XENO: INTERFACE-ENRICHED NEURAL OPERATORS FOR STABLE ROLLOUTS ON SHARP MOVING INTERFACES

Anonymous authors
Paper under double-blind review

ABSTRACT

One-step accuracy does not predict autoregressive rollout stability for neural operators on sharp moving interfaces. We study this failure in Stefan phase-change problems and find that FNO and Transolver, despite one-step interface errors below 0.2 px, reach 20-step errors of 32.1 and 22.4 px, respectively, in the sharp R4 regime, while remaining stable in the smooth R3 regime. We introduce XENO, which augments a spectral neural operator with an Operator Enrichment Block (OEB): interface-conditioned basis functions modulated by a learned signed-distance-like field and gated by ReZero. On R4, XENO reduces the median 20-step Hausdorff error from 32.1 px for FNO to 1.09 px under held-out validation-rollout checkpoint selection. Matched ablations show that this improvement is not explained by rollout-noise training, parameter count, gating, phase access, or alternative geometric conditioning. The improvement is regime-specific: XENO is worse than the strongest baseline on R3 and Burgers, and 36.9% of R4 trajectories still exceed 5 px. We also find that spectral-energy collapse accompanies rollout failure, but do not establish it as a cause.

1 INTRODUCTION

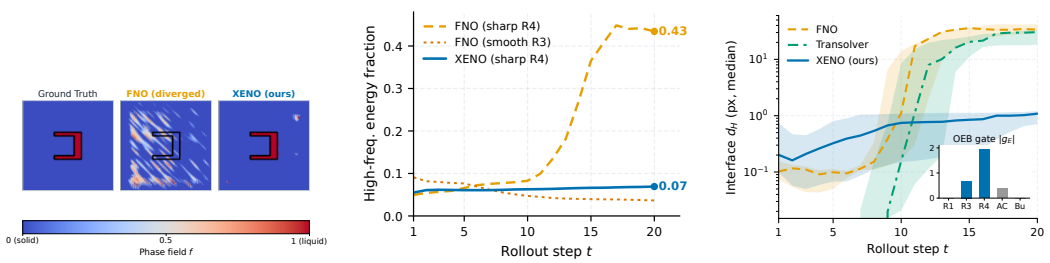
Neural operators learn solution operators for parametric PDEs, achieving orders-of-magnitude speedups over classical solvers at inference time (Li et al., 2020b; Wu et al., 2024). This success, however, rests on an implicit smoothness assumption: the learned basis functions (Fourier modes, attention keys) are globally smooth, and the underlying solution is expected to lie in a compatible function space. In the Stefan phase-change problem (Stefan, 1891), the liquid-fraction field f is discontinuous across a moving interface $\Gamma(t)$, and the temperature field has a kink at $\Gamma(t)$ due to latent heat release. A neural operator deployed as a simulator is applied autoregressively, feeding its own output back as the next input.

We find that autoregressive rollout stability of spectral operators is regime-dependent: the operators remain stable on smooth interfaces but diverge on the sharp R4 regime. We compare two 2D Stefan regimes (Section 4): a convection-dominated regime with smooth interfaces (R3), and a heterogeneous-conductivity regime whose interfaces are geometrically complex with steep temperature kinks (R4; the interface-region temperature gradient is $15\times$ larger than in R3, and the discrete interface has $2\times$ more boundary cells). On the smooth regime, FNO and Transolver are stable over a 20-step rollout (median interface error $\sim$1–2 px). On the sharp regime, the *same* models, despite low one-step error (interface error < 0.2 px), diverge: each step injects error at the sharp kink that compounds, and by 20 steps the predicted interface is at $\sim$30 px, worse than the persistence baseline (Figure 1C). The failure is thus not an artifact of weak baselines (they are stable elsewhere and accurate one step) but a property of how smooth bases represent sharp, evolving geometry. Figure 1 summarizes the rollout failure, its measured spectral signature, and the effect of OEB.

We report this spectral-energy collapse (Fig. 1B) as a measured *correlate* of the divergence, not a proven cause: a controlled single-variable sharpness sweep does not reproduce a "sharper-is-harder" trend, so we read it only through the cross-regime R3-vs-R4 comparison (full analysis in Appendix A).

These observations motivate an interface-localized correction rather than a domain-wide enrichment. We introduce an *Operator Enrichment Block* (OEB), added to each spectral layer, which learns a set

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.



(a) (b) (c)

Figure 1: **Rollout divergence, spectral-energy collapse, and XENO stabilization** (R4, seed 42). (A) FNO, accurate one step ahead, diverges by rollout step 20; XENO tracks the true interface (black). (B) Spectral-energy collapse accompanies divergence: FNO's low-frequency spectral energy collapses under rollout ($\approx 27\%$ of truth by step 20) while XENO holds it near truth (§A). (C) Median rollout interface error (log scale; bands = seed min–max): XENO $\sim$1 px vs. $\sim$30 px for baselines. Inset: the OEB gate is large on R4 and small where enrichment is not beneficial.

of interface-anchored basis functions. The OEB takes as input a predicted signed-distance-like field φ and produces a localized additive correction u_{enriched} which is fused into the spectral output via a learnable scalar gate:

$$u_{\text{out}} = \text{GeLU}(\text{LN}(u_{\text{spectral}} + g_E \cdot u_{\text{enriched}})).$$

XENO reduces to the standard spectral operator when $g_E = 0$, which holds at initialization; the gate grows during training only if the gradient signal requires it, providing training stability. Matched ablations isolate the contribution of the enrichment from the training protocol, parameter count, gating, phase access, and alternative geometric conditioning.

The method has two components: (a) **interface localization**: φ is predicted by a lightweight InterfaceFieldNet supervised by eikonal and isotherm losses, so no ground-truth signed distance is required at test time; (b) **training stability**: the ReZero gate (Bachlechner et al., 2021) makes the enrichment a no-op at initialization that grows only with gradient support, avoiding instability when the M enrichment atoms are activated at Stage C.

Contributions. We establish three results. First, one-step accuracy does not predict autoregressive rollout stability: spectral, wavelet, attention, and branch–trunk operators can be accurate one step ahead yet diverge under rollout. Second, this failure is concentrated in the localized, heterogeneity-induced sharp-interface regime R4, while the smooth-interface regime remains stable. Third, an interface-localized enrichment improves rollout stability in R4: the proposed Operator Enrichment Block reduces the median 20-step interface error from approximately 30 px to 1 px. Matched ablations separate this improvement from the training protocol and other architectural ingredients, while experiments on R3, Allen–Cahn, and Burgers define the operating regime of the method. We also measure spectral-energy collapse during rollout and treat it as a correlate of failure rather than a causal explanation. OEB combines a learned interface field with a local basis correction inside an autoregressive spectral operator; unlike XFEM, the enrichment is learned and shared across problem instances. The conservation projection (CCP) is analyzed in Appendix B and is not claimed as a contribution.

2 Background

Neural operators. Neural operators learn mappings between function spaces (Kovachki et al., 2023). The Fourier Neural Operator (FNO) (Li et al., 2020b) applies a spectral convolution in Fourier space, retaining the lowest K modes. Transolver (Wu et al., 2024) replaces the spectral projection with a physics-informed attention mechanism but retains global smoothness. Both achieve state-of-the-art accuracy on smooth PDEs (Navier-Stokes, Darcy flow) but, as we show, are inadequate for moving-interface problems.

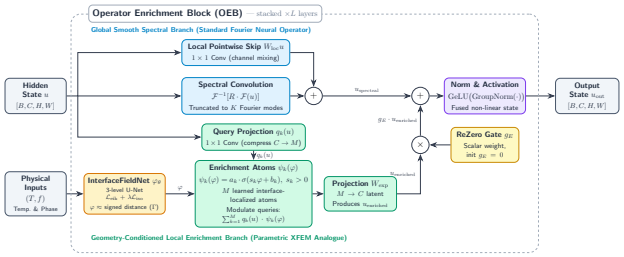


Figure 2: **One Operator Enrichment Block (OEB).** The spectral (FNO) branch runs in *parallel* with the enrichment branch: (T, f) drive INTERFACEFIELDNET φ_θ, and $E_\theta(u, \varphi)$ forms M interface-localized atoms modulated by query maps $q_k(u)$, fused via the ReZero gate g_E (Eq. 3). At $g_E{=}0$ the block is exactly a standard FNO block; the full model stacks $\times L$ blocks.

The Stefan problem. The one-phase Stefan problem models a solid-liquid phase transition (Stefan, 1891):

$$\rho c_p \partial_t T = \nabla \cdot (\lambda \nabla T) \qquad \text{in } \Omega \setminus \Gamma(t), \tag{1}$$

$$\rho \mathcal{L}_{\text{lat}} V_n = [\![\lambda \nabla T]\!] \cdot \hat{n} \qquad \text{on } \Gamma(t), \tag{2}$$

where T is temperature, $\Gamma(t)$ is the moving interface, V_n its normal velocity, $\mathcal{L}_{\text{lat}}$ latent heat, and $[\![\cdot]\!]$ the jump operator. We use the enthalpy formulation (Voller & Prakash, 1987): the conserved variable is $H = \rho c_p T + \rho \mathcal{L}_{\text{lat}} f$ where $f \in \{0, 1\}$ is the liquid fraction. Interface Hausdorff distance $d_H(\hat{\Gamma}, \Gamma)$ is the primary metric.

Extended FEM (XFEM). Classical XFEM (Moës et al., 1999) enriches the FEM basis with interface-aware functions, e.g., Heaviside or abs-value enrichment centered on a level-set field. This enables sharp interface representation without remeshing, but requires solving a new FEM system with new enrichment degrees of freedom for each problem instance. Our OEB is the operator-learning analogue: it learns a *parametric* enrichment that applies across the entire distribution of ICs/BCs, amortizing the enrichment cost.

3 METHOD

We learn a one-step operator $G_\theta : (T_t, H_t, f_t) \mapsto (T_{t+1}, H_{t+1}, f_{t+1})$ from data on a regular spatial grid; at rollout time G_θ is applied autoregressively, evaluated on the Stefan benchmark (Section 4).

3.1 OPERATOR ENRICHMENT BLOCK (OEB)

Each XENO block is a standard FNO layer with spectral convolution and the usual local linear (pointwise) branch, as in Li et al. (2020b), augmented by an enrichment branch fused in parallel:

$$u_{\text{out}} = \text{GeLU}\left(\text{LN}\left(\underbrace{\mathcal{F}^{-1}[R \cdot \mathcal{F}[u]] + W_{\text{loc}}\, u}_{\text{spectral (FNO)}} + g_E \cdot \underbrace{E_\theta(u, \varphi)}_{\text{enrichment (OEB)}} \right) \right), \tag{3}$$

where $\mathcal{F}$ is the Fourier transform, $R \in \mathbb{C}^{K \times d_v \times d_v}$ are learnable per-mode spectral weights, W_{loc} is a 1×1 convolution (the local/skip branch present in every standard FNO layer, applied to u in parallel with the spectral term, not to its output), $g_E \in \mathbb{R}$ is a scalar ReZero gate initialized to zero, and E_θ is the enrichment branch. Setting the enrichment atom count $M{=}0$ (so $E_\theta \equiv 0$; M is distinct from the retained Fourier mode count K above) recovers a standard FNO block exactly, which is the No-OEB ablation (Section 4.3). Figure 2 diagrams one block.

