# BETTER SIMULATIONS DO NOT GUARANTEE BETTER SURROGATES: A PAIRED STUDY OF NUMERICAL FIDELITY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Numerical simulations approximate PDE solutions, but their usefulness as training data also depends on the dynamics learned from them. We investigate this distinction with PDE-FIDELITY, a paired benchmark that varies the polynomial degree of one, three, and five in the discontinuous Galerkin (DG) method while keeping physical instances and the observation interface fixed. This design separates numerical label accuracy, prediction from common reference histories, and autonomous rollout. Across four systems and seven architectures, degree-five supervision improves label accuracy and aggregate one-step prediction over degree one, yet the rollout benefit is system-dependent and can reverse. In GLM–MHD, the aggregate degree-five/degree-one rollout error ratio is roughly three; the reversal persists under the tested training controls and is confirmed for U-Net on fresh physical instances. An exact DG construction shows that such a reversal can arise even at the unique training-loss minimizers. Within this construction, enforcing mean preservation removes the reversal without adding learned parameters, demonstrating how model structure can change the value of a numerical source. Finite-horizon propagation yields conditional vector-error bounds that retain defect directions and temporal interactions. Corresponding reconstruction experiments show close agreement with neural rollout errors over short windows, with model-dependent limits at longer horizons. Complementary finer-grid experiments also produce substantial U-Net gains, showing that accurate supervision can help without implying a universal source ranking. These results motivate evaluating numerical supervision through the autonomous dynamics learned from it, rather than inferring its value from label or one-step accuracy alone.

## 1 INTRODUCTION

Neural surrogates offer a way to reduce the cost of repeatedly simulating physical systems governed by partial differential equations (PDEs). They learn from numerical trajectories and are then used to predict how a system evolves. Benchmarks such as PDEBench, The Well, and APEBench have made diverse simulation data and evaluation tools available for this purpose (Takamoto et al., 2022; Ohana et al., 2024; Koehler et al., 2024). For these models, the numerical trajectories define the supervision from which the dynamics are learned.

This suggests a natural expectation: more accurate simulations should help train better forecasters. Obtaining such data can require substantial additional computation (Ohana et al., 2024), even when all trajectories are stored on the same observation grid. Yet an autonomous forecaster repeatedly uses its own predictions as inputs. Its trajectory depends not only on the errors it introduces, but also on how it responds to earlier deviations. More accurate labels—and even better one-step predictions—therefore need not guarantee better closed-loop forecasts.

We investigate this distinction with PDE-FIDELITY, a paired PDE dataset spanning four systems. For each physical identity, trajectories are evolved independently at discontinuous Galerkin (DG) degrees $p \in \{1, 3, 5\}$ and stored through a common $128^2$ observation interface. Against common finer-grid references, we compare numerical-label error, one-step error from reference histories, and closed-loop forecast error, with seven architectures evaluated under matched training and evaluation



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.

protocols. Numerical accuracy is measured rather than inferred from polynomial degree, allowing us to track whether its benefits survive learning and feedback.

Relative to degree-one supervision, degree-five supervision improves aggregate one-step reference accuracy across all four systems, but closed-loop outcomes differ. Allen–Cahn retains the benefit, GLM–MHD reverses the aggregate ranking, and aggregate rollout gains remain unresolved for Richtmyer–Meshkov instability (RMI) and shallow water. In GLM, the U-Net/FNO reversal persists after equal-budget second-stage adaptation. Yet a finer-grid source $H$, restricted to the same observation interface, benefits U-Net, with confirmation on new physical identities and training seeds. These findings do not imply a general preference for less accurate data.

Our contributions are:

- **A paired PDE dataset and benchmark.** We introduce PDE-FIDELITY, pairing independently evolved numerical trajectories by physical identity across four systems at a fixed observation resolution. Matched protocols and seven-architecture baselines provide a framework for studying how numerical accuracy translates into learning value.
- **Controlled studies of numerical supervision.** We document system-dependent closed-loop outcomes despite aggregate one-step improvements, and use controlled GLM studies to examine feedback, training, and source choice.
- **A constructive mechanism and rollout-error analysis.** We give an exactly analyzable DG construction in which numerical sources and matched finite training produce reversal. Within this construction, a one-scalar, mean-preserving architectural control restores source ordering under exact fitting. This establishes a possible mechanism without attributing the GLM observations to it. Finite-horizon analysis further supports error reconstruction in trained models with explicit nonlinear remainders and measured limits.

## 2 RELATED WORK

**Numerical PDE data.** PDEBench, PDEArena, The Well, and APEBench benchmark learned dynamics (Takamoto et al., 2022; Gupta & Brandstetter, 2023; Ohana et al., 2024; Koehler et al., 2024). We instead pair independently evolved numerical sources by physical identity at a common stored resolution. This differs from multifidelity or multiresolution methods that combine sources or optimize acquisition (Howard et al., 2023; Li et al., 2024), and from work showing that an emulator can outperform its training solver against a finer reference (Koehler & Thuerey, 2025). DG degree changes both numerical approximation and work (Cockburn & Shu, 2001).

**Recursive prediction.** Noise injection, pushforward training, and differentiable unrolling address feedback-generated inputs (Sanchez-Gonzalez et al., 2020; Brandstetter et al., 2022; List et al., 2025). We use related objectives as controls and analyze the complete forecast history through finite-time input–response propagation (Dahleh et al., 2011), with explicit nonlinear remainders. Ratios are accompanied by absolute errors, persistence, and physical failures (McGreivy & Hakim, 2024).

## 3 PAIRED NUMERICAL DATA AND EVALUATION

### 3.1 DATASET SCOPE AND PAIRED NUMERICAL SOURCES

Figure 1 summarizes the paired construction, the fixed learning interface, and the three common-reference evaluation endpoints. The benchmark contains 1,191 physical identities across Allen–Cahn, GLM–MHD, Richtmyer–Meshkov instability (RMI), and shallow water with topography (SWE). Each identity has three independently evolved numerical trajectories at DG degrees $p \in \{1, 3, 5\}$, giving 3,573 label trajectories. All branches are observed on a common $128^2$ grid, so changing the numerical source does not change the model's spatial input resolution.

A physical identity $\xi_i$ specifies equation parameters and an analytic initial condition. A degree-$p$ solver evolves $U_p(\xi_i, t)$, and the stored label is

$$y_{p,t}^{(i)} = \mathcal{O}_{128} U_p(\xi_i, t), \qquad p \in \{1, 3, 5\}. \tag{1}$$

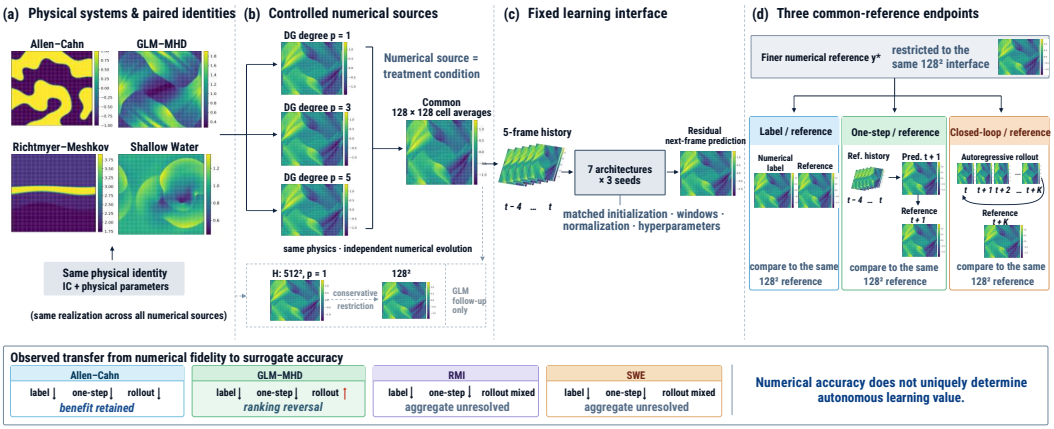



Figure 1: **Paired design and evaluation.** Numerical branches share physical identities, observation shape, and training controls. One-step evaluation refreshes reference histories; rollout initializes once.

Here, $\mathcal{O}_{128}$ conservatively averages onto the common observation grid. SWE velocities are recovered from averaged depth and momenta. The degree branches start independently from the same analytic initial condition; they are not interpolations of one stored trajectory. Finer-grid references $y_t^{\star(i)}$ are restricted to the same observation interface. Numerical source is the treatment condition, whereas reference error is a measured property of that source: raising the degree can change dissipation, dispersion, limiting, and computational work together.

Table 1: Benchmark scope. All models receive five observed frames on a common $128^2$ grid. GLM includes the recurrent cleaning field $\psi$; SWE includes static topography. References are finite numerical solutions.

| System | Channels | Identities | Formal reference | Forecast steps |
|---|---|---|---|---|
| Allen–Cahn | 1 | 300 | $256^2, p = 3$ | 56 |
| GLM–MHD | $6 + 1$ | 291 | $512^2, p = 1$ | 56 |
| RMI | 4 | 300 | $512^2, p = 1$ | 95 |
| SWE–topography | $3 + 1$ | 300 | $384^2, p = 3$ | 46 |

Allen–Cahn, RMI, and SWE use 260/20/20 training/validation/test identities; GLM uses 251/20/20. Splitting by identity prevents windows from the same trajectory from appearing in different splits. GLM retains the intersection of successful degree branches after nine incomplete $p = 5$ attempts, so its conclusions are conditional on successful generation. Appendix A gives the equations, generation costs, and reference checks.

### 3.2 Controlled learning and the prediction interface

We train ResNet, U-Net, DeepONet, FNO, Transolver, DPOT, and GNOT from scratch (He et al., 2016; Ronneberger et al., 2015; Lu et al., 2021; Li et al., 2021; Wu et al., 2024; Hao et al., 2024; 2023). The four systems, seven architectures, three degrees, and three seeds give $4 \times 7 \times 3 \times 3 = 252$ baseline models. Within an architecture and seed, degree branches share initialization, sampled windows, normalization, hyperparameters, and effective batch size. The baseline recipe uses 40 epochs, AdamW, an initial learning rate of $3 \times 10^{-4}$, and an effective batch size of 32. Checkpoints are selected using validation identities. Training details appear in Appendix B.

Each model predicts a residual next frame from five states. With $h_t = (z_{t-4}, \ldots, z_t)$, the prediction and recursive history update are

$$\widehat{z}_{t+1} = z_t + f_\theta(h_t), \qquad h_{t+1} = (z_{t-3}, \ldots, z_t, \widehat{z}_{t+1}). \tag{2}$$

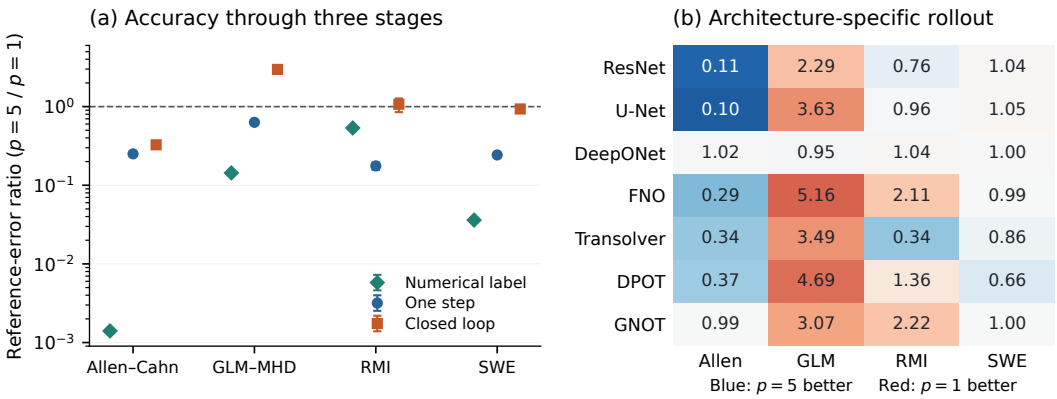



Figure 2: **Local numerical gains do not determine rollout rankings.** Ratios are $p = 5/p = 1$; values below one favor degree five. Left: aggregate endpoints. Right: architecture-specific rollouts. Intervals are paired 95% bootstrap intervals.

This interface is shared across the degree branches. SWE topography is static conditioning, whereas all seven GLM channels participate in the recurrence.

### 3.3 Common-reference evaluation and aggregation

We evaluate numerical labels, one-step predictions, and autonomous rollouts against the same reference (Figure 1). One-step evaluation refreshes five common reference frames before every prediction. Rollout receives the reference history only at initialization and then feeds back predictions; later references enter only scoring. This fixes one-step inputs and rollout initialization across models, rather than giving each model its own numerical source as input. Degree-specific histories are a supplementary endpoint. Main evaluation uses CPU float32 throughout.

For forecast times $\mathcal{T}$ and primary physical channels $\mathcal{C}$, the trajectory-level error is

$$\mathcal{E}(z, y^\star) = \left[ \frac{\sum_{t\in\mathcal{T}} \sum_{c\in\mathcal{C}} \|z_{t,c} - y^\star_{t,c}\|_2^2}{\sum_{t\in\mathcal{T}} \sum_{c\in\mathcal{C}} \|y^\star_{t,c}\|_2^2} \right]^{1/2}. \tag{3}$$

The same full-space–time endpoint applies to labels, stacked one-step predictions, and recursive forecasts. It does not average per-frame relative errors or stop at the first physical failure. GLM scores the six physical conservative channels, but its recurrent $\psi$ channel also enters the solver-consistent pressure diagnostic. Nonpositive density or pressure, or a non-finite state or pressure, constitutes a GLM physical failure. Finite but physically inadmissible trajectories remain scored to the fixed terminal time; non-finite outcomes remain flagged and must not be dropped to form finite-only aggregates.

## 4 Baseline results across systems and architectures

We first test whether the numerical advantage of degree five survives one-step learning and autonomous rollout. Figure 2 reports degree-five/degree-one error ratios under the same common-reference evaluation; absolute errors are given in Appendix E.

### 4.1 Numerical gains consistently transfer to one-step prediction

The degree-five/degree-one label-error ratios are $0.0014$, $0.1436$, $0.5352$, and $0.0361$ for Allen–Cahn, GLM, RMI, and SWE, respectively. The corresponding aggregate one-step ratios are $0.251$, $0.631$, $0.176$, and $0.243$, with all intervals below one. Degree-five supervision therefore improves local prediction across all four systems.

### 4.2 CLOSED-LOOP TRANSFER IS SYSTEM DEPENDENT

Allen–Cahn retains the advantage, with an aggregate rollout ratio of $0.327$ $[0.293, 0.359]$. More accurate supervision can therefore produce substantially better autonomous forecasts.

GLM reverses the ranking. Its aggregate rollout ratio is $2.971$ $[2.819, 3.121]$, despite lower label and one-step errors. Six of seven architectures perform worse at degree five, so the reversal is not confined to one model.

RMI and SWE have rollout ratios of $1.071$ $[0.851, 1.27]$ and $0.934$ $[0.873, 1.034]$, respectively. Their aggregate effects remain unresolved, although individual architectures can benefit.

Thus, the local advantage of degree-five supervision transfers strongly for Allen–Cahn, reverses for GLM, and does not establish an aggregate gain for RMI or SWE.

### 4.3 THE GLM REVERSAL IS NOT A WEAK-BASELINE ARTIFACT

Removing models that improve by less than $1\%$ over persistence in either one-step or rollout evaluation retains all seven GLM architectures and leaves its aggregate ratio unchanged at $2.971$. Applying the same filter to SWE leaves a rollout ratio of $0.908$ $[0.826, 1.047]$, which remains unresolved (Appendix F).

Physical validity further sharpens the GLM result. Degree-one U-Net and FNO fail in $1/60$ and $27/60$ evaluated rollouts, whereas both degree-five branches fail in all $60$. The degree-five local advantage therefore coexists with systematic closed-loop failure.

## 5 GLM CONTROLS: FEEDBACK AND RECURSIVE TRAINING

The GLM baseline raises two questions: when does the local degree-five advantage turn into a rollout reversal, and can recursive training remove it? We first vary the duration of autonomous feedback and then separate two effects of recursive training.

### 5.1 SUSTAINED FEEDBACK EXPOSES THE REVERSAL

For the original checkpoints, we periodically restore the reference history and vary the interval of uninterrupted prediction. As the feedback interval increases, the $p = 5/p = 1$ error ratio rises from $0.493$ to $3.63$ for U-Net and from $0.679$ to $5.156$ for FNO (Figure 3, left). The degree-five advantage is therefore local: it disappears and reverses as model predictions are fed back for longer.

History restoration changes both the states encountered and the errors subsequently generated, so this experiment identifies feedback dependence rather than a fixed-input response. Extending training to 120 epochs also retains the reversal (Appendix G).

### 5.2 RECURSIVE TRAINING REDUCES BUT DOES NOT REMOVE THE REVERSAL

K1 supervises one step, whereas K4 optimizes four recursively generated steps. For degree-five FNO, K4 lowers rollout error from $0.3503$ to $0.1251$; for U-Net, it lowers error from $0.0979$ to $0.0713$ (Figure 3, right). Degree-one models also improve, however, leaving degree five $3.45\times$ worse for FNO and $2.08\times$ worse for U-Net.

The same benefit appears when K1 and K4 continue from identical, physically valid generated histories: the K4/K1 remaining-horizon error ratio is $0.3996$ $[0.2595, 0.5846]$. Recursive training therefore improves continuation from a shared input, but does not restore the degree-five ordering.

### 5.3 THE RESOLVED GAIN COMES FROM PREDICTED-HISTORY EXPOSURE

We next separate predicted-history exposure from cross-step gradient flow. TF4 remains teacher-forced, SG4 uses recursive histories with stopped cross-step gradients, and BP4 differentiates through those histories. SG4 and BP4 have the same forward loss at fixed parameters but different gradients (Proposition 7).

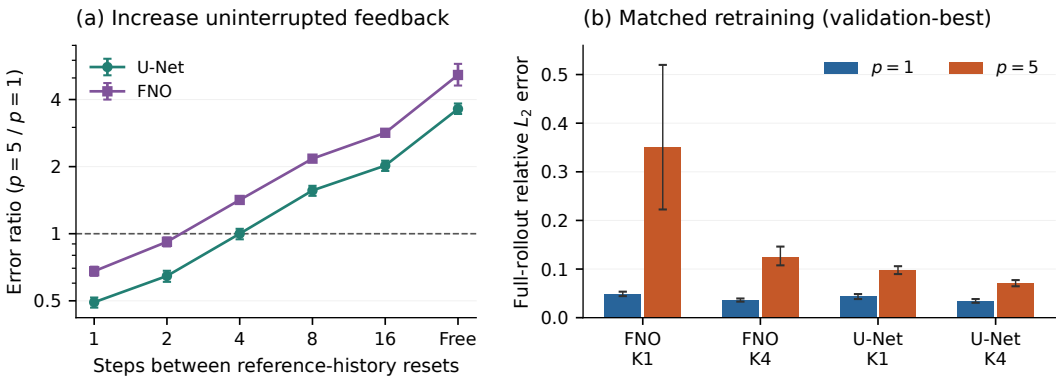



Figure 3: **Feedback reveals the reversal; recursive training reduces its magnitude.** Left: ratios as uninterrupted feedback grows. Right: matched K1/K4 retraining. Intervals are paired 95% bootstrap intervals.

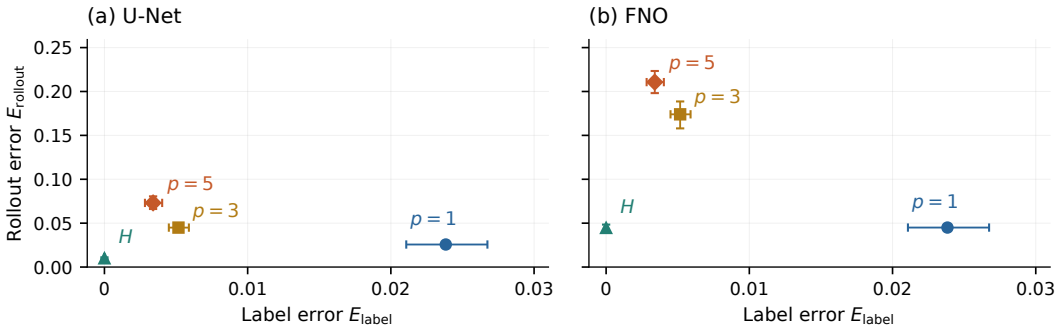



Figure 4: **Label error does not determine source value.** Under matched TF4-to-BP4 training, degree elevation lowers label error but worsens rollout, whereas $H$ improves U-Net; the FNO gain is unresolved. Bars show 95% identity-bootstrap intervals.

For degree-five FNO, the SG4/TF4 rollout ratio is $0.383$ $[0.2335, 0.5732]$, showing a clear benefit from training on predicted histories. The BP4/SG4 ratio is $0.9539$ $[0.8674, 1.0919]$, so an additional benefit from cross-step differentiation is unresolved.

Thus, recursive training can substantially reduce the GLM rollout penalty, primarily through the measured effect of predicted-history exposure in this control. It changes the magnitude of the reversal, but does not recover degree-five superiority.

## 6 Numerical Source Choice and Conditional Gains

The degree comparison leaves a narrower question: is the GLM reversal a general consequence of lowering numerical error, or does it depend on how the numerical source interacts with training? We address this by comparing two routes to lower label error, allowing bounded source-specific adaptation, and separating source choice from recursive continuation.

### 6.1 Different routes to lower label error produce different rollouts

Alongside $128^2, p = 1, 3, 5$, we introduce

$$H: \quad 512^2, p = 1 \longrightarrow 128^2,$$

obtained by restricting a finer-grid trajectory to the common observation interface. Its discrepancy from the formal reference is zero by construction.

