WHEN DO PHYSICAL CONSTRAINTS CORRECT NEURAL PDE ROLLOUTS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive neural PDE surrogates are routinely regularized and corrected with physical constraints, yet the relationship between enforcing a constraint and reducing prediction error remains poorly understood. We study this relationship for gradient-flow rollouts, where the free energy must decrease along the trajectory. Enforcing this temporal condition on fixed pretrained surrogates yields a discrete Lyapunov certificate, and the same certificate produces sharply different field correction: on Allen–Cahn it removes 99.9% of the structural error and 75.6% of the field error; on Cahn–Hilliard it removes 96.5% of the structural error and 1.1% of the field error. We explain this contrast through the certificate-to-error ratio $R = \|\delta_{\text{cert}}\|/\|e\|$. The same-step field benefit obeys $B_{\text{local}} \leq R$, and for small corrections $B_{\text{local}} \approx R\cos\theta$, where $\cos\theta$ is the correction–error alignment.The bound holds exactly across 15,412 activated steps, and the alignment sign predicts whether enforcement corrects or degrades a rollout across systems, distribution shifts, backbones, and independent training seeds. Autoregressive feedback can amplify a small local correction into a complete long-horizon rescue, while lower field error alone does not ensure correct dynamics. Constraint satisfaction is therefore an incomplete proxy for prediction quality: certification, field correction, and dynamical fidelity are distinct measurable properties of a learned rollout.

1 INTRODUCTION

Learned PDE surrogates replace repeated numerical integration with a learned map between physical states (Li et al., 2021; Lu et al., 2021; Brandstetter et al., 2022; Pfaff et al., 2021). Autoregressive deployment composes that map with itself, so each prediction becomes the input distribution for the next, and long rollouts can drift even when one-step predictions look plausible (Stachenfeld et al., 2021; McCabe et al., 2023; Lippe et al., 2023). A standard response is to impose physics: residuals, conservation laws, or energies enter the training objective (Raissi et al., 2019; Beucler et al., 2021; Wu et al., 2023), or a projection corrects each emitted state at inference (Valente et al., 2025; Rochman Sharabi & Louppe, 2025; Huang & Perdikaris, 2026).

Between these two activities sits a question that neither answers. Enforcing a constraint makes the constraint hold; what does that enforcement tell us about the prediction error? A rollout can satisfy a physical law exactly and still be far from the reference solution, and an enforcement step can move the field toward the reference, leave it unchanged, or push it away. Constraint satisfaction is used in practice as a proxy for prediction quality. Whether that proxy carries information, and how much, has not been quantified.

This paper starts from an empirical contrast that makes the question concrete. We take two phase-field systems, Allen–Cahn and Cahn–Hilliard, train comparable neural surrogates, and enforce the same temporal condition on their rollouts: the discrete free energy may not increase between accepted states. The certificate works in both cases; the field correction could not be more different. On out-of-distribution Allen–Cahn, enforcement removes 99.9% of the structural (energy) error and 75.6% of the field error. On Cahn–Hilliard, it removes 96.5% of the structural error and 1.1% of the field error (Table 1, Figure 2). Successful temporal certification substantially repairs one rollout and barely touches another.

We explain the contrast through what the certificate can see. A temporal energy enforcement returns a correction δ_{cert}; the quantity that bounds its use is the certificate-to-error ratio $R = \|\delta_{\text{cert}}\|/\|e\|$,

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

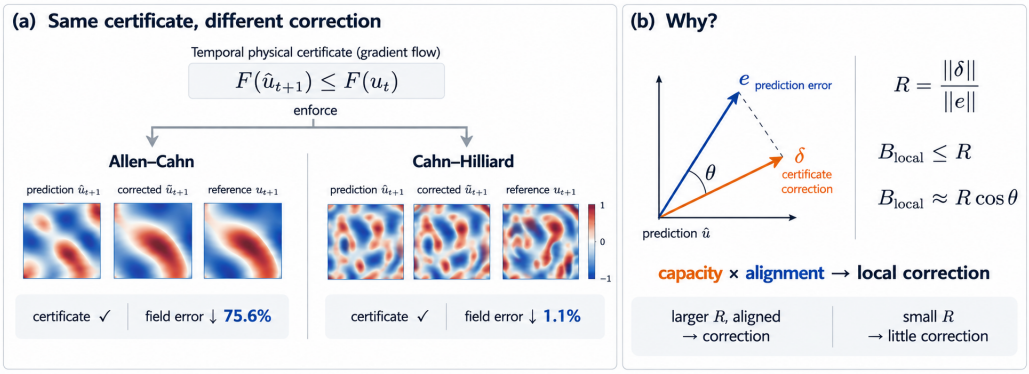

Figure 1: **The same temporal certificate can have very different corrective content. (a)** Enforcing non-increasing free energy substantially repairs an out-of-distribution Allen–Cahn prediction but barely changes Cahn–Hilliard, despite successful certification in both systems. **(b)** For prediction error e and certificate-restoring correction δ, the certificate-to-error ratio $R = \|\delta\|/\|e\|$ bounds the same-step fractional field improvement, $B_{\text{local}} \le R$. For small corrections, $B_{\text{local}} \approx R\cos\theta$. Correction therefore depends on both the available displacement and its alignment with the field error.

where e is the prediction error against the reference. The same-step field benefit obeys $B_{\text{local}} \le R$ (a triangle inequality), and for small corrections $B_{\text{local}} \approx R\cos\theta$, where $\cos\theta$ is the alignment between the correction and the true error direction. Capacity and alignment therefore separate: Cahn–Hilliard sits at small R, so certification cannot correct; mis-trained Allen–Cahn seeds can sit at negative $\cos\theta$, so enforcement degrades the field by exactly the predicted amount. Local correction also differs from long-horizon benefit, because autoregressive feedback amplifies sustained corrections. Figure 1 summarizes this chain: detect a temporal violation, correct it within a bounded local budget, amplify through the rollout, and evaluate the dynamics separately.

The evidence covers Allen–Cahn, conserved Cahn–Hilliard, and eight-field grain growth; FNO, periodic U-Net, and DeepONet backbones; weak, standard, and penalty-trained surrogates; interface-width, mobility, composition, horizon, and resolution shifts; and five independent training seeds for the central mechanism. The controls include per-state projection, generic energy descent, a mis-specified energy functional, and single-step interventions, each isolating one degree of freedom that certification leaves open.

Our contributions are:

- **Temporal certification.** We formalize the gap between state feasibility and trajectory-level Lyapunov structure (Proposition 1) and construct nested free-energy sublevels around fixed pretrained neural PDE surrogates, yielding a discrete Lyapunov certificate for gradient-flow rollouts (Proposition 2).
- **Certificate correctability.** We introduce the certificate-to-error ratio $R = \|\delta_{\text{cert}}\|/\|e\|$, derive the local bound $B_{\text{local}} \le R$ and the alignment identity $B_{\text{local}} \approx R\cos\theta$, and show that magnitude and alignment together explain local correction and degradation across systems, shifts, architectures, and independent training seeds.
- **Rollout and dynamics.** We separate local correction from autoregressive amplification, and use generic-descent and mis-specified-energy controls to show that temporal certification, field accuracy, and dynamical fidelity require distinct evaluation.

2 Related Work

Learned PDE surrogates and autoregressive error. Neural operators learn maps between function spaces and serve as standard surrogates for parametric PDEs (Li et al., 2021; Lu et al., 2021; Kovachki et al., 2023); graph and mesh simulators play a similar role (Pfaff et al., 2021; Sanchez-Gonzalez et al., 2020). Autoregressive error growth is documented and mitigated architecturally (McCabe et al., 2023), through training noise and temporal downsampling (Stachenfeld et al., 2021), and through

multistep refinement of the prediction itself (Lippe et al., 2023). These works primarily target rollout accuracy and stability; the information that an enforced physical condition carries about the remaining field error is less well characterized.

Physics-constrained learned dynamics. Physics-informed training adds residuals, conservation laws, or energy structure to the loss (Raissi et al., 2019; Beucler et al., 2021). For phase-field models specifically, energy dissipation has been built into PINN losses, operator architectures, and energy-splitting training objectives (Wu et al., 2023; Huang & Greenberg, 2025). Hard-constraint approaches embed the constraint in the architecture so compliance holds by construction. These methods impose physical structure during learning. Our setting is complementary: the surrogate is already trained and fixed, and we study what an inference-time certificate says about its errors.