Enrichment branch E_θ. The branch takes the hidden state u and the interface field φ. It forms M *learned* interface-localized basis functions (atoms)

$$\psi_k(\varphi) = a_k\, \sigma(s_k\, \varphi + b_k)\,, \qquad s_k = \text{softplus}(\tilde{s}_k) > 0, \qquad k = 1, \ldots, M, \tag{4}$$

so each ψ_k is a smooth *step* in φ whose transition is centred near the interface $\varphi = 0$, with learnable amplitude a_k, slope s_k, and offset b_k. A 1×1 convolution compresses u into M spatial query maps $q_k(u)$, which are modulated by the atoms and projected back to the hidden width by a 1×1 convolution W_{exp}:

$$E_\theta(u, \varphi) = W_{\text{exp}}[\, q_k(u)\, \psi_k(\varphi)\,]_{k=1}^{M}, \qquad [\,\cdot\,]_{k=1}^{M} := \text{stack along the channel dim } (M \text{ channels}), \tag{5}$$

A single sigmoid atom is not itself localized; the atoms enter as a *signed* learned combination (amplitudes a_k), forming bump-like, interface-concentrated responses, while the query maps $q_k(u)$ further gate where the correction is applied (Section A verifies the net correction activates at the interface, $2.3\times$ stronger within 3 px of the front). Because E_θ is not zero-initialized but the gate g_E is, the gate receives a nonzero gradient from step one while the enrichment atoms begin training only once it moves off zero.

ReZero gate. The scalar g_E is a learnable parameter initialized to 0 (one per block), implementing the ReZero scheme (Bachlechner et al., 2021): at initialization $u_{\text{out}} = \text{GeLU}(\text{LN}(u_{\text{spectral}}))$, so XENO starts as a pure spectral operator, and g_E grows selectively during staged training (Section 3.3) in blocks that benefit from enrichment. If all $g_E = 0$, XENO is *exactly* the No-OEB ablation baseline (Section 4.3), making the contribution directly measurable.

3.2 Interface Field φ

The interface field φ is predicted by a lightweight convolutional network InterfaceFieldNet taking (T_t, f_t) as input: a 3-level U-Net with base channel width C=32 (encoder $2 \to C \to 2C \to 4C$, decoder $4C \to 2C \to C$ with skip connections, final 1×1 convolution to 1 output channel, no output activation). We supervise φ with two losses:

$$\mathcal{L}_{\text{eik}} = \mathbb{E}_x \left[\left(|\nabla \varphi(x)| - 1 \right)^2 \right], \tag{6}$$

$$\mathcal{L}_{\text{iso}} = \mathbb{E}_x [\text{BCE}(\sigma(\varphi(x)),\ \mathbf{1}\{f_t(x) > 0.5\})]\,. \tag{7}$$

$\nabla\varphi$ in Eq. 6 is a forward finite difference on the discrete grid (one-sided at boundaries); φ is predicted on the same grid as the input fields, so no separate interpolation is needed. $\mathcal{L}_{\text{eik}}$ encourages φ to approximate a signed distance (unit-gradient) field. $\mathcal{L}_{\text{iso}}$ treats φ as a logit and fixes its *sign* convention by weak supervision against the physical phase label at the *same* timestep the network conditions on: $\varphi(x) > 0$ where the liquid fraction $f_t(x) > 0.5$ (liquid side) and $\varphi(x) < 0$ otherwise (solid side), which jointly with $\mathcal{L}_{\text{eik}}$ drives $\varphi \to 0$ near the f_t=0.5 isotherm. This directly fixes the sign of φ from data rather than leaving it arbitrary. φ is *not* claimed as a contribution; it is a mechanism that localizes OEB.

3.3 Staged Training

Activating enrichment ($M > 0$) before φ is well-trained causes instability: the enrichment branch receives noisy geometric signals and injects random perturbations. We address this with four training stages over $N_{\text{ep}} = 200$ epochs:

Stage	Epochs	Enrichment	λ_{iso}	λ_{eik}	λ_{iface}	Notes
A	0–10%	$M = 0$ (off)	0	0	0	Spectral warm-up
B	10–20%	$M = 0$	1.0	0	0	φ-net supervised
C	20–40%	$M = 8$ (on)	$1.0 \to 0.1$	0.1	0.5	OEB on; rollout noise
D	40–100%	$M = 8$	$0.1 \to 0.01$	0.1	0.5	early-stop window

λ_{eik} and λ_{iface} are constant once active (Stage C onward); only λ_{iso} is annealed within a stage. Rollout noise (std $= 0.02 \times \hat{\sigma}_{\text{channel}}$) is injected in stages C/D to expose the model to the error distribution it encounters at autoregressive rollout time; $\hat{\sigma}_{\text{channel}}$ is that channel's own empirical standard deviation, recomputed per training batch (not a fixed dataset-wide constant) over the corresponding input tensor (T_{in}, H_{in}, or f_{in}); only the inputs are perturbed, targets stay clean. Early stopping (patience 30 from Stage D entry) is supported but disabled for the main XENO models (patience set to 250, beyond the 200-epoch budget) because rollout stability is acquired late (App. O).

3.4 Training Loss

The total loss is:

$$\mathcal{L} = \underbrace{\|T_{\text{pred}} - T_{\text{true}}\|_2^2 + \|H_{\text{pred}} - H_{\text{true}}\|_2^2 + \|f_{\text{pred}} - f_{\text{true}}\|_2^2}_{\text{field, equal-weighted}} + \lambda_{\text{eik}}\mathcal{L}_{\text{eik}} + \lambda_{\text{iso}}\mathcal{L}_{\text{iso}} + \lambda_{\text{iface}}\mathcal{L}_{\text{iface}}, \tag{8}$$

where H is pre-scaled by a per-regime constant (the 99th percentile of $|H|$ over the training split, or a fixed dataset value) before this loss and throughout training/evaluation, bringing its otherwise much larger natural magnitude (enthalpy $\propto \rho\mathcal{L}_{\text{lat}}$) down to the same order as T and $f \in [0, 1]$; T and f are used in their natural physical/bounded units without further standardization. The three field terms are summed with equal weight (no learned or manually-tuned per-channel weighting). $\mathcal{L}_{\text{iface}}$ is a soft interface-alignment term,

$$\mathcal{L}_{\text{iface}} = \|\sigma(10(f_{\text{pred}} - 0.5)) - \sigma(10(f_{\text{true}} - 0.5))\|_2^2, \tag{9}$$

the squared distance between sharpened-sigmoid soft interface masks, penalizing predicted-interface location directly on the phase field f (unlike $\mathcal{L}_{\text{iso}}$, which supervises φ). Coefficients λ follow the staged schedule above.

4 Experiments

4.1 Stefan Benchmark and Regimes

We use three Stefan regimes spanning a range of interface sharpness: **R1** (1D, 256 points, uniform λ; an easy control where the interface barely moves), **R3** (2D, 128^2, convection-dominated, *smooth* interfaces), and **R4** (2D, 128^2, *heterogeneous* λ; geometrically complex, *sharp* interfaces). The distinction is quantitative: the R4 temperature field carries sharp interior features: the 99th-percentile interior $|\nabla T|$ (excluding a boundary ring so the Dirichlet walls do not dominate) is 2.6 vs. 0.17 on R3 ($\sim 15\times$), whereas R3's convective field is smooth. R4 is the sharp-interface regime our analysis predicts will challenge smooth bases. Metrics: interface Hausdorff distance d_H in pixels (primary); we report the *median* over the 200-trajectory test set because d_H is a max-based statistic with heavy tails, and additionally report failure rate (fraction with $d_H > 5$ px) and a **persistence** baseline (predict f unchanged from the initial condition) as a reference baseline. The 5 px threshold is $\sim 3.9\%$ of the 128-px domain width (App. S): well above the < 0.2 px one-step error every model achieves, but small enough that a trajectory crossing it has visibly lost the interface, not merely accumulated numerical noise.

Baselines: **FNO** (Li et al., 2020b), **Transolver** (Wu et al., 2024), **PINO** (Li et al., 2024), all trained on the same data with one-step supervision. PINO diverges on Stefan ($d_H > 20$) and is deferred to Appendix G. All reported numbers come from a single frozen-checkpoint rollout harness on held-out test trajectories, rather than training-time validation metrics, which we find are systematically unreliable proxies for rollout stability (protocol note in App. O; hyperparameters in App. H; data generation in App. T).

4.2 Failure Case Analysis

On a representative moving R4 trajectory, FNO and the No-OEB ablation progressively smear and detach from the front while XENO tracks it (Fig. 5, App. I).

Performance across regimes (Table 1, Fig. 6). The instability is *regime-specific*, and so is the benefit of the fix: existing operators are already stable on smooth interfaces (R3) and on a 1D shock (Burgers), collapse on the localized-sharp regime (R4, where XENO delivers its $\sim 30\times$ improvement), and are best- but-modest on the distributed-sharp Allen–Cahn regime (App. M). **XENO is not uniformly beneficial outside the target regime**: it is worse than Transolver on R3 and worse than FNO on Burgers even where the learned gate is near zero (Table 1; full discussion in Section 6).

R4 is bimodal, not uniformly stable (Table 5, App. I). Splitting the test set by true interface motion shows the aggregate median hides two different outcomes: on near-static trajectories the baselines

Table 1: Median 20-step rollout interface d_H (px) by regime, with the learned OEB gate magnitude (discussion below). Bold = best; (†) XENO-R3 uses the reduced-data schedule. All entries are seed 42 at its held-out validation-rollout selected checkpoint (only R4 has a multi-seed study; see Table 2 for FNO/Transolver/XENO cross-seed medians on R4). Complete per-regime results incl. failure rates: App. I.

Regime	persistence	FNO	Transolver	XENO	OEB gate $\overline{\lvert g_E \rvert}$
R3 (2D, smooth interface)	26.1	1.8	1.0	1.6†	0.68
R4 (2D, localized sharp interface)	8.4	30.7	25.7	**1.0**	1.94
Allen–Cahn (2D, distributed sharp)	10.7	8.6	9.6	**7.3**	0.38
Burgers (1D, shock)	23.0	**0.8**	9.2	3.0	0.00

corrupt an interface that should not move at all (FNO 40.4 px) while XENO leaves it essentially in place (0.42 px); on moving trajectories XENO genuinely tracks, beating persistence on 82% of cases versus 52–60% for baselines (i.e. the baselines are no better than predicting no motion). XENO still exceeds 5 px on 36.9% of R4 trajectories overall (95% CI $[32.9, 41.0]\%$, 8-seed mean; App. I), concentrated on the *fastest-moving* fronts, where small per-step tracking errors accumulate over a long displacement even though OEB still corrects the underlying spectral smoothing (representative failure case, $d_H^{20} = 21$ px on a front advancing 29.6 px over 20 steps: Fig. 7, App. I).

Quantitative comparison. Table 2 reports the median 20-step interface Hausdorff distance across all test trajectories and seeds (per-seed values in App. P; complementary metrics (IoU, surface distance, field L_1/L_2, energy drift) in App. S).

