$ for AC and $1/18$ for CH), so conservation explains a substantial part of the AC-CH gap, but a residual gap to CH remains, attributable to some combination of operator order, free-energy curvature, and branch stability. Isolating these residual axes requires further intermediate systems – a non-conservative fourth-order phase-field model or a CH variant with reduced free-energy stiffness – and is left for future work. CAC also lands at $R_i^{(T)}/\epsilon_m = 5.1$, somewhat below CH's 7.8; some of the CAC-CH gap may reflect bifurcation-regime differences rather than conservation alone.

5. Conclusion

We proposed bifurcation preservation as a physics-level diagnostic for neural phase-field surrogates and demonstrated it on the Cahn-Hilliard critical droplet boundary. A droplet-aware FNO shifts the threshold-stable $m = 8$ dissolve/persist boundary by 16.2% despite substantially improving rollout error over a spinodal-only baseline. The failure persists across seeds, non-critical droplet budgets up to 50 trajectories, longer training, exact zero-mode conservation, and a U-Net architectural baseline; a 2D (m, R_0) sweep extends the dissociation to a full transition curve, including out-of-training interface regimes. A controlled mechanism sweep across AC, conservative AC, and CH (differing along one structural axis at a time under a matched epoch-checkpoint protocol) shows that AC preserves the bifurcation at convergence but fails when undertrained into the CH rollout-error band, and that adding mass conservation to AC reduces preservation rate from $13/24$ to $4/24$ checkpoints. Conservation is implicated as a contributing CH-specific feature, with operator order, free-energy curvature, and branch stability as remaining candidates. The diagnostic therefore functions not only as a pass/fail evaluation but as a mechanism-isolation tool when combined across physical systems.

Impact Statement

This paper presents work whose goal is to improve the evaluation of neural surrogates for physical systems. The proposed diagnostic is intended to help identify qualitative failures that may be hidden by aggregate error metrics. We do not foresee direct negative societal impacts from this work beyond those generally associated with deploying inaccurate scientific machine-learning models in high-stakes physical simulations.

References

Alikakos, N. D., Bates, P. W., and Chen, X. Convergence of the Cahn–Hilliard equation to the Hele–Shaw model.

Archive for Rational Mechanics and Analysis, 128(2): 165–205, 1994.

Allen, S. M. and Cahn, J. W. A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. *Acta Metallurgica*, 27(6):1085–1095, 1979.

Brandstetter, J., Worrall, D., and Welling, M. Message passing neural PDE solvers. In *International Conference on Learning Representations*, 2022.

Bronsard, L. and Kohn, R. V. Motion by mean curvature as the singular limit of Ginzburg–Landau dynamics. *Journal of Differential Equations*, 90(2):211–237, 1991.

Cahn, J. W. and Hilliard, J. E. Free energy of a nonuniform system. i. interfacial free energy. *Journal of Chemical Physics*, 28(2):258–267, 1958.

Chen, L.-Q. Phase-field models for microstructure evolution. *Annual Review of Materials Research*, 32(1):113–140, 2002.

Dai, S. and Pego, R. L. Universal bounds on coarsening rates for mean-field models of phase transitions. *SIAM Journal on Mathematical Analysis*, 42(6):2362–2407, 2005.

Greydanus, S., Dzamba, M., and Yosinski, J. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Groves, S. M., Lu, M.-J., Alvarez-Yela, A. C., Gerardo-Ramírez, M., Stukenberg, P. T., Lowengrub, J. S., and Janes, K. A. Cahn–hilliard dynamical models for condensed biomolecular systems. *bioRxiv*, 2025. bioRxiv 2025.07.13.664571.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021.

Kovachki, N., Li, Z., Liu, B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., and Anandkumar, A. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2020.

Lippe, P., Veeling, B., Perdikaris, P., Turner, R., and Brandstetter, J. PDE-Refiner: Achieving accurate long rollouts with neural PDE solvers. In *Advances in Neural Information Processing Systems*, 2023.