Inference-time projection and correction. Output projection onto physical manifolds enforces algebraic laws on pretrained models (Valente et al., 2025); training-free projections enforce governing-equation residuals on neural PDE solvers (Rochman Sharabi & Louppe, 2025); residual-based per-step correction stabilizes long rollouts of a fixed surrogate (Huang & Perdikaris, 2026); hard-constrained generation modifies pretrained samplers (Cheng et al., 2025); and constraint-projected training places conservation and entropy sets inside learning (Singha, 2025). The closest of these works enforce PDE residuals at inference. We differ in the constrained object—a temporal free-energy sublevel, not an instantaneous residual—and in the question: we analyze the information an enforced constraint carries about field error, including when enforcement fails to correct or actively harms.

Rollout stabilization and correction amplification. Learned correctors in hybrid solvers can amplify autoregressive error, and indirect correction reduces that amplification (Wei et al., 2025). We observe the mirror image of this effect: a physically certified correction, sustained over a rollout, can amplify a small local budget into a complete rescue (Section 5.1). The sign of the effect is decided by the correction–error alignment, which a learned corrector does not expose.

Gap. Across these literatures, methods *enforce* physical structure or *improve* rollout behavior. This paper focuses on an analytical question that remains less explored across these literatures: how much information does an enforced temporal constraint carry about field error, and how does that information control local correction and downstream rollout behavior?

3 Temporal Certification

3.1 Setting

Let S_θ be a pretrained one-step surrogate and ϕ_n the accepted state at rollout step n, with the bare prediction $\widehat{\phi}_{n+1} = S_\theta(\phi_n)$. We study gradient-flow systems with evaluable free energy,

$$\partial_t \phi = -\mathcal{G}\frac{\delta F}{\delta \phi}, \qquad \frac{dF}{dt} = -\left\langle \frac{\delta F}{\delta \phi}, \mathcal{G}\frac{\delta F}{\delta \phi} \right\rangle = -D(\phi) \le 0, \tag{1}$$

where $\mathcal{G}$ is positive semidefinite, discretized as

$$F_h[\phi] = \sum_h \left[\frac{\varepsilon^2}{2} |\nabla_h \phi|^2 + W(\phi) \right] \Delta x^d, \tag{2}$$

with a discrete mass constraint for conserved flows. Conservation laws and one-step residuals are properties of individual states. The inequality $dF/dt \le 0$ instead orders successive states, and it is the condition that autoregressive composition can silently break.

3.2 State feasibility does not imply trajectory stability

A state-level corrector returns $\phi_{n+1} \in \mathcal{A}_n$, where $\mathcal{A}_n$ may encode conservation or a one-step equation residual. Such a set can reference the previous state and still leave the energy sequence uncontrolled.

Proposition 1 (Per-state feasibility does not imply Lyapunov stability)**.** *Suppose that for each n, the per-state feasible set $\mathcal{A}_n$ contains states with arbitrarily large F_h. Then membership $\phi_n \in \mathcal{A}_n$ for every n imposes no uniform upper bound on $F_h[\phi_n]$: there exist admissible rollouts with $F_h[\phi_n] \to \infty$.*

Proof sketch. Choose $\psi_n \in \mathcal{A}_n$ with $F_h[\psi_n] \geq F_h[\phi_0] + n$. The sequence is feasible by construction while its energy diverges. The premise holds for common unbounded feasibility sets such as affine conservation constraints and divergence-free subspaces. Residual-defined feasible sets require separate boundedness analysis and are not covered by the proposition in general. Appendix A gives details.

Empirically, on out-of-distribution Allen–Cahn with a weak surrogate, per-state projection reduces the energy-violation rate only from 0.541 to 0.423, while the temporal construction below removes every violation. Improving frame-level admissibility does not impose a temporal ordering.

3.3 Temporal sublevel enforcement provides a certificate

For the monotone target, define the moving feasible set

$$\mathcal{C}_n = \{\phi : F_h[\phi] \leq \tau_n\} \cap \{\phi : M_h(\phi) = M_0\}, \qquad \tau_n = F_h[\phi_n], \tag{3}$$

with the conservation factor omitted for non-conserved flows. The target depends on the previously accepted state, so nested sublevels impose a temporal ordering. We realize the constraint in two stages: restore the affine conservation law exactly when required, then apply an unconditionally energy-decreasing convex-splitting step until the state enters $\mathcal{C}_n$, with bisection on the segment to the previous accepted state as a feasible fallback. The convex-splitting step is established numerical machinery (Eyre, 1998; Shen et al., 2018; 2019); we use it as a controlled intervention around a fixed surrogate, and the procedure is a constructive relaxation of the ideal metric projection.

Proposition 2 (Discrete Lyapunov certificate)**.** *Assume $W \geq 0$ and use the monotone target in equation 3 with the feasible fallback above. Then every returned state satisfies $\phi_{n+1} \in \mathcal{C}_n$. Consequently $F_h[\phi_{n+1}] \leq F_h[\phi_n]$, the discrete free-energy sequence is non-increasing and bounded below, and its interfacial and bulk contributions remain bounded by $F_h[\phi_0]$ along the rollout.*

The certificate certifies the encoded energy law. It is not a bound on field error and not a convergence statement for the learned trajectory; Sections 4 and 5 measure those separately. Empirically, the certificate behaves as stated across surrogate qualities and backbones: enforcement drives the measured violation rate to zero in every tested configuration, and it is a bitwise no-op on rollouts that are already compliant, so compliant predictions pass through unchanged (Appendix E).

3.4 A stricter rate target is a dynamical probe

Monotonicity constrains the sign of the energy change but not its rate. Let $\mu_n = \delta F_h / \delta \phi$ denote the discrete chemical potential at ϕ_n, and let Δt_S denote the physical time interval between successive surrogate emissions. The physical dissipation $D_n = \langle \mu_n, \mathcal{G}\mu_n \rangle$ suggests the stricter target

$$\tau_n^{\mathrm{rate}} = F_h[\phi_n] - \gamma \Delta t_S D_n, \qquad \gamma \in (0, 1]. \tag{4}$$

This target can be infeasible for a finite correction budget and can over-dissipate an already healthy rollout, so we use it as a dynamical probe in Section 5, outside the unconditional certificate.

3.5 Evaluation definitions

We distinguish quantities that are observable at deployment from quantities used for analysis. *Activation* and the *energy-violation rate* are reference-free rollout observables: they require only the emitted states and the known functional. The *field error* $e = \phi_{\mathrm{ref}} - \widehat{\phi}$ and everything derived from it require the reference trajectory and serve for evaluation. On an activated step, the enforcement returns a correction δ, and we define the certificate-to-error ratio

$$R = \frac{\|\delta\|}{\|e\|}, \tag{5}$$

the normalized same-step field benefit $B_{\mathrm{local}} = (\|e\| - \|e - \delta\|)/\|e\|$, and the rollout-level benefit B_{rollout}, the relative change in final field error between the enforced and bare rollouts. B_{local} is a same-step quantity; B_{rollout} is a trajectory quantity, and the two are never compared against the same bound.

Experiments use Allen–Cahn (L^2 flow), conserved Cahn–Hilliard (H^{-1} flow), and eight-field grain growth on periodic grids, with reference trajectories from an energy-stable semi-implicit Fourier spectral solver. Surrogates are FNO, periodic U-Net, and DeepONet one-step models trained on 160 trajectories and rolled out beyond the training horizon, under shifts in interface width, mobility, composition, horizon, and resolution. Dispersion is reported across trajectories or training seeds; steps within one rollout are not treated as independent replicates. Full protocols are in Appendices B–D.

4 Certificate Correctability

4.1 The discovery experiment: certification succeeds, correction varies

Enforce the temporal certificate of Section 3.3 on OOD rollouts of comparable surrogates for the two gradient flows, and measure two reductions: the drop in structural (energy) error and the drop in field error, both against the reference. Table 1 records the contrast. Allen–Cahn converts certification into substantial correction; conserved Cahn–Hilliard converts the same certification into almost none; grain growth lies between. Figure 2 shows the corresponding energy and field-error trajectories. This is the puzzle: *why can successful temporal certification substantially repair one rollout and barely change another?*

4.2 The certificate-to-error ratio bounds local correction

The answer starts with how much the certificate can move the field at all. On an activated step, the enforcement returns a correction δ against a prediction with error e. The corrected error is $\|e - \delta\|$,