Table 2: Stefan R4 benchmark, interface Hausdorff distance (px, 128×128 grid), median over 200 test trajectories. The key distinction is between 1-step accuracy and 20-step rollout stability: all baselines except DeepONet have low 1-step error, yet all diverge under rollout; XENO is slightly behind 1-step but far more rollout-robust at 20 steps (bimodality: §4.2). Median reported because d_H is heavy-tailed. XENO/FNO/Transolver 20-step values are cross-seed medians under held-out-rollout checkpoint selection (3+ seeds); other baselines are single-seed; fail rates and full checkpoint-selection protocol in App. P–I. Necessity ablations in Table 3. † PINO diverges (App. G).

Model	class	1-step d_H	20-step $d_H \downarrow$	fail (>5px)	$\times$XENO
FNO (Li et al., 2020b)	global spectral	0.11	32.1	1.00	29$\times$
WNO (Tripura & Chakraborty, 2023)	wavelet	**0.02**	35.1	—	32$\times$
Transolver (Wu et al., 2024)	attention	< 0.01	22.4	0.94	21$\times$
DeepONet (Lu et al., 2021)	branch–trunk	11.0	23.6	—	22$\times$
CNO (Raonic et al., 2023)	alias-aware conv	0.08	21.5	—	20$\times$
F-FNO (Tran et al., 2021)	factorized spectral	0.08	13.5	—	12$\times$
U-Net (Gupta & Brandstetter, 2022)	local conv	0.14	9.4	—	9$\times$
PINO (Li et al., 2024)	physics-informed	—	$\gg$1†	—	—
XENO (ours)	spectral + OEB	0.15–0.57	**1.09**	**0.37**	1$\times$

We report the 1-step row alongside rollout because XENO's advantage is not lower per-step error (Transolver's and WNO's 1-step interface errors are essentially zero, as good as or better than XENO's) but that its per-step errors do not *compound*. The comparison set spans global-spectral (FNO), alias-aware (CNO), factorized-spectral (F-FNO), physics-informed (PINO), attention (Transolver), branch–trunk (DeepONet), a space-frequency-localized wavelet basis (WNO), and a local, hierarchical convolutional architecture (U-Net), and **every one of them diverges under rollout**. Ordering the baselines by how localized their representation is traces a weak, non-monotonic trend (App. J.1): the wavelet WNO, designed for space-frequency localization, still fails worst (35.1 px) alongside FNO (32.1); a convolutional architecture (U-Net) is best among baselines (9.4) yet still $\sim 9\times$ worse than XENO (1.09). Generic locality is therefore insufficient; the improvement comes from placing the enrichment at the predicted interface. This is also consistent with the also-convolutional CNO (21.5), designed against spectral aliasing, still diverging. A plain FNO with more Fourier modes ($K \in \{4, 8, 16, 32\}$; App. D) also remains unstable (20–60 px). A mechanism-level account of why each baseline behaves as it does is given in App. J.

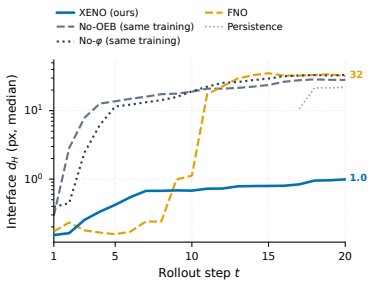


Figure 3: **OEB, not noise training, causes rollout stability.** Median d_H vs. rollout step, R4. **No-OEB** gets XENO's identical staged + rollout-noise training with enrichment disabled; it diverges on top of FNO (step 20: $\sim$31 px, Table 3), while full XENO stays at $\sim$1 px.

4.3 Ablations

This experiment isolates our contribution from the training procedure: XENO trains with input-noise injection (Section 3.3) while FNO/Transolver do not, so the stability gain could in principle come from *training*, not architecture. We control for this directly with **No-OEB**, in which XENO uses the *identical* staged schedule and rollout-noise injection with the enrichment disabled ($g_E \equiv 0$, $M = 0$), i.e. an FNO block trained exactly like XENO (rollout curves in Fig. 3; per-seed values in App. P).

Table 3: Necessity ablation on R4, all variants *trained identically*. Removing OEB recovers the FNO failure; removing φ is comparably damaging (App. K); neither component alone suffices.

Variant	1-step d_H	20-step d_H (median) $\downarrow$
XENO (full: OEB + φ)*	**0.15–0.57**	1.09 **(mean** 1.03 ± 0.17**)**
− OEB ($g_E \equiv 0$, $M{=}0$)	0.27–0.29	31.0 (mean 31.2 ± 2.5)
− φ (enrichment without interface field)	0.34–0.47	32.2 (mean 31.4 ± 1.8)

*Full XENO is an 8-seed median; ablation variants are 3 seeds each (App. P); both gaps far exceed seed-to-seed variation in either group.

All three variants are trained identically (three seeds each; per-seed values in App. P). Removing OEB reproduces the FNO failure profile despite XENO's full rollout-stability training, so the enrichment, not the noise injection, is the discriminating component: **OEB produces a $\sim$29$\times$ median improvement under the matched ablation protocol**. Removing φ instead is comparably damaging: the enrichment atoms are inert on a degenerate geometry field with no interface to localize on. The two components are therefore mutually necessary (App. K); removing *either* collapses the model back to the $\sim$33 px baseline failure (CCP ablation: App. B).

Rollout-stabilization baseline: pushforward-trained FNO (full protocol: App. R). The No-OEB ablation controls for our own training protocol but not for established rollout-stabilization *training* methods (§5). A standard FNO trained with a pushforward loss (Brandstetter et al., 2022) (3 seeds, same checkpoint selection) improves markedly over plain FNO ($32.1 \rightarrow 8.50/10.07/13.09$ px, mean 10.55) but does not close the gap to XENO ($\sim$8–13$\times$ remaining), indicating the gain is architectural rather than a training-stabilization trick. This improvement comes with a computational cost: XENO's per-step inference cost is $\sim 4\times$ a matched-depth plain FNO (App. F), mostly from evaluating the InterfaceFieldNet every step, while remaining orders of magnitude cheaper than the classical solver it targets.

Specificity of the OEB contribution. The No-OEB ablation shows that *some* enrichment is necessary; it does not show that the *specific* combination (a learned, atom-localized enrichment conditioned on a trained interface field) is what matters, as opposed to extra parameters, gating, raw phase access, or a geometric feature computed some other way. We isolate each ingredient with two families of R4 controls under the identical 20-step rollout harness (full protocol and numbers in App. L). **Five free controls**, applied post hoc to the main XENO model with no retraining (gating

off, spatially shifting or shuffling the learned φ, or substituting a deterministic signed-distance transform (SDT) for it), all collapse rollout stability to 35–65 px, the same failure range as No-OEB (full XENO: 1.09 px). **Five trained controls** (A1–A5), each retrained from scratch removing one ingredient of OEB while matching everything else, confirm this is not a training-protocol artifact: none recovers stability, collapsing to 10–34$\times$ worse than full XENO. This includes **A5**, which tests information asymmetry: since the SDT controls see only the phase field f while the learned φ_θ also sees temperature T, we gave a deterministic control the same (T, f) access; it narrows the gap to full XENO (A1 $27.8 \to 19.5$ px) but does not close it, so this asymmetry is not the primary explanation; the learned geometry accounts for most of the remaining difference (one control, A2, is inconclusive due to a train/inference mismatch; App. L). Together these ten controls indicate the learned, interface-conditioned localization itself, not any single ingredient in isolation, is the critical component of the enrichment.

Spectral behavior during rollout (full analysis, Appendix A). We measure the spectral behavior of the baseline during rollout. On the sharp regime it progressively fails to *sustain* the interface: by step 20 its predicted phase field retains only $\approx 27\%$ of the ground-truth spectral energy (median over seeds); the front therefore smears, whereas XENO's matches the truth ($\approx 100\%$). The resulting error is large ($\approx 40\times$ XENO's residual) and *low-frequency* dominated, i.e. a gross front displacement, not high-frequency ringing; more Fourier modes or a space-frequency-localized wavelet basis (WNO) do not recover (App. D). We keep this measured collapse as the mechanism claim and are more cautious about *why* sharpness triggers it: a controlled sharpness-only sweep does not support a monotonic sharper-is-harder account (App. A), so we confine the claim to the R3-vs-R4 comparison. The enrichment itself is measurably interface-anchored ($2.3\times$ stronger within 3 px of the front) and the gate engages only where beneficial (Fig. 6); robustness to feedback noise, longer rollouts, and locality are in App. E.

5 Related Work

Neural operators for smooth PDEs. FNO (Li et al., 2020b) established the spectral operator paradigm; geometry-aware variants (DAFNO (Liu et al., 2023), GINO (Li et al., 2023), GNO (Li et al., 2020a)) all retain globally smooth bases and are not designed for discontinuous solutions.

Discontinuity- and interface-aware operator learning. Closest to our setting are operators built for discontinuous or sharp solutions. Lanthaler et al. (2022) show a *nonlinear* reconstruction is necessary for operator learning of PDEs with discontinuities; interface-conditioned DeepONets (XI-DeepONet, φ-DeepONet) (Bi et al., 2025; Roy et al., 2026) feed the branch–trunk network an interface or level-set field; piecewise/cut reconstruction schemes (Dang et al., 2026) partition the domain to represent sharp transitions; IANO (Wang et al., 2026) similarly conditions on an interface field but targets single-query super-resolution on a static field, not autoregressive rollout, and does not predict its own field end-to-end. Long et al. (2026) solve free-boundary problems via a flow map/homeomorphism to a fixed conjugate domain, whereas OEB keeps the grid fixed and enriches the basis locally at the predicted interface. Cavallazzi et al. (2026)'s Walsh-Hadamard Neural Operator (WHNO) replaces the Fourier basis itself with piecewise-constant Walsh functions, representing coefficient discontinuities exactly without Gibbs ringing; this targets one-step accuracy on discontinuous-coefficient problems rather than rollout stability, and is a full basis replacement rather than a local enrichment on an existing spectral backbone. OEB differs from all of the above in three ways: it is a drop-in enrichment for a *spectral* backbone rather than a branch–trunk model, domain-deformation map, or basis replacement; it *predicts* its own interface field φ; and it targets *autoregressive rollout* stability, not single-query accuracy.

Rollout stabilization. Autoregressive PDE surrogates are often stabilized by training-time interventions, such as multi-step / pushforward training (Brandstetter et al., 2022) and denoising refinement (PDE-Refiner) (Lippe et al., 2023), or by architectural changes to how local and global information are mixed, as in the Dual-Scale Neural Operator (DSO) (Dong et al., 2026), which separates local-detail and global-trend processing into two branches to curb blurring and drift over long rollouts on smooth flows; this is a different mechanism from OEB's interface- *localized* enrichment and is evaluated on smooth flows rather than sharp moving interfaces. Training-time interventions are complementary to an architectural fix: our No-OEB ablation controls for our own input-noise protocol, and a

pushforward-trained FNO comparison (§4.3) shows the gain is architectural. Relative to PDE-Refiner specifically, its diffusion-style multi-step denoising repeats the forward pass several times per output step, while OEB adds a fixed per-step overhead (App. F) and preserves the $O(1)$ single-step inference cost characteristic of spectral operators.

Neural methods for Stefan and phase-change. PINNs have been applied to Stefan problems (Cai et al., 2021) but require per-instance optimization; PINO (Li et al., 2024) combines FNO with physics losses and we show it diverges on Stefan. None of the above incorporate interface-aware enrichment.