Pego, R. L. Front migration in the nonlinear cahn–hilliard equation. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 422(1863): 261–278, 1989. doi: 10.1098/rspa.1989.0027.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Rubinstein, J. and Sternberg, P. Nonlocal reaction–diffusion equations and nucleation. *IMA Journal of Applied Mathematics*, 48(3):249–264, 1992.

Strogatz, S. H. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. CRC Press, 2nd edition, 2018.

Voorhees, P. W. The theory of ostwald ripening. *Journal of Statistical Physics*, 38:231–252, 1985. doi: 10.1007/BF01017860.

Yue, P., Zhou, C., and Feng, J. J. Spontaneous shrinkage of drops and mass conservation in phase-field simulations. *Journal of Computational Physics*, 223:1–9, 2007.

A. The mechanism: drift of the zero Fourier mode

To find out what FNO-B is actually doing past the dissolution event, we look at snapshots of the predicted field $\phi(x, t)$ at five times. Figure 3 shows these for the two cases that fail most dramatically, $R_0 = 0.100$ and $R_0 = 0.106$.

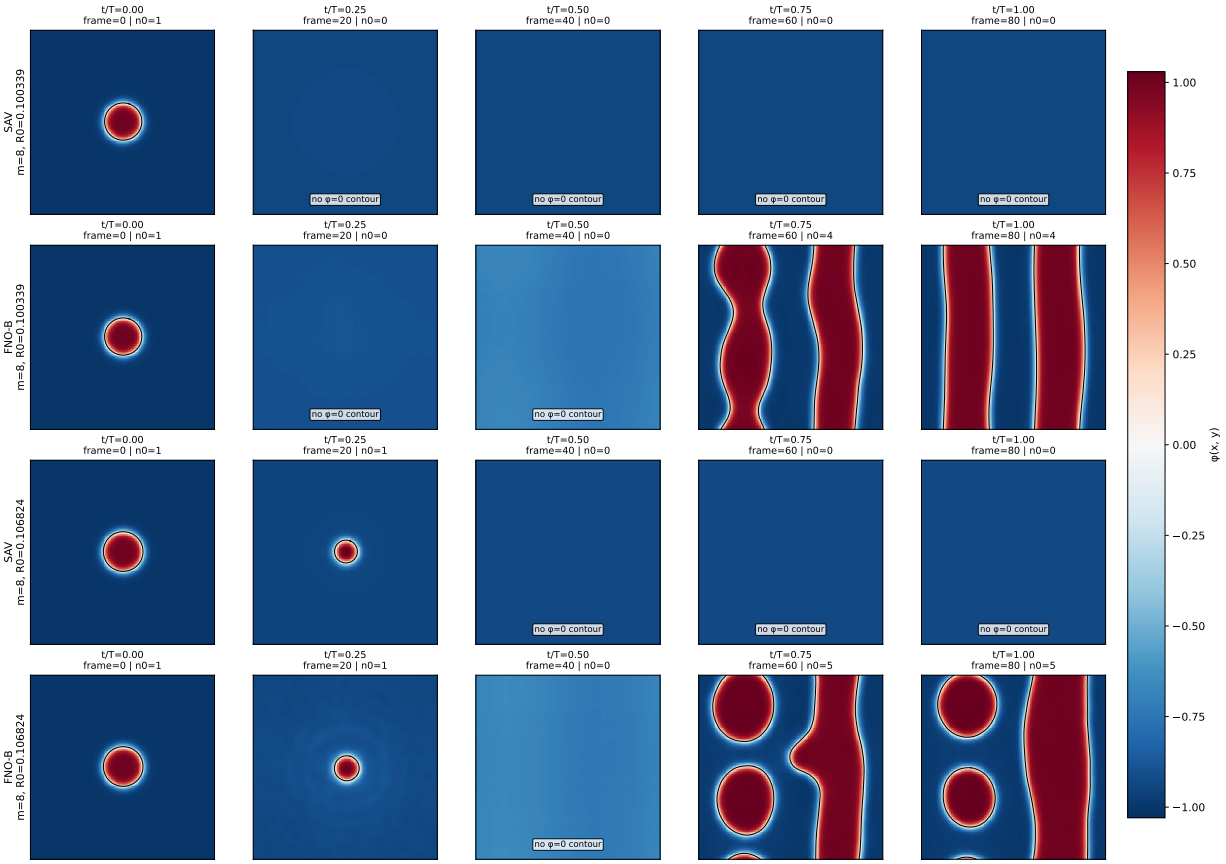

Figure 3. Field snapshots at $t/T \in \{0, 0.25, 0.5, 0.75, 1.0\}$ for $m = 8$, $R_0 = 0.100$ (top two rows) and $R_0 = 0.106$ (bottom two rows). SAV (rows 1, 3) dissolves the droplet and remains in a uniform negative phase (mean ≈ -0.93) for the rest of the trajectory. FNO-B (rows 2, 4) initially follows the dissolution, but by $t/T = 0.5$ the field's spatial mean has drifted from -0.93 toward -0.7, and by $t/T = 0.75$ it has reached the spinodal-decomposition regime (mean near zero) and produced multiple stripe-like positive-phase bands. The contour-based classifier counts these as persistent droplets, hence the 16.2% boundary shift.

The snapshots tell a clear story. SAV reaches the dissolved state by $t/T = 0.25$ (top row, $R_0 = 0.100$) or $t/T = 0.5$ (third row, $R_0 = 0.106$) and stays there: a uniform field at $\phi \approx -0.93$ with no $\phi = 0$ contour anywhere. FNO-B initially does the same thing. But then something interesting goes wrong.