Under the same TF4-to-BP4 recipe, the label errors of $p = 1, 3, 5$ decrease from $0.0238$ to $0.0052$ and $0.0034$, while rollout error increases for both U-Net and FNO (Figure 4). The spatially refined

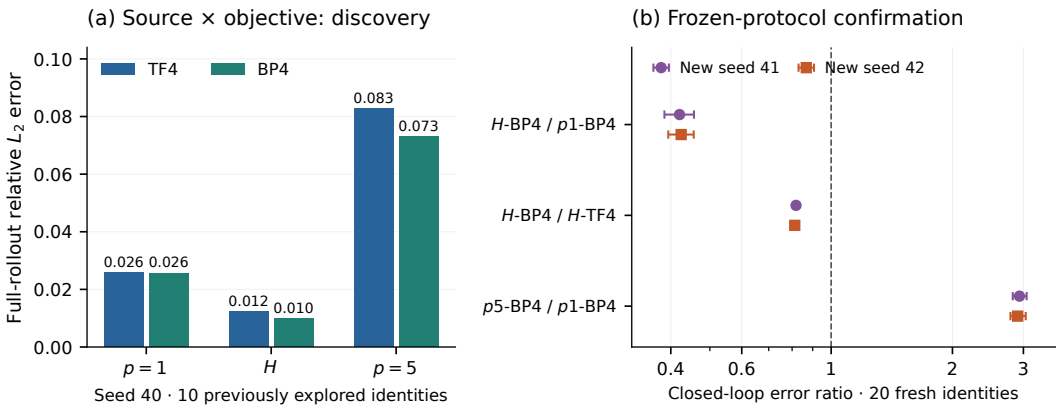



Figure 5: **The $H$ advantage precedes BP4 and is strengthened by it.** Left: source × continuation. Right: frozen confirmation on fresh identities. Ratios below one favor the numerator.

source behaves differently: $H$ gives U-Net a rollout error of $0.0099$. For FNO, the $H/p = 1$ ratio is $0.991$ $[0.949, 1.031]$, so its benefit remains unresolved.

Thus, lower label error does not induce a common ordering of training sources. Degree elevation and spatial refinement both improve the numerical labels, but they need not produce the same learned dynamics.

### 6.2 Bounded adaptation changes magnitude, not ordering

A common training recipe could disadvantage degree five. We therefore allow each architecture–source pair to select among six matched second-stage configurations that vary learning rate and recursive unroll length.

Adaptation lowers degree-five rollout error by $5.4\%$ for U-Net and $13.8\%$ for FNO. Nevertheless, adapted degree five remains $2.796$ $[2.62, 2.993]$ and $3.977$ $[3.72, 4.242]$ times worse than adapted degree one, respectively, despite lower one-step error.

The contrast with $H$ also remains. Relative to adapted degree one, adapted $H$ lowers U-Net rollout error by $60.5\%$; the corresponding FNO point estimate is $5.3\%$ lower. The reversal is therefore not explained by the tested learning-rate or unroll-length mismatch.

### 6.3 The $H$ gain comes from the source and is amplified by BP4

We next distinguish the effect of $H$ from the effect of BP4. For each source, TF4 and BP4 continuations start from the same first-stage checkpoint and use the same additional-update budget.

$H$ already improves U-Net under TF4 continuation, so its advantage is not created entirely by BP4. BP4 then further reduces rollout error within $H$. The source therefore provides the advantage, while recursive continuation strengthens it. Because BP4 changes both predicted-history exposure and cross-step gradient flow, this comparison does not separate those two effects.

The frozen confirmation reproduces both gains. The $H$-BP4/$p = 1$-BP4 rollout ratios are $0.4205$ and $0.4244$, while BP4 reduces error within $H$ by $18.18\%$ and $18.78\%$. In contrast, degree five has $85.15\%$ lower label error than degree one and one-step ratios of $0.474$ and $0.483$, but rollout ratios of $2.934$ and $2.904$. Its predictions fail physically on all evaluated trajectories, whereas $H$-BP4 fails on only 0 and 1 trajectories, respectively.

Finally, $H$-BP4 has $7.66\%$ and $7.51\%$ higher one-step error than $H$-TF4, despite producing lower rollout error. Together, the degree-five and $H$ comparisons show that one-step improvement is neither sufficient nor necessary for rollout improvement. The value of a numerical source is determined by its interaction with closed-loop training, not by label or one-step error alone.

# 7 SOURCE–TRAINING MECHANISM AND ROLLOUT RECONSTRUCTION

One-step error need not order rollout performance. We separate source-induced learning from closed-loop propagation: an exact DG construction treats the first, and full-history GLM reconstructions treat the second.

## 7.1 A CONSTRUCTIVE DG REVERSAL

Consider periodic $u_t + u_x = 0$ on $[0, N]$, with an even number $N$ of unit cells and initial cell values $1 + \frac{1}{4}(-1)^j$. We use upwind DG at $p = 1, 5$, exact weak-form integration, no limiter, and SSPRK33 with step $1/16$. Let $x_{p,n}$ denote the saved integer-time cell averages. The exact averages are $x^\star_n = \phi_0 + \frac{1}{4}(-1)^n\phi_\pi$, with $(\phi_0)_j = 1$, $(\phi_\pi)_j = (-1)^j$, and $\|v\|^2_N = N^{-1}\sum_j v_j^2$.

The five-frame learner $g_a(h) = aLh$ uses only the newest frame and forces the mean and alternating modes to share one coefficient:

$$\mathcal{L}_p(a) = \frac{1}{16}\sum_{n=4}^{11} \|ax_{p,n} - x_{p,n+1}\|^2_N. \tag{4}$$

We score indices $5, \dots, 12$ with the common normalization; one-step evaluation refreshes the reference history, whereas rollout is autonomous from the exact history $0, \dots, 4$.

**Theorem 1** (DG reversal under matched finite training)**.** *Starting from $a_p^{(0)} = 1$, apply twenty updates $a_p^{(k+1)} = a_p^{(k)} - \frac{1}{2}\mathcal{L}_p'(a_p^{(k)})$. For $\widehat{a}_p = a_p^{(20)}$,*

$$E_{\mathrm{label},5} < E_{\mathrm{label},1}, \qquad E_1(\widehat{a}_5) < E_1(\widehat{a}_1), \qquad E_R(\widehat{a}_5) > E_R(\widehat{a}_1).$$

*The same ordering holds at the exact minimizers.*

Writing $x_{p,n} = \phi_0 + \frac{1}{4}c_{p,n}\phi_\pi$ gives

$$a^\star_p = \frac{8 + \beta N_p}{8 + \beta S_p}, \qquad \beta = \tfrac{1}{16}, \quad S_p = \sum_{n=4}^{11} c^2_{p,n}, \quad N_p = \sum_{n=4}^{11} c_{p,n}c_{p,n+1}.$$

Hence the coefficient is source-determined. Degree-one data have nearly lost the alternating mode, so fitting mainly preserves the mean. Degree-five data retain negatively correlated fluctuations, forcing the shared coefficient to compromise between exact multipliers $+1$ and $-1$ and attenuate the mean repeatedly. Rational arithmetic certifies the gaps (Appendix C).

A mean-preserving one-scalar control removes this compromise without adding a parameter and favors degree five at every finite integer horizon $m \geq 1$ (Proposition 3). This establishes one source–architecture mechanism, not the cause of the GLM reversal.

## 7.2 FULL-HISTORY DEFECT PROPAGATION

Let $G_t$ be the complete normalized history update and $x^\star_t$ a shift-consistent reference. With $\eta_t = x_t - x^\star_t$, let $B$ insert the newest frame and $O$ map history errors to scored physical channels. With fixed raw reference energy $Q > 0$, the rollout score is $E = \|\operatorname{stack}_{t=1}^{T} O\eta_t\|/\sqrt{Q}$. Define

$$B\beta_t = G_t(x^\star_t) - x^\star_{t+1}, \qquad n_t(v) = G_t(x^\star_t + v) - G_t(x^\star_t) - A_t v,$$

where $A_t$ contains the exact shift blocks and is ordinarily $DG_t(x^\star_t)$. Then

$$\eta_{t+1} = A_t\eta_t + B\beta_t + n_t(\eta_t), \qquad \xi_{t+1} = A_t\xi_t + B\beta_t. \tag{5}$$

For a reference start, $\eta_0 = \xi_0 = 0$. With $\Phi(t, s) = A_{t-1}\cdots A_s$,

$$(\mathcal{T}\boldsymbol{\beta})_t = \sum_{j<t} O\Phi(t, j+1)B\beta_j, \qquad \widehat{E} = \|\mathcal{T}\boldsymbol{\beta}\|/\sqrt{Q}. \tag{6}$$

This is exact for the linear DG learners. In general, $Q\widehat{E}^2 = \boldsymbol{\beta}^\top\mathcal{T}^\top\mathcal{T}\boldsymbol{\beta}$, so propagation depends on defect direction and temporal interaction, not label error alone (Dahleh et al., 2011).

**Theorem 2** (Finite-horizon control)**.** *Suppose $\|n_j(v)\| \le h_j(\|v\|)$ on certified reference balls of radii $R_j$, where each $h_j : [0, R_j] \to [0, \infty)$ is finite and nondecreasing. Suppose valid bounds for $\|\Phi(t, j+1)B\|$ and $\|O\Phi(t, j+1)B\|$ are used in the recursions $r_t, s_t$ of Appendix D. If $r_j \le R_j$ for all $j < T$, then the rollout exists through $T$ and*

$$|E - \widehat{E}| \le \frac{\|\operatorname{stack}_{t=1}^{T} O(\eta_t - \xi_t)\|}{\sqrt{Q}} \le \Delta, \qquad \Delta := \frac{(\sum_{t=1}^{T} s_t^2)^{1/2}}{\sqrt{Q}}. \tag{7}$$

The recursion has two roles: $r_t$ bounds the full-history deviation over all stored frames and recurrent channels so that each nonlinear remainder is evaluated within its stated ball, while $s_t$ bounds the propagated remainder in the physical channels used for scoring after restoring their frozen scales. The nondecreasing envelope makes the same remainder bound valid throughout the enclosed history ball. Stacking these observed bounds over time and dividing by $\sqrt{Q}$ gives $\Delta$ in the same units as the reported rollout score.

Thus $\widehat{E}_A + \Delta_A < \widehat{E}_B - \Delta_B$ certifies $E_A < E_B$. Appendix D gives the precise hypotheses, recursion, and proof; sampled JVPs alone do not certify the required operator bounds. We additionally test a once-corrected path with $q_t = x_t^\star + \xi_t$, $p_t = q_t + d_t$, $d_0 = 0$, and $d_{t+1} = DG_t(q_t)d_t + G_t(q_t) - q_{t+1}$.

The correction uses the defect of the proxy path as a new forcing. It is propagated along that path, allowing contributions introduced at different times to reinforce or offset one another before their final norm is taken. The reconstruction tests below therefore compare the resulting error vectors as well as the scalar forecast scores.

### 7.3 GLM RECONSTRUCTION TESTS

We evaluate both proxies on the fixed-final models, not the adapted models: BP4 U-Net/FNO at $p = 1, p = 5, H$ and U-Net $H$-TF4. Ten reused identities and five reference-started windows yield 350 eight-step paths. Local defects and directional gains, considered separately, do not order all models. Let $\delta$ be the observed proxy-to-rollout vector discrepancy, normalized by $\sqrt{Q}$. We record a mean contrast as separated when proxy and actual gaps have the same sign and the proxy gap exceeds $\overline{\delta}_A + \overline{\delta}_B$.

At the point estimates, both proxies separate all seven contrasts at four and eight steps. Correction reduces the discrepancy budget relative to the actual $p = 5$–$p = 1$ gap from 35% to 2.5% for U-Net and from 41% to 4.2% for FNO. Here the budget is the sum of the two models' mean vector discrepancies. A smaller budget relative to the observed gap means that reconstruction resolves more of the measured difference between the two numerical sources.

At 56 steps, both proxies separate all four U-Net contrasts and none of the three FNO contrasts; correction can worsen agreement. U-Net's $p = 5$–$p = 1$ gap remains separated despite a 150.4% corrected full-history relative vector discrepancy, so correct ranking need not imply accurate reconstruction. For the within-$H$ seed-41/42 comparison, only the corrected proxy separates both contrasts, with mean relative six-channel vector discrepancies of 0.296%–0.961%.

These are empirical reconstructions, not independent rollout certificates or a causal explanation of GLM. The observed $\delta$ uses the actual rollout and differs from the theoretical bound $\Delta$. Appendix L reports failures, margin intervals, and sensitivity checks.

## 8 CONCLUSION

The learning value of numerical fidelity depends on its interaction with source, architecture, and training. Equal-budget adaptation preserves the GLM reversal, whereas frozen confirmation shows that the finer-grid source $H$, combined with BP4, lowers U-Net rollout error by about 58% on fresh identities and seeds. The exact DG construction exhibits one source–architecture mechanism and shows that mean preservation can restore the high-order advantage. Joint reconstruction recovers all short-window contrasts and corrected long-horizon U-Net comparisons, while long-horizon FNO marks its current limit. Overall, higher fidelity becomes useful when source choice and rollout-aware training induce compatible autonomous dynamics.

## AI USE STATEMENT

Generative AI tools were used as auxiliary aids under the authors' direction. The authors led the research questions, experimental design, analysis, and core manuscript writing. AI assisted with research ideation, drafting selected passages, and polishing author-written text. For implementation, the authors specified the tasks and methods and used AI to help draft and revise code. For mathematical claims, the authors supplied the proof directions and key ideas, and used AI to help develop detailed arguments. The authors made the final research and presentation decisions and take responsibility for the proofs, code, results, citations, and final manuscript.

## REPRODUCIBILITY STATEMENT

Section 3 defines the paired comparison and endpoints. Appendices A and B give numerical configurations, training objectives, shared recipes, and aggregation rules; Appendices E and F give the complete primary scores and persistence audit. Appendices C and D contain the DG construction, rational certificate, sensitivity grid, and full proofs. Training controls, four-source comparisons, bounded adaptation, and frozen confirmation appear in Appendices G, H, I, and J. Appendices K and L specify the response and reconstruction diagnostics, including identity-level aggregation and precision checks. The processed evaluation records and display-generation code reproduce the reported figures and summary statistics; the full simulation and checkpoint archives are separate resources. We commit to publicly releasing PDE-FIDELITY upon publication, including paired trajectories and references, frozen splits, simulation-generation code and configurations, training and evaluation pipelines, and environment specifications (Appendix O).

## ETHICS STATEMENT

The study uses simulated physical fields and contains no personal or human-subject data. Its purpose is methodological evaluation of learned PDE forecasts. Physical-admissibility diagnostics are reported alongside prediction errors because low average error alone does not establish suitability for physical deployment.

## REFERENCES

Mohammad Alhawwary and Z.J. Wang. Fourier analysis and evaluation of dg, fd and compact difference methods for conservation laws. *Journal of Computational Physics*, 373:835–862, 2018. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2018.07.018. URL https://www.sciencedirect.com/science/article/pii/S0021999118304790.

Johannes Brandstetter, Daniel E. Worrall, and Max Welling. Message passing neural PDE solvers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=vSix3HPYKSU.

Noel Chalmers and Lilia Krivodonova. Spatial and modal superconvergence of the discontinuous Galerkin method for linear equations. *Journal of Scientific Computing*, 72:128–146, 2017. doi: 10.1007/s10915-016-0349-8. URL https://doi.org/10.1007/s10915-016-0349-8.

Bernardo Cockburn and Chi-Wang Shu. Runge–Kutta discontinuous Galerkin methods for convection-dominated problems. *Journal of Scientific Computing*, 16(3):173–261, 2001. doi: 10.1023/A:1012873910884.

Mohammed Dahleh, Munther A. Dahleh, and George Verghese. Lectures on dynamic systems and control, chapter 10: Discrete-time linear state-space models. MIT OpenCourseWare, 6.241J, 2011. URL https://ocw.mit.edu/courses/6-241j-dynamic-systems-and-control-spring-2011/bab3ecdb64baf945acd206d8f370167f_MIT6_241JS11_chap10.pdf.

A. Dedner, F. Kemm, D. Kröner, C.-D. Munz, T. Schnitzer, and M. Wesenberg. Hyperbolic divergence cleaning for the mhd equations. *Journal of Computational Physics*, 175(2):645–673, 2002. ISSN 0021-9991. doi: https://doi.org/10.1006/jcph.2001.6961. URL https://www.sciencedirect.com/science/article/pii/S002199910196961X.

Jayesh K Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized PDE modeling. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=dPSTDbGtBY.

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. GNOT: A general neural operator transformer for operator learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12556–12569. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/hao23c.html.

Zhongkai Hao, Chang Su, Songming Liu, Julius Berner, Chengyang Ying, Hang Su, Anima Anandkumar, Jian Song, and Jun Zhu. DPOT: Auto-regressive denoising operator transformer for large-scale PDE pre-training. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17616–17635. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/hao24d.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Amanda A. Howard, Mauro Perego, George Em Karniadakis, and Panos Stinis. Multifidelity deep operator networks for data-driven and physics-informed problems. *Journal of Computational Physics*, 493:112462, 2023. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2023.112462. URL https://www.sciencedirect.com/science/article/pii/S0021999123005570.

Felix Koehler and Nils Thuerey. Neural emulator superiority: When machine learning for pdes surpasses its training data. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 165660–165702. Curran Associates, Inc., 2025. doi: 10.52202/085713-5521. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f200119a40846e508954abcd61f5f3fd-Paper-Conference.pdf.

Felix Koehler, Simon Niedermayr, Rüdiger Westermann, and Nils Thuerey. Apebench: A benchmark for autoregressive neural emulators of pdes. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 120252–120310. Curran Associates, Inc., 2024. doi: 10.52202/079017-3822. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d9875ebcf74bccdc5076acab0dbee62c-Paper-Datasets_and_Benchmarks_Track.pdf.

Shibo Li, Xin Yu, Wei Xing, Robert Kirby, Akil Narayan, and Shandian Zhe. Multi-resolution active learning of Fourier neural operators. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2440–2448. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/li24k.html.

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

Bjoern List, Li-Wei Chen, Kartik Bali, and Nils Thuerey. Differentiability in unrolled training of neural physics simulators on transient dynamics. *Computer Methods in Applied Mechanics and Engineering*, 433:117441, 2025. ISSN 0045-7825. doi: https://doi.org/10.1016/j.cma.2024.117441. URL https://www.sciencedirect.com/science/article/pii/S0045782524006960.

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

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2113–2122, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/maclaurin15.html.

Nick McGreivy and Ammar Hakim. Weak baselines and reporting biases lead to overoptimism in machine learning for fluid-related partial differential equations. *Nature Machine Intelligence*, 6:1256–1269, 2024. doi: 10.1038/s42256-024-00897-5.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond B. Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich R. Kerswell, Suryanarayana Maddu, Jonah Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam Parker, Miles Cranmer, and Shirley Ho. The well: a large-scale collection of diverse physics simulations for machine learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 44989–45037. Curran Associates, Inc., 2024. doi: 10.52202/079017-1430. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4f9a5acd91ac76569f2fe291b1f4772b-Paper-Datasets_and_Benchmarks_Track.pdf.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8459–8468. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sanchez-gonzalez20a.html.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. Pdebench: An extensive benchmark for scientific machine learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 1596–1611. Curran Associates, Inc., 2022. doi: 10.52202/068431-0117. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/0a9747136d411fb83f0cf81820d44afb-Paper-Datasets_and_Benchmarks.pdf.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for PDEs on general geometries. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53681–53705. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wu24r.html.

# A NUMERICAL DATA, IDENTITIES, AND REFERENCE HIERARCHY

## A.1 GOVERNING SYSTEMS AND STORED STATES

All systems are two-dimensional. Allen–Cahn, GLM–MHD, and SWE use periodic boundaries. RMI is periodic in the transverse direction, with fixed post-shock inflow at the left boundary and transmissive outflow at the right boundary. A physical parameter row and analytic initial condition define each identity independently of the DG degree. Numerical evolution starts separately for each degree; paired observations are taken at the same physical output times.

**Allen–Cahn.** On $\Omega = [-1, 1]^2$, the phase field satisfies

$$\partial_t u = \varepsilon \Delta u + \varepsilon^{-1}(u - u^3), \qquad \varepsilon = 6.25 \times 10^{-3}. \tag{8}$$

Analytic Fourier initial conditions vary the mean, amplitude, spectral cutoff, and maximum mode. The numerical labels use $128^2$ elements at degrees 1, 3, and 5. Exact DG cell averages provide the stored observations. There are 61 saved frames through $t = 0.5$, of which the first five form the observed history. The reference uses $256^2$ elements at degree 3 followed by $2 \times 2$ conservative restriction. The associated physical energy is

$$\mathcal{F}(u) = \int_\Omega \left[ \frac{\varepsilon}{2} |\nabla u|^2 + \frac{(1 - u^2)^2}{4\varepsilon} \right] dx. \tag{9}$$

**GLM–MHD.** The state is $U = (\rho, m_x, m_y, E, B_x, B_y, \psi)$, with $m_x = \rho u$, $m_y = \rho v$, and adiabatic index $\gamma = 5/3$. The generalized Lagrange multiplier (GLM) cleaning variable $\psi$ is used to control magnetic divergence (Dedner et al., 2002). Parameterized multimode Orszag–Tang initial conditions vary sonic and Alfvénic Mach numbers, modes, amplitudes, and phases. The numerical generator uses Lax–Friedrichs surface fluxes, central DG volume fluxes, positivity limiting, and SSPRK33 time stepping. Its 61 saved frames are spaced by $\Delta t = 0.005$ and end at $t = 0.3$. The formal reference is $512^2$, degree one, restricted to $128^2$. The network predicts all seven fields; the primary accuracy endpoint uses the first six. Seven-channel recurrence is essential: removing $\psi$ only from the main score does not remove its effect on future predictions or on the pressure diagnostic.

**Richtmyer–Meshkov instability.** The compressible Euler state is $U = (\rho, \rho u, \rho v, E)$ and satisfies $\partial_t U + \nabla \cdot F(U) = 0$. A shock interacts with a perturbed material interface. Identities vary with shock Mach number, density contrast, perturbation amplitude, integer mode, and phase. The paired branches share these values. There are 100 saved frames through $t = 0.36$ and a 95-frame forecasting interval. The formal reference uses $512^2$ elements at degree one. The degree-three and degree-five label errors are close but not monotonically ordered; the observed degree-five error is slightly larger.

**Shallow water with topography.** For bottom elevation $b$ and velocity $\mathbf{u} = (u, v)$, the dynamics are

$$\partial_t h + \nabla \cdot (h\mathbf{u}) = 0, \tag{10}$$

$$\partial_t (h\mathbf{u}) + \nabla \cdot \left( h\mathbf{u}\mathbf{u}^\top + \tfrac{1}{2} g h^2 I \right) = -gh\nabla b. \tag{11}$$

The solver evolves conservative variables. The stored dynamic state is $(h, u, v)$ after conservative averaging and conversion of momenta to velocities. The bottom field is static conditioning and is excluded from the dynamic loss. The 51 saved frames end at $t = 0.5$; the forecast contains 46 steps. The formal reference is $384^2$, degree three, restricted to $128^2$.