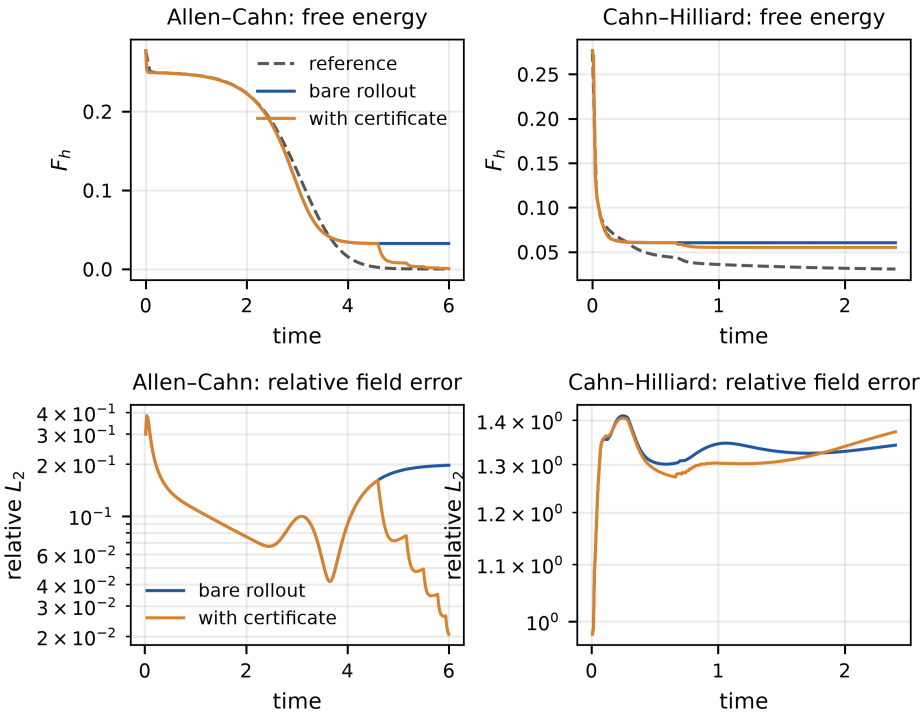

Figure 2: **Successful temporal certification does not determine field correction.** Free energy (top) and relative field error (bottom) for representative out-of-distribution Allen–Cahn (left) and Cahn–Hilliard (right) rollouts. Temporal enforcement restores the non-increasing free-energy relation in both systems, yet its effect on field accuracy differs sharply: Allen–Cahn improves substantially, while Cahn–Hilliard remains close to the bare rollout. This contrast motivates the correctability analysis in Sec. 4.

Table 1: **The same temporal certificate yields very different field correction.** Structural and field reduction are the decreases in energy error and field error after monotone enforcement, relative to the reference, on OOD (thin-interface) rollouts of the weak-surrogate lead setting. Section 4.2 explains the ordering through the ratio R; field reduction is not guaranteed by certification.

System	Structural reduction	Field reduction	Empirical role
Allen–Cahn	99.9%	75.6%	certificate + correction
Grain growth	87.3%	15.8%	intermediate response
Cahn–Hilliard	96.5%	1.1%	certificate, little correction

so by the triangle inequality the normalized same-step benefit obeys

$$B_{\text{local}} = \frac{\|e\| - \|e - \delta\|}{\|e\|} \leq \frac{\|\delta\|}{\|e\|} = R, \tag{6}$$

with equality when δ aligns with e. The certificate's correction capacity is bounded by the ratio of what it moves to what is wrong. Where the prediction error is large relative to any energy-consistent correction—as for conserved Cahn–Hilliard, whose errors can lie along energy level sets—R is small and certification cannot correct, whatever the enforcement budget.

The bound is exact in practice. Across 15,412 activated steps spanning 21 surrogate/regime combinations, $B_{\text{local}} \leq R$ is never violated (maximum $B_{\text{local}} - R = -6 \times 10^{-7}$). The ratio also orders regimes by realized benefit better than the raw energy overshoot (Spearman $\rho = 0.895$ versus 0.740). This ordering is preserved across the five held-out regimes (Spearman $\rho = 1.000$), supporting R as a compact summary of when certification translates into field correction.

4.3 Alignment decides how the budget is spent

The bound controls the available correction magnitude. Its effect on the field also depends on direction. For small corrections, expanding

$$\|e - \delta\|^2 = \|e\|^2 - 2\langle e, \delta\rangle + \|\delta\|^2$$

gives

$$B_{\text{local}} \approx R\cos\theta, \qquad \cos\theta = \frac{\langle \delta, e\rangle}{\|\delta\|\,\|e\|}.$$

The two factors describe different aspects of the correction. The ratio R measures the displacement made available by certificate restoration relative to the existing prediction error. The alignment $\cos\theta$

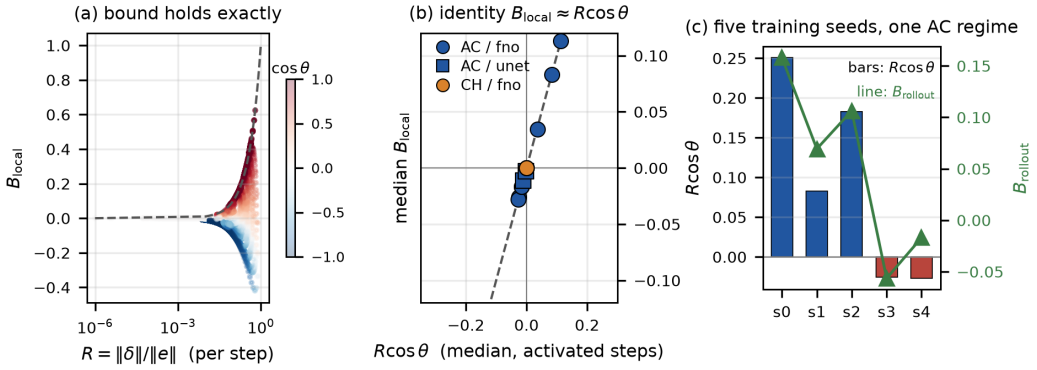

Figure 3: **Certificate correctability.** (a) Per-step local benefit B_{local} against the certificate-to-error ratio R on 15,412 activated steps; the dashed line is the bound $B_{\text{local}} = R$, never violated. Color encodes the alignment $\cos\theta$. (b) Model-level medians of B_{local} against $R\cos\theta$; points lie on the identity line, with Cahn–Hilliard at the origin. Systems are colors, backbones are marker shapes. (c) Five independent FNO training seeds on one Allen–Cahn regime: the sign of $R\cos\theta$ (bars) predicts corrective versus degrading enforcement, matching the sign of the rollout benefit B_{rollout} (line).

measures how much of that displacement points toward the reference. Large R can produce little benefit when the correction is poorly aligned, while a well-aligned correction has limited immediate effect when R is small. Negative alignment predicts degradation even though the temporal certificate is restored. All three cases appear in the experiments below: small correction capacity, useful alignment, and harmful alignment.

4.4 The mechanism survives independent training seeds

If magnitude and alignment explain correctability, the alignment sign should predict the direction of the rollout effect for independently trained surrogates. On the moderate-drift Allen–Cahn regime across five FNO training seeds, it does: the sign of $R\cos\theta$ predicts whether enforcement corrects or degrades the rollout in every seed (Figure 3c). Two seeds produce systematically anti-aligned corrections ($\cos\theta < 0$), and enforcement degrades those rollouts by the predicted amount ($B_{\mathrm{rollout}} = -0.06$). A larger certificate budget cannot rescue this case, because the bound constrains magnitude, not direction. Detection, capacity, and alignment are three separate facts about a rollout: the certificate can fire often (detection), move little (small R), or move against the error (negative $\cos\theta$). Appendix F reports the full seed-level analysis.

We use R strictly as an evaluation quantity computed against reference trajectories. What a deployment can monitor without the reference is the certificate's activation and the size of its own correction; the analysis above quantifies how far that activation can be trusted as a correction signal.

5 Rollout Consequences and Dynamics

5.1 Autoregressive feedback amplifies sustained correction

Local correctability does not determine the rollout outcome. On an unstable long-horizon Allen–Cahn rollout (U-Net, thin-interface shift), the certificate fires on most steps with a small local budget ($R \approx 0.034$), yet sustained enforcement rescues the trajectory: the bare rollout diverges while the corrected rollout remains close to the reference, giving $B_{\mathrm{rollout}} \approx 1.0$ (Figure 4).

A single-intervention control locates this gain in the autoregressive feedback loop. Applying the same certificate correction at one early step and then releasing the surrogate changes the final error by only $B_{\mathrm{rollout}} \approx 0.002$. Under sustained enforcement, each corrected state becomes the input to the next surrogate evaluation, altering the subsequent trajectory. The local bound describes the immediate correction; the long-horizon gain reflects repeated interaction with the autoregressive dynamics.

5.2 Monotonicity does not determine kinetics

A generic backtracking line search reaches the same monotone energy target as the temporal corrector and earns the same empirical certificate. Its fitted coarsening exponent is 0.001 against the reference 0.35: Lyapunov-stable and dynamically frozen. Following the gradient-flow descent direction instead produces exponents 0.41 (monotone) and 0.35 (rate target) on the same setting. The sign of ΔF constrains neither the mechanism nor the rate of the evolution.