The spatial mean of the FNO-B field drifts upward over time:

	$t/T = 0$	0.25	0.5	0.75	1.0
SAV mean ($R_0 = 0.100$)	-0.93	-0.93	-0.93	-0.93	-0.93
FNO-B mean ($R_0 = 0.100$)	-0.93	-0.90	-0.72	-0.06	$+0.01$
SAV mean ($R_0 = 0.106$)	-0.93	-0.93	-0.93	-0.93	-0.93
FNO-B mean ($R_0 = 0.106$)	-0.93	-0.89	-0.70	-0.08	-0.03

The Cahn-Hilliard equation conserves total mass, which on a periodic square is the same as conserving the spatial mean of ϕ. SAV gets this right exactly. FNO-B does not. The mean drifts toward zero.

This matters because $\phi \approx 0$ is precisely where Cahn-Hilliard is unstable: it is the spinodal point of the double-well free energy. Once the FNO-B field reaches that regime, the natural dynamics are spinodal decomposition, and FNO-B happens to be very well trained on spinodal trajectories. So the model produces what it knows, namely phase-separated stripes, in a

regime where the reference solver would correctly do nothing.

In Fourier language, the spatial mean of ϕ is just the zero Fourier/Laplacian mode $\hat{a}_0$. Under the periodic square boundary conditions we use, the basis FNO operates in is the Fourier basis, which coincides with the Laplacian eigenbasis. Vanilla FNO does not enforce any constraint on $\hat{a}_0$, so the network is free to drift in this mode. The bifurcation diagnostic is what makes the drift visible. Drift of the zero Fourier mode is a sufficient mechanism for FNO-B's most dramatic failures ($R_0 = 0.100$ and $R_0 = 0.107$): once the spatial mean enters the spinodal regime, the network produces stripe-like patterns whose $\phi = 0$ contours the classifier records as persistence. The milder $R_0 = 0.113$ case does not show the dramatic stripe regeneration seen at smaller radii, but the post-hoc correction experiment below shows that it also involves zero-mode drift: raw FNO-B drifts from mean -0.917 to -0.855, while the mass-corrected rollout dissolves by $t/T = 0.5$. Aggregate MSE averages over the whole field and dilutes both the contour-regeneration and delayed-dissolution failures; the bifurcation classification surfaces them.