## A.2 COUNTS, SPLITS, AND REFERENCE CALIBRATION

The primary collection has 3,573 label trajectories and 244,053 saved label states. Reference and parameter-extrapolation trajectories are not included in these totals. GLM's population is conditional on successful generation of all three degrees: nine incomplete degree-five attempts are excluded at every degree before assigning the paired population.

Formal references are finite numerical approximations. Additional checks use ten Allen–Cahn cases at $512^2, p = 3$, ten GLM training/validation cases at $1024^2, p = 1$, a 20-case RMI $1024^2$ reference

Table 2: Identity-level composition. Each retained physical identity has three degree trajectories; reference trajectories are additional assets.

| System | Train | Validation | Test | Identities | Frames/trajectory |
|---|---|---|---|---|---|
| Allen–Cahn | 260 | 20 | 20 | 300 | 61 |
| GLM–MHD | 251 | 20 | 20 | 291 | 61 |
| RMI | 260 | 20 | 20 | 300 | 100 |
| SWE | 260 | 20 | 20 | 300 | 51 |
| Total | 1031 | 80 | 80 | 1191 | — |

Table 3: Mean label/reference errors over the 20 original test identities. Each identity is counted once per degree.

| System | $p = 1$ | $p = 3$ | $p = 5$ | $p = 5/p = 1$ |
|---|---|---|---|---|
| Allen–Cahn | 0.1539 | 0.0081 | 0.0002 | 0.0014 |
| GLM–MHD | 0.0239 | 0.0052 | 0.0034 | 0.1436 |
| RMI | 0.0373 | 0.0198 | 0.0200 | 0.5352 |
| SWE | 0.0290 | 0.0026 | 0.0010 | 0.0361 |

audit, and ten SWE cases at $512^2, p = 3$. Allen–Cahn and SWE retain their degree ordering on the checked cases; GLM's six-channel ordering agrees on its ten-case audit. RMI's audit retains channel-level reference sensitivity. These checks are distinct from the five-identity, same-prediction rescoring in Appendix J.

**Generation work and stored size.** The degree-five/degree-one median wall-time ratios are 5.83, 34.29, 20.09, and 19.65 in the system order above. Production contexts use 16 threads for Allen–Cahn, 32 for RMI, 24 for SWE, and predominantly 32 for GLM (two retained trajectories used 24). These measurements compare numerical work within each system.

# B Learning configurations and measurement definitions

## B.1 Original architecture panel

All 252 baseline models are trained from scratch with seeds 40, 41, and 42. Input/output projections adapt to the system's channel count, while an architecture's internal configuration stays fixed across degrees. The optimizer is AdamW with initial learning rate $3 \times 10^{-4}$, minimum learning rate $1 \times 10^{-6}$, weight decay $1 \times 10^{-4}$, and $\epsilon = 1 \times 10^{-8}$. The schedule uses two warm-up epochs and cosine decay over 40 epochs. Betas are $(0.9, 0.95)$ except for SWE, which uses $(0.9, 0.999)$. Gradient clipping uses norm 1, and effective batch size is 32. Micro-batches and accumulation preserve that effective size.

Each training trajectory contributes 32 sampled legal windows per epoch. The same architecture/seed pair shares initialization and sampling streams across degrees. Standardization uses training data and is shared within each system. GLM and RMI use the frozen 64-trajectory cap for estimating normalization. FNO and DPOT train in float32; other architectures generally use bfloat16. The primary common-history re-evaluation uses CPU float32 consistently. Training predicts the next-state residual in normalized coordinates; headline evaluation is the unnormalized physical-channel score in Eq. 3. Validation uses held-out identities and the recorded one-step criterion, not test error.

These are fixed architecture configurations; in particular, DPOT is trained from scratch.

## B.2 Four-target training objectives

Let $g_\theta(h) = \mathrm{last}(h) + f_\theta(h)$ be the full next-state map in Eq. 2, $h_j^{\mathrm{true}}$ the true history preceding target $y_{t+j}$, and $\widehat{h}_j$ its recursively updated counterpart, initially equal to the same true history. For

Table 4: Frozen architecture configurations used for every system and degree. Input/output projections adapt only to channel count.

| Architecture | family | frozen configuration |
|---|---|---|
| ResNet | convolutional residual | width 160; 10 residual blocks |
| U-Net | encoder–decoder | base width 48 |
| DeepONet | branch–trunk | 32 sensors; basis 128; branch/trunk widths 512/256 |
| FNO | spectral operator | hidden width 64; 24 modes; 4 layers |
| Transolver | transformer operator | hidden 192; 6 layers; 8 heads; 32 slices |
| DPOT | operator transformer | patch 8; embedding 512; depth 4; 4 blocks; 32 modes |
| GNOT | operator transformer | hidden 128; 4 layers; 4 heads; inner width 256 |

four consecutive targets, the losses are

$$\mathcal{L}_{\mathrm{TF4}} = \tfrac{1}{4}\sum_{j=1}^{4} \ell\big(g_\theta(h_j^{\mathrm{true}}), y_{t+j}\big), \tag{12}$$

$$\mathcal{L}_{\mathrm{BP4}} = \tfrac{1}{4}\sum_{j=1}^{4} \ell\big(g_\theta(\widehat{h}_j), y_{t+j}\big). \tag{13}$$

Here $\ell$ is mean squared error on the seven standardized training channels. SG4 uses the same recursive construction as BP4 but detaches the complete input history before each update. Thus TF4 and SG4 differ in their training histories, while SG4 and BP4 also distinguish the availability of gradients through earlier predictions. K1 uses the first target; K4 is the four-step BP4 objective in the factorial controls.

### B.3 Shared two-stage GLM protocol

For GLM follow-ups, $H$ denotes existing $512^2, p = 1$ training-identity trajectories restricted to $128^2$. The held-out formal reference uses the same numerical source: $H$ has zero label discrepancy against $512^2$ by construction and nonzero discrepancy against $1024^2$.

All U-Net branches use 7,053,607 parameters (base width 48), residual prediction, the same seven-channel normalizer, five history frames, and the original 251/20 training/validation split. FNO retains its original capacity and the same normalizer and split. Each epoch samples 32 windows per training identity, giving 251 updates at an effective batch size of 32; four-target windows have 53 legal starts. Table 5 specifies the U-Net stages. Within source and seed, TF4/BP4 continuations share their fixed first-stage checkpoint, identities, window stream, normalizer, and additional-update budget.

Table 5: The shared two-stage GLM/U-Net recipe. Both stages use AdamW with $\beta = (0.900, 0.950)$, $\epsilon = 1.000 \times 10^{-8}$, weight decay $1.000 \times 10^{-4}$ and gradient-norm clipping at 1.

| Setting | First stage | Continuation |
|---|---|---|
| Epochs / updates | 40 / 10,040 | 40 / 10,040 additional |
| Objective | TF4 | TF4 or BP4 |
| Effective batch | 32 | 32 |
| Microbatch | 1 for $H$ | 32 |
| Learning rate | $3.000 \times 10^{-4} \to 10^{-6}$ | $1.000 \times 10^{-5} \to 10^{-6}$ |
| Schedule | 2-epoch warmup, cosine | Cosine; fresh optimizer |
| Arithmetic | BF16 autocast | FP32; TF32 disabled |
| Primary endpoint | Fixed epoch-40 last | Fixed 10,040-update final |
| Sensitivity | Own-target validation best | Full-rollout validation best |

Stage one uses gradient accumulation from microbatches of one, with FNO's spectral core in FP32 during BF16 autocast. Continuation uses FP32 with TF32 disabled and a fresh AdamW optimizer. Stage-one $p = 1$ and $p = 5$ starts are existing matched TF4 runs; the seed-41/42 confirmation $H$ starts are independently trained.

Continuation validation uses the original 20 validation identities, common-reference initial histories, and full 56-step six-channel error every 502 updates, including update zero. The primary endpoint is the fixed final checkpoint; validation-best is a sensitivity endpoint. All seven channels recur, and scoring and physical-failure rules follow the definitions below.

### B.4 Endpoint and aggregation

Write $e^q_{m,p,s,i}$ for Eq. 3 evaluated for architecture $m$, degree $p$, seed $s$, identity $i$, and mode $q$ (one step or closed loop). Within each model, the degree ratio is a ratio of arithmetic means,

$$r^q_m = \frac{\frac{1}{3\cdot 20}\sum_{s,i} e^q_{m,5,s,i}}{\frac{1}{3\cdot 20}\sum_{s,i} e^q_{m,1,s,i}}, \qquad R^q_{5/1} = \exp\left(\frac{1}{7}\sum_{m=1}^{7}\log r^q_m\right). \tag{14}$$

**Alternative histories and targets.** The completed common-history archive evaluates each degree-trained model with histories from $p = 1$, $p = 3$, $p = 5$, and the formal reference, in both one-step and free modes, and scores against both its own numerical target and the common reference. This produces $252 \times 20 \times 4 \times 2 \times 2 = 80{,}640$ scoring rows. The paper's main endpoint selects common reference histories and reference targets. Own-target fit describes agreement with supervision, whereas common-reference one-step error measures local accuracy.

**Physical diagnostics and non-finite values.** GLM pressure follows the numerical solver's convention,

$$P = (\gamma - 1)\left[E - \frac{1}{2}\left(\frac{m_x^2 + m_y^2}{\rho} + B_x^2 + B_y^2 + \psi^2\right)\right], \qquad \gamma = 5/3. \tag{15}$$

An ever-failed trajectory has any non-finite state/pressure, nonpositive density, or nonpositive pressure during prediction. The first flagged step and affected-cell fractions provide additional diagnostics. The primary six-channel error includes every finite trajectory through the fixed endpoint even after failure. Thus, a numerical metric can remain finite for a physically inadmissible forecast. Non-finite errors are kept flagged rather than removed by a finite-only reduction. The confirmation scores contain no non-finite primary errors; their nonzero failure counts concern physical admissibility.

## C An exact DG source–training mechanism and its range

This construction derives the learned coefficient from numerical trajectories and a specified training rule. It establishes an existence result for a restricted learner. The baseline is exactly certified; the surrounding finite grid is a numerical sensitivity check.

### C.1 Numerical sources from the DG weak form

Consider $u_t + u_x = 0$ on the periodic interval $[0, N]$, where $N$ is a positive even integer, each cell has width one, and

$$u(x, 0) = 1 + \tfrac{1}{4}(-1)^j, \qquad x \in [j, j+1), \quad j = 0, \ldots, N-1.$$

In $\|v\|_N^2 = N^{-1}\sum_j v_j^2$, the mean vector $\phi_0$ and alternating-cell vector $\phi_\pi$ are orthonormal. Translation by an integer number of cells gives the exact observed trajectory $x^\star_n = \phi_0 + \frac{1}{4}(-1)^n\phi_\pi$. Use upwind DG of degree $p = 1$ or 5, Legendre basis $P_k$ on each mapped cell, exact spatial integration, no limiter, and SSPRK33 with step $1/16$. The initial projection is exact. The constant spatial mode is preserved; the alternating-cell part evolves within its invariant native DG subspace.

Testing with $P_\ell$ gives cell mass $1/(2\ell + 1)$. The volume integral is

$$\int_{-1}^{1} P_k(\xi)P'_\ell(\xi)\,d\xi = 2\,\mathbf{1}_{\{k<\ell,\ \ell-k \text{ odd}\}}.$$

The right upwind trace contributes $-1$ per coefficient; the left upwind trace contributes $-(-1)^\ell$ because the neighboring cell has opposite sign. Thus all native modal coefficients evolve, without

fitting a scalar decay surrogate:

$$(A_p)_{\ell k} = (2\ell+1)\big(2\mathbf{1}_{\{k<\ell,\ \ell-k \text{ odd}\}} - 1 - (-1)^{\ell}\big), \quad R_p = I + \frac{A_p}{16} + \frac{A_p^2}{2\cdot 16^2} + \frac{A_p^3}{6\cdot 16^3}. \tag{16}$$

The saved averages satisfy

$$c_{p,n} = e_0^{\top} R_p^{16n} e_0, \qquad x_{p,n} = \phi_0 + \tfrac{1}{4} c_{p,n} \phi_\pi. \tag{17}$$

In particular,

$$A_1 = \begin{pmatrix} -2 & -2 \\ 6 & 0 \end{pmatrix}, \qquad A_5 = \begin{pmatrix} -2 & -2 & -2 & -2 & -2 & -2 \\ 6 & 0 & 0 & 0 & 0 & 0 \\ -10 & 0 & -10 & -10 & -10 & -10 \\ 14 & 0 & 14 & 0 & 0 & 0 \\ -18 & 0 & -18 & 0 & -18 & -18 \\ 22 & 0 & 22 & 0 & 22 & 0 \end{pmatrix}.$$

The cubic is the SSPRK33 stability polynomial applied to this linear ODE. DG modal analysis and the distinction between spatial and temporal dispersion/dissipation are established tools (Chalmers & Krivodonova, 2017; Alhawwary & Wang, 2018).

### C.2 Training and the exact three-stage scores

The learner $g_a(h) = aLh$ has a five-frame interface but uses only the newest frame. Set $\beta = 1/16$ and

$$S_p = \sum_{n=4}^{11} c_{p,n}^2, \qquad N_p = \sum_{n=4}^{11} c_{p,n} c_{p,n+1}.$$

Differentiating Eq. 4 yields

$$\mathcal{L}_p'(a) = \frac{8(a-1) + \beta(aS_p - N_p)}{8}, \quad h_p = \mathcal{L}_p'' = \frac{8+\beta S_p}{8} > 0, \quad a_p^\star = \frac{8+\beta N_p}{8+\beta S_p}. \tag{18}$$

Hence, the unique minimizer follows from source energy and lagged correlation.

Each prediction uses the exact history at saved indices $0, \dots, 4$ initially. Common-reference one-step prediction refreshes the full history before every prediction, whereas autonomous prediction feeds back its own outputs. At $t = 1, \dots, 8$ relative to saved index four, the reference-history defect and actual autonomous error are

$$r_{p,t} = (a-1)\phi_0 + \tfrac{1}{4}(a+1)(-1)^{t-1}\phi_\pi, \tag{19}$$

$$e_{p,t} = (a^t - 1)\phi_0 + \tfrac{1}{4}(a^t - (-1)^t)\phi_\pi. \tag{20}$$

For $Q = 8(1+\beta) = 17/2$, the root relative errors are

$$E_{\text{label},p}^2 = \frac{\beta}{Q}\sum_{t=1}^{8}(c_{p,4+t} - (-1)^t)^2, \tag{21}$$

$$E_1(a)^2 = \frac{(a-1)^2 + \beta(a+1)^2}{1+\beta}, \tag{22}$$

$$E_R(a)^2 = \frac{1}{Q}\sum_{t=1}^{8}\big[(a^t-1)^2 + \beta(a^t - (-1)^t)^2\big]. \tag{23}$$

These are deterministic scalar-physics scores for this construction.

*Proof of Theorem 1.* The matrices, projected initial state, saved-state recursion, minimizers, and squared metrics are rational. Table 6 encloses the exact values with rational decimal endpoints. The disjoint intervals establish all three strict squared-error comparisons and $0 < a_5^\star < a_1^\star < 1$; taking positive square roots preserves their directions.

For twenty common full-batch updates with step $1/2$, the quadratic objective gives

$$a_p^{(k)} - a_p^\star = (1 - h_p/2)^k (1 - a_p^\star).$$

Table 6: Outward-rounded rational enclosures for the specified DG construction. These are deterministic bounds, not confidence intervals. The coefficient row displays $1 - a_p^\star$ to retain strict separation from zero. Finite-training comparisons use the derivative margins in the proof.

| Quantity | Degree one | Degree five |
|---|---|---|
| $S_p$ | $[1.1834, 1.1835] \times 10^{-4}$ | $[5.1711, 5.1712]$ |
| $N_p$ | $[-3.0991, -3.0990] \times 10^{-5}$ | $[-5.1524, -5.1523]$ |
| $1 - a_p^\star$ | $[1.1666, 1.1667] \times 10^{-6}$ | $[0.0775, 0.0776]$ |
| $E^2_{\mathrm{label},p}$ | $[0.0587, 0.0588]$ | $[0.0023, 0.0024]$ |
| $E_1(a_p^\star)^2$ | $[0.2352, 0.2353]$ | $[0.2230, 0.2231]$ |
| $E_R(a_p^\star)^2$ | $[0.1176, 0.1177]$ | $[0.1891, 0.1892]$ |
| $\gamma_p^\star$ | $[-0.2619, -0.2618]$ | $[-0.9964, -0.9963]$ |
| $E_R(\widetilde{g}_{\gamma_p^\star})^2$ | $[0.0541, 0.0542]$ | $[1.9521, 1.9522] \times 10^{-5}$ |

The exact computation also verifies $1 < h_p < 2$, so each iterate remains between $a_p^\star$ and one and $|a_p^{(20)} - a_p^\star| < 2^{-20}$. On $a \in [0, 1]$, differentiation of Eqs. 22–23 gives

$$\left|\frac{dE_1(a)^2}{da}\right| \le 4, \qquad \left|\frac{dE_R(a)^2}{da}\right| \le \frac{4}{8}\sum_{t=1}^{8} t = 18.$$

Exact arithmetic verifies that the one-step squared gap exceeds $8 \cdot 2^{-20}$ and the rollout squared gap exceeds $36 \cdot 2^{-20}$. Perturbing both coefficients by less than $2^{-20}$ therefore cannot close either gap. The label comparison is unchanged by training. □

### C.3 Architecture intervention and exact propagation

Let $\Pi_0 v = \langle v, \phi_0 \rangle \phi_0$ denote projection onto the spatial mean and define the one-scalar control $\widetilde{g}_\gamma(h) = \Pi_0 Lh + \gamma(I - \Pi_0)Lh$.

**Proposition 3** (Mean-preserving control)**.** *Fit $\widetilde{g}_\gamma$ by exact minimization of the same source loss. Then $\gamma_p^\star = N_p/S_p$, and for every finite integer horizon $m \ge 1$, both its aggregate common-reference one-step error and its autonomous rollout error over steps $1, \dots, m$ are strictly smaller for degree five than for degree one.*

*Proof of Proposition 3.* The model $\widetilde{g}_\gamma$ preserves the spatial mean exactly. Its only variable loss term has derivative $\beta(\gamma S_p - N_p)/8$ and positive curvature $\beta S_p/8$. Therefore, its unique minimizer is $\gamma_p^\star = N_p/S_p$. Exact rational arithmetic gives

$$\gamma_1^\star \in [-0.2619, -0.2618], \qquad \gamma_5^\star \in [-0.9964, -0.9963].$$

Let $\rho_p = -\gamma_p^\star$, so $0 < \rho_1 < \rho_5 < 1$. The mean error vanishes. At any integer step $t \ge 1$, the absolute alternating coefficient error is $\frac{1}{4}(1 - \rho_p^t)$, strictly smaller for degree five. Summing its squared values over any positive finite forecast horizon, with the common positive reference denominator, proves both the autonomous and one-step comparisons. □

The two model families each have one learned scalar. The control changes the fixed treatment of the mean, rather than increasing the number of learned parameters. Exact fitting of this control is a different optimization specification from twenty unrescaled updates; the latter is not asserted to reach its minimizer.

Both complete five-frame updates are linear. With their exact derivatives, $n_t(v) \equiv 0$ and $\eta_t = \xi_t$ in Eq. 5. The source and loss determine the injected defects and the response together; their forced propagation is exact. The numerical full-history check agrees to $3.89 \times 10^{-16}$ on the original sensitivity grid. This supplies a direct connection to the propagation analysis, while nonlinear neural models require the separate remainder and reconstruction checks.

The degree-one fitted multiplier is nearly one, and the degree-five newest-frame response is smaller: $(a_5^\star)^k < (a_1^\star)^k$. Both baseline full-horizon signed cross-terms are negative, whereas GLM's measured degree-five cross-term is positive.

Table 7: Time refinement at the original geometry. Shared-family predictions use twenty matched updates; the mean-preserving family is exactly fitted. Persistence has one-step error $0.4851$ and eight-step error $0.3430$ at all three time steps. These decimal summaries supplement the rational baseline certificate.

| $\Delta t$ | $p$ | Label | Shared one | Shared roll | Mean-pres. one | Mean-pres. roll |
|---|---|---|---|---|---|---|
| $1/16$ | 1 | 0.2424 | 0.4851 | 0.3430 | 0.1790 | 0.2327 |
| $1/16$ | 5 | 0.0483 | 0.4723 | 0.4349 | 0.0009 | 0.0044 |
| $1/32$ | 1 | 0.2424 | 0.4851 | 0.3430 | 0.1794 | 0.2328 |
| $1/32$ | 5 | 0.0468 | 0.4722 | 0.4376 | 0.0012 | 0.0058 |
| $1/64$ | 1 | 0.2424 | 0.4851 | 0.3430 | 0.1794 | 0.2328 |
| $1/64$ | 5 | 0.0466 | 0.4722 | 0.4379 | 0.0012 | 0.0060 |

### C.4 Time-step and training-geometry sensitivity without neural training

Before evaluating the sensitivity grid, we fixed time steps $\Delta t \in \{1/16, 1/32, 1/64\}$, alternating amplitudes $\varepsilon \in \{1/8, 1/4, 1/2\}$ around mean one, training newest-frame starts $s \in \{4, 6, 8, 12\}$, and horizons $T \in \{1, 2, 4, 8, 16, 32\}$: 216 settings, each with both degrees. Saving remains at integer physical times. Each source uses eight training pairs $s, \dots, s+7$, the original loss normalization, and the same twenty-update closed form. The exact shared-coefficient minimizer and exactly fitted mean-preserving control are evaluated alongside persistence.

For these settings, replace $\beta$ by $\varepsilon^2$, $4$ by $s$, and $8$ by $T$ in the scoring expressions, but retain eight pairs in the training loss. The initial exact history is $s-4, \dots, s$. All selected starts are even; varying cell patterns or saving phases is outside this check. The label and rollout denominators are $T(1+\varepsilon^2)$, and

$$E^2_{\mathrm{pers,R}} = \frac{\varepsilon^2}{T(1+\varepsilon^2)} \sum_{t=1}^{T} (1-(-1)^t)^2, \qquad \text{Skill} = 1 - E_R / E_{\mathrm{pers,R}}.$$

All metrics, mean/fluctuation error components, and persistence scores are retained for every setting.

The three-stage reversal survives both time-step reductions at the baseline geometry (Table 7). On the full initial grid, it occurs in all 180 tested multi-step settings and none of the 36 one-step settings, where autonomous and common-reference one-step prediction coincide. The mean-preserving control gives lower degree-five one-step and autonomous errors in all 216 settings.