XFEM and enriched bases. Classical XFEM (Moës et al., 1999; Belytschko & Black, 1999) enriches FEM shape functions with Heaviside or distance-function terms; OEB is inspired by this but operates in the operator-learning regime, learning enrichment bases that generalize across initial and boundary conditions. We do not compare against XFEM-learning methods and claim no generalization advantage over them.

ReZero and gated initialization. ReZero (Bachlechner et al., 2021) and FixUp (Zhang et al., 2019) show zero-initialized residual branches improve training stability; we apply this to the OEB gate so XENO starts as a pure spectral operator and activates enrichment only with gradient support.

6 Limitations

Our evidence maps a specific operating regime rather than a universal advantage. OEB helps most on localized, heterogeneity-induced sharp interfaces (R4); its benefit is modest on distributed sharpness (Allen–Cahn) and it is unneeded, and worse than the best baseline, where the baseline is already stable (R3, 1.6 vs. 1.0 px; Burgers, 3.0 vs. 0.8 px). The ReZero gate is a no-op only at initialization, not a property of the trained model, so XENO is not uniformly beneficial outside its target regime, and it does not eliminate failure: it still exceeds 5 px on 36.9% of R4 trajectories (95% CI $[32.9, 41.0]\%$) and on $\sim 64\%$ of the harder Allen–Cahn trajectories (§4.2).

Two limitations are methodological. Rollout stability is recovered by selecting checkpoints on a held-out validation *rollout*, not one-step validation (App. O, P), which is a deployment-time criterion, not an architectural guarantee. Our mechanism evidence is a rollout-level spectral comparison plus the No-OEB ablation, without a layer-by-layer spectral-transfer measurement, so we cannot establish that spectral-energy loss *precedes* rather than accompanies divergence. Results are at $\leq 128^2$ resolution in 2D; higher resolution, 3D, and a PDE-Refiner (Lippe et al., 2023) comparison are future work.

OEB also has two costs beyond the amortized-vs-FEM comparison (App. F). Relative to a plain FNO of matched depth/modes, it adds a $\sim 4\times$ per-step inference overhead (8.5 vs. 2.0 ms at batch 1, 128×128 grid), mostly from the InterfaceFieldNet evaluated every autoregressive step, while remaining orders of magnitude cheaper than a classical solver; this is the relevant number against other *neural* baselines. And φ's isotherm loss (Eq. 7) requires an explicit phase indicator f at training time; our four regimes provide this natively, but a PDE without one (e.g. compressible shocks defined by a discontinuity in T/ρ/p rather than a phase label) would require manually constructing a pseudo-phase indicator before φ can be supervised.

7 Conclusion

We identify a failure mode of neural operators in autoregressive simulation of localized, heterogeneity-induced sharp interfaces. The proposed Operator Enrichment Block improves rollout stability in this regime by adding an interface-conditioned correction to a spectral operator. Matched ablations show that the improvement is not explained by the rollout-noise training protocol or by generic parameterization and conditioning changes. The method is regime-specific: it improves R4 substantially, provides smaller gains on distributed sharpness, and is worse than the best baseline on smooth R3 and Burgers. Spectral-energy collapse accompanies rollout failure, but our experiments do not establish it as its cause.

AI USE STATEMENT

We used large language models for copy-editing, literature suggestions, boilerplate code, and LaTeX assistance. Related-work suggestions were independently verified by the authors. We did not use generative AI to design experiments, generate results, or produce data. All reported results were generated by the authors using the released training and evaluation code. We reviewed all AI-assisted text and code and take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

All main results come from a single frozen-checkpoint evaluation harness described in Section 4.1; hyperparameters are listed in Appendix H and data generation (solver, resolution, IC/BC sampling, and the disjoint-trajectory train/val/test split) in Appendix T. The checkpoint-selection protocol that the main R4 result depends on is detailed in Appendix O and P, including per-seed values and the applied held-out validation-rollout rule. Code, checkpoints, and data-generation scripts will be made available upon publication to reproduce every table and figure end-to-end.

REFERENCES

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth. In *Uncertainty in artificial intelligence*, pp. 1352–1361. PMLR, 2021.

Ted Belytschko and Tom Black. Elastic crack growth in finite elements with minimal remeshing. *International journal for numerical methods in engineering*, 45(5):601–620, 1999.

Ran Bi, Jingrun Chen, and Weibing Deng. Xi-deeponet: An operator learning method for elliptic interface problems. *Journal of Computational Physics*, 538:114164, 2025.

Johannes Brandstetter, Daniel Worrall, and Max Welling. Message passing neural pde solvers. *arXiv preprint arXiv:2202.03376*, 2022.

Shengze Cai, Zhicheng Wang, Sifan Wang, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks for heat transfer problems. *Journal of Heat Transfer*, 143(6):060801, 2021.

Giorgio M Cavallazzi, Miguel Pérez Cuadrado, and Alfredo Pinelli. Walsh-hadamard neural operators for solving pdes with discontinuous coefficients. *Journal of Computational Physics*, pp. 115124, 2026.

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.

Huanshuo Dong, Hao Wu, Hong Wang, Qin-Yi Zhang, and Zhezheng Hao. Dso: Dual-scale neural operators for stable long-term fluid dynamics forecasting, 2026. URL https://arxiv.org/abs/2603.26800.

Jayesh K Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized pde modeling. *arXiv preprint arXiv:2209.15616*, 2022.

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.

Samuel Lanthaler, Roberto Molinaro, Patrik Hadorn, and Siddhartha Mishra. Nonlinear reconstruction for operator learning of pdes with discontinuities. *arXiv preprint arXiv:2210.01074*, 2022.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Graph kernel network for partial differential equations. *arXiv preprint arXiv:2003.03485*, 2020a.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020b.

Zongyi Li, Nikola Kovachki, Chris Choy, Boyi Li, Jean Kossaifi, Shourya Otta, Mohammad Amin Nabian, Maximilian Stadler, Christian Hundt, Kamyar Azizzadenesheli, et al. Geometry-informed neural operator for large-scale 3d pdes. *Advances in Neural Information Processing Systems*, 36: 35836–35854, 2023.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM/IMS Journal of Data Science*, 1(3):1–27, 2024.

Phillip Lippe, Bas Veeling, Paris Perdikaris, Richard Turner, and Johannes Brandstetter. Pde-refiner: Achieving accurate long rollouts with neural pde solvers. *Advances in Neural Information Processing Systems*, 36:67398–67433, 2023.

Ning Liu, Siavash Jafarzadeh, and Yue Yu. Domain agnostic fourier neural operators. *Advances in neural information processing systems*, 36:47438–47450, 2023.

Zongjia Long, Qi Zhou, Aiqing Zhu, Dong Dai, and Yiqun Liu. Deep neural operator for free boundary problems. *Nature Machine Intelligence*, pp. 1–12, 2026.

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(3):218–229, 2021.

Nicolas Moës, John Dolbow, and Ted Belytschko. A finite element method for crack growth without remeshing. *International journal for numerical methods in engineering*, 46(1):131–150, 1999.

Bogdan Raonic, Roberto Molinaro, Tim De Ryck, Tobias Rohner, Francesca Bartolucci, Rima Alaifari, Siddhartha Mishra, and Emmanuel De Bézenac. Convolutional neural operators for robust and accurate learning of pdes. *Advances in Neural Information Processing Systems*, 36: 77187–77200, 2023.

Sumanta Roy, Stephen T Castonguay, Pratanu Roy, and Michael D Shields. ϕ-*deeponet* : $Adiscontinuitycapturingneuraloperator.$ *arXiv e-prints*, $pp.\ arXiv - -2604, 2026.$

Johan Stefan. Über die theorie der eisbildung, insbesondere über die eisbildung im polarmeere. *Annalen der Physik*, 278(2):269–286, 1891.

Alasdair Tran, Alexander Mathews, Lexing Xie, and Cheng Soon Ong. Factorized fourier neural operators. *arXiv preprint arXiv:2111.13802*, 2021.

Tapas Tripura and Souvik Chakraborty. Wavelet neural operator for solving parametric partial differential equations in computational mechanics problems. *Computer Methods in Applied Mechanics and Engineering*, 404:115783, 2023.

Vaughan R Voller and Chander Prakash. A fixed grid numerical modelling methodology for convection-diffusion mushy region phase-change problems. *International journal of heat and mass transfer*, 30 (8):1709–1719, 1987.

Zhenzhong Wang, Xin Zhang, Jun Liao, and Min Jiang. Cross-field interface-aware neural operators for multiphase flow simulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26769–26777, 2026.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for pdes on general geometries. *arXiv preprint arXiv:2402.02366*, 2024.

Hongyi Zhang, Yann N Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. *arXiv preprint arXiv:1901.09321*, 2019.

A MECHANISM: THE SPECTRAL OPERATOR CANNOT SUSTAIN THE INTERFACE UNDER ROLLOUT

Full mechanism analysis referenced from §4.3 (the summary paragraph in the main text).

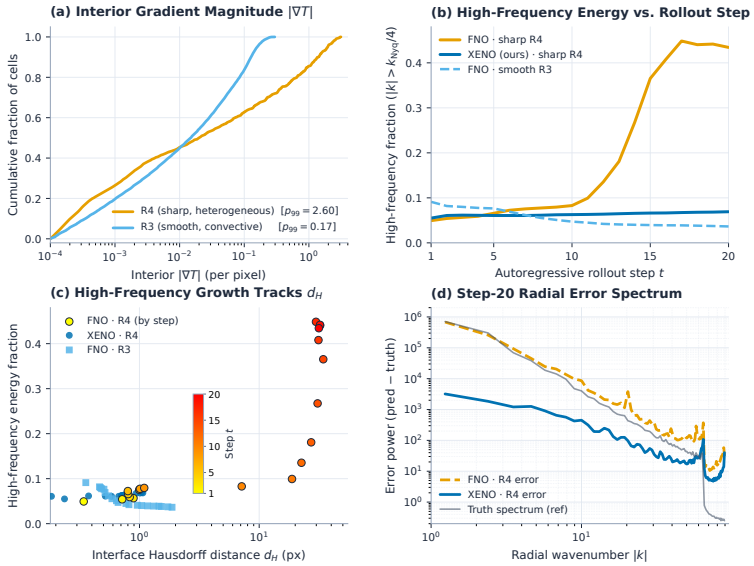


Figure 4: **Measured mechanism.** (A) R4's temperature field has sharp interior features ($\sim 15\times$ smooth R3). (B) FNO's high-freq. *fraction* rises under rollout on R4 ($0.05 \to 0.43$) but stays flat on R3 and for XENO, a symptom of low-freq. collapse (FNO retains $\approx 27\%$ of truth's spectral energy by step 20; XENO $\approx 100\%$). (C) tracks the interface divergence; (D) step-20 *error* spectrum: FNO's is large and low-wavenumber-concentrated ($\approx 40\times$ XENO's), not a high-frequency tail.