B. Per-trajectory predicted radius traces

Figure 4 shows $R(t)/R(0)$ for SAV and FNO-B at four near-critical $m = 8$ initial radii, providing the per-trajectory detail underlying the left panel of Fig. 5. The reference dissolves cleanly for $R_0 \in \{0.100, 0.107, 0.113\}$ and persists for $R_0 = 0.120$. FNO-B disagrees in three different ways: classical over-stabilization at $R_0 = 0.113$ ($R(T)/R(0) = 0.56$, barely above threshold); contour re-emergence at $R_0 = 0.107$ climbing to ≈ 1.5; and extreme contour re-emergence at $R_0 = 0.100$ reaching ≈ 2.5, which is not a physical droplet radius. These three distinct failure modes are diluted by aggregate MSE but surfaced by the bifurcation classifier.

C. Droplet-data budget ablation

To test whether the $m = 8$ failure is simply an artifact of the small non-critical droplet training set, we retrained FNO-B with larger non-critical droplet budgets while keeping the same held-out near-boundary evaluation set. Near-critical radii were intentionally excluded from training so that the diagnostic tests extrapolation to the transition region rather than interpolation over examples already sampled on both sides of the boundary.

Increasing the non-critical droplet budget did not remove the failure. With the original 17-droplet budget, all three seeds produced valid held-out $m = 8$ brackets, but the boundary error remained large ($22.0\% \pm 5.0\%$). With a nominal 30-droplet budget, only one of three seeds produced a valid bracket, and that seed had 24.9% boundary error. With 50 non-critical droplets, again only one of three seeds produced a valid bracket, with 24.9% boundary error.

Notably, the direction of failure shifts with the non-critical droplet budget. At 17 droplets, all seeds produce a valid bracket but over-persist subcritical near-boundary droplets. At the larger budgets, the majority of seeds fail to form a valid bracket because they classify all held-out radii as dissolving. When a bracket does form, the boundary error remains pinned at 24.9% for both larger budgets. This rotation of failure mode without reduction in boundary error suggests that non-critical droplet exposure shapes the model's outcome bias, but does not resolve the near-critical interface dynamics that determine the bifurcation boundary.

Table 9. Non-critical droplet budget ablation for FNO-B on the held-out $m = 8$ boundary at threshold 0.5. A missing bracket means the model classified all held-out near-boundary droplets on the same side of the transition.

Droplets	Valid / seeds	BE (bracket)	rel. L^2
17	3/3	$22.0\% \pm 5.0\%$	0.522 ± 0.166
30	1/3	24.9%	0.648 ± 0.098
50	1/3	24.9%	0.519 ± 0.011

D. Allen-Cahn cross-PDE control details

Solver. We use a semi-implicit Fourier scheme for AC,

$$\hat{\phi}_k^{n+1} = \frac{\hat{\phi}_k^n - \Delta t\, 4\phi^n \widehat{((\phi^n)^2 - 1)}_k}{1 + \Delta t\, \epsilon_m^2 |k|^2}.$$