The initial multi-step grid contains only even horizons. After inspecting that grid, we specified an additional complete parity check at $T = 3, 5, 7$, retaining the same 36 time-step/amplitude/start combinations per horizon. This additional analysis is identified separately in the protocol and Figure 6. It finds reversal in 96/108 settings: all $T = 5, 7$ settings, and 24/36 at $T = 3$. At $T = 3$ and amplitude $1/2$, all twelve time-step/start settings instead give better degree-five rollout accuracy. The mean-preserving control favors degree five in all 108 added settings.

All 324 settings are included, including the 48 without three-stage reversal. The twenty-update and exact-fit shared families agree on the classification throughout. FP64 substep evaluation and independently computed 80-digit saved-interval matrix powers agree in all reversal decisions; the maximum absolute difference across the evaluated scalar quantities is $1.96 \times 10^{-14}$. A fixed $1 \times 10^{-12}$ squared-score guard separates numerical near-ties; the baseline proof remains the separate rational certificate. Figure 6 shows the entire setting grid, and the source package retains all coefficients and scores.

At the original eight-step setting, the shared degree-one rollout skill is only about $2.33 \times 10^{-4}\%$, while degree five has negative skill, approximately $-26.8\%$. The benchmark's separate persistence-skill audit is in Appendix F.

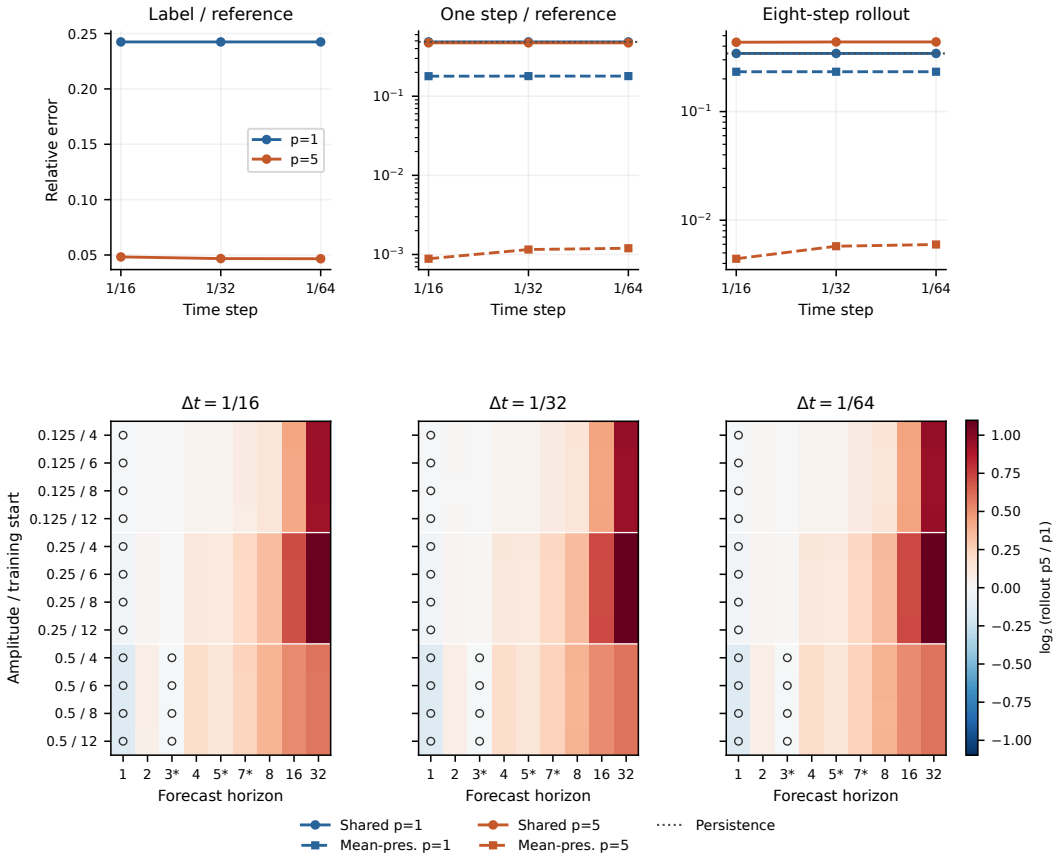



Figure 6: **A finite neighborhood of the DG construction, with its boundaries.** Top: label, common-reference one-step, and eight-step autonomous errors at amplitude $1/4$ and training start four as the time step decreases; the learned shared family uses twenty matched updates, the mean-preserving family is exactly fitted, and persistence is shown on prediction panels. Bottom: shared-family rollout $p = 5/p = 1$ ratios on every amplitude/start/horizon setting. Columns marked $*$ belong to the additional parity analysis; the remaining 216 settings were fixed before evaluation. The color boundary is ratio one; all settings retain the label and one-step degree-five advantage. The $T = 1$ column and the amplitude-$1/2$, $T = 3$ cells do not reverse. The mean-preserving family favors degree five in all 324 settings.

# D SOURCE FITTING, FINITE-HORIZON ANALYSIS, AND PROOFS

We use the normalized full-history coordinates of Theorem 2, writing $b_t := B\beta_t$ and $n_t^\star := n_t$. Here $X = (\mathbb{R}^d)^m$ has its Euclidean norm, stacked outputs use the Euclidean product norm, and all norm balls are closed. Empty products satisfy $\Phi(s, s) = I$. The insertion $B$ is an isometry, $L$ extracts the newest frame, and $O = PSL$ restores channel scales before selecting the scored channels. The reference is only required to be shift-consistent; it need not satisfy a closed finite-memory evolution law. Normalization means canceling in differences, but the denominator $Q > 0$ uses raw reference frames. All state bounds retain every recurrent channel. The maps are deterministic: forward randomness and mutable auxiliary state must be frozen or incorporated in the state being analyzed. The reference and normalization are shared when comparing receivers, whereas their defects and derivatives may differ.

## D.1 SOURCE FITTING, HISTORY SHIFT, AND THE REFERENCE DEFECT

The reference sequence need not be an orbit of a closed observation-space PDE. Let $x_t^s$ be any paired, shift-consistent source history in the same frozen normalized coordinates as $x_t^\star$. Write $g_\theta = LG_\theta$ and define the source fitting defect and source–reference displacement by

$$a_t = g_\theta(x_t^s) - Lx_{t+1}^s, \qquad e_t^s = x_t^s - x_t^\star. \tag{24}$$

Here $a_t \in \mathbb{R}^d$ contains all training channels; the corresponding full-history defect is $Ba_t$. Let $A_t$ have the exact shift blocks and put $n_t^\star(v) = G_\theta(x_t^\star + v) - G_\theta(x_t^\star) - A_t v$.

**Proposition 4** (Source–reference bridge)**.** *Where these evaluations exist, the reference defect obeys the exact identity*

$$b_t = Ba_t + e_{t+1}^s - A_t e_t^s - n_t^\star(e_t^s). \tag{25}$$

*Proof.* Shift consistency gives $G_\theta(x_t^s) = x_{t+1}^s + Ba_t$. Substitute $x_t^s = x_t^\star + e_t^s$ into the definition of $n_t^\star$ and subtract $x_{t+1}^\star$. □

Thus reference defects combine source fitting, label displacement, and the model's response to displaced histories, with their signs and directions retained. A smaller scalar label error alone does not order $\|b_t\|$, even before propagation. Furthermore, source fitting on a paired evaluation history is not the empirical training loss: transferring a training-loss bound to those histories requires additional coverage or generalization assumptions. In GLM, the six-channel label score also leaves the seventh recurrent channel uncontrolled. Equation (25) is an identity on paired sequences, not a statistical domain-adaptation guarantee or an assumption of finite-memory closure.

## D.2 SUPPORTING LOCAL SOURCE-TO-TRAINING IDENTITIES

The construction is finite and explicit. For a nonlinear learner, differentiating through a finite optimization process is an established technique (Maclaurin et al., 2015); the following identities clarify the quantities that an upstream diagnostic would have to measure.

Freeze normalization, identities, windows, initialization, random streams, and the finite update schedule. A paired data path $u_i(\alpha) = u_i^L + \alpha(u_i^H - u_i^L)$ changes histories and targets together. Its interior is a data intervention, not an intermediate DG degree, and physical admissibility is not automatic. For a teacher-forced batch, use the actual differentiable objective $\mathcal{L}(\theta, \alpha) = \frac{1}{2}\sum_i w_i \|g_\theta(h_i(\alpha)) - y_i(\alpha)\|^2 + R(\theta)$ with fixed weights and a source-independent regularizer.

**Proposition 5** (Local source-to-training and rollout sensitivity)**.** *Suppose the required state/parameter mixed derivatives of $g$ and all finite update maps exist and are continuous locally. At a fixed parameter let $e_i = g_\theta(h_i) - y_i$, $J_i = D_\theta g_\theta(h_i)$, and $T_i = D_h J_i[\dot{h}_i]$. Then the source forcing of the teacher-forced gradient is*

$$B_{\mathrm{src}} = \partial_\alpha \nabla_\theta \mathcal{L} = \sum_i w_i \left[ T_i^\top e_i + J_i^\top (D_h g_\theta(h_i)\dot{h}_i - \dot{y}_i) \right]. \tag{26}$$

*If $s_{k+1} = U_k(s_k, \alpha)$ includes all required optimizer state and has source-independent $s_0$, its tangent satisfies*

$$v_0 = 0, \qquad v_{k+1} = D_s U_k\, v_k + \partial_\alpha U_k. \tag{27}$$

*For the final parameter component $v_\theta$, fixed-reference defects and responses satisfy*

$$\dot{b}_t = D_\theta G_\theta(x_t^\star) v_\theta, \qquad \dot{A}_t = D_\theta D_x G_\theta(x_t^\star)[v_\theta].$$

*On the actual autonomous path, the exact local derivative instead satisfies*

$$\zeta_0 = 0, \qquad \zeta_{t+1} = D_x G_\theta(x_t)\zeta_t + D_\theta G_\theta(x_t) v_\theta.$$

*Proof.* At fixed $\theta$, differentiating $\sum_i w_i J_i^\top e_i + \nabla R$ in $\alpha$ gives Eq. 26 by the product rule. Applying the chain rule to the finite update composition gives Eq. 27; the reference quantities then differentiate with respect to the same final parameter direction. Differentiating the generated-state recursion gives the last equation, evaluated on $x_t$, not $x_t^\star$. □

Writing $M_k = D_s U_k$, $f_k = \partial_\alpha U_k$, and $\Pi(b, a) = M_{b-1} \cdots M_a$ gives $v_K = \sum_{j<K} \Pi(K, j+1) f_j$. For gradient descent, $M_k = I - \eta_k \nabla_\theta^2 \mathcal{L}_k$ and $f_k = -\eta_k B_{\mathrm{src},k}$. No convergence, commuting Hessians, or inverse-Hessian approximation is required. The common $v_\theta$ links the defect and response changes; prescribing their changes independently is not a source-to-training explanation.

For stacked observed error $\mathbf{e} = \operatorname{stack} O(x_t - x_t^\star)$ and stacked reference defect $\mathbf{r}$, the squared-score derivatives are $F_R' = 2\langle \mathbf{e}, \operatorname{stack} O\zeta_t\rangle/Q$ and $F_1' = 2\langle \mathbf{r}, \dot{\mathbf{r}}\rangle/Q$. The cross-term derivative is $C' = 2(\langle \dot{\mathbf{r}}, \mathbf{e} - \mathbf{r}\rangle + \langle \mathbf{r}, \dot{\mathbf{e}} - \dot{\mathbf{r}}\rangle)/Q$. If at $\alpha_0$, $F_1' \le -a < 0$, $F_R' \ge b > 0$, and on $[\alpha_0, \alpha_0 + \delta]$ the second derivatives obey $|F_1''| \le M_1$, $|F_R''| \le M_R$, with $M_1\delta \le a$, $M_R\delta \le b$, Taylor's theorem gives a local reversal. These assumed derivative signs are a sufficient condition, not a consequence of DG degree. Means of root errors require differentiating those means, rather than substituting a pooled squared score.

AdamW requires moments, decoupled decay, clipping, and resets inside $U_k$; the smoothness assumptions must hold on the branch used. The teacher-forced formula cannot be substituted for a BP4 recursive objective. SG4 requires differentiating the recomputed update field, not taking a naive higher derivative through stopped edges. Discrete selection rules and bitwise finite-precision execution are not covered by a smooth real-arithmetic assertion.

**A finite source exchange for fixed features.** For $g_W(h) = Lh + W\varphi(h)$, define columns $X_s[:, i] = \sqrt{w_i}\varphi(h_{s,i})$ and $Y_s[:, i] = \sqrt{w_i}(y_{s,i} - Lh_{s,i})$. If $C_s = X_s X_s^\top + \lambda I$ is positive definite for $s = L, H$ and fixed $\lambda \ge 0$, the unique ridge minimizer is $W_s = Y_s X_s^\top C_s^{-1}$. With $\Delta X = X_H - X_L$, $\Delta Y = Y_H - Y_L$, subtracting the two normal equations gives the exact identity

$$W_H - W_L = \big[(\Delta Y - W_L \Delta X) X_H^\top + (Y_L - W_L X_L)\Delta X^\top\big] C_H^{-1}.$$

Indeed, $C_H - C_L = \Delta X X_H^\top + X_L \Delta X^\top$ and $Y_H X_H^\top - Y_L X_L^\top = \Delta Y X_H^\top + Y_L \Delta X^\top$. This explains how feature energy and lagged targets jointly determine a finite coefficient change in the specified class. It is not a fixed-feature approximation justified for the trained U-Net or FNO.

### D.3 The dynamic weighting of a short unroll

Fix parameters, forward randomness, and a source window $x_0^s, \ldots, x_K^s$. Let $\widehat{x}_0 = x_0^s$, $\widehat{x}_{t+1} = G_\theta(\widehat{x}_t)$, and $\eta_t = \widehat{x}_t - x_t^s$. Choose source-path linearizations $A_t^s$ with exact shift blocks and write

$$\eta_{t+1} = A_t^s \eta_t + B a_t + B\nu_t, \qquad \eta_0 = 0, \tag{28}$$

where $B\nu_t$ is the exact linearization remainder. Define $\Phi_s(t, j) = A_{t-1}^s \cdots A_j^s$ for $t > j$, with $\Phi_s(t, t) = I$. The block map $\mathcal{T}_s : \mathbb{R}^{Kd} \to \mathbb{R}^{Kd}$ has entries

$$(\mathcal{T}_s)_{t,j} = \begin{cases} L\Phi_s(t, j+1)B, & 0 \le j < t \le K, \\ 0, & j \ge t, \end{cases} \qquad W_s = \mathcal{T}_s^\top \mathcal{T}_s. \tag{29}$$

Rows index forecast times $1, \ldots, K$; columns index injections $0, \ldots, K-1$. Stack the $a_t$ and $\nu_t$ as $\mathbf{a}$ and $\boldsymbol{\nu}$.

**Proposition 6** (Short-unroll weighting with a remainder)**.** *For mean squared loss over all $d$ standardized training coordinates,*

$$L_{\mathrm{TF,K}} = \frac{\|\mathbf{a}\|^2}{Kd}, \qquad L_{\mathrm{rec,K}} = \frac{\|\mathcal{T}_s(\mathbf{a}+\boldsymbol{\nu})\|^2}{Kd}. \tag{30}$$

*If $\|\mathcal{T}_s\boldsymbol{\nu}\| \le \varepsilon_s$, then*

$$\left| L_{\mathrm{rec,K}} - \frac{\mathbf{a}^\top W_s \mathbf{a}}{Kd} \right| \le \frac{2\|\mathcal{T}_s\mathbf{a}\|\varepsilon_s + \varepsilon_s^2}{Kd}. \tag{31}$$

*Proof.* Iterating Eq. (28) and extracting the newest frame gives $\mathrm{stack}_{t=1}^{K} L\eta_t = \mathcal{T}_s(\mathbf{a}+\boldsymbol{\nu})$. Teacher forcing has errors $a_t$. Expand the squared norm and apply Cauchy–Schwarz to its signed cross term. □

A non-circular sufficient bound uses remainder envelopes $\|n_t^s(v)\| \le \mu_t$ on prescribed balls $\|v\| \le R_t$, where $n_t^s(v) = G_\theta(x_t^s + v) - G_\theta(x_t^s) - A_t^s v$. If $R_0 = 0$ and $\|A_t^s\|R_t + \|a_t\| + \mu_t \le R_{t+1}$, induction encloses the recursive path, and one may take

$$\varepsilon_s = \left[ \sum_{t=1}^{K} \left( \sum_{j<t} \|L\Phi_s(t, j+1)B\|\mu_j \right)^2 \right]^{1/2}. \tag{32}$$

This sufficient envelope can be conservative; a measured discrepancy alone does not certify it. We impose neither independence nor zero mean; temporal cross terms are retained. Because $LB = I$, $\mathcal{T}_s$ is block triangular with identity diagonal blocks, so $W_s$ is positive definite for this full-channel training loss.

The weighting can reverse local-loss preferences without changing propagation. For a two-step, two-coordinate example, take a constant source state $r$ and the affine maps $g_c(x) = r + A(x-r) + c$, with $A = \mathrm{diag}(2, 0)$. Compare $c_A = (1, 0)$ with $c_B = (0, \sqrt{2})$. Their teacher-forced squared-error sums are 2 and 4, whereas their recursive sums are 10 and 4. All remainders vanish, and $A$ is identical. More generally, for fixed positive-definite $W_s$ with unequal extreme eigenvalues, defects on the minimum- and maximum-eigenvalue directions can have oppositely ordered Euclidean and weighted norms. This establishes a possible tradeoff, not a prediction of an optimizer's outcome.

When parameters change, both $\mathbf{a}_\theta$ and $W_{s,\theta}$ generally change. Where differentiable,

$$\partial_{\theta_j}(\mathbf{a}_\theta^\top W_{s,\theta}\mathbf{a}_\theta) = 2(\partial_{\theta_j}\mathbf{a}_\theta)^\top W_{s,\theta}\mathbf{a}_\theta + \mathbf{a}_\theta^\top(\partial_{\theta_j}W_{s,\theta})\mathbf{a}_\theta. \tag{33}$$

The proposition therefore does not identify a fixed weighting throughout training or describe a complete AdamW trajectory. Its seven-channel source loss is also distinct from the six-channel physical reference score $O = PSL$: reference transfer additionally involves Proposition 4, state coverage, and the longer scoring horizon.

### D.4 Forward equality and the stopped-gradient update

**Proposition 7** (Gradient paths)**.** *At identical parameters, windows, and forward randomness, detaching each complete input history gives SG and BP identical recursive predictions and numerical losses. At differentiable points, let*

$$J_t = DG_\theta(\widehat{x}_t), \qquad H_t = \partial_\theta G_\theta(\widehat{x}_t)\big|_{\widehat{x}_t \text{ fixed}}, \qquad U_{t+1} = J_t U_t + H_t, \quad U_0 = 0. \tag{34}$$

*If $v_t$ is the derivative of the $t$th weighted loss with respect to $\widehat{x}_{t+1}$, the updates satisfy*

$$g_{\mathrm{BP}} = \nabla_\theta L_{\mathrm{rec,K}} = \underbrace{\sum_t H_t^\top v_t}_{g_{\mathrm{SG}}} + \underbrace{\sum_t (J_t U_t)^\top v_t}_{g_{\mathrm{cross}}}. \tag{35}$$

*Proof.* Detachment preserves forward values. Ordinary differentiation gives $U_t = \partial_\theta \widehat{x}_t$ and the displayed sensitivity recurrence. The chain rule sums $U_{t+1}^\top v_t$. Detaching the input retains only $H_t^\top v_t$ at each update. □

Here $g_{\mathrm{SG}}$ is a computational surrogate gradient, not a different ordinary derivative of the same scalar function. BP is the exact gradient of the recursive training objective; this does not order eventual reference error, physical validity, or results under finite optimizer steps. Neither the added term nor unrolling implies that Jacobians become contractions. For nonsmooth maps, these ordinary derivative statements require differentiability on the evaluated path; an autodifferentiation branch at a kink is not automatically a Fréchet derivative. The defect identities and remainder bounds remain meaningful for chosen linear maps, but quadratic remainder claims require additional local regularity. Stateful forward operations must be frozen, or their state included explicitly.

### D.5 Forced propagation and a non-circular remainder bound

For completeness, $G_j$ is defined on $x_j^\star + \overline{B}_X(0, R_j)$ for every $j < T$. Each envelope $h_j : [0, R_j] \to [0, \infty)$ is finite and nondecreasing. With these envelopes and the independently valid operator bounds of Theorem 2, set $r_0 = \|\eta_0\|$ and compute, without using the unknown rollout,

$$r_t = \|\xi_t\| + \sum_{j<t} M_{tj} h_j(r_j), \qquad s_t = \sum_{j<t} N_{tj} h_j(r_j). \tag{36}$$

The required test is $r_j \le R_j$ for all $j < T$; no terminal domain radius is needed.

*Proof of Theorem 2.* The exact history error satisfies

$$\eta_{t+1} = A_t \eta_t + b_t + n_t^\star(\eta_t). \tag{37}$$

Set $w_t = \eta_t - \xi_t$. Since $\xi_0 = \eta_0$, $w_0 = 0$. Exact shifting implies $n_j^\star(v) \in \operatorname{im} B$, so whenever the preceding evaluations exist,

$$w_t = \sum_{j<t} \Phi(t, j+1) B\nu_j, \qquad B\nu_j = n_j^\star(\eta_j). \tag{38}$$

Existence and containment follow together by induction. Initially $\|\eta_0\| = r_0 \le R_0$. Suppose states through $t-1$ exist and $\|\eta_j\| \le r_j \le R_j$ there. The next update exists. Since $B$ is an isometry and $h_j$ is nondecreasing, $\|\nu_j\| = \|n_j^\star(\eta_j)\| \le h_j(\|\eta_j\|) \le h_j(r_j)$. Consequently,

$$\|\eta_t\| \le \|\xi_t\| + \sum_{j<t} M_{tj} h_j(r_j) = r_t, \qquad \|Ow_t\| \le s_t. \tag{39}$$

For $t < T$, the radius test permits the next evaluation. No radius $R_T$ is needed. Stacking the observed bounds first proves $\|\operatorname{stack} Ow_t\|/\sqrt{Q} \le \Delta$; applying the reverse triangle inequality then proves $|E - \widehat{E}| \le \Delta$.