5.3 Lower field error can favor the wrong dynamics

The mis-specified-energy control changes only the functional inside the corrector (the interface width is scaled by 2.5). On the weak-surrogate Allen–Cahn setting it attains final relative $L_2 = 0.215$, lower than every physically specified correction, while over-coarsening with exponent 0.64 against the reference 0.35 (Figure 5). Snapshot accuracy alone would rank this trajectory above the physically faithful ones. Field error and dynamical fidelity measure different properties, and a correction can trade one against the other.

5.4 Evaluating dynamics alongside certification and accuracy

The controls above give a concrete measurement rule. State-level constraints should be checked with state-level quantities; temporal laws should be checked along the rollout, since snapshot error does not reveal energy ordering; and dynamical fidelity needs its own observable, here the coarsening

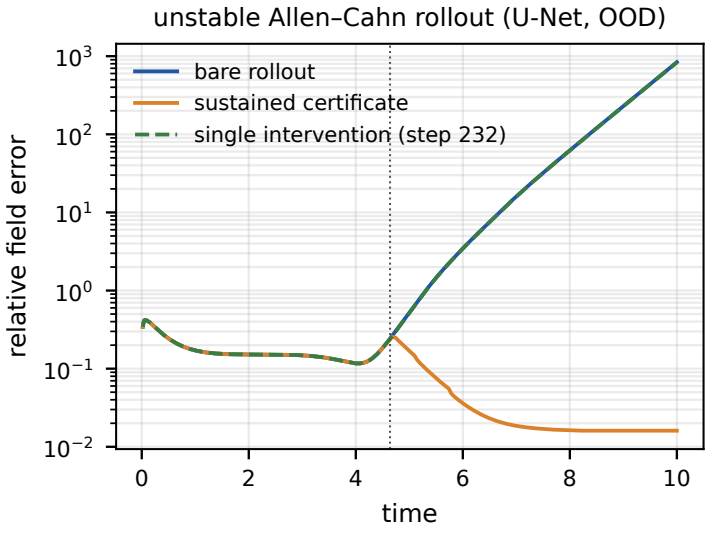

Figure 4: **Amplification requires sustained correction.** Field error against the reference along an unstable Allen–Cahn rollout (U-Net, thin-interface shift). The bare rollout diverges; sustained certificate enforcement tracks the reference ($B_{\text{rollout}} \approx 1.0$) despite a small per-step budget ($R \approx 0.034$); a single certified intervention at an early step leaves the rollout to re-diverge ($B_{\text{rollout}} \approx 0.002$).

rate. The wrong-functional and generic-descent controls show that any one of field accuracy, energy monotonicity, and kinetics can rank a physically wrong trajectory favorably, so the three measurements coexist in our evaluation. The dissipation-rate probe adds a calibration lesson: it recovers the reference coarsening exponent on the weak-surrogate setting, but over-dissipates a stronger Allen–Cahn rollout and is frequently infeasible for Cahn–Hilliard. A rate target is compatible dynamical information in some regimes and a mismatched prior in others.

6 Scope and Discussion

Deployment boundary. On the corresponding shorter-horizon 2D benchmark, a coarse-step energy-stable solver at matched wall-clock reaches relative $L_2 = 0.006$ with exponent 0.469, more accurate and dynamically closer than surrogate plus enforcement ($L_2 \approx 2.18$, exponent 0.305). Where a classical solver is an acceptable forward model, it should be preferred. The construction addresses the complementary deployment condition: a learned surrogate is already the forward model—inside a larger pipeline, an amortized design loop, or a setting without a trusted solver—and a trajectory-level certificate and a quantified correction signal are required.

Formal scope. The Lyapunov certificate applies to gradient-flow systems with an evaluable free energy and is enforced around a fixed pretrained surrogate. Other temporal structures require guarantees matched to their governing physics. The correctability result has two levels of scope. The inequality $B_{\text{local}} \leq R$ is a deterministic geometric statement for any correction under the chosen norm. The empirical relation between correction capacity, alignment, and realized benefit is validated here for temporal free-energy enforcement across the systems, shifts, architectures, and training seeds studied above.

Implications for constrained surrogates. Constraint violation, field error, and a dynamics-relevant observable answer different questions. Violation measures whether the encoded physical relation holds along the rollout. Field error measures agreement with a reference state. A kinetic observable tests whether the trajectory evolves at the correct rate. Reporting only constraint violation can therefore hide inaccurate certified rollouts, while field error alone can favor a trajectory generated using the wrong physical functional, as in Fig. 5.

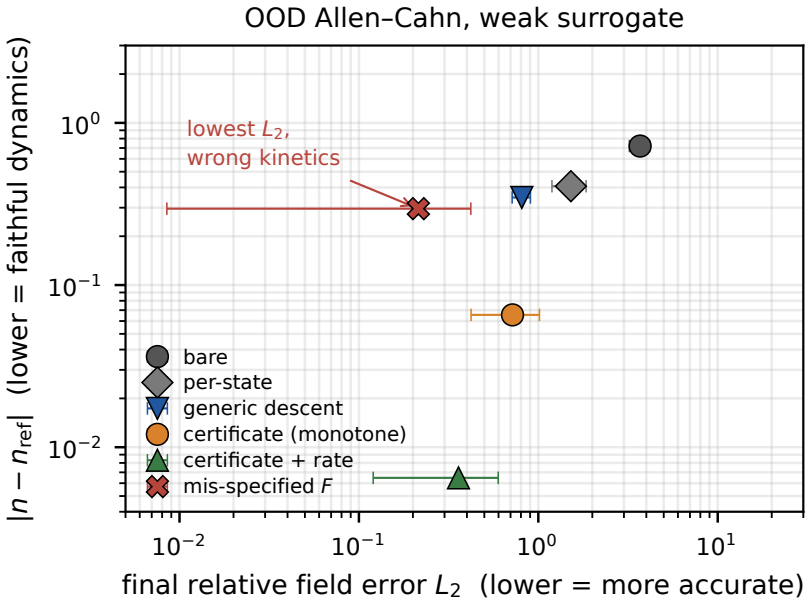

Figure 5: **Certification and accuracy do not determine dynamics.** Final field error against coarsening-exponent deviation on OOD Allen–Cahn (weak surrogate). The physically specified corrections (monotone, rate target) achieve low field error with near-reference kinetics; generic energy descent satisfies the same energy certificate while nearly freezing coarsening; the mis-specified functional attains the lowest field error of the comparison while over-coarsening. Exponents are fitted over the growth window; error bars are standard deviations across trajectories.

Certificate activation remains observable without a reference trajectory. We use R, B_{local}, and B_{rollout} as evaluation quantities when reference solutions are available. The geometric bound and alignment decomposition apply to any correction-returning constraint under the chosen norm; their empirical informativeness is established here for temporal free-energy constraints in gradient-flow rollouts. For constraint design, the results suggest examining the correction itself: useful corrective constraints should induce displacements that are sufficiently large and aligned with relevant prediction errors. A certificate with consistently small R can provide a strong structural guarantee while offering little direct field correction.

7 Conclusion

We asked when physical constraint enforcement corrects autoregressive neural PDE rollouts. For gradient flows, temporal sublevel enforcement supplies a discrete Lyapunov certificate that state feasibility cannot. Its correction content is quantified by two numbers: the certificate-to-error ratio R, bounding the local benefit ($B_{\mathrm{local}} \leq R$), and the correction–error alignment, deciding its sign ($B_{\mathrm{local}} \approx R\cos\theta$). Autoregressive feedback can amplify a small sustained local correction into a complete rescue; dynamical fidelity remains a separate measurement. Constraint satisfaction is an incomplete proxy for prediction quality: certification, correction, and dynamics are distinct measurable properties.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

The formal assumptions and proof details for Propositions 1–2 are given in Appendix A. Appendix B specifies the PDEs, discrete energies, reference solvers, data generation, and distribution shifts; Appendix C records surrogate training and the distinction between weak, strong, and penalty-trained models. Appendix D gives the complete inference-time algorithm, correction budgets, fallback, and the fact that correction iterations do not advance physical time. Appendices E–I report full quantitative, seed-level correctability, negative-control, sensitivity, and equal-compute results.