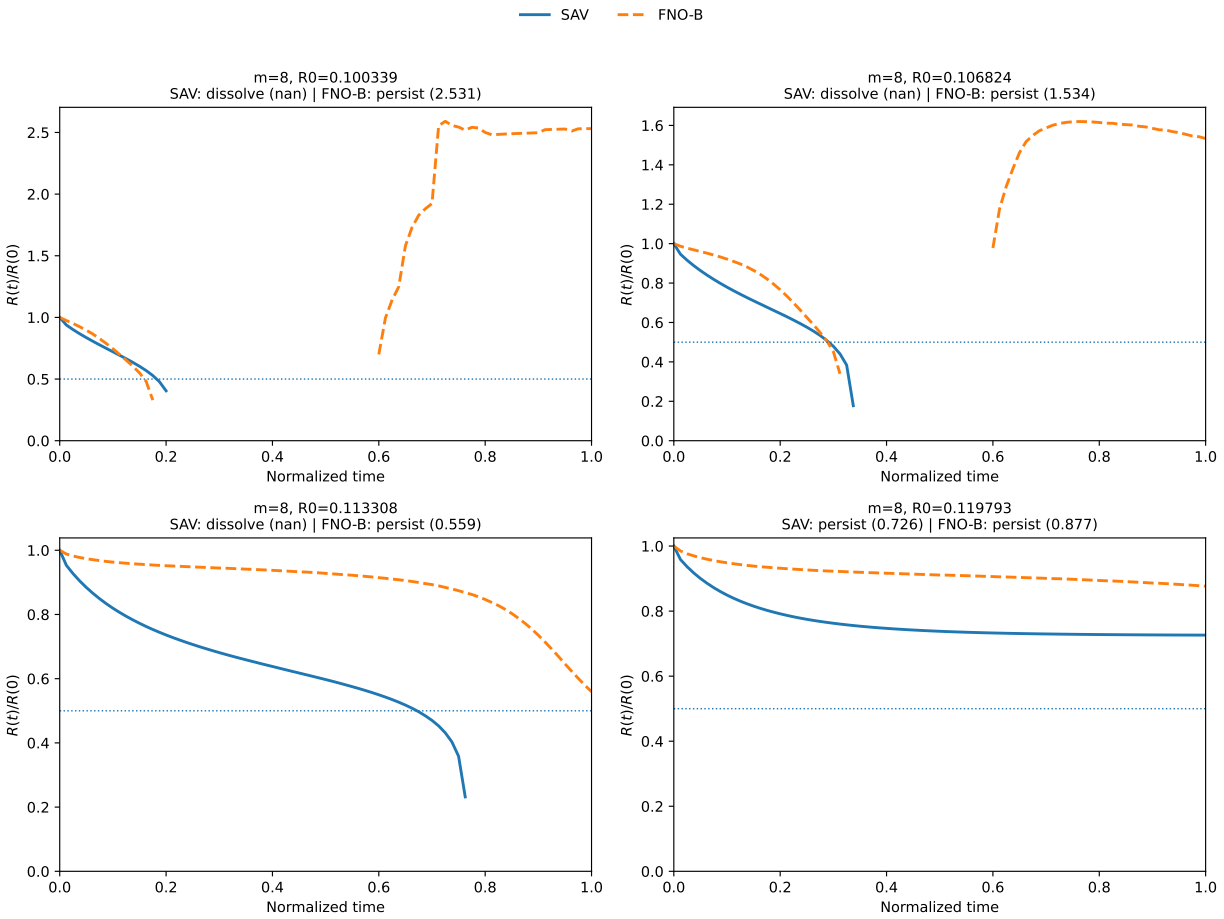

Figure 4. Predicted radius over time for SAV (blue, solid) and FNO-B (orange, dashed) at four near-critical $m = 8$ initial radii. Persist threshold $R(t)/R(0) = 0.5$ marked. SAV dissolves at $R_0 \in \{0.100, 0.107, 0.113\}$ and persists at $R_0 = 0.120$. FNO-B classifies the three subcritical cases as persistent.

At $T = 1$ we use $\Delta t = 2.5 \times 10^{-5}$ (matching the CH protocol); at $T = 20$ we use $\Delta t = 10^{-4}$, which the second-order AC operator permits stably. As a sanity check we verified that the spatial mean of every AC reference trajectory drifts (it should, since AC does not conserve mass), distinguishing the AC solver from a CH solver.

Training data. For each horizon we generated 25 spinodal AC trajectories (mean-zero noise initialization) and 17 non-critical droplet AC trajectories with R_0 outside a $\pm 30\%$ band around the AC reference $R_i^{(T)}$. Non-critical radii are recomputed for each horizon to remain non-critical relative to the corresponding reference.