For two models, chain the intervals:

$$\widehat{E}_A + \Delta_A < \widehat{E}_B - \Delta_B \quad \Longrightarrow \quad E_A < E_B. \tag{40}$$

For a fixed paired collection with weights $w_i \ge 0$, $\sum_i w_i = 1$, the analogous bound is $|\sum_i w_i E_i - \sum_i w_i \widehat{E}_i| \le \sum_i w_i \Delta_i$. These are deterministic statements for the checked collection, rather than confidence intervals for unseen identities. □

The radii are constructed without the unknown rollout. A failed test means that this bound does not certify the trajectory; it does not establish instability. The envelopes may be nonquadratic, and $A_t$ need only retain the exact history shift. Thus the enclosure itself does not require global smoothness.

A usable envelope is a separate mathematical requirement. When $A_j = DG_j(x_j^\star)$ and the derivative is Lipschitz with constant $\mathcal{L}_j$ throughout the reference ball, integration along the segment gives

$$n_j^\star(v) = \int_0^1 [DG_j(x_j^\star + sv) - DG_j(x_j^\star)]v \, ds, \qquad h_j(r) = \tfrac{1}{2}\mathcal{L}_j r^2. \tag{41}$$

Here the ball is convex and the derivative bound must hold along every segment used. For a merely locally Lipschitz map, a valid Lipschitz constant $\ell_j$ and a chosen linear $A_j$ instead give the less informative envelope $h_j(r) = (\ell_j + \|A_j\|)r$.

The argument also applies to prescribed reset blocks. At an exact reference reset, initialize that block with $\eta_0 = 0$ and retain all defects injected inside it. For a shared donor history, retain its common nonzero $\eta_0$. Each block has its own products, remainder budget and raw reference energy. Combining disjoint blocks reproduces a squared-error score only with the corresponding energy weights; it is not an unweighted average of block-relative errors. Changing block length changes both the retained injections and their interactions. The theorem imposes no monotone ordering as the feedback length increases.

### D.6 Directional and temporal structure of the defects

Writing $b_j = B\beta_j$, the finite-horizon operator in the main text has blocks

$$\mathcal{T}_{t,j} = \begin{cases} O\Phi(t, j+1)B, & j < t, \\ 0, & j \geq t, \end{cases} \qquad 1 \leq t \leq T, \quad 0 \leq j < T. \tag{42}$$

For a shared nonzero initial deviation, let $a_t = O\Phi(t, 0)\eta_0$. Variation of constants gives

$$\operatorname{stack}_{t=1}^{T} O\xi_t = a + \mathcal{T}\boldsymbol{\beta}, \qquad Q\widehat{E}^2 = \|a + \mathcal{T}\boldsymbol{\beta}\|^2. \tag{43}$$

At a reference initialization, $a = 0$ and $W = \mathcal{T}^*\mathcal{T}$ yields $Q\widehat{E}^2 = \boldsymbol{\beta}^* W \boldsymbol{\beta}$. Off-diagonal blocks retain interactions between defects injected at different times. We impose neither independence nor zero mean and retain these interactions. At a donor handoff, Eq. 43 also retains the interaction between propagated initial deviation and subsequent defects. The full normalized $\beta_j$ and the raw scored defect $Ob_j$ are different quantities when some recurrent channels are unscored.

**Proposition 8** (Smaller defects need not imply smaller cumulative error)**.** *Even for two autonomous affine maps with the same derivative, strictly smaller reference defects at every step need not yield a smaller finite-horizon score.*

*Proof.* Take a two-dimensional state with $m = 1$, $O = I$, constant reference $x_t^\star = (1, 1)$, and $T = 2$, so $Q = 4$. Let

$$G^{(k)}(x) = x^\star + A(x - x^\star) + \beta^{(k)}, \tag{44}$$

$$A = \operatorname{diag}(1, 0), \quad \beta^{(1)} = (3/4, 0), \quad \beta^{(2)} = (0, 1). \tag{45}$$

Both maps have derivative $A$ and zero Taylor remainder. The defect norms are $3/4 < 1$ at both times. Starting at the reference, the two error sequences are $((3/4, 0), (3/2, 0))$ and $((0, 1), (0, 1))$. Their cumulative squared errors are $45/16 > 2$, respectively. Division by the common $Q$ preserves this reversal. □

Temporal structure alone can also suffice: in the scalar forced recurrence with $A_t = 1$, zero initial error and two steps, defect sequences $(3/4, 3/4)$ and $(1, -1)$ give cumulative squared responses $45/16$ and 1. These elementary examples establish possibilities within the propagation law.

More generally, suppose the restriction of $W$ to an admissible defect subspace has eigenvalues $\lambda_{\text{high}} > \lambda_{\text{low}} \geq 0$. Unit eigenvectors in those directions have different cumulative costs despite equal defect norms. Scaling the high-cost vector by any $0 < \alpha < 1$ with $\alpha^2\lambda_{\text{high}} > \lambda_{\text{low}}$ gives a smaller defect norm but a larger propagated score. This observation holds with the propagation operator fixed; it does not require a change in its largest gain. If $W$ were proportional to the identity on that subspace, this particular reversal would be impossible.

### D.7 A proxy-centered correction

Whenever the indicated evaluations exist, the reference-linear path $q_t = x_t^\star + \xi_t$ has a directly evaluable update defect,

$$c_t = G_t(q_t) - q_{t+1} = n_t^\star(\xi_t). \tag{46}$$

Its size can be assessed without the autonomous rollout, although small defects still require controlled propagation. Recentring the enclosure at this known path can reduce the required neighborhood; it is not guaranteed to do so.

More generally, let $q_0, \dots, q_T$ be any prescribed shift-consistent path. Choose $K_t$ with the exact upper-slot shift, ordinarily $K_t = DG_t(q_t)$, and define

$$n_t^q(v) = G_t(q_t + v) - G_t(q_t) - K_t v, \tag{47}$$

$$d_0 = x_0 - q_0, \qquad d_{t+1} = K_t d_t + c_t, \qquad p_t = q_t + d_t. \tag{48}$$

Both $c_t$ and $n_t^q(v)$ lie in $\operatorname{im} B$. Put $\Psi(t,s) = K_{t-1} \cdots K_s$ for $t > s$ and $\Psi(s,s) = I$. For any path $v$, write $E(v) = \| \operatorname{stack}_{t=1}^T O(v_t - x_t^\star) \| / \sqrt{Q}$.

**Theorem 9** (Proxy-centered corrected enclosure)**.** *Suppose $\mathcal{R}_j > 0$ and $G_j$ is defined on $q_j + \overline{B}_X(0, \mathcal{R}_j)$, and $\|n_j^q(v)\| \le h_j^q(\|v\|)$ there, with finite nonnegative nondecreasing $h_j^q$ and $h_j^q(0) = 0$. Let independently justified bounds satisfy*

$$\widetilde{M}_{tj} \ge \|\Psi(t, j+1)B\|, \qquad \widetilde{N}_{tj} \ge \|O\Psi(t, j+1)B\|. \tag{49}$$

*Construct*

$$\rho_0 = \|d_0\|, \qquad \rho_t = \|d_t\| + \sum_{j<t} \widetilde{M}_{tj} h_j^q(\rho_j), \tag{50}$$

$$\widetilde{s}_t = \sum_{j<t} \widetilde{N}_{tj} h_j^q(\rho_j), \qquad \Delta_+ = \frac{\left(\sum_{t=1}^T \widetilde{s}_t^2\right)^{1/2}}{\sqrt{Q}}. \tag{51}$$

*If $\rho_j \le \mathcal{R}_j$ for every $j < T$, the rollout exists through $T$ and*

$$\|x_t - q_t\| \le \rho_t, \qquad \|O(x_t - p_t)\| \le \widetilde{s}_t, \qquad |E(x) - E(p)| \le \Delta_+. \tag{52}$$

*The same construction bounds the uncorrected path by*

$$|E(x) - E(q)| \le \frac{\left[\sum_{t=1}^T (\|Od_t\| + \widetilde{s}_t)^2\right]^{1/2}}{\sqrt{Q}}. \tag{53}$$

*Proof.* Set $e_t = x_t - q_t$ and $w_t = e_t - d_t = x_t - p_t$. Then

$$e_{t+1} = K_t e_t + c_t + n_t^q(e_t), \qquad w_0 = 0, \quad w_{t+1} = K_t w_t + n_t^q(e_t). \tag{54}$$

Therefore $w_t = \sum_{j<t} \Psi(t, j+1) B \nu_j$, with $B\nu_j = n_j^q(e_j)$. The same simultaneous existence and containment induction as above gives $\|\nu_j\| \le h_j^q(\rho_j)$, the state radii and the observed bounds. Stack and apply the reverse triangle inequality. Finally, $x_t - q_t = d_t + w_t$ gives the uncorrected bound. □

**Corollary 10** (A quadratic local regime)**.** *Suppose $h_j^q(r) \le \kappa r^2$ on the stated balls. Let $a = \max_{t \le T} \|d_t\|$ and $\Gamma = \max_{t \le T} \sum_{j<t} \widetilde{M}_{tj}$. If $2a \le \min_{j<T} \mathcal{R}_j$ and $4\kappa\Gamma a \le 1$, all tests pass, $\rho_t \le 2a$, and*

$$\Delta_+ \le \frac{4\kappa a^2}{\sqrt{Q}} \left[ \sum_{t=1}^T \left( \sum_{j<t} \widetilde{N}_{tj} \right)^2 \right]^{1/2}. \tag{55}$$

*Proof.* If earlier radii are at most $2a$, then $\rho_t \le a + 4\kappa\Gamma a^2 \le 2a$. The initial radius is at most $a$. Induction proves the tests; substitute $h_j^q(\rho_j) \le 4\kappa a^2$ into the observed bound. When $a = 0$, every correction and remainder is zero. □

**Proposition 11** (Correction defect and Newton interpretation)**.** *Whenever the indicated evaluations exist,*

$$G_t(p_t) - p_{t+1} = n_t^q(d_t). \tag{56}$$

*For $K_t = DG_t(q_t)$, $q \mapsto p$ is one Newton step for the finite trajectory equations, including the initial condition.*

*Proof.* Equation 48 gives $p_{t+1} = G_t(q_t) + K_t d_t$, proving the identity. The trajectory residual is $(q_0 - x_0, q_1 - G_0(q_0), \dots, q_T - G_{T-1}(q_{T-1}))$. Its derivative has identity diagonal blocks and subdiagonal blocks $-K_t$. The Newton equation is exactly Eq. 48; triangular invertibility requires neither invertible nor contractive $K_t$. □

### D.8 FIXED-HORIZON ORDER DOUBLING IN A DEFECT AMPLITUDE

**Proposition 12** (One correction of the reference-linear path)**.** *Fix $T < \infty$. Suppose each $G_t$ is continuously differentiable with locally Lipschitz derivative on an open neighborhood of $x_t^\star$, and take $A_t = DG_t(x_t^\star)$. Holding $b_t$ fixed, define*

$$G_t^\epsilon(x) = G_t(x) - (1-\epsilon)b_t, \qquad x_0^\epsilon = x_0^\star + \epsilon\eta_0, \tag{57}$$

$$q_t^\epsilon = x_t^\star + \epsilon\xi_t, \qquad K_t^\epsilon = DG_t(q_t^\epsilon). \tag{58}$$

*Apply Eq. 48 to this family, with $d_0^\epsilon = 0$. For sufficiently small $|\epsilon|$, all evaluations exist and, uniformly over the fixed horizon,*

$$c_t^\epsilon = O(\epsilon^2), \qquad x_t^\epsilon - q_t^\epsilon = O(\epsilon^2), \qquad x_t^\epsilon - p_t^\epsilon = O(\epsilon^4). \tag{59}$$

*Proof.* Choose closed balls inside the finitely many open reference neighborhoods. The locally Lipschitz derivatives supply a common quadratic Taylor bound on smaller balls about every sufficiently nearby $q_t^\epsilon$, with fixed positive allowable radii. Direct substitution yields $c_t^\epsilon = n_t^\star(\epsilon\xi_t) = O(\epsilon^2)$. The matrices $K_t^\epsilon$ and their finite products remain uniformly bounded as $\epsilon \to 0$. Variation of constants therefore gives $a_\epsilon = \max_t \|d_t^\epsilon\| = O(\epsilon^2)$. For sufficiently small amplitudes, both tests of Corollary 10 hold uniformly. Its containment bound gives $x^\epsilon - q^\epsilon = O(\epsilon^2)$, and the propagated-remainder representation gives $x^\epsilon - p^\epsilon = O(a_\epsilon^2) = O(\epsilon^4)$. The zero-defect case is exact. □

Here $\epsilon$ scales continuously injected reference defects and the initial deviation. It is not a mesh width, polynomial degree, timestep or donor-only probe amplitude. The original map occurs at $\epsilon = 1$, which need not lie in the guaranteed neighborhood. This statement concerns approximation of a learned trajectory, not improvement of its PDE prediction. The constants may depend strongly on the fixed horizon; the argument makes no uniform claim as $T$ grows. It also gives upper orders, rather than requiring an observed log–log slope of exactly four: a leading coefficient can vanish, and finite precision can dominate sufficiently small errors. Automatic differentiation at a nonsmooth activation boundary does not establish the local regularity assumed here; Theorem 9 can instead use a valid nonquadratic envelope.

### D.9 SIGNED INTERACTION AND NUMERICAL CERTIFICATION

For $t = 1, \dots, T$, let $r_t = Ob_{t-1}$, $y_t = O(x_t - x_t^\star)$ and $y_t^+ = O(p_t - x_t^\star)$. Stack these vectors and define $C = 2\langle r, y - r\rangle/Q$ and $C^+ = 2\langle r, y^+ - r\rangle/Q$. Under Theorem 9, form the adjoint

$$\lambda_T = O^* r_T, \qquad \lambda_t = O^* r_t + K_t^* \lambda_{t+1}, \quad t = T-1, \dots, 1. \tag{60}$$

Then

$$C - C^+ = \frac{2}{Q} \sum_{j=0}^{T-1} \langle B^* \lambda_{j+1}, \nu_j \rangle, \tag{61}$$

$$|C - C^+| \le \frac{2}{Q} \sum_{j=0}^{T-1} \|B^* \lambda_{j+1}\| h_j^q(\rho_j). \tag{62}$$

Indeed, insert the propagated-remainder representation for $x - p$ into the stacked inner product and interchange the two finite sums. The resulting backward weights are exactly $\lambda_{j+1}$; Cauchy–Schwarz gives the bound. This controls a signed quantity directly. It need not improve on $2\|r\|\Delta_+/\sqrt{Q}$; when both bounds apply, take their minimum.

For a computed correction $\widehat{d}$, its recurrence residual $\sigma_t = \widehat{d}_{t+1} - K_t \widehat{d}_t - c_t$ adds an exact defect:

$$G_t(q_t + \widehat{d}_t) - (q_{t+1} + \widehat{d}_{t+1}) = n_t^q(\widehat{d}_t) - \sigma_t. \tag{63}$$

An initial mismatch $\iota_0 = \widehat{d}_0 - (x_0 - q_0)$ also contributes $-\Psi(t, 0)\iota_0$ to $x_t - (q_t + \widehat{d}_t)$. Validated bounds for these residuals and model/JVP evaluation roundoff must enter a numerical certificate. Residuals computed with an approximate JVP do not alone bound its error relative to the declared $K_t$. If approximate histories lose shift consistency, their discrepancy requires whole-state propagation bounds until accounted for. Real-arithmetic proofs alone do not certify recorded floating-point rollouts.

Finally, sampled directional JVPs are not upper bounds for $M_{tj}$ or $N_{tj}$, and finite probe discrepancies are not uniform remainder envelopes on a ball. Appendix L evaluates the forced and corrected paths against actual neural trajectories. The measured vector discrepancies assess empirical approximation; they do not establish uniform remainder envelopes or worst-case operator bounds for the PDE models. In the Euclidean full-history norm with $m \geq 2$, the exact shift already implies $\|DG_t\| \geq 1$: a direction with zero oldest frame retains its entire norm in the shifted upper slots. No strict one-step contraction is assumed in these finite-horizon results.

A direct, less directional alternative is available when only a radial Lipschitz bound about $q$ is justified. If $\|G_t(q_t + v) - G_t(q_t)\| \leq \ell_t \|v\|$ on the stated ball and $\|c_t\| \leq \bar{c}_t$, construct $a_0 = \|x_0 - q_0\|$ and $a_{t+1} = \ell_t a_t + \bar{c}_t$. Passing the same domain tests proves $\|x_t - q_t\| \leq a_t$ by induction.

### D.10 Physical admissibility and the size of a local neighborhood

**Proposition 13** (A pointwise admissibility margin)**.** *Suppose every reference cell has $\rho^\star \geq \rho_{\min} > 0$ and $p(u^\star) \geq p_{\min} > 0$. On a full seven-channel raw-state neighborhood $\|u - u^\star\|_\infty \leq a$ with density bounded below by $\rho_{\min}/2$, let the solver-consistent pressure satisfy $|p(u) - p(u^\star)| \leq C_p \|u - u^\star\|_\infty$. If*

$$\|u - u^\star\|_\infty < \min\{a, \rho_{\min}/2, p_{\min}/(2C_p)\}, \tag{64}$$

*with the last restriction omitted when $C_p = 0$, then $\rho > \rho_{\min}/2$ and $p(u) > p_{\min}/2$.*

*Proof.* The density bound follows componentwise and places the perturbed state inside the stipulated positive-density neighborhood. Apply the pressure Lipschitz bound there. Such a finite $C_p$ follows from a bounded pressure gradient on a bounded neighborhood away from zero density. □

A full normalized-history bound can imply the required pointwise bound through $\|SL\eta\|_\infty \leq \|SL\|_{2\to\infty}\|\eta\|_2$. The six-channel observation norm alone cannot: keeping those channels exact while increasing $|\psi|$ eventually makes the pressure in Eq. 15 negative. This is a logical limitation of the score, not an attribution of observed failures to $\psi$. Physical admissibility and a useful linearization neighborhood are separate conditions; neither guarantees the other.

## E Complete primary absolute errors

Table 8 contains all degrees and architectures for both primary prediction modes. Its entries are generated from the archived common-reference-history rows, independently of the older degree-matched-history summaries.

Table 8: Common-reference-history errors for all four systems. Entries average three seeds and 20 identities; T denotes stacked one-step prediction and R denotes full closed loop.

| | | One-step error (T) | | | Closed-loop error (R) | | |
|---|---|---|---|---|---|---|---|
| System | Model | $p=1$ | $p=3$ | $p=5$ | $p=1$ | $p=3$ | $p=5$ |
| Allen–Cahn | ResNet | 0.0052 | 0.0005 | 0.0003 | 0.1526 | 0.0178 | 0.0167 |
| | U-Net | 0.0046 | 0.0009 | 0.0004 | 0.1528 | 0.0355 | 0.0152 |
| | DeepONet | 0.0100 | 0.0100 | 0.0101 | 0.1554 | 0.1552 | 0.1586 |
| | FNO | 0.0034 | 0.0014 | 0.0013 | 0.1545 | 0.0484 | 0.0452 |
| | Transolver | 0.0034 | 0.0005 | 0.0005 | 0.1517 | 0.0506 | 0.0514 |
| | DPOT | 0.0053 | 0.0013 | 0.0013 | 0.1494 | 0.0552 | 0.0546 |
| | GNOT | 0.0099 | 0.0099 | 0.0099 | 0.1545 | 0.1537 | 0.1531 |
| GLM–MHD | ResNet | 0.0188 | 0.0055 | 0.0050 | 0.0405 | 0.0725 | 0.0928 |
| | U-Net | 0.0202 | 0.0084 | 0.0100 | 0.0311 | 0.0904 | 0.1128 |
| | DeepONet | 0.0560 | 0.0561 | 0.0560 | 0.4073 | 0.3918 | 0.3873 |
| | FNO | 0.0259 | 0.0150 | 0.0176 | 0.0494 | 0.2002 | 0.2545 |
| | Transolver | 0.0198 | 0.0073 | 0.0077 | 0.1417 | 0.5057 | 0.4939 |
| | DPOT | 0.0224 | 0.0203 | 0.0250 | 0.1039 | 0.4070 | 0.4879 |
| | GNOT | 0.0328 | 0.0333 | 0.0339 | 0.1258 | 0.2904 | 0.3857 |
| RMI | ResNet | 0.0255 | 0.0080 | 0.0019 | 0.0555 | 0.0428 | 0.0422 |
| | U-Net | 0.0274 | 0.0090 | 0.0029 | 0.1527 | 0.1908 | 0.1463 |
| | DeepONet | 0.0651 | 0.0652 | 0.0652 | 0.2705 | 0.2690 | 0.2809 |
| | FNO | 0.0348 | 0.0137 | 0.0040 | 0.0646 | 0.0945 | 0.1361 |
| | Transolver | 0.0414 | 0.0127 | 0.0024 | 0.1840 | 0.1447 | 0.0617 |
| | DPOT | 0.0314 | 0.0116 | 0.0043 | 0.0551 | 0.0624 | 0.0750 |
| | GNOT | 0.0422 | 0.0316 | 0.0302 | 0.0753 | 0.0962 | 0.1674 |
| SWE | ResNet | 0.0164 | 0.0026 | 0.0015 | 0.0553 | 0.0578 | 0.0576 |
| | U-Net | 0.0175 | 0.0034 | 0.0024 | 0.0576 | 0.0572 | 0.0604 |
| | DeepONet | 0.0396 | 0.0396 | 0.0396 | 0.2381 | 0.2382 | 0.2384 |
| | FNO | 0.0198 | 0.0043 | 0.0038 | 0.0334 | 0.0331 | 0.0331 |
| | Transolver | 0.0250 | 0.0036 | 0.0022 | 0.1565 | 0.1719 | 0.1346 |
| | DPOT | 0.0193 | 0.0055 | 0.0046 | 0.1340 | 0.0909 | 0.0890 |
| | GNOT | 0.0396 | 0.0396 | 0.0396 | 0.2383 | 0.2382 | 0.2383 |

# F PERSISTENCE SKILL AND ARCHITECTURE-SUBSET SENSITIVITY

**Definition and audit population.** For each system, architecture, degree, and prediction mode, we define

$$\text{Skill} = 1 - \frac{\overline{\mathcal{E}}_{\text{model}}}{\overline{\mathcal{E}}_{\text{persistence}}}, \tag{65}$$

where the bars average the same per-identity relative $L^2$ errors over the original three seeds and 20 test identities. Persistence receives the same common-reference history as the model. In one-step evaluation it copies the last true frame at each refreshed query; in free rollout it repeats the fifth initial frame throughout the forecast. Zero residual and last-frame prediction therefore give the same persistence baseline under the residual interface. Persistence is repeated across seeds only to align the paired observations, not to create additional independent numerical samples.