The supplementary code contains the data generators, reference solvers, surrogate backbones, physical correctors, experiment configurations, evaluation scripts, and archived measurements used for the reported analyses. The supplied scripts and measurements regenerate the principal correctability, alignment, amplification, sensitivity, and control analyses, together with the reported tables and figures.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for providing feedback on mathematical arguments and experimental design, and assisting with the interpretation and presentation of experimental results. We have not used generative AI tools for generating or modifying experimental measurements, and other required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for literature discovery, manuscript drafting and language editing, LaTeX and code assistance, figure layout and presentation, and simulated peer-review feedback. We have reviewed all AI-assisted work. We checked LLM-generated research ideas for potential plagiarism through a manual literature survey, LLM-generated code was verified and tested for correctness by 2 authors. AI-assisted scientific suggestions were checked against the underlying mathematical derivations and experimental artifacts; references identified with AI assistance were verified against the original publications; and AI-assisted code, figures, and LaTeX were inspected for correctness. Experimental figures and quantitative results were produced from verified experimental measurements rather than AI-generated data. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

T. Beucler, M. Pritchard, S. Rasp, J. Ott, P. Baldi, and P. Gentine. Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters*, 126(9):098302, 2021.

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

Chaoran Cheng, Boran Han, Danielle Maddix, Abdul Fatir Ansari, Andrew Stuart, Michael W. Mahoney, and Bernie Wang. Gradient-free generation for hard-constrained systems. In *International Conference on Learning Representations*, 2025.

D. J. Eyre. Unconditionally gradient stable time marching the cahn–hilliard equation. In *MRS Proceedings*, volume 529, pp. 39–46, 1998.

Xinquan Huang and Paris Perdikaris. Physicscorrect: A training-free approach for stable neural pde simulations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22057–22065, 2026. doi: 10.1609/aaai.v40i26.39360.

Yunfei Huang and David S. Greenberg. Geometric and physical constraints synergistically enhance neural pde surrogates. In *International Conference on Machine Learning*, 2025.

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

Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. Pde-refiner: Achieving accurate long rollouts with neural pde solvers. In *Advances in Neural Information Processing Systems*, 2023.

L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3): 218–229, 2021.

Michael McCabe, Peter Harrington, Shashank Subramanian, and Jed Brown. Towards stability of autoregressive neural operators. *Transactions on Machine Learning Research*, 2023.

T. Pfaff, M. Fortunato, A. Sanchez-Gonzalez, and P. W. Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021.

M. Raissi, P. Perdikaris, and G. E. Karniadakis. 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.

O. Rochman Sharabi and G. Louppe. Enforcing governing equation constraints in neural pde solvers via training-free projections. *arXiv preprint arXiv:2511.17258*, 2025.

A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, and P. W. Battaglia. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning*, 2020.

J. Shen, J. Xu, and J. Yang. The scalar auxiliary variable (sav) approach for gradient flows. *Journal of Computational Physics*, 353:407–416, 2018.

J. Shen, J. Xu, and J. Yang. A new class of efficient and robust energy stable schemes for gradient flows. *SIAM Review*, 61(3):474–506, 2019.

M. Singha. Learning under laws: A constraint-projected neural pde solver that eliminates hallucinations. *arXiv preprint arXiv:2511.03578*, 2025.

K. Stachenfeld, D. B. Fielding, D. Kochkov, M. Cranmer, T. Pfaff, J. Godwin, C. Cui, S. Ho, P. Battaglia, and A. Sanchez-Gonzalez. Learned coarse models for efficient turbulence simulation. *arXiv preprint arXiv:2112.15275*, 2021.

Matilde Valente, Tiago C. Dias, Vasco Guerra, and Rodrigo Ventura. Physics-consistent machine learning with output projection onto physical manifolds. *Communications Physics*, 8:433, 2025.

Hao Wei, Aleksandra Franz, Bjoern List, and Nils Thuerey. Inc: An indirect neural corrector for auto-regressive hybrid pde solvers. In *Advances in Neural Information Processing Systems*, 2025.

Yue Wu, Tianyu Jin, Chuqi Chen, Ganghua Fan, Yuan Lan, Luchan Zhang, and Yang Xiang. Energy stable neural network for gradient flow equations. *arXiv preprint arXiv:2309.10002*, 2023.

A Formal Results and Proof Details

A.1 Notation and discrete setting

Let $\Omega = [0,1]^d$ be discretized on a periodic uniform grid with spacing Δx. For the scalar phase-field systems, we use the quartic double-well potential

$$W(\phi) = \frac{1}{4}(\phi^2 - 1)^2,$$

and the discrete free energy

$$F_h[\phi] = \sum_{x_i} \left(\frac{\varepsilon^2}{2} |\nabla_h \phi(x_i)|^2 + W(\phi(x_i)) \right) \Delta x^d. \tag{7}$$

The discrete mass is $M_h(\phi) = \sum_{x_i} \phi(x_i) \Delta x^d$. The monotone construction uses the target $\tau_n = F_h[\phi_n]$ and, for conserved flows, intersects the energy sublevel set with $M_h(\phi) = M_h(\phi_0)$. All statements below concern these discrete quantities.

A.2 Per-state feasibility and the missing implication

Proposition 1. Assume each feasible set $\mathcal{A}_n$ contains states of arbitrarily large F_h. For every n, select $\psi_n \in \mathcal{A}_n$ satisfying $F_h[\psi_n] \geq F_h[\phi_0] + n$. Then ψ_n is feasible at every step while $F_h[\psi_n] \to \infty$. The premise holds for common unbounded feasibility sets such as affine conservation constraints and divergence-free subspaces. Residual-defined feasible sets require separate boundedness analysis; the proposition applies to them only when their feasible sets remain unbounded in energy. The result does not cover a state set that explicitly includes the desired energy bound. Its role is to formalize that state feasibility alone supplies no temporal Lyapunov implication.

A.3 Constructive feasibility and Lyapunov certificate

For the monotone target, the implementation first restores the affine conservation constraint when needed and then applies an unconditionally energy-decreasing convex-splitting update. If the capped descent budget does not enter $\{F_h \leq F_h[\phi_n]\}$, the implementation bisects the segment between the descended state and ϕ_n. The previous accepted state is a feasible endpoint because $F_h[\phi_n] = \tau_n$; continuity of F_h therefore guarantees a feasible crossing on the segment. The bisection maintains feasible and infeasible endpoints and returns the bracketed feasible point closest to the descended state after the fixed bisection budget. Hence the returned state obeys $F_h[\phi_{n+1}] \leq F_h[\phi_n]$. Induction gives $F_h[\phi_n] \leq F_h[\phi_0]$ for all n. When $W \geq 0$, the interfacial and bulk contributions are separately bounded by the initial discrete energy. This is the certificate asserted in Proposition 2; it is not a bound on field error or a convergence theorem for the learned trajectory.

A.4 Rate-target feasibility

The rate target $\tau_n^{\mathrm{rate}} = F_h[\phi_n] - \gamma \Delta t_S D_n$ is stricter than the monotone target. The previous accepted state can therefore be infeasible, so the monotone fallback cannot certify this target. The implementation returns the best descended iterate and records the residual when the target is not reached. This distinction is why the rate intervention is treated as a dynamical probe and not as part of the unconditional certificate.

B PDE Systems, Data, and Reference Solvers

B.1 Allen–Cahn and Cahn–Hilliard data generation

The supplied generator uses periodic 128×128 grids by default. Allen–Cahn trajectories use $\varepsilon = 0.02$, solver step 10^{-4}, 20,000 reference steps, and save every 200 steps; Cahn–Hilliard uses $\varepsilon = 0.02$, solver step 2×10^{-5}, 40,000 reference steps, and save every 400 steps. Both use mobility 1.0 in distribution. Random spinodal initial conditions are generated as $\phi_0 = \bar{\phi} + \eta$, with independent $\eta \sim \mathrm{Uniform}(-0.1, 0.1)$ at each grid point. In distribution, $\bar{\phi}$ is sampled uniformly from $[-0.05, 0.05]$ for Allen–Cahn and from $[-0.2, 0.2]$ for Cahn–Hilliard. All draws use explicit seeds. The default training set contains 160 trajectories.

B.2 Distribution shifts

The archived generator defines thinner interfaces by $\varepsilon = 0.013$, thicker interfaces by $\varepsilon = 0.030$, high mobility by mobility 2.0, and low mobility by mobility 0.5. Separate scripts evaluate asymmetric composition, longer horizons, and resolution changes. Each generated trajectory stores the physical parameters and seed alongside the state sequence.

B.3 Reference-solver checks

Reference trajectories are produced by the supplied Fourier-spectral solvers. Evaluation records the discrete energy and, for conserved flows, mass. The main experiments use the same reference trajectory, physical horizon, and metric definition across compared interventions. Correction iterations operate between surrogate emissions at fixed rollout index; they do not increment the physical time used for the reference comparison.