T-screen. Before committing to $T = 20$, we ran a horizon screen at $T \in \{1, 4, 9, 20\}\, t_{\text{char}}$. The transition from unresolved ($R_i/\epsilon \approx 1.7$, bracket failure) to resolved ($R_i/\epsilon \approx 7.3$, bracket preservation) occurs between $T = 4$ and $T = 9$.

Reference refinement triggered by dense evaluation. At $T = 20$, the original four-radius held-out set $[0.88, 0.94, 1.06, 1.12]\, R_i^{(T)}$ gave label match across all seeds. Adding the dense pair at $[0.97, 1.03]\, R_i^{(T)}$ revealed that $0.97\, R_i^{(T)}$ actually persists, refining the threshold-0.5 reference from $\delta R_i/R_i = 3.1\%$ to 0.22%. We re-ran evaluation against the refined reference and re-confirmed label preservation across all three seeds. The diagnostic protocol therefore catches its own reference imprecision when the held-out grid is densified near the predicted boundary.

Per-seed details.

Table 10. FNO-B-AC per-seed held-out classifications. $T = 20$ uses six radii $[0.88, 0.94, 0.97, 1.03, 1.06, 1.12]\, R_i^{(T)}$; $T = 1$ uses four radii $[0.88, 0.94, 1.06, 1.12]\, R_i^{(T)}$. Reference labels at threshold 0.5: DDDPPP for $T = 20$, DDPP for $T = 1$.

horizon	seed	labels	rel. L^2	max $\lvert\Delta\langle\phi\rangle\rvert$	outcome
$T = 20$	123	DDDPPP	0.0029	$\sim 10^{-3}$	match (BE $< 3\%$)
$T = 20$	124	DDDPPP	0.0042	$\sim 10^{-3}$	match (BE $< 3\%$)
$T = 20$	125	DDDPPP	0.0147	$\sim 10^{-3}$	match (BE $< 3\%$)
$T = 1$	123	PPPP	0.0115	0.00129	over-persist
$T = 1$	124	PPPP	0.0229	0.00432	over-persist
$T = 1$	125	DDDD	0.0290	0.00128	over-dissolve