We measure the associated spectral behaviour directly (Fig. 4), separating the spectra of the prediction, the ground truth, and the *error* (prediction minus truth). On the sharp regime a spectral operator progressively fails to *sustain* the interface: by step 20 its predicted phase field carries only $\approx 27\%$ of the ground-truth spectral energy (median over seeds; 22–65%); the front therefore smears, whereas XENO's predicted energy matches the truth ($\approx 100\%$, every seed). The resulting error is large ($\approx 40\times$ XENO's residual) and *low-frequency* dominated: only $\sim 15\%$ of the FNO error energy lies above a quarter-Nyquist cutoff (vs. $\sim 54\%$ for XENO's much smaller residual), i.e. the error is a gross displacement of the front, not high-frequency ringing. The often-cited rise in the high-frequency *fraction* of the prediction ($0.05 \to 0.43$, in lock-step with divergence) is therefore a *symptom* of this collapse, not its cause: as the low-frequency interface content decays, the small residual energy is proportionally more high-wavenumber. A plain FNO with more modes (K up to 32) does not recover, and even a spatially-localized wavelet basis (WNO) still diverges, so the instability is not remedied by adding modes or by generic basis localization (Appendix D).

OEB acts where the sharpness is. The enrichment injects an interface-anchored, non-band-limited basis that represents the kink locally. Fig. 8 (Appendix N) shows the learned enrichment magnitude is $2.3\times$ stronger within 3 px of the interface than far from it (Pearson correlation with proximity 0.44), and Section 4 (Fig. 6) shows the ReZero gate engages only in the regime where this is beneficial.

Scope of the mechanism claim. We keep the measured collapse above as the paper's mechanism claim and are more cautious about *why* the sharp regime triggers it. A natural candidate, that a truncated global basis cannot represent a near-discontinuity with error compounding under re-projection, predicts rollout error should rise *monotonically* with interface sharpness. A controlled single-variable test rejects this: training XENO and FNO on five Allen–Cahn datasets at fixed interface half-widths $\epsilon \in \{0.008, 0.017, 0.028, 0.034, 0.045\}$ (smaller ϵ = sharper; IC/BC, domain, grid, and budget held fixed) gives the *opposite* trend: rollout error *rises* with ϵ for both models, so the sharpest interface is the easiest (ϵ=0.008: XENO 2.35, FNO 3.03 px; ϵ=0.045: 13.96, 14.22 px).

We flag a confound in interpreting even this controlled result: Allen–Cahn's bistable reaction term $u - u^3$ is an *active restorative force* pulling the field back to the bulk phases $u = \pm 1$, and this force strengthens as ϵ shrinks (the reaction term dominates the $\epsilon^2\nabla^2 u$ diffusion term at small ϵ in Eq. 10); it may actively counteract spectral smearing near the interface, independent of anything an operator learns. The one-phase Stefan problem (R3/R4) has no such term: its dynamics are purely diffusion-driven, with the interface position set by an external jump condition rather than an internal restoring force. The AC sweep's finding could therefore reflect this reaction-term self-correction dominating at small ϵ, rather than demonstrate that sharpness alone does not cause divergence. We therefore confine the mechanism claim to the R3-vs-R4 comparison, where sharpness covaries with heterogeneity, rather than to sharpness in isolation, and do not treat the AC sweep as resolving the sharpness question for diffusion-driven interfaces.

B CCP ANALYSIS: WHY CONSERVATION PROJECTION FAILS FOR OPEN-BOUNDARY STEFAN

The Conservation Correction Projection (CCP) pins the mean enthalpy $\bar{H}_{\mathrm{pred}}$ to the initial condition mean $\bar{H}_0 = \bar{H}_{\mathrm{IC}}$ at every step. For a *closed* adiabatic system this is exact. However, R1 and R4 use Dirichlet/Robin boundary conditions that drive latent heat in or out: $\bar{H}(t) \neq \bar{H}_0$. In practice, we observe enthalpy rises of 200–760% over 20 steps for the Stefan regimes studied here; pinning to $\bar{H}_0$ therefore over-corrects and *increases* drift. A with-CCP ablation confirms this: enabling CCP increases energy drift by $\sim 50\%$ versus leaving it off. CCP remains a sound mechanism for closed-system Stefan with adiabatic boundaries, and we report it here for completeness.

C 1D BURGERS: A BOUNDARY CASE WHERE THE BASELINE IS ALREADY STABLE

We include a 1D Burgers shock as a non-Stefan test of a *localized* sharp front, training XENO, FNO and Transolver on the same $800/100/200$ split and evaluating with the unified rollout harness (the shock mapped to the $f{=}0.5$ level set). The result is a negative result: a trained FNO is *already* rollout-stable here (20-step median $d_H = 0.77$ px), Transolver diverges (9.2), and XENO (3.0) is *worse than FNO*, not merely neutral. XENO's learned OEB gate idles to near-zero on Burgers ($\overline{|g_E|} = 0.00$), which argues against a large, actively-engaged enrichment contribution accounting for this gap. We have not run a No-OEB ablation on Burgers, so whether the residual gap comes from the (unused) φ-network or the staged training schedule is unresolved. The rollout instability we target is a higher-dimensional, heterogeneity-driven phenomenon (R4), not a property of every sharp feature; a 1D shock does not trigger it for FNO, and XENO is not the right choice there.

D FOURIER-MODE SWEEP

To check whether the R4 instability is merely a matter of too few retained modes, we train FNO with $K \in \{4, 8, 16, 32\}$ Fourier modes (same reduced-data recipe; single seed, exploratory) and evaluate 20-step rollout on R4. Every setting remains unstable (median d_H of $20/62/36/25$ px for $K = 4/8/16/32$); more modes do not restore stability, consistent with the mechanism in Section A (band-limited layers cannot represent the moving kink, and additional modes trade spectral smoothing for aliasing). We flag a limitation of this sweep: it uses a *reduced*-data training recipe (half the standard training set) to keep the $4\times$ configuration budget tractable, and App. O separately shows that halving the training data makes rollout stability markedly more fragile even for the stabilized XENO model (seed-dependent divergence up to 20.0 px under reduced data, vs. stability under the full protocol). Higher-mode configurations have more parameters and are plausibly more data-hungry, so we cannot rule out that some of the observed divergence at $K{=}16/32$ reflects data starvation rather than an intrinsic failure of higher-mode spectral bases; re-running at least one higher-mode configuration under the full-data protocol would directly resolve this confound and is left to future work.

E ROBUSTNESS: INPUT NOISE AND LONGER ROLLOUTS

Two checks that the R4 result (referenced from §A) is not an artifact of the specific rollout length or of noise-free feedback. **Input-noise robustness:** injecting Gaussian noise into the autoregressive feedback (std $\sigma \cdot \hat{\sigma}_{\text{channel}}$, $\sigma \in \{0, .01, .02, .05\}$) leaves XENO's 20-step interface error essentially unchanged ($0.98 \rightarrow 1.00$ px), while FNO stays at $\sim$32 px throughout; the stability is a property of the operator, not of clean feedback. **Longer rollouts:** extending to 50 steps, XENO reaches 19.7 px versus FNO 34 and Transolver 147; XENO is far more stable well beyond the 20-step horizon, though we note it too degrades eventually: the fix delays, and greatly slows, divergence rather than eliminating it.

F INFERENCE COST / AMORTIZATION

The R4 training and evaluation data in this paper are generated by a GPU lattice-Boltzmann (LBM D2Q9) solver (App. T). For the amortization comparison below we instead time a separate, independent reference solver – a 2D enthalpy finite-element method (FEM) implementation – as a representative classical numerical cost, since it is the more common baseline for diffusion/enthalpy-formulation Stefan problems in the literature; we do not have a comparable per-simulation wall-clock figure for the LBM generator itself, and did not optimize the FEM reference for speed. XENO's one-time training cost amortizes over N queries: $T_{\text{XENO}}(N) = T_{\text{train}} + N\,T_{\text{inf}}$ vs. $T_{\text{FEM}}(N) = N\,T_{\text{FEM}}$, with $T_{\text{FEM}} = 782\,\text{s}$ per R4-sized simulation (median over 24 runs, `baselines/classical/fem_timing.py`) and $T_{\text{inf}} \ll T_{\text{FEM}}$ (inference is $\sim 10^4\times$ cheaper). Break-even occurs at $N \approx 363$ queries; beyond that a campaign of many phase-change simulations is substantially lower in wall-clock cost with XENO. We report this for completeness; it is not a contribution of this paper, and the comparison is against this FEM reference specifically, not the LBM solver used for training data.

Per-step overhead relative to plain FNO. The above amortizes against a classical solver; separately, we measure the overhead OEB adds *relative to a standard FNO* of matched depth/modes on the same 128×128 R4 grid (single RTX A5000, wall-clock over 50 iterations after 10 warmup calls). At batch size 1 (one autoregressive step), XENO takes 8.5 ms vs. FNO's 2.0 ms ($4.3\times$, $+334\%$); at batch size 32, 220.6 ms vs. 51.5 ms ($4.3\times$, $+328\%$); the overhead is a roughly constant multiplicative factor, not a fixed cost that amortizes away at larger batch. INTERFACEFIELDNET alone costs ≈ 1.6 ms per call and is invoked once per OEB block (4 blocks), accounting for most of the added latency; XENO has 4.01M parameters vs. FNO's 2.12M ($1.9\times$). This overhead is still $\sim 10^3$–$10^4\times$ cheaper than the FEM reference above, so it does not change the amortization conclusion, but it is the correct number for comparing against other *neural* baselines rather than a classical solver.

G PINO DIVERGENCE ON STEFAN

PINO (Li et al., 2024) was trained on both R1 and R4 with identical hyperparameters to FNO and Transolver. Final val Hausdorff: 21.1 (R1), 147.0 (R4). We attribute this to the physics residual loss in PINO being poorly calibrated for the enthalpy formulation with latent heat; the loss landscape becomes dominated by the physics term early in training, preventing field-level convergence. We include PINO in Table 2 with its divergence flagged ($\gg$ 1 px, †) rather than a numeric rollout comparison, since a Hausdorff distance is not a meaningful summary once the field itself has not converged.

H HYPERPARAMETERS

Parameter	R1	R4
Params (millions)	1.05	4.01
FNO modes K	8	8
Enrichment atoms M	8	8
Hidden width d_v	64	128
Blocks	4	4
Optimizer	AdamW, weight decay 10^{-4}	
Learning rate	3e-4	3e-4
LR schedule	cosine decay, $T_{\max}$ = total epochs	
Batch size	32	8
Total epochs	200	200
Early stopping	disabled for main XENO models (App. O)	
Rollout noise std	0.02	0.02
Seeds	42, 123, 456	42, 123, 456

I COMPLETE PER-REGIME RESULTS

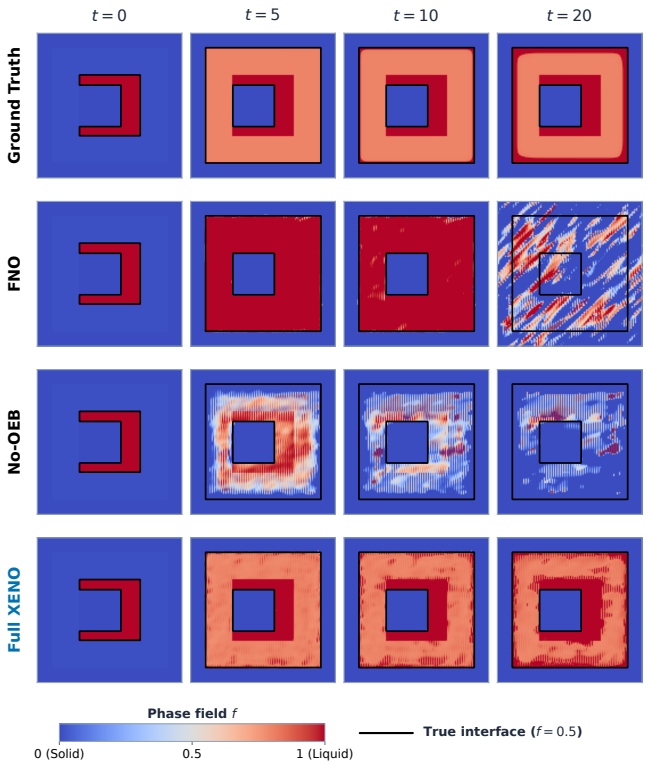


Figure 5: **What rollout divergence looks like.** A representative moving R4 test trajectory: predicted interface (colour) vs. the true interface (black, f=0.5) at $t = 0, 5, 10, 20$. FNO and the No-OEB ablation progressively smear and detach from the front; full XENO tracks it.