The audit covers all $4 \times 7 \times 3 \times 3 \times 20 \times 2 = 10{,}080$ primary scoring cells. Reaggregating the model errors reproduces the original all-seven results to numerical precision. Table 9(a) reports degree-one skill, the baseline used to define the sensitivity subset.

Within each system, the descriptive subset contains architectures whose degree-one skill exceeds $1\%$ in both one-step and free-rollout modes. Its membership is then fixed across the degree-five comparison and all bootstrap draws. The aggregate remains the geometric mean of architecture-specific ratios of mean errors, with the original 5,000 paired crossed seed–identity draws.

The subset retains all seven models for GLM and RMI; ResNet, U-Net, Transolver, and DPOT for Allen–Cahn; and all except DeepONet and GNOT for SWE. SWE's excluded models have one-step skill of $0.066\%$ and $0.014\%$, respectively. Allen–Cahn also has weak or negative skill for

some models, particularly in rollout. The sensitivity preserves the qualitative findings: Allen–Cahn benefits, GLM reverses, and RMI/SWE aggregate gains remain unresolved. In particular, all seven GLM architectures pass the joint threshold, so the primary reversal is unchanged.

Table 9: Predictive skill and aggregate sensitivity on the original test identities. (a) Degree-one skill in percent, defined as one minus the ratio of mean model error to mean persistence error; T and R denote one step and rollout. (b) All seven architectures remain the primary analysis. The post-hoc subset requires degree-one skill above 1.000% in both modes, with membership then fixed for the degree-five comparison. Intervals use the original crossed seed–identity bootstrap, conditional on each fixed subset.

**(a) Degree-one skill (%)**

| | Allen–Cahn | | GLM–MHD | | RMI | | SWE | |
|---|---|---|---|---|---|---|---|---|
| Model | T | R | T | R | T | R | T | R |
| ResNet | 47.520 | 1.327 | 67.394 | 92.111 | 61.977 | 90.573 | 58.692 | 76.823 |
| U-Net | 53.903 | 1.228 | 64.941 | 93.947 | 59.113 | 74.062 | 55.858 | 75.858 |
| DeepONet | -0.122 | -0.423 | 2.803 | 20.703 | 2.827 | 54.052 | 0.066 | 0.236 |
| FNO | 66.052 | 0.161 | 54.992 | 90.387 | 48.119 | 89.028 | 50.102 | 85.984 |
| Transolver | 66.111 | 1.960 | 65.663 | 72.413 | 38.245 | 68.752 | 36.950 | 34.435 |
| DPOT | 46.401 | 3.397 | 61.067 | 79.763 | 53.112 | 90.637 | 51.193 | 43.826 |
| GNOT | 0.126 | 0.121 | 43.120 | 75.503 | 37.044 | 87.217 | 0.014 | 0.132 |

**(b) Aggregate $p = 5/p = 1$ error ratios**

| System | Subset | $N$ | One step [95% CI] | Rollout [95% CI] |
|---|---|---|---|---|
| Allen–Cahn | All seven | 7 | 0.251 [0.236, 0.265] | 0.327 [0.293, 0.359] |
| Allen–Cahn | Skill $> 1\%$ | 4 | 0.112 [0.101, 0.122] | 0.192 [0.164, 0.221] |
| GLM–MHD | All seven | 7 | 0.631 [0.610, 0.651] | 2.971 [2.819, 3.121] |
| GLM–MHD | Skill $> 1\%$ | 7 | 0.631 [0.610, 0.651] | 2.971 [2.819, 3.121] |
| RMI | All seven | 7 | 0.176 [0.157, 0.195] | 1.071 [0.851, 1.270] |
| RMI | Skill $> 1\%$ | 7 | 0.176 [0.157, 0.195] | 1.071 [0.851, 1.270] |
| SWE | All seven | 7 | 0.243 [0.228, 0.256] | 0.934 [0.873, 1.034] |
| SWE | Skill $> 1\%$ | 5 | 0.138 [0.126, 0.149] | 0.908 [0.826, 1.047] |

# G Complete GLM training controls and checkpoint sensitivity

The K1/K4 and capacity analysis uses 5,000 crossed seed–identity bootstrap draws. The main text reports validation-best checkpoints; fixed epoch 40 provides a separate sensitivity endpoint.

## G.1 Feedback resets and shared generated histories

The reset diagnostic uses the original production checkpoints. After every $K \in \{1, 2, 4, 8, 16\}$ predicted frames it restores the five-frame reference history at the corresponding time. Error always covers 56 predictions, including any physically invalid frames. The degree-five/one ratios for U-Net are $0.493, 0.646, 1, 1.563, 2.021$, followed by $3.63$ without resets; for FNO they are $0.679, 0.92, 1.417, 2.172, 2.837$, followed by $5.156$. The $K = 1$ endpoint matches common-reference teacher forcing. Reset intervals in Figure 3 are recomputed from archived paired cells using 5,000 crossed draws. Reset length and training-unroll length are different experimental variables.

Generated-history continuation instead transfers a naturally reached donor history once, then lets each receiver run autonomously to the original terminal time. A degree-five base-K1 FNO supplies the same five-frame input to base K1 and K4 receivers. At forecast step eight, the remaining suffix has 48 frames, and all 60 donor histories are physically valid; K4/K1 suffix error is $0.3996$ $[0.2595, 0.5846]$. At step 32 the suffix has 24 frames, 36/60 donor histories have already failed, and the ratio is $0.5057$ $[0.3127, 0.7952]$. These FNO comparisons show a benefit even before donor failure.

Table 10: Complete capacity–objective comparison. Entries are mean full-horizon reference errors. "Best" denotes validation-best and "final" fixed epoch 40. The failure column counts degree-five seed–identity cells with any physical failure under best selection (out of 60). K4 does not recover the degree ordering in any row.

| | | | Best | | Final | | Best $p5$ |
|---|---|---|---|---|---|---|---|
| Model | Capacity | Loss | $p1$ | $p5$ | $p1$ | $p5$ | failed |
| FNO | base | K1 | 0.0489 | 0.3503 | 0.0489 | 0.3503 | 60/60 |
| FNO | base | K4 | 0.0362 | 0.1251 | 0.0362 | 0.1251 | 60/60 |
| FNO | large | K1 | 0.0359 | 0.1440 | 0.0359 | 0.1440 | 58/60 |
| FNO | large | K4 | 0.0286 | 0.0740 | 0.0286 | 0.0740 | 43/60 |
| U-Net | base | K1 | 0.0435 | 0.0979 | 0.0423 | 0.1054 | 60/60 |
| U-Net | base | K4 | 0.0342 | 0.0713 | 0.0342 | 0.0713 | 58/60 |
| U-Net | large | K1 | 0.0325 | 0.1076 | 0.0325 | 0.0984 | 59/60 |
| U-Net | large | K4 | 0.0320 | 0.0689 | 0.0320 | 0.0689 | 47/60 |

### G.2 Matched objectives, capacity, and full results

The factorial study trains 48 models: two architectures, two degrees, two capacities, two objectives, and three seeds. Base/large U-Net widths are 48/96; FNO widths are 64/128 with 24 modes and four layers in both cases. The two objectives share initialization within the architecture, capacity, and seed, as well as the normalizer, physical identities, window streams, 40-epoch schedule, and an effective batch size of 32. Both sample 32 windows per identity per epoch from the same 53 legal four-step starts. Original production K1 permits 56 one-step starts, so its checkpoint is not substituted for the factorial K1 control.

K4 averages four per-step mean squared losses on standardized seven-channel states, recursively feeding each prediction into the next five-frame history and retaining gradients through the unroll. K1 scores its first next-frame prediction from the same sampled starts. Training uses AdamW with learning rate $3 \times 10^{-4}$, two warm-up epochs, cosine decay to $1 \times 10^{-6}$, weight decay $1 \times 10^{-4}$, betas $(0.9, 0.95)$, and gradient clipping at norm one. Validation-best selects the full one-step normalized validation MSE evaluated every two epochs.

All 24 FNO best checkpoints equal their epoch-40 endpoints. Across both architectures, 45 of 48 best/final states are identical; three U-Net configurations differ. FNO base degree-five K4/K1 is $0.3573$ $[0.2194, 0.5359]$ under either choice. U-Net base changes from $0.7283$ $[0.6639, 0.7876]$ at best to $0.6764$ $[0.6224, 0.7493]$ at final. Both preserve the degree-five error reduction.

Define the degree–objective interaction as the degree-five K4/K1 ratio divided by the corresponding degree-one ratio. For base FNO it is $0.4817$ $[0.3033, 0.7129]$. For base U-Net it changes from $0.9252$ $[0.7606, 1.1381]$ at best to $0.8356$ $[0.6885, 0.9484]$ at final. Thus the absolute degree-five benefit persists.

One-step and autonomous gains differ even within these controls. For degree-five FNO, the common-reference one-step error changes from $0.0172$ to $0.0169$, ratio $0.9809$ $[0.9557, 1.0087]$, alongside a 64.3% rollout reduction. For U-Net it increases from $0.01$ to $0.0117$ at best while rollout error decreases; this direction persists at final. The one-step improvement is neither an adequate measure of the rollout gain nor necessary for it in this example.

### G.3 The specified 120-epoch extension

The extension comprises 24 continuations for Allen/GLM, FNO/U-Net, degrees one/five, and three seeds. It starts from each original epoch-40 last checkpoint with optimizer state retained. An additional 80-epoch cosine schedule decreases learning rate from $1 \times 10^{-4}$ to $1 \times 10^{-6}$. Table 11 uses fixed epoch 120; a best checkpoint selected during the extension is a different endpoint.

### G.4 Four-target objective controls

The TF4/BP4 study uses the objectives in Eq. 13 and the SG4 definition in Appendix B.2. Each arm receives four targets at the same 53 legal starts, 10,040 optimizer updates, and effective batch size

Table 11: Fixed epoch-120 results with common reference histories. Intervals are nominal 95% crossed-bootstrap intervals; both GLM degree-five models physically fail in all 60 seed–identity cells.

| System | Model | $p1$ error | $p5$ error | $p5/p1$ [95% interval] |
|---|---|---|---|---|
| Allen | U-Net | 0.153 | 0.020 | 0.128 [0.079, 0.194] |
| Allen | FNO | 0.156 | 0.042 | 0.268 [0.202, 0.340] |
| GLM | U-Net | 0.033 | 0.091 | 2.747 [2.533, 2.960] |
| GLM | FNO | 0.043 | 0.212 | 4.928 [4.402, 5.457] |

Table 12: FNO four-target controls on the original cohort. Every degree-five arm has physical failure in all 60 cells.

| Arm | $p1$ error | $p5$ error | $p1$ failed | $p5$ failed |
|---|---|---|---|---|
| TF4 | 0.049 | 0.343 | 29/60 | 60/60 |
| SG4 | 0.038 | 0.131 | 23/60 | 60/60 |
| BP4 | 0.036 | 0.125 | 19/60 | 60/60 |

32. Twelve TF4/SG4 jobs are new; six compatible factorial BP4 jobs are reused. All 18 best/final checkpoint states are identical.

Each primary objective contrast uses 97.5% pointwise percentile intervals across its two degrees, yielding Bonferroni simultaneous 95% coverage within that contrast. At degree one, SG4/TF4 is $0.7674$ $[0.7248, 0.7997]$ and BP4/SG4 is $0.9555$ $[0.9357, 0.9752]$. At degree five these are $0.383$ $[0.2335, 0.5732]$ and $0.9539$ $[0.8674, 1.0919]$. Seed 41 reverses the latter direction.

### G.5 Single-stage high-fidelity controls

The single-stage $H$-TF4/SG4 controls use the objectives in Eq. 13 and Appendix B.2. On seed 40 and the original 20 exploratory test identities, FNO $H$ error fell from $0.0575$ to $0.0365$, but matched $H$-SG4/$p = 1$-SG4 was $0.981$ $[0.944, 1.017]$ with 9/20 versus 6/20 physical failures. U-Net $H$ error increased from $0.0495$ to $0.1179$, and $H$-SG4/$p = 1$-SG4 was $2.06$ $[1.895, 2.231]$, with 19/20 versus 6/20 failures.

### G.6 Additional receiver and history checks

These original-cohort diagnostics examine receiver dynamics, spectral capacity, and history length.

**Residual exchange and receiver dynamics.** K1 and K4 receivers produce different rollouts under the same injected K1 residual sequence measured along a common reference path. Conversely, supplying degree-five K1 receivers with K4 residuals gives residual-swap/self-rollout error ratios of approximately $0.959$ for FNO and $1.072$ for U-Net, without reproducing the main K4 benefit. Together, these comparisons support a role for receiver dynamics alongside the reference-path residual sequence. The interventions require future reference information and serve as oracle diagnostics.

**Spectral capacity, history, and field controls.** For degree-five K1 FNO, widening from 64 to 128 at 24 modes yields error $0.144$, compared with $0.1944$ from increasing modes from 24 to 48 at width 64. A nine-slot full-history/masked-five comparison over an aligned 52-step horizon gives degree-five ratios of $1.1$ for FNO and $0.988$ for U-Net. Spectral amplitude, phase, radial, and angular interventions affect selected oracle rollouts. Cleaning-field replacements improve some late rollouts with pronounced seed dependence, while the direct $\psi^2$ pressure contribution does not explain most observed negative pressure.

# H FOUR NUMERICAL SOURCES UNDER A COMMON TWO-STAGE PROTOCOL

## H.1 SOURCES AND FIXED-RECIPE COMPARISON

The comparison uses the four existing GLM sources $128^2, p = 1, 3, 5$ and $H$. U-Net and FNO use seed 40 and the shared TF4-to-BP4 protocol in Appendix B.3. The primary endpoint is the fixed final checkpoint; validation-best is a sensitivity check. The completed comparison contains eight source–architecture cells.

**Scoring and uncertainty.** Evaluation uses discovery identities 900001–900010, common $512^2$ reference histories, and the full 56-step six-channel score. Errors are arithmetic means of per-identity relative $L^2$ errors. Physical failures follow Eq. 15 and remain included.

The paired bootstrap resamples the ten identities 20,000 times. The same draws apply across sources and architectures. Against $p = 1$, final $p = 3$ rollout ratios are $1.75$ $[1.62, 1.92]$ for U-Net and $3.868$ $[3.61, 4.147]$ for FNO; $p = 5$ ratios are $2.846$ $[2.654, 3.071]$ and $4.683$ $[4.466, 4.925]$. The $H/p = 1$ ratios are $0.387$ $[0.348, 0.436]$ and $0.991$ $[0.949, 1.031]$, respectively. Table 13 gives all three errors and physical failures for final and best checkpoints.

Table 13: Four-source comparison on the ten explored transfer identities: seed 40, common $512^2$ reference histories and scores. All errors are per-identity full-horizon relative $L_2$ means over six physical channels. Failed counts refer to autonomous predictions. $H$ has zero label discrepancy by construction. Best checkpoints are selected only on the old validation set.

| | | U-Net | | | FNO | | |
|---|---|---|---|---|---|---|---|
| Source | $E_{\text{label}}$ | One step | Rollout | Failed | One step | Rollout | Failed |
| *Fixed final: primary endpoint* | | | | | | | |
| $p = 1$ | 0.0238 | 0.0199 | 0.0257 | 0/10 | 0.0266 | 0.0450 | 3/10 |
| $p = 3$ | 0.0052 | 0.0082 | 0.0449 | 7/10 | 0.0148 | 0.1740 | 10/10 |
| $p = 5$ | 0.0034 | 0.0094 | 0.0731 | 10/10 | 0.0172 | 0.2107 | 10/10 |
| $H$ | 0.0000 | 0.0017 | 0.0099 | 0/10 | 0.0045 | 0.0446 | 5/10 |
| *Validation-best: sensitivity* | | | | | | | |
| $p = 1$ | 0.0238 | 0.0199 | 0.0250 | 0/10 | 0.0266 | 0.0447 | 3/10 |
| $p = 3$ | 0.0052 | 0.0082 | 0.0449 | 6/10 | 0.0148 | 0.1740 | 10/10 |
| $p = 5$ | 0.0034 | 0.0094 | 0.0721 | 9/10 | 0.0172 | 0.2096 | 10/10 |
| $H$ | 0.0000 | 0.0017 | 0.0099 | 0/10 | 0.0045 | 0.0445 | 5/10 |

## H.2 SOURCE EXCHANGE BETWEEN TRAINING STAGES

The completed ABC comparison evaluates three routes: A, $p = 1$ first stage followed by $p = 1$ BP4; B, $p = 1$ first stage followed by $H$ BP4; and C, $H$ first stage followed by $H$ BP4. All use the same additional-update budget. Table 14 uses the same ten discovery identities.

Table 14: Completed ABC boundary: fixed final, ten explored transfer identities, common $512^2$ reference histories. The two-stage route appears in the second column; every continuation uses BP4.

| Model | Route | One-step error | Closed-loop error | Failed |
|---|---|---|---|---|
| FNO | A: $p = 1 \to p = 1$ | 0.0266 | 0.0450 | 3/10 |
| FNO | B: $p = 1 \to H$ | 0.0084 | 0.0515 | 8/10 |
| FNO | C: $H \to H$ | 0.0045 | 0.0446 | 5/10 |
| U-Net | A: $p = 1 \to p = 1$ | 0.0199 | 0.0257 | 0/10 |
| U-Net | B: $p = 1 \to H$ | 0.0034 | 0.0123 | 2/10 |
| U-Net | C: $H \to H$ | 0.0017 | 0.0099 | 0/10 |

FNO C/A is $0.991$ $[0.95, 1.031]$ in closed loop despite a one-step ratio of $0.168$ $[0.159, 0.177]$. Its B/A closed-loop ratio is $1.145$ $[1.115, 1.177]$, and all ten identities worsen. The fixed-recipe FNO comparison does not reproduce the clear U-Net benefit. Appendix I separately reports a small FNO

$H$ benefit after bounded second-stage adaptation on the original test cohort; its models and identities differ from this fixed-recipe ABC comparison.

### H.3 Finer-reference sensitivity

Identities 900001–900005 also have an existing $1024^2, p = 1$ reference. Table 15 compares scores against $512^2$ and $1024^2$ on exactly these five identities. The $512^2$ initial histories and forecast arrays remain fixed; only the scoring reference changes. This is distinct from the five-identity sensitivity subset of the later confirmation population.

$H$'s zero $512^2$ label error is a construction identity, not zero discretization error. Against $1024^2$, its label error is nonzero ($0.002$), and the four-source label ordering and architecture-dependent rollout pattern remain. For the fixed final endpoint, the $1024^2$ U-Net rollout means are $0.0243$, $0.0432$, $0.0716$, and $0.0105$ for $p = 1, 3, 5, H$; FNO means are $0.0439$, $0.157$, $0.208$, and $0.0445$.

Table 15: Same five identities under two scoring references. Within each checkpoint selection, the initial $512^2$ histories and predicted arrays are identical in the 512 and 1024 rows; only the scoring reference changes. Numerical label error is also rescored. These are identities 900001–900005, distinct from the fresh confirmation subset.

| | | U-Net | | | FNO | | |
|---|---|---|---|---|---|---|---|
| Source | $E_{\text{label}}$ | One step | Rollout | Failed | One step | Rollout | Failed |
| *Fixed final: five identities, $512^2$ reference* | | | | | | | |
| $p = 1$ | 0.0215 | 0.0176 | 0.0234 | 0/5 | 0.0232 | 0.0433 | 1/5 |
| $p = 3$ | 0.0047 | 0.0071 | 0.0429 | 3/5 | 0.0130 | 0.1568 | 5/5 |
| $p = 5$ | 0.0030 | 0.0081 | 0.0714 | 5/5 | 0.0152 | 0.2079 | 5/5 |
| $H$ | 0.0000 | 0.0015 | 0.0101 | 0/5 | 0.0039 | 0.0442 | 2/5 |
| *Fixed final: five identities, $1024^2$ reference* | | | | | | | |
| $p = 1$ | 0.0224 | 0.0181 | 0.0243 | 0/5 | 0.0236 | 0.0439 | 1/5 |
| $p = 3$ | 0.0053 | 0.0077 | 0.0432 | 3/5 | 0.0134 | 0.1570 | 5/5 |
| $p = 5$ | 0.0029 | 0.0085 | 0.0716 | 5/5 | 0.0156 | 0.2080 | 5/5 |
| $H$ | 0.0020 | 0.0025 | 0.0105 | 0/5 | 0.0045 | 0.0445 | 2/5 |
| *Validation-best: five identities, $512^2$ reference* | | | | | | | |
| $p = 1$ | 0.0215 | 0.0176 | 0.0229 | 0/5 | 0.0232 | 0.0431 | 1/5 |
| $p = 3$ | 0.0047 | 0.0071 | 0.0432 | 2/5 | 0.0130 | 0.1568 | 5/5 |
| $p = 5$ | 0.0030 | 0.0081 | 0.0697 | 5/5 | 0.0152 | 0.2072 | 5/5 |
| $H$ | 0.0000 | 0.0015 | 0.0101 | 0/5 | 0.0039 | 0.0442 | 2/5 |
| *Validation-best: five identities, $1024^2$ reference* | | | | | | | |
| $p = 1$ | 0.0224 | 0.0181 | 0.0238 | 0/5 | 0.0236 | 0.0437 | 1/5 |
| $p = 3$ | 0.0053 | 0.0077 | 0.0435 | 2/5 | 0.0134 | 0.1570 | 5/5 |
| $p = 5$ | 0.0029 | 0.0085 | 0.0699 | 5/5 | 0.0156 | 0.2073 | 5/5 |
| $H$ | 0.0020 | 0.0025 | 0.0105 | 0/5 | 0.0045 | 0.0444 | 2/5 |

## I Bounded, equal-budget source adaptation

This experiment adapts the shared second-stage recipe to each numerical source while reusing its TF4 first stage (Appendix B.3).

### I.1 Frozen search and selection protocol

The six groups are U-Net/FNO $\times\{p = 1, p = 5, H\}$, all at the same $128^2$ interface and seed 40. Each reuses its existing 40-epoch TF4 first stage. The second-stage grid crosses peak learning rates $\{1\times10^{-5}, 3\times10^{-6}, 3\times10^{-5}\}$ with recursive unroll lengths $K \in \{2, 4\}$, with differentiation through predicted histories. Each of the six candidates per group receives 40 epochs, 10,040 optimizer updates, and an effective batch size of 32.