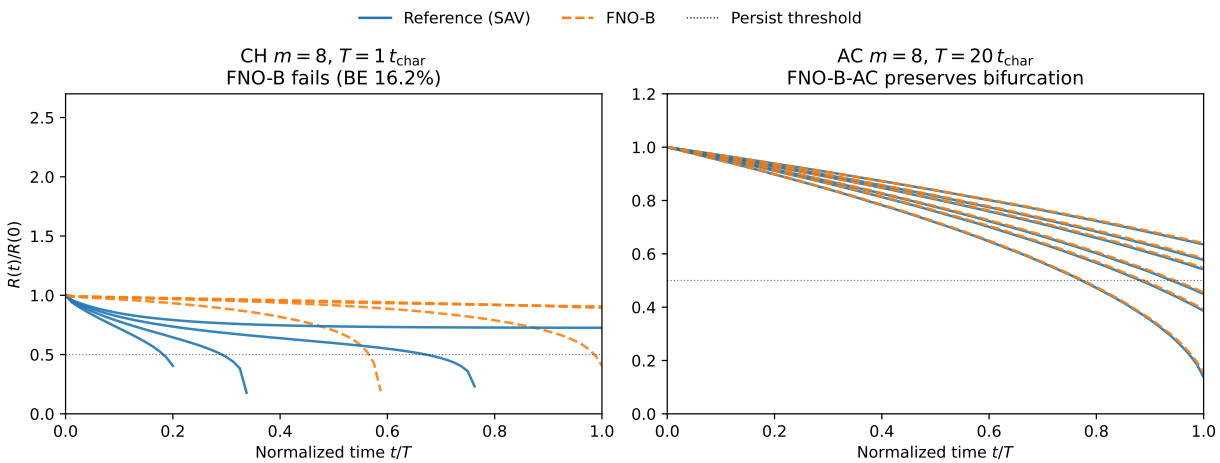

Figure 5. Per-trajectory illustration of the CH-failure regime and the AC-at-convergence regime. *Left*: CH $m = 8$ at $T = 1\, t_{\text{char}}$. FNO-B classifies three subcritical droplets as persistent through delayed dissolution ($R_0 = 0.113$) and contour re-emergence ($R_0 = 0.100, 0.107$), shifting the boundary by 16.2%. *Right*: Allen-Cahn $m = 8$ at $T = 20\, t_{\text{char}}$, $R_i^{(T)}/\epsilon_m \approx 7.3$. FNO-B-AC tracks the reference across all six held-out near-boundary radii. The matched-protocol mechanism scatter (Fig. 2, main text) extends this comparison to a controlled epoch sweep across AC, CAC, and CH.

E. Post-hoc zero-mode correction

To test whether mass conservation alone explains the observed failure, we applied a post-hoc correction after each FNO-B step:

$$\phi^{\text{corr}}_{t+\Delta t} = \phi_{t+\Delta t} - \langle \phi_{t+\Delta t} \rangle + \langle \phi_0 \rangle.$$

For the mild near-critical case $R_0 = 0.113308$, raw FNO-B drifts in the spatial mean from -0.917 to -0.855 and incorrectly remains just above the persist threshold, $R(T)/R(0) = 0.559$. The corrected rollout keeps the mean fixed at -0.917 and dissolves by $t/T = 0.5$, matching the qualitative SAV outcome. This is consistent with known sensitivity of phase-field droplet dynamics to mass-conservation errors (Yue et al., 2007). However, applying the same correction to the full $m = 8$ held-out sweep makes every trajectory dissolve, including the reference-persisting case $R_0 = 0.119793$. Thus zero-mode drift is a causal component of the failure, but zero-mode conservation alone is not sufficient for bifurcation preservation; the nonzero-mode/interface dynamics near the critical radius must also be learned correctly.

Table 11. Effect of post-hoc zero-mode correction on $m = 8$ held-out classification at threshold 0.5. The correction removes raw FNO-B false persistence for subcritical droplets, but also incorrectly dissolves the reference-persisting case.

R_0	SAV	FNO-B raw	FNO-B corrected
0.100339	dissolve	persist	dissolve
0.106824	dissolve	persist	dissolve
0.113308	dissolve	persist	dissolve
0.119793	persist	persist	dissolve

F. Trained conservation and architecture baselines

To test whether the identified failure can be removed during learning, we trained a projected FNO baseline, FNO-B+P, that enforces conservation of the zero Fourier/Laplacian mode in both training and autoregressive rollout. The projection replaces the raw prediction by

$$\phi^{P}_{t+\Delta t} = \phi^{\mathrm{raw}}_{t+\Delta t} - \langle \phi^{\mathrm{raw}}_{t+\Delta t} \rangle + \langle \phi_t \rangle.$$

This model preserves the spatial mean to numerical precision (mean drift $\approx 10^{-7}$), but it fails to produce a valid dissolve/persist bracket for the held-out $m = 8$ boundary in all three seeds. In all three seeds, FNO-B+P classified every held-out $m = 8$ radius as dissolving, including the reference-persisting case $R_0 = 0.119793$. Thus, the projection removes the false-persistence failure caused by zero-mode drift, but over-corrects the dynamics and loses the persisting branch. Enforcing mass conservation during training and rollout is therefore not sufficient for bifurcation preservation.