Table 4 collects the unified-harness rollout results for every regime studied. All values are median 20-step interface Hausdorff distance (px) over the held-out test set; failure rate is the fraction of trajectories exceeding 5 px at the final step.

The pattern is the paper's scope statement in numbers: on smooth interfaces (R3) existing operators are already stable and XENO adds nothing; on localized heterogeneity-driven sharpness (R4) baselines

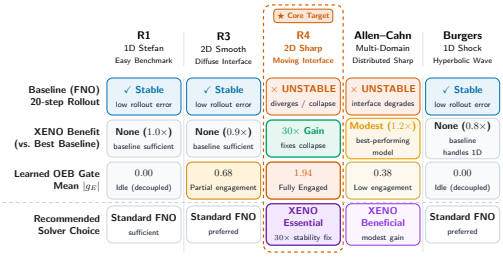


Figure 6: **Where does XENO help?** Regimes by difficulty (columns) vs. baseline (FNO) stability, XENO's benefit, and the OEB gate magnitude (§4.2).

Table 4: All regimes, all models (median 20-step d_H, px / failure rate, $d_H > 5$ px). "—" = not run (1D-only or 2D-only architecture) or failure rate not defined for this benchmark (Burgers). R4's XENO cell is the mean failure rate across all 8 main seeds (36.9%, 95% CI $[32.9, 41.0]\%$); every other cell (including XENO on R3/AC) is single-seed (s42), since only R4 has a multi-seed study. Persistence = copy the initial condition.

	R3 (smooth)	R4 (localized sharp)	AC (distributed sharp)	Burgers (1D shock)
XENO	**1.62** / 0.03	**1.09** / 0.37	**7.29** / 0.64	3.03
FNO	∼1–2 / low	30.7 / 1.00	8.62 / 0.73	**0.77**
Transolver	∼1–2 / low	25.7 / 0.94	9.62 / 0.84	9.23
Persistence	26.1 / —	8.4 / 0.50	10.7 / 0.90	—

diverge and XENO does not; on distributed sharpness (AC) XENO is best on every metric but by a modest margin (App. M); on a 1D shock the baseline is already stable and the OEB idles (App. C).

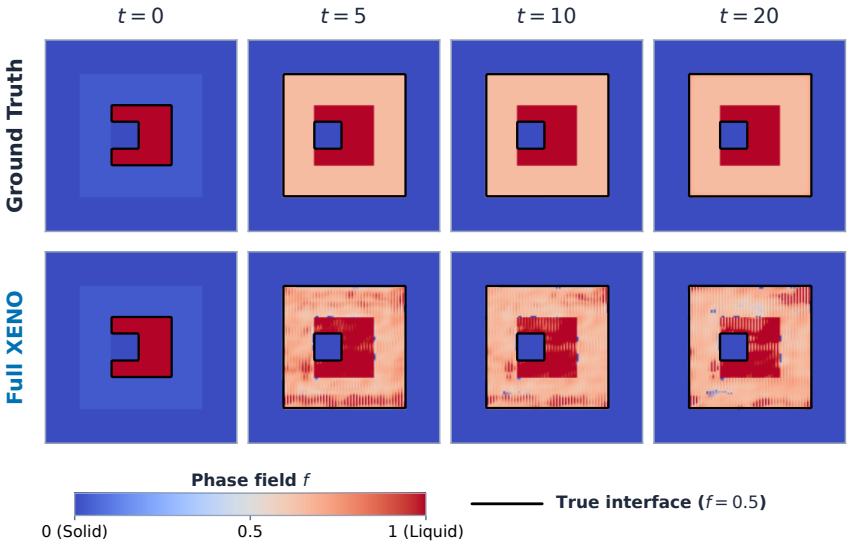


Figure 7: A XENO failure case (referenced from §4.2): the true interface (a single sharp front) advances ∼30 px over 20 steps; XENO lags on this fast motion ($d_H^{20} = 21$ px). Such failures are dominated by rapid interface displacement rather than topology change.

J WHY EACH BASELINE BEHAVES AS IT DOES

These architectures differ substantially in representation: FNO and F-FNO are genuinely band-limited (a fixed, truncated set of retained Fourier modes); Transolver mixes information through

Table 5: R4 rollout d_H (px, median) stratified by true interface motion (referenced from §4.2), XENO's representative seed 42 (checkpoint-selection App. P); failure rate is fraction with $d_H > 5$ px over all 200 trajectories. "beats persist" = fraction of *moving* trajectories with error below the persistence baseline. XENO's 32% fail rate here matches the 8-seed cross-seed mean of 37% (above) to within seed-to-seed spread.

Model	static (n=100)	moving (n=100)	fail rate (>5px)	beats persist (moving)
persistence	0.00	27.1	50%	—
FNO	40.4	25.0	100%	52%
Transolver	36.8	25.5	94%	60%
XENO	**0.42**	**12.4**	**32%**	**82%**

learned global attention slices, which is not a band-limited basis in the Fourier sense; WNO is a multiresolution wavelet method whose own literature emphasizes joint space-frequency localization, not global support; CNO is a convolutional architecture that applies its nonlinearities at an anti-aliased resolution specifically to avoid introducing spurious high frequencies. We report what is empirically true for each rather than assume a shared cause.

FNO / F-FNO (diverge on R4). Both are literally band-limited: each layer retains a fixed set of K Fourier modes. The predicted interface progressively loses spectral energy and smears under rollout, and the resulting error is large and low-frequency-dominated (Section A). F-FNO's factorized mixing and residual layers diverge less severely (13.5 px vs. FNO's 32.1) but still well above XENO, so a better-engineered spectral operator delays but does not prevent the failure. One-step accuracy is unaffected for either; divergence is a rollout phenomenon.

Transolver (diverges on R4). Transolver is not Fourier-band-limited; it mixes information through learned attention over a small number of global slices. It nonetheless diverges (22.4 px), so whatever explains FNO's failure is not the whole story: some other property of *global* information mixing, not band-limiting specifically, may be implicated. We do not have a mechanism-level account for Transolver and do not claim one.

WNO (near-perfect one step, still diverges). WNO reaches the best one-step error of any model on R4 (0.02 px) yet ends at 35 px after 20 steps, despite being explicitly designed for space-frequency localization rather than global support. It is the clearest single case that one-step fidelity does not imply rollout stability, and its failure despite localization argues against a simple "global band-limiting causes this" account: localization in the representation is not by itself sufficient for rollout stability.

DeepONet (degrades smoothly). The fixed global trunk basis acts as a strong smoother; predictions blur rather than sharpen, yielding large but non-explosive error (24 px).

CNO (alias-aware, still diverges). CNO's anti-aliased nonlinearities mitigate one specific failure mode (spurious high-frequency content from pointwise nonlinearities) without preventing rollout divergence (21.5 px). This argues against aliasing being a complete explanation, but, since CNO is convolutional rather than a truncated global spectral basis, it does not by itself support a band-limiting account either; we report the empirical failure without attributing a cause we have not measured for this architecture.

U-Net (markedly better but still diverges). A convolutional encoder-decoder with deep down-sampling/upsampling paths reaches the most stable rollout of any non-XENO baseline (9.4 px), well below every other baseline we test, including CNO, which is also convolutional yet fails at 21.5 px. This gap between two convolutional architectures suggests U-Net's relative robustness is better attributed to its hierarchical multi-scale receptive field (pooling gives it access to large, non-local spatial context despite using only local convolutional kernels) and its lack of spectral truncation, rather than to convolutional locality *per se*: strict spatial locality alone does not explain why CNO fails where U-Net does not. U-Net still exceeds the 5 px failure threshold, so whatever advantage it has is not sufficient to fully close the gap to XENO on this benchmark (Section J.1); we did not test a

wider family of hierarchical/local architectures, and it remains possible that a different design would do better.

PINO. See App. G: the physics residual is ill-conditioned across a latent-heat discontinuity and dominates training.

More Fourier modes do not help (FNO specifically). App. D: every $K \in \{4, 8, 16, 32\}$ diverges (20–62 px) for FNO; simply retaining more modes does not restore stability for this architecture.

J.1 Does locality alone suffice? A discussion

Section 4's comparison set includes, deliberately, architectures spanning several representational choices (spectral: FNO, F-FNO; wavelet: WNO; alias-aware convolutional: CNO; attention: Transolver; branch–trunk: DeepONet) and a hierarchical convolutional architecture (U-Net) whose kernels are local but whose downsampling/upsampling path gives it a large effective receptive field. The U-Net result is notable: it is markedly more stable than every other baseline (9.4 px vs. 13–35 px) while still diverging well above XENO (1.1 px), and better than CNO (21.5 px), another convolutional architecture, which argues against attributing U-Net's advantage to convolutional locality alone (see App. J's U-Net paragraph). Whatever U-Net's advantage over CNO is (candidates include its multi-scale hierarchy or its lack of spectral truncation) it is not sufficient to close the gap on this benchmark; OEB's interface-anchored, geometry-conditioned enrichment, local in a more targeted sense and activated only where φ indicates an interface, does more than generic locality or hierarchy. WNO's failure despite its own explicit space-frequency localization (above) shows that locality in a representation's basis functions is not by itself sufficient either; what distinguishes OEB is locality *anchored to the predicted interface itself*, not locality of the basis in general. A systematic comparison against a wider family of local and hybrid local/spectral architectures is left to future work.

K Why removing φ is as damaging as removing the OEB

In the component ablation (Table 3) removing the interface field φ degrades the 20-step rollout to a median 32.2 px, statistically indistinguishable from removing the entire enrichment block (31.0 px), and back at the baseline failure level. Concretely, this ablation does not zero out or bypass φ: InterfaceFieldNet's output is replaced by a fixed, non-learned ramp field $x+y$ (linearly increasing across the spatial grid, normalized to $[-1, 1]$ per axis, independent of the input state), so φ is always defined and non-constant but carries no information about the true interface location. This is the expected outcome of a *mutually necessary* pair: φ is the coordinate that the enrichment atoms $\psi_k(\varphi)$ act on, so with this degenerate, input-independent φ the atoms have no interface to localize on and the enrichment does not recover interface-localized behavior, yielding performance comparable to No-OEB. Symmetrically, φ is computed but unused without the OEB to consume it. Neither component is useful in isolation; only the combination localizes a non-band-limited correction at the moving interface. (An earlier reduced-data run had suggested No-φ was *worse* than No-OEB; at the full-data protocol the two are equal, and we report the full-data result.)