Every candidate starts a fresh AdamW optimizer with betas $(0.9, 0.95)$, $\epsilon = 1 \times 10^{-8}$, weight decay $1 \times 10^{-4}$, norm-one gradient clipping, and cosine decay to one tenth of its peak rate. The

Table 16: **All 36 candidate validation scores.** Each entry is the minimum mean 56-step validation error over the same 21 scheduled evaluations; bold marks the selected candidate in each column. Every trial completes 10,040 updates without numerical failure.

| Unroll | Peak rate | U-Net | | | FNO | | |
|---|---|---|---|---|---|---|---|
| | | $p=1$ | $p=5$ | $H$ | $p=1$ | $p=5$ | $H$ |
| 4 | $10^{-5}$ | 0.0277 | 0.0791 | 0.0127 | 0.0514 | 0.2157 | 0.0515 |
| 2 | $10^{-5}$ | 0.0283 | 0.0849 | 0.0142 | 0.0544 | 0.3047 | 0.0572 |
| 4 | $3\times10^{-6}$ | 0.0284 | 0.0845 | 0.0143 | 0.0538 | 0.2833 | 0.0559 |
| 2 | $3\times10^{-6}$ | 0.0284 | 0.0884 | 0.0154 | 0.0559 | 0.4237 | 0.0597 |
| 4 | $3\times10^{-5}$ | **0.0274** | **0.0735** | **0.0116** | **0.0481** | **0.1876** | **0.0461** |
| 2 | $3\times10^{-5}$ | 0.0277 | 0.0807 | 0.0131 | 0.0521 | 0.2236 | 0.0534 |

balanced seed-40 sampling stream and legal four-target windows are shared; $K = 2$ uses the first two targets from the same windows. The first stage, architecture, normalization, numerical source, and observation interface are not retuned.

Validation uses the original 20 validation identities, common $512^2$ reference histories, and the full 56-step raw six-channel error of Eq. 3, averaged equally over identities. There are 21 evaluation opportunities: update zero and every 502 updates through 10,040. Each group's candidate and endpoint minimize this validation mean; ties follow the frozen candidate order and then earliest update. All 36 trials finish without numerical failure. Selection is frozen for all six groups before this search accesses the original 20 test identities.

Table 16 retains every candidate's best validation score. All six groups select $K = 4$ and peak rate $3 \times 10^{-5}$, although source-specific choices were allowed. The selected update is 10,040 for all U-Net sources and FNO-$H$, 9,538 for FNO-$p = 1$, and 9,036 for FNO-$p = 5$.

### I.2 Test endpoints and fixed-recipe controls

All selected models use the same original 20 test identities, five reference input frames, seven-channel history recurrence, and 56-step evaluation horizon. The primary score includes six physical channels. One-step predictions use the common reference history at every step. The fixed-recipe comparator is the newly run $K = 4$, peak-rate $1 \times 10^{-5}$ candidate at its final update; a separate sensitivity selects its best validation endpoint. Thus, adaptation and both controls use the same test cohort, unlike the ten-identity four-source display. All three policies are reported in Table 17.

Relative to fixed-final controls, adaptation reduces $p = 5$ rollout error from $0.0742$ to $0.0702$ for U-Net (5.4%) and from $0.2039$ to $0.1758$ for FNO (13.8%). Adapted $p = 1$ also improves, retaining the source reversal. The common-reference one-step $p = 5/p = 1$ ratios remain below one: approximately $0.479$ for U-Net and $0.652$ for FNO. Comparing against validation-best fixed controls also retains the rollout reversal (Table 18).

### I.3 Paired comparisons and statistical scope

For each architecture and selection policy, the source ratio divides the arithmetic mean error for $p = 5$ or $H$ by the corresponding $p = 1$ mean. Marginal 95% percentile intervals use 5,000 paired identity-bootstrap draws; every draw applies the same identity weights to numerator and denominator. The intervals are conditional on the already selected seed-40 models.

Adapted $H$ reduces rollout error relative to adapted $p = 1$ by 60.5% for U-Net and 5.3% for FNO. On these same test identities, the fixed-final FNO $H/p = 1$ ratio is $1.01$ $[0.977, 1.042]$, while the adapted ratio is $0.947$ $[0.917, 0.974]$. This supports a small FNO benefit under the selected recipe.

Physical admissibility follows Appendix B, including the cleaning field in pressure. Both adapted $p = 5$ models fail physically on 20/20 trajectories. Adapted U-Net has no failures for $p = 1$ or $H$; adapted FNO has 7/20 and 10/20, respectively. All predictions remain finite and all failing trajectories remain in the full-horizon means.

Table 17: **All test endpoints for bounded adaptation and fixed controls.** Every row uses the same 20 identities, reference histories, and 56-step horizon. Fixed final uses four-step continuation at peak learning rate $10^{-5}$; fixed val-best selects the endpoint within that recipe. Physical failures are ever-failed trajectories under Eq. 15; all remain in the error means. No row contains non-finite forecasts.

| Model | Source | Selection | One-step | Rollout | Failures |
|---|---|---|---|---|---|
| U-Net | $p = 1$ | Fixed final | 0.0199 | 0.0260 | 0/20 |
| U-Net | $p = 1$ | Fixed val-best | 0.0199 | 0.0253 | 0/20 |
| U-Net | $p = 1$ | Adapted | 0.0198 | 0.0251 | 0/20 |
| U-Net | $p = 5$ | Fixed final | 0.0096 | 0.0742 | 20/20 |
| U-Net | $p = 5$ | Fixed val-best | 0.0096 | 0.0749 | 20/20 |
| U-Net | $p = 5$ | Adapted | 0.0095 | 0.0702 | 20/20 |
| U-Net | $H$ | Fixed final | 0.0018 | 0.0109 | 1/20 |
| U-Net | $H$ | Fixed val-best | 0.0018 | 0.0109 | 1/20 |
| U-Net | $H$ | Adapted | 0.0017 | 0.0099 | 0/20 |
| FNO | $p = 1$ | Fixed final | 0.0264 | 0.0468 | 8/20 |
| FNO | $p = 1$ | Fixed val-best | 0.0264 | 0.0467 | 8/20 |
| FNO | $p = 1$ | Adapted | 0.0262 | 0.0442 | 7/20 |
| FNO | $p = 5$ | Fixed final | 0.0173 | 0.2039 | 20/20 |
| FNO | $p = 5$ | Fixed val-best | 0.0173 | 0.2028 | 20/20 |
| FNO | $p = 5$ | Adapted | 0.0171 | 0.1758 | 20/20 |
| FNO | $H$ | Fixed final | 0.0046 | 0.0473 | 12/20 |
| FNO | $H$ | Fixed val-best | 0.0046 | 0.0473 | 12/20 |
| FNO | $H$ | Adapted | 0.0045 | 0.0419 | 10/20 |

Table 18: **Paired rollout ratios on the same original test cohort.** Entries are ratios of identity means with marginal 95% paired identity-bootstrap intervals, conditional on each set of seed-40 models. Each numerator and denominator use the same selection policy.

| Model | Selection | $p = 5/p = 1$ | $H/p = 1$ |
|---|---|---|---|
| U-Net | Fixed final | 2.852 [2.681, 3.046] | 0.418 [0.381, 0.455] |
| U-Net | Fixed val-best | 2.957 [2.783, 3.152] | 0.429 [0.392, 0.465] |
| U-Net | Adapted | 2.796 [2.620, 2.993] | 0.395 [0.360, 0.431] |
| FNO | Fixed final | 4.353 [4.082, 4.664] | 1.010 [0.977, 1.042] |
| FNO | Fixed val-best | 4.343 [4.073, 4.652] | 1.012 [0.981, 1.042] |
| FNO | Adapted | 3.977 [3.720, 4.242] | 0.947 [0.917, 0.974] |

The fidelity reversal persists under this equal-budget search over second-stage training configurations.

# J HIGH-FIDELITY SOURCE CONTROLS AND FROZEN CONFIRMATION

The source-by-objective and confirmation branches use the shared two-stage protocol in Appendix B.3.

## J.1 SOURCE-BY-OBJECTIVE DISCOVERY MATRIX

The completed seed-40 source-by-objective matrix was evaluated on identities 900001–900010. They are distinct from both the original 20 test identities and the later 20 fresh confirmation identities. Table 19 reports the fixed-final matrix using common $512^2$ reference histories.

Table 19: Discovery-stage U-Net matrix: seed 40, ten previously explored transfer identities, fixed final, common $512^2$ reference. Failure counts refer to the closed-loop predictions.

| Source | Objective | One-step error | Closed-loop error | Failed |
|---|---|---|---|---|
| $p = 1$ | TF4 | 0.0197 | 0.0259 | 0/10 |
| $p = 1$ | BP4 | 0.0199 | 0.0257 | 0/10 |
| $H$ | TF4 | 0.0016 | 0.0122 | 1/10 |
| $H$ | BP4 | 0.0017 | 0.0099 | 0/10 |
| $p = 5$ | TF4 | 0.0091 | 0.0827 | 10/10 |
| $p = 5$ | BP4 | 0.0094 | 0.0731 | 10/10 |

Under TF4 and BP4, $H/p = 1$ closed-loop ratios are respectively $0.472$ $[0.43, 0.522]$ and $0.387$ $[0.348, 0.437]$. For a fixed seed, let $\mathcal{E}_{a,q}$ denote the arithmetic mean of identity errors for source $a$ and objective $q$, and define

$$R_H = \frac{\mathcal{E}_{H,\mathrm{BP4}}}{\mathcal{E}_{1,\mathrm{BP4}}}, \quad I_H = \frac{\mathcal{E}_{H,\mathrm{BP4}}/\mathcal{E}_{1,\mathrm{BP4}}}{\mathcal{E}_{H,\mathrm{TF4}}/\mathcal{E}_{1,\mathrm{TF4}}}, \quad R_5 = \frac{\mathcal{E}_{5,\mathrm{BP4}}}{\mathcal{E}_{1,\mathrm{BP4}}}. \tag{66}$$

Here $I_H = 0.82$ $[0.799, 0.844]$ directly measures the enhancement of the relative $H$ benefit. The original $p = 5$ branch remains worse than $p = 1$: its TF4 and BP4 ratios are $3.189$ $[3.007, 3.38]$ and $2.846$ $[2.65, 3.073]$. Its analogous interaction, $0.892$ $[0.867, 0.923]$, represents a smaller disadvantage, not recovered superiority.

## J.2 FRESH-IDENTITY CONFIRMATION AND REFERENCE SENSITIVITY

After freezing the recipe and checkpoint rules, 20 identities, 910001–910020, were generated from the original continuous Latin-hypercube ranges. A parameter-tuple audit at tolerance $1 \times 10^{-12}$ found no overlap. These new identities were unused in training, model selection, and preceding diagnostic development. The confirmation includes five branches per seed: $\{p = 1, H\} \times \{TF4, BP4\}$, plus $p = 5$-BP4.

Table 20: All new-population closed-loop means, with physical failures in parentheses out of 20. Final is the fixed-budget endpoint; best is selected on old validation identities. Seed 40 is a discovery training seed; 41 and 42 are additional repeats.

| Seed | Selection | $p = 1$-TF4 | $p = 1$-BP4 | $H$-TF4 | $H$-BP4 | $p = 5$-BP4 |
|---|---|---|---|---|---|---|
| 40 | final | 0.0262 (0) | 0.0259 (0) | 0.0126 (0) | 0.0102 (0) | 0.0756 (20) |
| 40 | best | 0.0257 (0) | 0.0252 (0) | 0.0126 (0) | 0.0102 (0) | 0.0750 (19) |
| 41 | final | 0.0263 (0) | 0.0257 (0) | 0.0132 (1) | 0.0108 (0) | 0.0755 (20) |
| 41 | best | 0.0254 (0) | 0.0251 (0) | 0.0134 (1) | 0.0108 (0) | 0.0749 (20) |
| 42 | final | 0.0263 (0) | 0.0256 (0) | 0.0134 (2) | 0.0109 (1) | 0.0744 (20) |
| 42 | best | 0.0254 (0) | 0.0253 (0) | 0.0134 (2) | 0.0109 (1) | 0.0744 (20) |

Every fresh identity in each new seed favors $H$-BP4 over $p = 1$-BP4 and $H$-TF4, whereas every identity disfavors $p = 5$-BP4 against $p = 1$-BP4. The within-$H$ BP4/TF4 one-step ratios are $1.0766$ $[1.0705, 1.0832]$ and $1.0751$ $[1.0695, 1.0808]$; closed-loop ratios are $0.8182$ $[0.8056, 0.8297]$ and $0.8122$ $[0.8025, 0.8214]$. Thus, both new seeds reproduce worse one-step accuracy with better

Table 21: Both new seeds under checkpoint and reference sensitivity. Entries give ratios with paired 95% identity-bootstrap intervals. $n/R$ specifies the number of identities/reference grid width; both five-identity rows use the same predetermined IDs and identical predictions. Definitions are in Equation 66.

| Seed | Selection | $n/R$ | $R_H$ | $I_H$ | $R_5$ |
|---|---|---|---|---|---|
| 41 | final | 20/512 | 0.421 [0.386, 0.457] | 0.836 [0.823, 0.849] | 2.934 [2.823, 3.056] |
| 41 | final | 5/512 | 0.421 [0.377, 0.488] | 0.831 [0.805, 0.856] | 3.066 [2.839, 3.371] |
| 41 | final | 5/1024 | 0.415 [0.373, 0.480] | 0.831 [0.806, 0.856] | 2.968 [2.746, 3.257] |
| 41 | best | 20/512 | 0.432 [0.397, 0.468] | 0.821 [0.809, 0.831] | 2.989 [2.891, 3.097] |
| 41 | best | 5/512 | 0.432 [0.387, 0.501] | 0.813 [0.798, 0.827] | 3.106 [2.879, 3.370] |
| 41 | best | 5/1024 | 0.426 [0.383, 0.491] | 0.816 [0.800, 0.832] | 3.002 [2.782, 3.250] |
| 42 | final | 20/512 | 0.424 [0.394, 0.456] | 0.832 [0.821, 0.844] | 2.904 [2.786, 3.040] |
| 42 | final | 5/512 | 0.431 [0.372, 0.521] | 0.844 [0.825, 0.862] | 2.978 [2.709, 3.340] |
| 42 | final | 5/1024 | 0.427 [0.369, 0.515] | 0.848 [0.831, 0.863] | 2.882 [2.626, 3.226] |
| 42 | best | 20/512 | 0.430 [0.400, 0.461] | 0.814 [0.801, 0.827] | 2.939 [2.823, 3.073] |
| 42 | best | 5/512 | 0.436 [0.377, 0.525] | 0.829 [0.799, 0.866] | 3.015 [2.742, 3.366] |
| 42 | best | 5/1024 | 0.432 [0.374, 0.519] | 0.834 [0.806, 0.867] | 2.915 [2.657, 3.250] |

rollout. All primary metrics are finite, but $H$-BP4 still has one physical failure in seed 42 and $p = 5$-BP4 fails on all 20 identities in each new seed (Table 20).

All intervals resample paired identities 20,000 times, preserving the pairing among all five conditions. Intervals are conditional on each trained seed, with the same 20 identities shared across seeds.

The first five new identities, 910001–910005, were specified for $1024^2/p = 1$ references before outcomes were observed. Table 21 compares the same five forecasts against $512^2$ and $1024^2$: both their initial histories and their prediction arrays remain unchanged, and only the scoring target differs. The $H$ benefit, enhancement and original degree-five reversal persist.

# K Realized-error decomposition and common-input responses

## K.1 Error cancellation on realized rollouts

The following diagnostics use exactly the eight fixed-final two-stage checkpoints in Table 13.

**An exact identity, with a signed interaction.** Let $g_\theta(h) = \mathrm{last}(h) + f_\theta(h)$ denote the complete next-frame map in physical coordinates, and let $P_6$ select its six physical channels. For reference history $h^\star_{i,t}$ and its deviation $\delta h_{i,t}$ along model $\theta$'s realized rollout, define

$$r_{i,t} = P_6\big[g_\theta(h^\star_{i,t}) - y^\star_{i,t+1}\big], \tag{67}$$

$$d_{i,t} = P_6\big[g_\theta(h^\star_{i,t} + \delta h_{i,t}) - g_\theta(h^\star_{i,t})\big]. \tag{68}$$

Then the actual next-frame error is $e_{i,t+1} = r_{i,t} + d_{i,t}$, irrespective of model nonlinearity. With $Q_i = \sum_t \|P_6 y^\star_{i,t+1}\|_2^2$ over the same 56 forecast steps, define

$$R_i = \frac{\sum_t \|r_{i,t}\|_2^2}{Q_i}, \quad D_i = \frac{\sum_t \|d_{i,t}\|_2^2}{Q_i}, \quad C_i = \frac{2\sum_t \langle r_{i,t}, d_{i,t}\rangle}{Q_i}. \tag{69}$$

Thus

$$E^2_{\mathrm{roll},i} = R_i + D_i + C_i, \qquad \overline{E^2_{\mathrm{roll}}} = \overline{R} + \overline{D} + \overline{C}. \tag{70}$$

The latter averages squared relative errors across identities; it is not the square of the paper's mean relative error. Nor does it add label, one-step, and rollout errors. $D_i$ depends on both the model and its realized history deviation, so this identity alone does not separate response differences from different inputs.

Common-reference one-step predictions were re-inferred because only autonomous arrays had been archived. Their mean errors reproduce the archived one-step scores within $1 \times 10^{-4}$ relative tolerance; checks at fixed rollout times also reproduce archived predictions within that tolerance. All 80

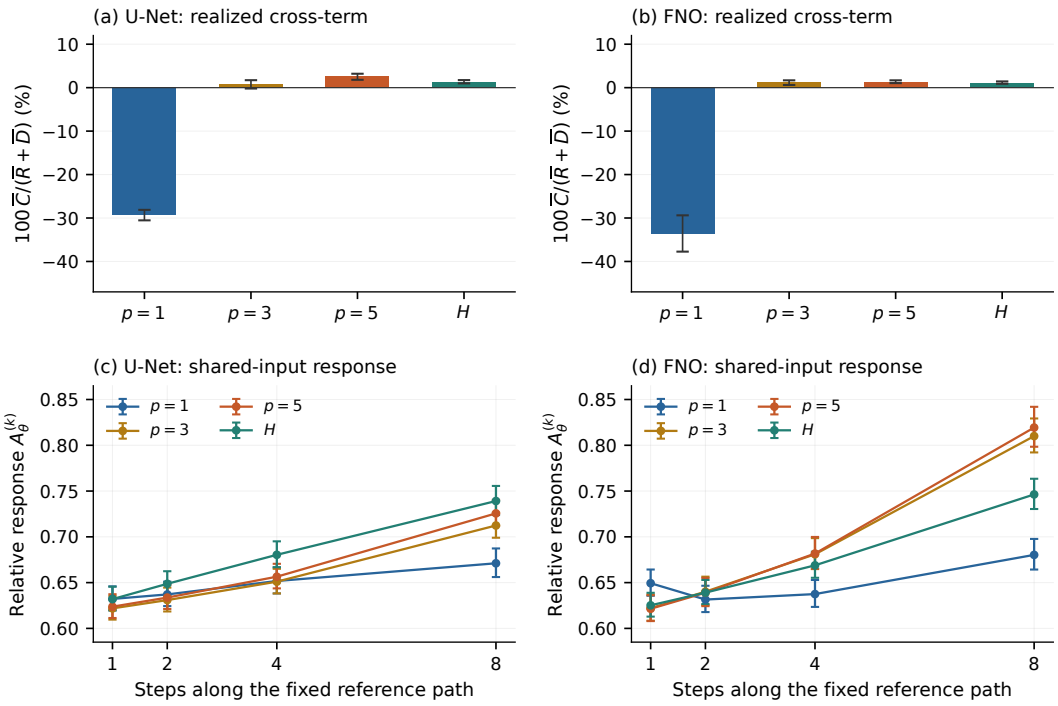



Figure 7: **Realized error cancellation and common-input response are distinct measurements.** Top: the signed ratio of mean cross-term to mean $R + D$ on ten explored identities; negative values denote cancellation, not a causal share of a source advantage. Bottom: directional response $A_\theta^{(k)}$ along common reference paths on 20 original validation identities, using the same eight donor-derived directions at each sampled state. All four sources, including $H$, are retained. Bars are marginal 95% identity-bootstrap intervals conditional on the fixed models. The derivative curves do not depend on finite-amplitude clipping; state-subset sensitivity appears in Table 25. The top and bottom panels use different cohorts.

identity-level equalities in Eq. 70 hold to numerical precision, and finite physically failing rollouts remain included. Complete per-identity values appear in Tables 26–27.

Table 22: Exact decomposition on the ten explored identities. The four energy columns show $10^3$ times their identity means; $\overline{E^2_{\text{roll}}} = \overline{R} + \overline{D} + \overline{C}$. The signed percentage is $100\overline{C}/(\overline{R} + \overline{D})$; it is not a causal share of the degree advantage. The last column counts negative full-horizon cross-terms.

| Model | Source | $10^3\overline{R}$ | $10^3\overline{D}$ | $10^3\overline{C}$ | $10^3\overline{E^2}$ | Signed % | $C_i < 0$ |
|---|---|---|---|---|---|---|---|
| U-Net | $p = 1$ | 0.407 | 0.557 | -0.283 | 0.681 | -29.300 | 10/10 |
| U-Net | $p = 3$ | 0.070 | 1.982 | 0.016 | 2.068 | +0.800 | 2/10 |
| U-Net | $p = 5$ | 0.094 | 5.242 | 0.139 | 5.475 | +2.600 | 0/10 |
| U-Net | $H$ | 0.003 | 0.097 | 0.001 | 0.102 | +1.400 | 0/10 |
| FNO | $p = 1$ | 0.734 | 2.375 | -1.050 | 2.060 | -33.800 | 10/10 |
| FNO | $p = 3$ | 0.231 | 30.306 | 0.352 | 30.889 | +1.200 | 1/10 |
| FNO | $p = 5$ | 0.310 | 43.892 | 0.598 | 44.800 | +1.400 | 0/10 |
| FNO | $H$ | 0.021 | 1.978 | 0.022 | 2.022 | +1.100 | 0/10 |

**Observed cancellation.** For both architectures, $C_i < 0$ on all ten identities at $p = 1$, while $C_i > 0$ on all ten at $p = 5$. The ratio $-\overline{C}/(\overline{R} + \overline{D})$ is 29.3% for degree-one U-Net and 33.8% for degree-one FNO (Figure 7). These percentages quantify cancellation within each model's decomposition. Degree-three U-Net has a positive mean cross-term with mixed identity-level signs.