Coarsening observable. We measure the characteristic domain length from the first moment of the structure factor,

$$L(t) = \frac{2\pi}{\langle k \rangle}, \qquad \langle k \rangle = \frac{\sum_{\mathbf{k}} |\mathbf{k}|\, |\hat{\phi}_{\mathbf{k}}|^2}{\sum_{\mathbf{k}} |\hat{\phi}_{\mathbf{k}}|^2},$$

where the spatial mean is removed before the Fourier transform. The coarsening exponent n is obtained from $L(t) \propto t^n$ by linear regression of $\log L(t)$ against $\log t$ over the final 60% of each finite trajectory. Exponents are fitted separately for each trajectory and then averaged across evaluated trajectories.

C Surrogate Models and Training

The supplied training driver supports FNO, periodic U-Net, and DeepONet backbones. Its default FNO uses 32 Fourier modes and width 64. The periodic U-Net has three resolution levels with channel widths 32, 64, and 128, using circular-padding convolutions, average pooling, bilinear upsampling, and skip connections. The DeepONet uses a fixed 16×16 sensor grid, branch and trunk hidden width 256, latent dimension 128, and eight Fourier frequencies for periodic coordinate features. Training uses Adam with learning rate 10^{-3}, cosine annealing, and batch size 32. The default training driver exposes 60 epochs, while the reported comparisons also include weak and 15-epoch checkpoints. Unless a physics penalty is explicitly enabled, training minimizes data loss only. The penalty-trained control adds a free-energy term through the same training driver; it is used only to distinguish training-time regularization from a per-rollout inference certificate.

DeepONet robustness experiments. The supplementary DeepONet experiments provide additional evidence across surrogate architectures. The zero-violation results reported in the main paper are evaluated using the monotone implementation described below.

D Inference-Time Enforcement

D.1 Complete construction

For each accepted state ϕ_n, the surrogate emits $\widehat{\phi}_{n+1} = S_\theta(\phi_n)$. For conserved systems, Stage 1 projects the prediction onto the mass hyperplane by a uniform scalar shift,

$$\phi \leftarrow \phi + \frac{M_0 - M_h(\phi)}{N \Delta x^2},$$

where N is the number of grid points; this restores the mass of the previous accepted state exactly.

Stage 2 evaluates the chosen energy target and applies at most $\kappa_{\max}$ convex-splitting descent steps. The implementation uses $\kappa_{\max} = 10$, a stabilization parameter of 6.0, 24 bisection steps for the monotone backstop, and a rate fraction of 0.5 for the rate-target probe. The wrong-F negative control uses the same monotone construction with ϵ multiplied by 2.5 in the corrector energy.

D.2 No-op behavior

If the emitted state already satisfies the mass and temporal energy conditions, the monotone intervention returns it without descent. This is the mechanism behind the no-op cases reported for compliant strong-FNO and U-Net rollouts.

D.3 TIME INTERPRETATION

The inner correction iterations are feasibility operations at the same surrogate rollout index. They do not represent additional increments of physical time. Runtime comparisons nevertheless include their wall-clock cost.

E FULL QUANTITATIVE AND ROBUSTNESS RESULTS

		violation rate ↓		final rel. L_2 ↓	
surrogate (training)	condition	bare	+certificate	bare	+certificate
Allen–Cahn					
FNO, 15 ep.	in-distribution	**0.000**	**0.000**	0.14 ± 0.09	0.14 ± 0.09
FNO, 15 ep.	OOD thin ε	0.393	**0.000**	12.22 ± 6.83	0.34 ± 0.46
FNO, 2 ep. (weak)	OOD thin ε	0.541	**0.000**	3.69 ± 0.48	0.72 ± 0.30
FNO, 15 ep. + phys. penalty	OOD thin ε	0.571	**0.000**	3.6×10^3	0.68 ± 0.66
U-Net, 15 ep.	OOD thin ε	**0.000**	**0.000**	0.32 ± 0.13	0.32 ± 0.13
DeepONet, 15 ep.†	OOD thin ε	0.308	0.039	8.8×10^9	0.78
Cahn–Hilliard					
FNO, 15 ep.	in-distribution	**0.000**	**0.000**	0.51 ± 0.42	0.52 ± 0.42
FNO, 15 ep.	OOD thin ε	0.187	**0.000**	1.38 ± 0.22	1.33 ± 0.15
FNO, 2 ep. (weak)	OOD thin ε	0.567	**0.000**	1.55 ± 0.17	1.52 ± 0.19
FNO, 15 ep. + phys. penalty	OOD thin ε	0.043	**0.000**	1.13 ± 0.24	1.14 ± 0.26
U-Net, 15 ep.	OOD thin ε	0.002	**0.000**	1.40 ± 0.03	1.35 ± 0.34
DeepONet, 15 ep.†	OOD thin ε	0.368	0.357	∞	1.41

Table 2: **The findings are not a weak-surrogate artifact.** Across surrogate qualities (2-epoch weak, 15-epoch plain, penalty-trained) and three backbones, out-of-distribution rollouts of any quality can violate the free-energy invariant, and temporal enforcement drives the violation rate to zero in every current-code case. Where the bare rollout is already stable (in-distribution FNO; U-Net), temporal enforcement is a bitwise no-op and accuracy is untouched. On Allen–Cahn, where the surrogate error is energy-visible, certification comes with a large accuracy gain; on conserved Cahn–Hilliard the same certificate leaves the iso-energetic field error largely untouched, consistent with the correctability analysis. †DeepONet rows were produced by an earlier corrector revision without the bisection backstop (their residual violation rates 0.039/0.357 are reported in the supplement and removed by the current code). n=4 trajectories per row, 300 OOD / 100 ID steps; mean ± s.d. across trajectories.

E.1 LEAD ALLEN–CAHN COMPARISON

Table 3 reports the lead Allen–Cahn comparison between the uncorrected surrogate and the inference-time enforcement variants.

Intervention	violation rate ↓	final relative L_2 ↓
Unconstrained	0.541 ± 0.045	3.69 ± 0.48
Per-state	0.423 ± 0.069	1.52 ± 0.33
Energy descent	0.000	0.81 ± 0.09
Lyapunov	0.000	0.72 ± 0.30
Dissipation target	0.000	0.36 ± 0.24
Wrong F	0.007 ± 0.010	0.21 ± 0.21

Table 3: Archived out-of-distribution Allen–Cahn comparison for the weak surrogate. Dispersion is across evaluated trajectories. The wrong-F row is a negative control and is not physically specified.

E.2 CERTIFICATION VERSUS CORRECTION

The error decomposition separates structural/energy error from field error. Relative reductions under Lyapunov enforcement are 99.9% versus 75.6% for Allen–Cahn, 87.3% versus 15.8% for grain

growth, and 96.5% versus 1.1% for Cahn–Hilliard. Section 4 explains this ordering quantitatively through the certificate-to-error ratio R and the alignment $\cos\theta$; Appendix F reports the seed-level confirmation.

E.3 Distribution and resolution stress tests

Additional evaluations cover interface-width, mobility, composition, horizon, and resolution shifts. The monotone construction retains the discrete-energy certificate in the current-code configurations reported in the paper. Field-error gains remain system-dependent, and the rate/kinetic conclusions degrade under some shifts. These stress tests support a broader scope for the certificate than for kinetic recovery.

E.4 Archived robustness figures

Figure 6 summarizes the before/after behavior across surrogate qualities and backbones; Figure 7 gives the controlled four-way comparison on the lead OOD Allen–Cahn setting.

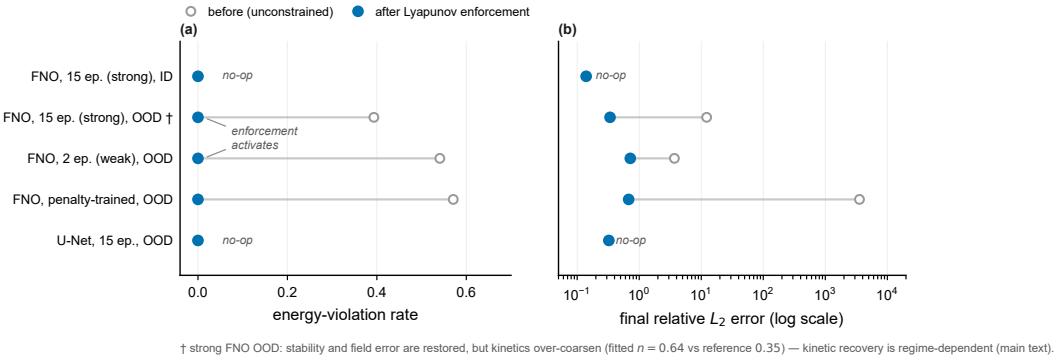

Figure 6: Before/after energy-violation rate and final relative L_2 error across surrogate regimes. Already compliant rollouts are unchanged; enforcement activates under structural drift. † The strong OOD FNO recovers the Lyapunov property and field accuracy but still over-coarsens, separating certification from kinetic fidelity.