L OEB specificity: full control results

Free controls (no retraining). Applied at test time to the trained main XENO model (8 seeds, n=200): *gate-off* sets $g_E \equiv 0$ post hoc (equivalent to inference-time No-OEB); *shift* rolls the learned φ field spatially by 16px before it enters OEB; *shuffle* permutes φ across the batch; *sdf_raw*/*sdf_scaled* substitute a deterministic signed-distance transform of the current phase field for the learned φ. Median-of-seed-medians over 8 seeds: gate-off 64.8 px, shift 45.4 px, shuffle 52.3 px, sdf_raw 37.3 px, sdf_scaled 35.2 px, versus 1.09 px for full XENO (Table 3). A learned φ that is present but wrong (shifted or shuffled) is exactly as damaging as having none; even the deterministic SDT, which supplies a geometrically correct interface location, does not recover stability.

Table 6 gives full per-control numbers for the trained-control study referenced from §4.3. Unlike the No-OEB/No-φ ablations (which use `use_ccp=False`, identical to the main XENO model), controls A1–A4 were originally trained with `use_ccp=True` (App. B discusses the mechanism,

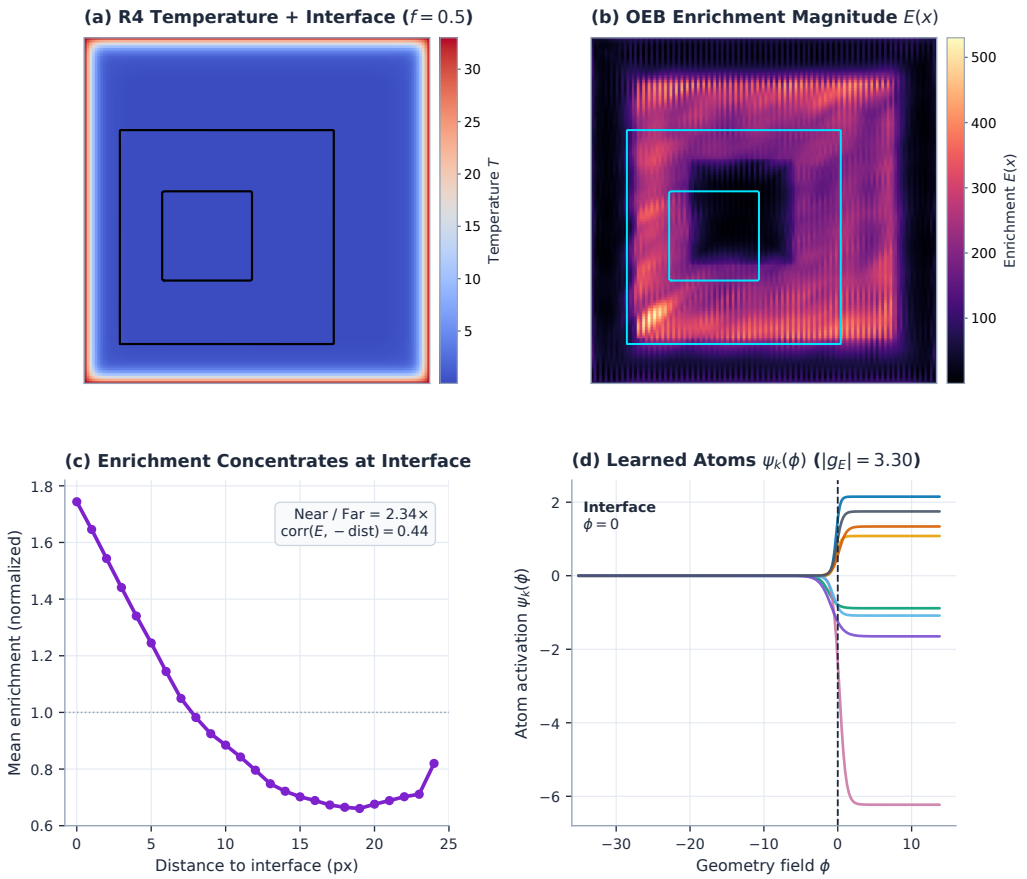


Figure 8: **OEB enrichment concentrates near the interface** (referenced from §A). (A) an R4 temperature field with the true interface; (B) the OEB enrichment magnitude for the same sample is stronger near that interface; (C) enrichment vs. distance-to-interface (2.3× stronger within 3 px of the front than far from it; Pearson correlation with proximity 0.44, i.e. a clear but not exclusive interface concentration); (D) the learned atom shapes $\psi_k(\varphi)$ are localized near φ=0.

though it does not itself train or evaluate these controls). Because `CCPLayer` has no learnable parameters, we could re-evaluate the existing A1–A3 checkpoints with `use_ccp=False` at inference time without retraining; A1 and A3 remain collapsed (`use_ccp=True`→`False`: A1 28.7 → 27.8; A3 35.7 → 33.9), so this confound does not explain their failure. A2, however, improves substantially under the same toggle (22.5 → 2.8); because this changes the autoregressive feedback distribution, we do not treat A2 as a clean ablation: an eval-time toggle changes what the autoregressive rollout actually feeds back into the model relative to what the checkpoint was trained under (the projected H is the next step's input), so this is a train/inference mismatch, not a clean ablation, and A2's true `use_ccp=False` behavior requires retraining from scratch (not done here). A4 was retrained with `use_ccp=False` from scratch (the rigorous fix), since it is the more architecturally load-bearing control (information access without the atom mechanism); see Table 6 for the result: 10.4 px (∼9.5× worse than full XENO), lower than the original `use_ccp=True` number (13.8 px) but still an order of magnitude short of full XENO, so the specificity conclusion for A4 is unchanged by the rigorous retrain: information access without the atom-localization mechanism is not sufficient.

We also address a distinct reviewer concern directly: the deterministic-SDF controls (A1, and CCP-toggle above) are computed from the phase field f alone, whereas the learned INTERFACEFIELDNET sees both T and f, an information asymmetry that could partly explain A1's collapse independent of whether φ is *learned*. We add **A5**: the same zero-parameter deterministic SDF as A1, but the enrichment atoms are additionally modulated by raw T (a new `SDFPlusTBranch`, $\psi_k(\varphi, T) = a_k\sigma(s_k\varphi + r_kT + b_k)$, r_k initialized to zero so A5 starts identical to A1 and can only use T if it

helps; +32 parameters over A1, i.e. effectively parameter-matched), trained `use_ccp=False` from scratch. Result: 19.5 px ($\sim 17.9\times$ worse than full XENO), lower than A1 alone ($27.8 \rightarrow 19.5$; T access helps somewhat) but far from closing the gap to full XENO or even to A4. This addresses the information-asymmetry concern: giving the deterministic-geometry control the same (T, f) input access as the learned φ_θ narrows but does not close the gap, so A1's original collapse is not primarily an information asymmetry; the learned (rather than deterministic) geometry accounts for most of the remaining difference.

Table 6: OEB specificity: no single-ingredient substitute recovers stability, with one control (A2) now inconclusive rather than resolved. A1/A3 numbers are `use_ccp=False` eval-time re-evaluations of the original checkpoints (a valid, exact re-evaluation since CCP has no learnable parameters); A4 and A5 are trained `use_ccp=False` from scratch (§L).

Control	Removes	Seeds	20-step d_H (median) $\downarrow$
Full XENO	—	8	1.09
A4 (concat)	atom-localization only	1	10.4 ($\sim 9.5\times$ worse)
A5 (sdf+T)	*learning* φ (SDT + raw T access)	1	19.5 ($\sim 17.9\times$ worse)
A1 (sdf-φ)	*learning* φ (uses SDT instead)	3	27.8 ($\sim 25\times$ worse)
A3 (FiLM)	φ entirely (raw phase conditioning)	2	33.9 ($\sim 31\times$ worse)
A2 (gated-MLP)[†]	all geometry (params+gating only)	3	2.8 (inconclusive, see text)

[†]A2's `use_ccp=False` number is an eval-time toggle of a checkpoint trained with `use_ccp=True`, not a from-scratch retrain; because the CCP projection is applied to the value fed back autoregressively, this is a train/inference mismatch rather than a clean ablation, and we do not treat it as resolving whether params+gating alone can substitute for OEB.

M Why the Allen–Cahn gain is modest

On AC, XENO is best on every metric (median error, failure rate, fraction beating persistence) but the median margin is small (7.29 across 3 seeds, vs. 8.62 for FNO). The mechanism explains this: OEB enrichment is spatially concentrated at the interface (the enrichment magnitude is $2.3\times$ larger within 3 px of the interface, Fig. 8), so its leverage is greatest when sharpness is *localized*. AC's phase field is sharp *everywhere* along dense, curvature-driven interfaces, so a local enrichment corrects a smaller fraction of the total error. The robust signals on AC are the tail metrics: XENO fails on 64% of trajectories versus 73–84% for baselines, and beats persistence on 91% versus 80–87%. Consistently, the learned gate engages only partially on AC ($\overline{|g_E|} = 0.38$, well below R4's 1.94 though above Burgers' 0.00). Across seeds, XENO's AC median is stable ($8.04/6.61/7.29$ px for s42/s123/s456, mean 7.31 ± 0.59); unlike the main R4 result, the AC result does not show a seed-outlier.

N The OEB gate across regimes

The learned gate magnitude correlates with regimes where enrichment is beneficial, but this is not a deployment-safety guarantee (see Section 6): on R3 and Burgers, where the gate is smallest, XENO is nonetheless worse than the best baseline in absolute terms. Validation should therefore be performed on the target regime rather than assuming the gate will neutralize the added complexity.

O Practical guidance: rollout stability emerges late in training

One methodological finding is that one-step validation metrics (including interface Hausdorff on validation *pairs*) are poor proxies for autoregressive stability: in our runs they plateau long before the rollout stabilizes. Concretely, seeds early-stopped at epochs 70–112 by a plateaued one-step validation metric produced 20-step rollouts 10–$20\times$ worse than the same seeds trained to the full 200 epochs, even though their one-step errors were nearly identical. Rollout stability is acquired *late*: in a representative run the 20-step median error was still 19 px at epoch 80 and 1.4 px at epoch 200. We therefore (i) disabled early stopping for all main XENO models, and (ii) recommend validating

Table 7: Mean absolute OEB gate $\overline{|g_E|}$ (across the four enrichment layers) after training on each regime, with the regime's baseline behavior. The gate is a scalar ReZero-style parameter initialized at 0; whatever value it reaches is learned from data alone.

Regime	$\overline{\lvert g_E \rvert}$	Baseline rollout	Interpretation
R1 (1D easy)	0.001	stable	enrichment unnecessary; gate stays at init
R3 (2D smooth)	0.68	stable	partial engagement; XENO worse than Transolver (1.6 vs 1.0)
R4 (localized sharp)	1.94	diverges	full engagement where enrichment is decisive
AC (distributed sharp)	0.38	degrades	partial engagement; local enrichment has less leverage
Burgers (1D shock)	0.00	stable	gate disengaged; XENO still worse than FNO (3.0 vs 0.8)

autoregressive surrogates with short *rollouts* rather than one-step metrics. All numbers in this paper come from a unified rollout harness evaluated on held-out test trajectories, never from training-time validation metrics.