We also trained a U-Net baseline on the same FNO-B training dataset. Unlike FNO-B+P, the U-Net does not enforce mass conservation and exhibits non-negligible mean drift. However, it produces a valid $m = 8$ boundary bracket in two of three seeds: seed 124 exactly matches the reference boundary, seed 126 shifts the boundary by 20.6%, and seed 125 produces no valid bracket because it classifies every held-out $m = 8$ radius as persisting. Across U-Net seeds, the exact-boundary seed also has the lowest held-out rollout error and lowest mean drift (relative $L^2 = 0.087$, mean drift 0.043), while the all-persist seed has the worst rollout error and largest mean drift (relative $L^2 = 0.378$, mean drift 0.105). However, these aggregate metrics do not fully determine the diagnostic: seed 126 forms a valid bracket but still shifts the boundary by 20.6%. Thus, rollout error and mean drift are informative but insufficient summaries of bifurcation behavior.

Table 12. Trained conservation and architecture baselines on the held-out $m = 8$ boundary at threshold 0.5. FNO-B+P enforces zero-mode conservation during training and rollout.

Model	valid / seeds	failure mode	BE (bracket)	mean drift
FNO-B	3 / 3	over-persist	$22.0\% \pm 5.0\%$	not conserved
FNO-B+P	0 / 3	all dissolve	–	$\sim 10^{-7}$
U-Net	2 / 3	seed-dependent	$10.3\% \pm 14.6\%$	0.070 ± 0.032

G. Training-horizon per-seed detail

Per-seed rollout reduction at 100 epochs: seed 124 dropped from rel. $L^2 = 0.524$ to 0.097, seed 125 from 0.687 to 0.332, and seed 126 from 0.356 to 0.137. Bifurcation behavior did not improve uniformly. Seed 124 – the lowest held-out rel. L^2 of any FNO-B model in this paper – produced no valid held-out $m = 8$ bracket, classifying every held-out near-boundary droplet as dissolving. Seed 125 retained a 20.6% boundary error. Seed 126 matched the reference boundary at threshold 0.5 but the bracket fails at stricter thresholds, indicating bracket fragility rather than robust recovery.

H. Mechanism-control protocol details

Conservative AC reference solver. The CAC equation (4) is integrated with the same semi-implicit Fourier scheme as AC, with an additional projection step that subtracts the spatial mean of the nonlinear term at every step. Mass is conserved to numerical precision ($|\langle\phi\rangle(t) - \langle\phi\rangle(0)| < 10^{-12}$) for every reference trajectory. We use $\Delta t = 10^{-4}$ at $T = 20\, t_{\mathrm{char}}$, matching the AC $T = 20$ protocol.

CAC reference boundary. A horizon screen at $T \in \{4, 9, 20\}\, t_{\mathrm{char}}$ selects $T = 20$ as the resolved regime, with reference boundary $R_i^{(T)} = 0.0769$ and $\delta R_i / R_i = 1.19\%$. Reference labels at threshold 0.5 across the six held-out radii are DDDPPP, identical to AC at $T = 20$.

Epoch-checkpoint sweep. For AC and CAC we train one 100-epoch FNO-B run per seed, saving intermediate checkpoints at epochs $\{1, 3, 5, 10, 20, 30, 50, 100\}$, yielding 24 checkpoints per system. Each checkpoint is evaluated on the six-radius held-out set at threshold 0.5 and classified as preserved (valid bracket, $\mathrm{BE} < 5\%$), shifted (valid bracket, $\mathrm{BE} \geq 5\%$), or no bracket (all six radii on the same side).

Failure-mode distribution within the CH rollout-error band. Inside the CH band (rel. $L^2 \in [0.097, 0.687]$), AC failures concentrate at early checkpoints (epochs 1, 3, plus seed-123 epoch 20) and are uniformly all-persist (PPPPPP) relative to a DDDPPP reference. CAC failures distribute more broadly across the rollout-error range, including failures at intermediate epochs (10, 20, 30) and shifted-but- bracketed failures at low rollout error (rel. $L^2 \approx 0.03$-0.05) where AC is uniformly preserved. This contrast indicates that the conservation-induced failure mode appears at lower aggregate error than the under-training-induced failure mode, supporting the interpretation that conservation contributes a distinct failure mechanism rather than simply degrading rollout fidelity.