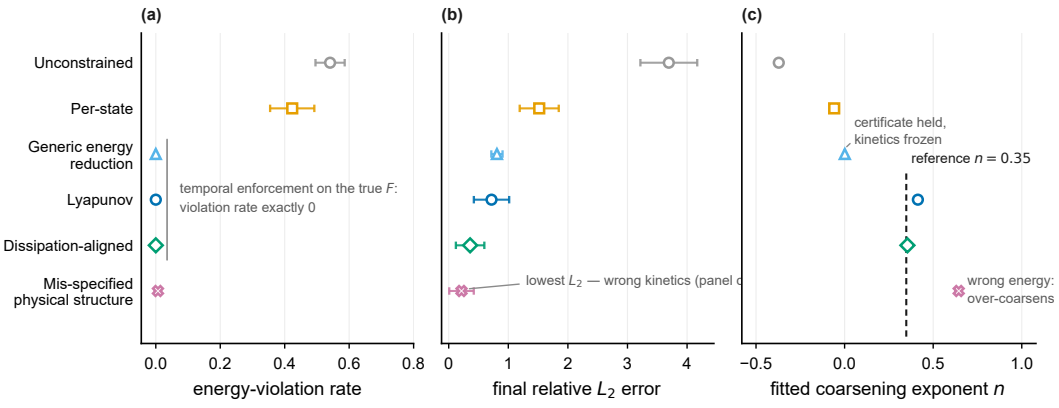

Figure 7: Controlled separation of feasibility, stability, accuracy, and kinetics on OOD Allen–Cahn: energy-violation rate, final relative L_2 error, and fitted coarsening exponent for each control.

F SEED-LEVEL CORRECTABILITY ANALYSIS

This appendix reports the confirmatory analysis behind Section 4. All quantities are computed per rollout step on the archived evaluation trajectories, then aggregated per trajectory, per seed, and per regime.

F.1 DEFINITIONS

For a surrogate prediction $\widehat{\phi}$ with reference ϕ_{ref}, let $e = \phi_{\mathrm{ref}} - \widehat{\phi}$ be the field error and δ the certificate correction returned by the monotone intervention. The per-step quantities are the certificate-to-error ratio $R = \|\delta\|/\|e\|$, the alignment $\cos\theta = \langle \delta, e\rangle/(\|\delta\|\,\|e\|)$, and the normalized same-step benefit $B_{\mathrm{local}} = (\|e\| - \|e - \delta\|)/\|e\|$. The activation rate A is the fraction of steps on which the certificate fires. The rollout-level benefit B_{rollout} is the relative change in final L_2 error versus the bare rollout. Steps with $A < 0.10$ are detection-negative and are excluded from the conditional R–benefit analysis.

F.2 BOUND VERIFICATION AND AGGREGATION

Across 15,412 activated steps spanning 21 surrogate/regime combinations, the bound $B_{\mathrm{local}} \le R$ is never violated (maximum $B_{\mathrm{local}} - R = -6 \times 10^{-7}$, within numerical tolerance). The first-order identity $B_{\mathrm{local}} \approx R\cos\theta$ holds with Spearman $\rho = 0.989$ per step and $\rho = 0.996$ after aggregation to one point per model. The ratio R orders regimes by realized benefit better than the raw energy overshoot ($\rho = 0.895$ versus 0.740), and the ordering transfers to five held-out regimes ($\rho = 1.000$).

F.3 SEED-LEVEL CONFIRMATION

Table 4 repeats the analysis across independent training seeds. On the moderate-drift Allen–Cahn regime (A1), the sign of $R\cos\theta$ predicts whether enforcement corrects or degrades the rollout across all five seeds ($\rho = 1.000$): seeds 0–2 are aligned and improve, seeds 3–4 are anti-aligned and degrade. Cahn–Hilliard (B) remains at $R \le 0.017$, with median $B_{\mathrm{local}} \le 3 \times 10^{-5}$ across all three seeds, confirming certification without appreciable local correction. The healthy in-distribution regime (E) is detection-negative: the certificate is silent and the field error is untouched.

System	Backbone	Regime	Seed	A	R	B_{local}	B_{rollout}
AC	FNO	A1 (thin ε)	0	0.093	0.257	0.249	0.158
AC	FNO	A1	1	0.158	0.084	0.083	0.069
AC	FNO	A1	2	0.064	0.183	0.183	0.106
AC	FNO	A1	3	0.591	0.033	-0.026	-0.056
AC	FNO	A1	4	0.501	0.047	-0.028	-0.017
AC	U-Net	A2 (thin ε)	0	0.947	0.037	-0.011	0.086
AC	U-Net	A2	1	0.565	0.034	-0.003	**1.000**
AC	U-Net	A2	2	0.000	—	—	0.000
CH	FNO	B (thin ε)	0	1.000	0.017	0.00002	0.008
CH	FNO	B	1	1.000	0.015	0.00002	0.022
CH	FNO	B	2	1.000	0.013	0.00003	-0.040
AC	FNO	D (high mobility)	0	0.000	—	—	0.000
AC	FNO	D	1	0.043	0.113	0.113	0.100
AC	FNO	D	2	0.000	—	—	0.000
AC	FNO	D	3	0.387	0.042	-0.017	-0.042
AC	FNO	D	4	0.309	0.069	-0.017	-0.004
AC	FNO	E (in-distribution)	0–4	0.000–0.033	—/0.062	—/0.034	0.000–0.020

Table 4: Seed-level confirmation of the correctability mechanism. A is the activation rate; R is the median certificate-to-error ratio on activated steps; B_{local} is the median per-step field correction; B_{rollout} is the relative final-L_2 change versus the bare rollout. Dashes mark detection-negative cases ($A < 0.10$), excluded from the conditional analysis. The A2 seed-1 row is the unstable long-horizon case in which sustained correction amplifies a small local budget into a full rescue.

F.4 DETECTION, LOCAL CORRECTABILITY, AND DYNAMICAL AMPLIFICATION

The three stages are separable. (i) *Detection*: a large field error can be invisible to the certificate (regime E and A2 seed 2: $A = 0$, $B = 0$). (ii) *Local correctability*: on activated steps, $B_{\mathrm{local}} \approx R\cos\theta \le R$; the certificate's capacity is bounded by R and signed by the alignment. (iii) *Dynamical*

amplification: sustained correction can amplify a small local budget into a full rescue (A2 seed 1: $R = 0.034$ per step, $B_{\text{rollout}} = 1.000$), whereas a single intervention on the same trajectories leaves $B_{\text{rollout}} \approx 0.002$. Local correctability therefore does not determine the rollout-level outcome; autoregressive feedback does.

G Dynamics and Negative Controls

Generic energy descent. The backtracking line-search control searches along the segment from the previous accepted state to the surrogate prediction,

$$\phi(\alpha) = \phi_n + \alpha(\hat{\phi}_{n+1} - \phi_n), \qquad \alpha \in [0, 1],$$

halving α until the same monotone energy target used by the certificate is reached. This direction does not follow the physical gradient flow. Its near-zero coarsening exponent on the lead setting demonstrates that satisfying the sign condition on ΔF does not determine the evolution rate.

Mis-specified energy. The wrong-F control changes only the energy used by the monotone corrector. It achieves low pointwise error while over-coarsening, separating projection-induced field improvement from fidelity to the intended physical functional.

Rate-target failures. The rate target recovers the reference exponent in the weak-surrogate lead Allen–Cahn experiment, but over-coarsens for the stronger Allen–Cahn surrogate and is frequently unreached for conserved Cahn–Hilliard. These cases are retained because they delimit what additional dynamical information provides beyond monotonicity.

H Sensitivity and Implementation Robustness

The archived sensitivity sweep varies correction budget, descent threshold, and stabilization. For the monotone intervention, $\kappa_{\max} = 1$ still yields zero violations through the fallback but triggers fallback on about 43% of evaluated steps in that sweep. Increasing the budget to 2 removes fallback and gives a mean of roughly 0.95 inner iterations; budgets 5, 10, and 20 then produce the same archived result. Across the tested descent-threshold and stabilization values, the monotone certificate remains intact while field error and kinetic observables vary, as expected for quantities not guaranteed by the certificate.

I Computational Cost and Solver Boundary

Table 5 reports wall-clock measurements obtained on a single NVIDIA RTX 5080 GPU using PyTorch 2.11 with CUDA 12.8. The surrogate and classical spectral solver are both GPU-resident. Each timed run is preceded by warm-up iterations, and CUDA is synchronized immediately before and after timing.