Relatedly, rollout stability is *data-hungry near its boundary*: with the training set reduced to half the trajectories (800 of 1600, stride-2 subsampling), XENO's 20-step R4 rollout becomes seed-dependent ($1.4/20.0/20.0$ px across three seeds, at indistinguishable one-step error and training loss), substantially worse than at the full budget (App. P). One-step metrics are blind to this transition. We report the ablation at the full-data budget for this reason, and flag the reduced-budget fragility as a caution for surrogate practitioners: an autoregressive surrogate can silently sit on the wrong side of a stability/data boundary while every standard training metric looks healthy. As App. P shows, one seed lands on an unstable *checkpoint* despite healthy training and short-horizon validation metrics, but this is recovered by selecting the checkpoint with a held-out validation rollout, under which no seed is unstable.

P PER-SEED RESULTS

Every seeded row in the main tables reports the median across seeds; this appendix lists the per-seed values, reported without excluding any run. For XENO we give each seed's 20-step error under two checkpoint-selection rules: the naive rule (lowest one-step validation Hausdorff) and a held-out *validation-rollout* rule (lowest 20-step Hausdorff on validation trajectories).

Table 8: R4, 20-step rollout d_H (px), XENO per seed under two checkpoint-selection rules (full training data, no early stopping; 8 seeds). The one-step-validation rule selects an early, rollout-fragile checkpoint for s456 and s2 (epochs 55 and 42); a held-out validation rollout selects each run's epoch-200 checkpoint instead, recovering both. We applied the identical held-out validation-rollout selection procedure to the baselines: every FNO and Transolver seed diverges under *both* available checkpoints (FNO 31–42 px across best/last on all 3 seeds; Transolver 18–33 px, 2 of 3 seeds, as s123 has only one saved checkpoint; App. Q). No-OEB and No-φ (28–34 px) are reported under the naive rule only, consistent with their seed-robust divergence under identical training.

XENO checkpoint rule	s42	s123	s456	s1	s2	s4	s5	s7	mean
one-step val. Hausdorff	0.98	1.11	13.6	0.89	16.0	0.71	1.13	1.40	4.47
held-out val. rollout	0.98	1.11	**1.19**	0.89	**1.21**	0.71	1.13	1.06	**1.03**

Under the naive one-step-validation rule, six of XENO's eight seeds roll out at 0.71–1.40 px while two (s456, s2) reach an early checkpoint whose training loss and short-horizon validation Hausdorff are healthy but whose 20-step rollout is not (13.6–16.0 px; App. O discusses why one-step and rollout metrics disagree). This is a checkpoint-*selection* artifact rather than intrinsic seed instability: the same runs' epoch-200 checkpoints, which the validation-rollout rule selects, roll out at 1.1–1.2 px. Applying the validation-rollout rule uniformly, all eight seeds land at 0.71–1.21 px (mean 1.03,

median 1.09; no seed's median is unstable; this is a seed-level statistic, distinct from the per-trajectory failure rate in App. I). The same 200 test trajectories are reused across all 8 seeds, so a naive t-based CI over seed medians understates uncertainty; a hierarchical bootstrap that jointly resamples trajectories *and* seeds (App. S) gives the more defensible interval, 95% CI $[0.87, 2.48]$, substantially wider on the upper bound, reflecting that a single unlucky seed-and-trajectory draw can look considerably worse than the point estimate suggests, even though the point estimate itself is unchanged. Whereas every baseline seed diverges to 18–42 px regardless of checkpoint. We therefore report XENO under validation-rollout selection and recommend short-rollout validation for autoregressive surrogates generally (App. O). The ablation variants (No-OEB, No-φ) diverge for every seed under either rule (28–34 px), confirming that instability is the seed-robust default behavior of a band-limited operator.

Q BASELINE CHECKPOINT-SELECTION EVIDENCE

The claim that baselines diverge *regardless* of checkpoint is, for XENO, established by construction (App. P). We verified it for the two main baselines rather than assume it: Table 9 applies the identical held-out validation-rollout selection procedure to every FNO and Transolver seed with more than one saved checkpoint.

Table 9: 20-step rollout d_H (px) for FNO and Transolver under both saved checkpoints, on the same seeds as Table 2. s123-Transolver has only `ckpt_best.pt` saved. Every case diverges under both checkpoints; no baseline seed is checkpoint-sensitive in the way s456/s2/s5-XENO were.

Model	seed	ckpt_best	ckpt_last	held-out-rollout-selected
FNO	42	33.65	30.66	30.66
FNO	123	42.08	41.81	41.81
FNO	456	32.09	34.01	32.09
Transolver	42	30.42	25.71	25.71
Transolver	123	18.11	—	18.11
Transolver	456	32.98	22.43	22.43

R PUSHFORWARD-TRAINED FNO: FULL PROTOCOL

Full protocol for the pushforward-trained FNO baseline referenced from §4.3. Each training step computes a one-step loss $\mathcal{L}_1$ on the ground-truth pair $(t, t{+}1)$, then feeds the model's own *stop-gradient* one-step prediction back in as input and computes a second-step loss $\mathcal{L}_2$ against $(t{+}2)$; the total loss is $\mathcal{L}_1 + \lambda\,\mathcal{L}_2$. An initial run at $\lambda{=}1$ collapsed to a conservative, rollout-lagging predictor (20-step median $d_H \approx 13$ px, worse than plain FNO, with one-step error also degraded to ≈ 11 px): the two-step term dominated and traded away one-step accuracy. We corrected this with a single, non-systematic adjustment to $\lambda{=}0.25$ (chosen to keep the one-step loss dominant, not tuned by a sweep) and retrained; both the $\lambda{=}1$ and $\lambda{=}0.25$ results are given here rather than only the better one. Optimizer, learning rate, and schedule (AdamW, 3×10^{-4}, cosine decay, weight decay 10^{-4}) match the FNO baseline exactly. Epochs (120 vs. FNO's 200) and batch size (24 vs. FNO's 8) differ: each pushforward step performs two forward/backward passes, so epochs were reduced and batch size increased to keep wall-clock training time comparable; we did not verify this trade-off is loss-neutral relative to matching FNO's exact budget. Evaluation uses the identical 20-step unified rollout harness and, where multiple checkpoints are available, the same held-out validation-rollout selection rule as XENO (App. O). Full metrics (IoU, ASSD, field L_1/L_2, energy drift) are in Table 10 below.

S ADDITIONAL EVALUATION METRICS

Hierarchical bootstrap. The main mean Hausdorff CI (§4) uses a nested bootstrap: resample the 8 seeds with replacement, then independently resample each drawn seed's 200 test trajectories with replacement, aggregate (*median* over trajectories per seed, matching the primary d_H metric, which is reported as a median because it is max-based and heavy-tailed; then *mean* over seeds), repeat $10{,}000\times$, and take the 2.5/97.5 percentiles. This accounts for the same 200 trajectories being reused across every seed, which a t-based CI over seed medians does not.

Table 10 reports the R4 20-step rollout under metrics complementary to the max-based pixel Hausdorff, on 200 test trajectories (median): domain-normalized Hausdorff (d_H as a percentage of the 128-px domain width), average symmetric surface distance (ASSD: the *mean* rather than the max nearest-neighbour contour distance), interface IoU, phase-field relative L_1 and L_2, and enthalpy (energy) drift. XENO is best on every metric; the pushforward-trained FNO (§5) is far better than a plain FNO but remains behind XENO throughout, corroborating that the gain is architectural rather than a training-stabilization trick.

Table 10: R4, 20-step rollout, median over 200 test trajectories (XENO seed 42 at its validation-rollout checkpoint). Lower is better except IoU (↑).

Model	d_H (px)	d_H (% dom)	ASSD (px)	IoU ↑	field L_1	field L_2	energy drift
FNO	30.7	24.0%	8.79	0.037	1.247	1.034	0.815
pushforward-FNO	8.50	6.64%	0.56	0.978	0.068	0.173	0.126
XENO (ours)	**0.98**	**0.76%**	**0.04**	**0.999**	**0.025**	**0.064**	**0.081**

T DATA GENERATION

Table 11 summarizes the solver, resolution, and initial/boundary-condition sampling for every regime used in the paper (full per-trajectory metadata, IC/BC/material parameters, and SHA256 checksums for every released file accompany the code). Train/validation/test splits are disjoint at the *trajectory* level: each generator partitions its sampled trajectories into three whole, non-overlapping sets before any step-pair is extracted, and each split is written to its own file (`train.h5`/`val.h5`/`test.h5`); no trajectory contributes frames to more than one split.

The Allen–Cahn (AC) regime solves, on a periodic 2D domain with a semi-implicit spectral (Fourier) scheme (reaction treated explicitly, diffusion implicitly),

$$\partial_t u = \epsilon^2 \nabla^2 u + (u - u^3) + h, \tag{10}$$

a bistable reaction–diffusion (Ginzburg–Landau) equation with double-well reaction $u - u^3$, interface half-width ϵ, and a constant external tilt h; the zero level set $\{u{=}0\}$ is the moving interface. The Burgers (1D shock) regime uses the exact inviscid entropy (Riemann) solution of

$$\partial_t u + \partial_x\left(\tfrac{1}{2}u^2\right) = 0, \tag{11}$$

with a single right-moving discontinuity of speed $s = (u_L + u_R)/2$ (Rankine–Hugoniot) between left/right states $u_L > u_R$ and initial shock location x_0; no numerical solver is used since the closed-form solution is exact.

Table 11: Data-generation summary. All splits are simulation-only, generated from the listed solver; $U[a, b]$ denotes a uniform draw per trajectory.

Regime	Solver	Grid × steps (Δt)	Boundary/initial conditions	Sampled parameters	Train/val/test
R1	Neumann analytical	256 (1D) $\times 200$	1D two-phase Stefan, exact solution	$T_i{\sim}U[-20,-2]$, $T_{\text{hot}}{\sim}U[2,25]$, domain $L{\sim}U[0.05,0.3]$	4000/500/500
R3	LBM (D2Q9)	$128^2 \times 100$ (200 LB steps/save)	Dirichlet hot/cold walls, adiabatic top/bottom	$T_{\text{hot}}, T_{\text{cold}}, g_\beta$, latent heat l_h (fixed $\nu{=}0.08, \alpha{=}0.04$)	1600/200/200
R4	LBM (D2Q9)	$128^2 \times 100$ (200 LB steps/save)	Convective walls (Biot), diurnal/const. ambient	Biot$\sim U[3,8]$; heterogeneous insulation/PCM/payload material presets	1600/200/200
AC	Spectral IMEX (Fourier)	$128^2 \times 50$ ($\Delta t{=}3{\times}10^{-3}$, 6 steps/save)	Random low-pass tanh IC, periodic BC	interface width $\epsilon{\sim}U[0.012,0.022]$, bias $h{\sim}U[-0.05,0.05]$	800/100/100
Burgers	Exact Riemann (entropy soln.)	128×20 ($\Delta t{=}0.02$)	Riemann shock/rarefaction IC	$u_L{\sim}U[0.6,1.4]$, $u_R{\sim}U[-0.4,0.2]$, $x_0{\sim}U[0.2,0.5]$	800/100/200

Every baseline is trained with the same optimizer, learning-rate schedule, batch size (per regime), and epoch budget as XENO (App. H); parameter counts are recorded per run and are matched to XENO within $\sim 10\%$ where architecturally possible (exact per-model counts are logged alongside every checkpoint and released with the code).