Method	wall time (s)	rel. $L_2 \downarrow$	exponent n
Reference solver 128^2	3.39	0	0.47
Coarse-step solver 128^2	0.34	0.01	0.47
Surrogate	0.21	3.07	0.24
Surrogate + Lyapunov	0.61	2.18	0.31
Surrogate + rate target	0.59	2.17	0.31

Table 5: Archived equal-compute Allen–Cahn comparison over the corresponding shorter horizon. The inexpensive classical solver Pareto-dominates surrogate plus enforcement on this benchmark.

This comparison defines the deployment boundary, not a speed claim. The construction is relevant when a learned surrogate is already the forward model and a structural rollout property is required.

Table 6: Evidence map for the principal experimental claims and their supporting supplementary results.

Claim	System	Setting	Evaluation	Supporting results
State–trajectory separation	AC	Weak FNO, OOD	Lyapunov enforcement and coarsening behavior	Lead comparison and coarsening results
Strong-model activation and no-op behavior	AC	15-epoch FNO, ID/OOD	Certificate activation and no-op behavior	FNO evaluation results
No-op behavior across architectures	AC	U-Net, OOD	Certificate activation and no-op behavior	U-Net evaluation results
Certification without field correction	CH	FNO	Energy-error and field-error decomposition	CH evaluation results
Incorrect-functional control	AC	Weak FNO	Enforcement under a misspecified energy functional	Wrong-functional control results
Rate-target overdissipation	AC	Strong FNO	Coarsening and rollout diagnostics	Rate-target evaluation results
Backbone robustness	AC/CH	DeepONet	Cross-architecture robustness evaluation	DeepONet evaluation results
Solver boundary	AC	Weak FNO	Matched wall-clock comparison	Solver comparison results

J Relationship to Prior Constraint Methods

Prior work spans training-time analytic constraints, post-training projection onto physical manifolds, training-free PDE projection, and hard-constrained generation. The closest distinction used in this paper is operational: whether membership in the enforced feasible set itself implies the temporal Lyapunov ordering being claimed. A constraint may reference the previous state and still fail to imply that ordering. The numerical projection mechanism is therefore not claimed as novel; the contribution is the state–trajectory separation, transfer of the discrete certificate to a fixed learned rollout, and the controlled evidence separating certification, field correction, and dynamics.

K Reproduction workflow and artifact map

The supplementary code is organized into `correctors/`, `solvers/`, `energy/`, `models/`, `eval/`, `scripts/`, and `results/`. A reproducer can (i) regenerate seeded AC/CH trajectories with `python -m ipc.scripts.gen data`, (ii) train FNO/U-Net/DeepONet surrogates with `python -m ipc.scripts.train`, (iii) execute the evaluation drivers with `ipc/scripts/run evals.sh`, and (iv) regenerate tables from the JSON measurements with `python -m ipc.scripts.make tables`. The bundled measurement files permit direct inspection of reported values without rerunning training. Data generation uses both `torch.Generator().manual seed` and `numpy.random.default rng`; each trajectory stores its seed and physical parameters.

L Evidence and Reproducibility Summary

Table 6 summarizes the experimental evidence supporting the principal claims and points to the corresponding supplementary results. The table is intended as a compact guide to the evaluations reported across the main paper and appendices.

M Additional Configuration and Diagnostic Tables

This appendix collects configuration details and diagnostic sweeps that support reproduction but are not needed for the main argument. We report the default data-generation settings, distribution shifts, correction parameters, and sensitivity to the inner correction budget.

M.1 Data-generation configuration

Table 7 records the default scalar-system simulation parameters used to generate the reference trajectories. The OOD evaluations modify individual physical parameters or rollout conditions while retaining the same generation and evaluation pipeline; these shifts are summarized in Table 8.

System	grid	ε	solver Δt	reference steps	save stride
Allen–Cahn	128^2	0.020	10^{-4}	20,000	200
Cahn–Hilliard	128^2	0.020	2×10^{-5}	40,000	400

Table 7: Default scalar-system data-generation settings in the supplied generator. OOD shifts override individual physical parameters while retaining the same generation code path.

Shift	in-distribution value	shifted value
thin interface	$\varepsilon = 0.020$	$\varepsilon = 0.013$
thick interface	$\varepsilon = 0.020$	$\varepsilon = 0.030$
high mobility	$M = 1.0$	$M = 2.0$
low mobility	$M = 1.0$	$M = 0.5$
long horizon	nominal horizon	generator multiplier
asymmetric composition	nominal mean distribution	dedicated asymmetric split
resolution	128^2	dedicated zero-shot resolutions

Table 8: Distribution-shift definitions recoverable from the supplied generator and evaluation scripts.

M.2 Correction defaults

The inference-time correction uses a small set of numerical parameters that control how the feasible state is reached after each surrogate prediction. Table 9 reports the defaults used in the archived experiments. These parameters affect the numerical realization of the correction, while the Lyapunov certificate itself follows from accepting only states that satisfy the temporal free-energy condition. In particular, the dissipation-rate target controlled by γ is an additional dynamical probe and is not required for the monotonicity guarantee.

Parameter	Default	Role
$\kappa_{\max}$	10	Maximum convex-splitting descent steps
Stabilization S	6.0	Convex-splitting stabilization
Rate fraction γ	0.5	Fraction of physical dissipation in rate target
Bisection iterations	24	Monotone feasibility backstop
Wrong-F scale	2.5ε	Mis-specified-energy negative control

Table 9: Default inference-time parameters used in the archived experiments. $\kappa_{\max}$ and the bisection backstop govern how the temporal feasible set is reached. The rate fraction γ is used only by the dissipation-target intervention and is not part of the unconditional Lyapunov certificate. The wrong-F scale is used exclusively for the negative control.

M.3 Sensitivity to correction budget

We next isolate the effect of the maximum number of convex-splitting descent steps, $\kappa_{\max}$, while keeping the remaining correction settings fixed. This sweep tests whether the reported behavior

$\kappa_{\max}$	Violation rate	Final L_2	Exponent	Fallback rate
1	0.000	0.882	0.191	0.431
2	0.000	0.865	0.414	0.000
5	0.000	0.865	0.414	0.000
10	0.000	0.865	0.414	0.000
20	0.000	0.865	0.414	0.000

Table 10: Sensitivity to the maximum correction budget $\kappa_{\max}$ in the archived diagnostic run. The energy-violation rate remains zero for every budget because the feasibility backstop preserves the temporal condition. With $\kappa_{\max} = 1$, the backstop is invoked on 43.1% of steps; two descent steps eliminate fallback in this run, after which the reported field error and coarsening exponent are unchanged through $\kappa_{\max} = 20$. Field accuracy and kinetics are empirical outcomes and are not implied by the Lyapunov certificate.

depends on a large inner correction budget. The temporal condition is satisfied throughout the sweep because the feasibility backstop remains available; increasing $\kappa_{\max}$ instead determines how often that backstop is needed and how closely the accepted update follows the descent construction before fallback.

The sweep therefore separates the correction budget from the structural guarantee. For this diagnostic trajectory, increasing the budget beyond two steps provides no measurable benefit in the reported observables. The $\kappa_{\max} = 1$ case also illustrates the role of the backstop: it can preserve feasibility when the descent budget is exhausted, without implying that the resulting field or kinetics are optimal.

N Archived Measurement Inventory

The anonymous supplement includes machine-readable measurements for the main Allen–Cahn and Cahn–Hilliard ID/OOD runs, weak-surrogate and strong-surrogate variants, U-Net and legacy DeepONet runs, coarsening fits, grain-growth kinetics, error decomposition, resolution stress tests, Swift–Hohenberg stress tests, sensitivity sweeps, diagnostics, wall-clock measurements, and the equal-compute Pareto comparison. Keeping these measurements alongside the scripts separates reported results from manuscript transcription and permits direct audit without retraining.

Evidence family	Archived measurement coverage
Allen–Cahn	ID, thin-interface OOD, high-mobility OOD, weak FNO, strong FNO, U-Net, legacy DeepONet, coarsening, lead falsification
Cahn–Hilliard	ID, thin-interface OOD, high-mobility OOD, asymmetric-composition OOD, weak FNO, U-Net, legacy DeepONet, coarsening
Other systems	grain-growth rollout/kinetics; Swift–Hohenberg stress test
Diagnostics	error decomposition, correction diagnostics, sensitivity, resolution
Compute	AC/CH wall-clock and equal-compute Pareto measurements

Table 11: Coverage of the archived JSON measurement layer shipped with the supplement. This table describes available evidence; not every archived run supports a main-paper claim.