FNO's near-parity $H$ result also has a precise descriptive accounting. Relative to $p = 1$, the mean changes are $\Delta R = -7.132 \times 10^{-4}$, $\Delta D = -3.969 \times 10^{-4}$, and $\Delta C = +1.072 \times 10^{-3}$, totaling $\Delta\overline{E^2_{\text{roll}}} = -3.824 \times 10^{-5}$. Approximately $96.6\%$ of the reduction in $R + D$ is offset by the cross-term change.

### K.2 Responses to shared deviations along common reference paths

**Identical inputs across receivers.** The 20 validation identities are 2, 7, 13, 18, 22, 24, 37, 40, 56, 76, 90, 100, 111, 123, 126, 150, 183, 191, 204, and 219. At forecast-history start indices $q \in \{8, 16, 32, 48\}$, the reference history is $y^\star_i[q : q + 5]$ using zero-based indexing; its next target is frame $q + 5$. Each of the eight frozen models supplies a donor history from its own autonomous prediction initialized by the same five reference frames. Subtracting the reference history preserves the donor deviation's spatial and temporal structure. Every receiver is tested on every donor's direction, not just its own.

Let $N$ denote the shared seven-channel training normalization. The donor deviation in these coordinates is $\delta z_{i,q,b} = N(h^{\text{donor}}_{i,q,b}) - N(h^\star_{i,q})$, with donor index $b$. Define

$$u_{i,q,b} = \frac{\delta z_{i,q,b}}{\text{RMS}(\delta z_{i,q,b})}, \qquad \widetilde{a}_{i,q} = \text{median}_b \, \text{RMS}(\delta z_{i,q,b}), \tag{71}$$

where RMS includes all $5 \times 7 \times 128 \times 128$ history entries. Each $u$ has unit RMS. The reference arrays, directions, normalizer, and amplitude scalars are identical across receivers and verified by hashes. There are 80 reference states and 640 donor directions, clustered within 20 physical identities.

**Derivative response, not a free-rollout Jacobian chain.** Write the normalized complete history update as

$$\mathcal{G}_\theta(z) = \big(z_{2:5},\, g^z_\theta(z)\big), \tag{72}$$

where $g^z_\theta$ includes the last-frame residual connection. Starting from the same $v_0 = u_{i,q,b}$ for every receiver, compute

$$v_{k+1} = D\mathcal{G}_\theta\big(N(h^\star_{i,q+k})\big)v_k, \qquad k = 0, \ldots, 7. \tag{73}$$

All seven channels participate in the shift and recursive direction update. Crucially, every derivative is evaluated at the same reference history, rather than at a model's generated base state. Responses at $k = 1, 2, 4, 8$ observe the newest frame's six physical channels after undoing channel standard deviations:

$$A^{(k)}_{\theta,i,q,b} = \frac{\|S_6 P_6 \,\text{last}(v_k)\|_2}{\|P_6 y^\star_{i,q+k+4}\|_2 \;\text{RMS}(v_0)}, \tag{74}$$

with $S_6$ the diagonal physical-channel standard deviations and $\text{RMS}(v_0) = 1$. This is a sampled directional response coefficient, not a worst-case operator norm or an autonomous stability criterion. It is also not $D_i$ in Eq. 70, and cannot be substituted there to reconstruct rollout error.

The implementation uses automatic-differentiation Jacobian–vector products, without forming a full Jacobian. For each checkpoint, centered finite differences at normalized steps $0.01, 0.003, 0.001, 3 \times 10^{-4}$ check the one-step JVP on validation identity 2 at $q = 8$, using a normalized difference of successive reference histories. The maximum relative discrepancy across these checks is $0.0361\%$. A last-frame-only map and the explicit residual-wrapper identity are checked separately. These are numerical implementation checks; finite responses to the actual donor directions are examined below.

**Identity-level aggregation and conditional intervals.** For each receiver and horizon, average $A^{(k)}_{\theta,i,q,b}$ over the four times and eight donors within identity, then average those 20 identity means. Source comparisons are ratios of these means, not means of probe-wise ratios. The bootstrap draws 20 identities with replacement 5,000 times, preserving all times, donors, and receiver pairings within each sampled identity. Marginal percentile intervals condition on the fixed models, donor directions, and validation cohort. The ten-identity decomposition uses its own identity-level resampling with the same count and RNG seed.

Degree-five responses are slightly smaller than degree one's at one step, but larger on average at eight steps (Table 23). The raw-six cosine between the $p = 5$ and $p = 1$ response directions decreases

Table 23: Common-reference-path directional responses on 20 original validation identities. $A^{(k)}$ averages over times and donors within identity, then equally over identities. The final column is a ratio of these means to the $p = 1$ model within the same architecture, with a paired 95% identity-bootstrap interval; it is not a ratio of individual probes or an operator norm.

| Model | Source | $\overline{A}^{(1)}$ | $\overline{A}^{(2)}$ | $\overline{A}^{(4)}$ | $\overline{A}^{(8)}$ | $\overline{A}^{(8)}/\overline{A}_1^{(8)}$ [95% CI] |
|---|---|---|---|---|---|---|
| U-Net | $p = 1$ | 0.6322 | 0.6371 | 0.6518 | 0.6711 | 1.0000 (baseline) |
| U-Net | $p = 3$ | 0.6219 | 0.6309 | 0.6510 | 0.7124 | 1.0620 [1.0500, 1.0740] |
| U-Net | $p = 5$ | 0.6237 | 0.6337 | 0.6565 | 0.7256 | 1.0810 [1.0710, 1.0930] |
| U-Net | $H$ | 0.6320 | 0.6487 | 0.6804 | 0.7391 | 1.1010 [1.0970, 1.1060] |
| FNO | $p = 1$ | 0.6494 | 0.6315 | 0.6375 | 0.6803 | 1.0000 (baseline) |
| FNO | $p = 3$ | 0.6221 | 0.6401 | 0.6809 | 0.8100 | 1.1910 [1.1790, 1.2050] |
| FNO | $p = 5$ | 0.6214 | 0.6390 | 0.6817 | 0.8194 | 1.2040 [1.1930, 1.2170] |
| FNO | $H$ | 0.6252 | 0.6391 | 0.6689 | 0.7463 | 1.0970 [1.0920, 1.1020] |

from $0.9665$ to $0.784$ for U-Net and from $0.9368$ to $0.6385$ for FNO between one and eight steps. Thus, the measured difference includes direction as well as magnitude. Across the full probe design, the eight-step $p = 3/p = 1$ and $p = 5/p = 1$ directions agree when aggregated separately by each donor or each identity; that does not require every individual probe to agree.

The paired eight-step $H/p = 5$ response ratio is $1.019$ $[1.01, 1.026]$ for U-Net and $0.911$ $[0.902, 0.919]$ for FNO. U-Net's higher $H$ response point estimate on the validation cohort contrasts with its much lower $H$ rollout error on the ten explored transfer identities ($0.0099$ versus $0.0731$). This is a discrepancy between cohort-level rankings, not a same-identity counterexample. Ratios to $p = 1$ use architecture-specific denominators; similar ratios across architectures do not establish similar absolute responses.

### K.3 Physical neighborhoods and finite-amplitude sensitivity

**A common amplitude rule, independent of the receiver.** For finite-amplitude probes, test whether $N(h^\star_{i,q}) \pm \widetilde{a}_{i,q} u_{i,q,b}$ is physically valid for every donor $b$, using positive density and the pressure convention in Eq. 15 at every history cell. If so, set $a_{i,q} = \widetilde{a}_{i,q}$. Otherwise, 24 bisection iterations find the largest valid symmetric amplitude below $\widetilde{a}_{i,q}$ across all eight directions, and $a_{i,q}$ is set to $90\%$ of that bound. The same $a_{i,q}$ is used for every donor and receiver at that reference state, before any response comparison.

This rule reduces amplitudes at 37/80 states. The reduction-factor minimum, 10th percentile, 25th percentile, median, and maximum are $0.0211, 0.0592, 0.166, 1, 1$; among reduced states, the median is $0.1218$. Both the 43 unreduced states and the 37 reduced states contain all 20 identities. The derivative response in Eq. 73 still uses unit directions at the original reference states: amplitude clipping does not change its JVP. It restricts the finite-amplitude physical neighborhood and identifies state subsets for sensitivity analysis.

**Finite response and local approximation.** Use initial deviations $\Delta_0 = s\, a_{i,q} u_{i,q,b}$ for $s \in \{0.01, 0.1, 1\}$. At one step, the output change is exactly the model's response to this common input minus its own reference-history output. To test longer local approximations, propagate the deviation with a common reference base at every step:

$$\Delta_{k+1} = \mathcal{G}_\theta(z^\star_{i,q+k} + \Delta_k) - \mathcal{G}_\theta(z^\star_{i,q+k}), \qquad z^\star_{i,q+k} = N(h^\star_{i,q+k}). \tag{75}$$

This reference-anchored nonlinear response is not a difference between two autonomous rollouts. Compare its endpoint with $s a_{i,q} v_k$ using

$$\eta^{(k)}_{\theta,i,q,b,s} = \frac{\|S_6 P_6 \operatorname{last}(\Delta_k - s a_{i,q} v_k)\|_2}{\|S_6 P_6 \operatorname{last}(s a_{i,q} v_k)\|_2}. \tag{76}$$

All initially constructed histories are valid. Subsequent invalid inputs and outputs are flagged and retained, with no non-finite diagnostic rows. Table 24 reports empirical probe quantiles and failure counts. At $1\%$ of the available amplitude, median eight-step discrepancies are approximately $0.13\%$–$0.22\%$; at the largest available amplitude, they rise to $5.5\%$–$10.9\%$, with 90th percentiles up to $37.4\%$.

Table 24: Finite-amplitude response checks on all 80 states and eight donor directions (640 correlated probes per model and amplitude). $s$ multiplies the common physically admissible amplitude $a_{i,q}$. Entries for $\eta$ are empirical median / 90th percentile percentages, not confidence intervals. Failure counts retain probes with any invalid input/output through step eight; these are diagnostic responses, not 56-step rollout failures.

| Model | Source | $s$ | $\eta^{(1)}$ (%) | $\eta^{(8)}$ (%) | Invalid input | Invalid output |
|---|---|---|---|---|---|---|
| U-Net | $p = 1$ | 0.010 | 0.050 / 0.150 | 0.160 / 0.490 | 0/640 | 0/640 |
| U-Net | $p = 1$ | 0.100 | 0.120 / 0.550 | 0.560 / 2.010 | 0/640 | 0/640 |
| U-Net | $p = 1$ | 1.000 | 1.190 / 5.280 | 5.580 / 18.500 | 5/640 | 3/640 |
| U-Net | $p = 3$ | 0.010 | 0.080 / 0.200 | 0.200 / 0.530 | 0/640 | 0/640 |
| U-Net | $p = 3$ | 0.100 | 0.150 / 0.740 | 0.770 / 2.040 | 0/640 | 0/640 |
| U-Net | $p = 3$ | 1.000 | 1.510 / 7.850 | 7.710 / 19.550 | 13/640 | 6/640 |
| U-Net | $p = 5$ | 0.010 | 0.080 / 0.230 | 0.220 / 0.710 | 0/640 | 0/640 |
| U-Net | $p = 5$ | 0.100 | 0.190 / 0.860 | 0.980 / 2.660 | 0/640 | 0/640 |
| U-Net | $p = 5$ | 1.000 | 1.880 / 8.900 | 9.700 / 23.990 | 11/640 | 7/640 |
| U-Net | $H$ | 0.010 | 0.070 / 0.240 | 0.200 / 0.530 | 0/640 | 0/640 |
| U-Net | $H$ | 0.100 | 0.250 / 1.650 | 1.080 / 4.370 | 0/640 | 0/640 |
| U-Net | $H$ | 1.000 | 2.500 / 15.270 | 10.930 / 37.380 | 3/640 | 4/640 |
| FNO | $p = 1$ | 0.010 | 0.050 / 0.120 | 0.130 / 0.380 | 0/640 | 0/640 |
| FNO | $p = 1$ | 0.100 | 0.140 / 0.720 | 0.550 / 2.450 | 0/640 | 0/640 |
| FNO | $p = 1$ | 1.000 | 1.350 / 7.250 | 5.460 / 23.420 | 3/640 | 3/640 |
| FNO | $p = 3$ | 0.010 | 0.050 / 0.180 | 0.150 / 0.510 | 0/640 | 0/640 |
| FNO | $p = 3$ | 0.100 | 0.220 / 0.820 | 0.990 / 4.000 | 0/640 | 0/640 |
| FNO | $p = 3$ | 1.000 | 2.260 / 8.330 | 9.470 / 32.650 | 12/640 | 3/640 |
| FNO | $p = 5$ | 0.010 | 0.060 / 0.170 | 0.160 / 0.510 | 0/640 | 0/640 |
| FNO | $p = 5$ | 0.100 | 0.260 / 0.970 | 1.110 / 4.200 | 0/640 | 0/640 |
| FNO | $p = 5$ | 1.000 | 2.620 / 9.620 | 10.930 / 33.480 | 25/640 | 3/640 |
| FNO | $H$ | 0.010 | 0.060 / 0.190 | 0.170 / 0.340 | 0/640 | 0/640 |
| FNO | $H$ | 0.100 | 0.290 / 1.080 | 1.060 / 2.620 | 0/640 | 0/640 |
| FNO | $H$ | 1.000 | 2.900 / 10.740 | 10.020 / 24.700 | 12/640 | 11/640 |

Table 25: State-subset sensitivity of the eight-step derivative response. All receivers retain the same 43 unmodified-amplitude or 37 reduced-amplitude states. Each subset contains all 20 identities, weighted equally after within-identity averaging. The amplitude rule does not change JVPs; this comparison changes which reference states enter the average. The last column gives median / 90th percentile finite-response linearization error at the requested actual-error scale on the 43 unreduced states only.

| Model | Source | Unreduced: $A^{(8)}/A_1^{(8)}$ | Reduced: $A^{(8)}/A_1^{(8)}$ | Unreduced $\eta^{(8)}$ (%) |
|---|---|---|---|---|
| U-Net | $p = 1$ | 1.000 [1.000, 1.000] | 1.000 [1.000, 1.000] | 3.360 / 8.440 |
| U-Net | $p = 3$ | 0.989 [0.983, 0.995] | 1.207 [1.184, 1.233] | 4.060 / 14.020 |
| U-Net | $p = 5$ | 0.991 [0.982, 0.999] | 1.263 [1.240, 1.289] | 4.550 / 15.440 |
| U-Net | $H$ | 1.051 [1.048, 1.054] | 1.211 [1.198, 1.224] | 4.560 / 17.820 |
| FNO | $p = 1$ | 1.000 [1.000, 1.000] | 1.000 [1.000, 1.000] | 3.100 / 9.280 |
| FNO | $p = 3$ | 1.052 [1.031, 1.077] | 1.439 [1.404, 1.470] | 4.380 / 20.340 |
| FNO | $p = 5$ | 1.040 [1.016, 1.067] | 1.501 [1.467, 1.535] | 4.430 / 20.150 |
| FNO | $H$ | 1.056 [1.051, 1.061] | 1.180 [1.164, 1.197] | 5.100 / 17.960 |

**State dependence remains visible.** The largest available amplitude reaches the requested median donor-error scale only on the 43 unreduced states. We report those states separately in Table 25, averaging their times and directions within each identity before equally averaging identities. On this subset, the eight-step $p = 5/p = 1$ derivative-response ratio is $0.991$ $[0.982, 0.999]$ for U-Net and $1.04$ $[1.016, 1.067]$ for FNO. On the reduced-state subset it is $1.263$ $[1.24, 1.289]$ and $1.501$ $[1.467, 1.535]$. Thus, U-Net's full-design excess does not persist on unreduced states.

These diagnostics establish realized error cancellation and response differences at controlled inputs, with explicit state and amplitude dependence.

Table 26: Complete identity-level U-Net decomposition for the explored transfer cohort. All four energy columns are multiplied by $10^3$; rounding can leave a small displayed sum discrepancy.

| Source | Identity | $10^3 R_i$ | $10^3 D_i$ | $10^3 C_i$ | $10^3 E^2_{\mathrm{roll},i}$ |
|---|---|---|---|---|---|
| $p = 1$ | 900001 | 0.1928 | 0.2633 | -0.1391 | 0.3169 |
| $p = 1$ | 900002 | 0.3235 | 0.4282 | -0.2226 | 0.5291 |
| $p = 1$ | 900003 | 0.5050 | 0.6984 | -0.3413 | 0.8621 |
| $p = 1$ | 900004 | 0.2597 | 0.4587 | -0.1796 | 0.5388 |
| $p = 1$ | 900005 | 0.3077 | 0.4518 | -0.2028 | 0.5567 |
| $p = 1$ | 900006 | 0.6059 | 0.7868 | -0.4214 | 0.9712 |
| $p = 1$ | 900007 | 0.3126 | 0.3359 | -0.2180 | 0.4305 |
| $p = 1$ | 900008 | 0.4798 | 0.5668 | -0.3318 | 0.7148 |
| $p = 1$ | 900009 | 0.6235 | 0.9680 | -0.4529 | 1.1386 |
| $p = 1$ | 900010 | 0.4622 | 0.6075 | -0.3173 | 0.7524 |
| $p = 3$ | 900001 | 0.0275 | 1.1823 | 0.0362 | 1.2459 |
| $p = 3$ | 900002 | 0.0462 | 1.4725 | 0.0394 | 1.5581 |
| $p = 3$ | 900003 | 0.0895 | 1.7100 | -0.0351 | 1.7644 |
| $p = 3$ | 900004 | 0.0426 | 2.9465 | 0.0487 | 3.0377 |
| $p = 3$ | 900005 | 0.0548 | 1.7513 | 0.0249 | 1.8309 |
| $p = 3$ | 900006 | 0.1228 | 2.5506 | -0.0445 | 2.6290 |
| $p = 3$ | 900007 | 0.0452 | 1.3029 | 0.0346 | 1.3827 |
| $p = 3$ | 900008 | 0.0817 | 2.2199 | 0.0215 | 2.3231 |
| $p = 3$ | 900009 | 0.1250 | 3.0090 | 0.0098 | 3.1438 |
| $p = 3$ | 900010 | 0.0680 | 1.6737 | 0.0204 | 1.7621 |
| $p = 5$ | 900001 | 0.0351 | 3.3187 | 0.0863 | 3.4402 |
| $p = 5$ | 900002 | 0.0576 | 5.0125 | 0.1546 | 5.2248 |
| $p = 5$ | 900003 | 0.1226 | 5.2941 | 0.1910 | 5.6077 |
| $p = 5$ | 900004 | 0.0583 | 6.5899 | 0.2421 | 6.8903 |
| $p = 5$ | 900005 | 0.0669 | 4.4713 | 0.1013 | 4.6396 |
| $p = 5$ | 900006 | 0.1691 | 6.4136 | 0.0118 | 6.5945 |
| $p = 5$ | 900007 | 0.0579 | 3.0190 | 0.0867 | 3.1637 |
| $p = 5$ | 900008 | 0.1093 | 5.4684 | 0.1056 | 5.6833 |
| $p = 5$ | 900009 | 0.1764 | 8.9344 | 0.3011 | 9.4119 |
| $p = 5$ | 900010 | 0.0859 | 3.8989 | 0.1090 | 4.0937 |
| $H$ | 900001 | 0.0010 | 0.0710 | 0.0016 | 0.0735 |
| $H$ | 900002 | 0.0018 | 0.0801 | 0.0016 | 0.0835 |
| $H$ | 900003 | 0.0039 | 0.0762 | 0.0009 | 0.0810 |
| $H$ | 900004 | 0.0020 | 0.1599 | 0.0027 | 0.1646 |
| $H$ | 900005 | 0.0027 | 0.1140 | 0.0026 | 0.1193 |
| $H$ | 900006 | 0.0053 | 0.1062 | 0.0007 | 0.1122 |
| $H$ | 900007 | 0.0018 | 0.0499 | 0.0009 | 0.0526 |
| $H$ | 900008 | 0.0042 | 0.0803 | 0.0004 | 0.0849 |
| $H$ | 900009 | 0.0057 | 0.1578 | 0.0014 | 0.1649 |
| $H$ | 900010 | 0.0024 | 0.0775 | 0.0007 | 0.0807 |

Table 27: Complete identity-level FNO decomposition for the explored transfer cohort. All four energy columns are multiplied by $10^3$; rounding can leave a small displayed sum discrepancy.

| Source | Identity | $10^3 R_i$ | $10^3 D_i$ | $10^3 C_i$ | $10^3 E^2_{\mathrm{roll},i}$ |
|---|---|---|---|---|---|
| $p = 1$ | 900001 | 0.3178 | 1.4531 | -0.5099 | 1.2610 |
| $p = 1$ | 900002 | 0.5964 | 1.9348 | -0.8771 | 1.6541 |
| $p = 1$ | 900003 | 0.9343 | 2.4244 | -1.2915 | 2.0673 |
| $p = 1$ | 900004 | 0.4035 | 2.8608 | -0.6595 | 2.6048 |
| $p = 1$ | 900005 | 0.5275 | 2.1634 | -0.7605 | 1.9304 |
| $p = 1$ | 900006 | 1.1432 | 2.6419 | -1.6019 | 2.1832 |
| $p = 1$ | 900007 | 0.5898 | 1.8343 | -0.8213 | 1.6029 |
| $p = 1$ | 900008 | 0.9012 | 2.3864 | -1.3223 | 1.9653 |
| $p = 1$ | 900009 | 1.1038 | 3.8078 | -1.4810 | 3.4306 |
| $p = 1$ | 900010 | 0.8249 | 2.2444 | -1.1704 | 1.8989 |
| $p = 3$ | 900001 | 0.0980 | 18.1331 | 0.0252 | 18.2562 |
| $p = 3$ | 900002 | 0.1573 | 15.9971 | 0.3783 | 16.5326 |
| $p = 3$ | 900003 | 0.2973 | 32.9990 | 0.6792 | 33.9755 |
| $p = 3$ | 900004 | 0.1389 | 27.8793 | 0.3534 | 28.3716 |
| $p = 3$ | 900005 | 0.1793 | 27.6955 | 0.2414 | 28.1161 |
| $p = 3$ | 900006 | 0.4008 | 33.3689 | 0.8477 | 34.6174 |
| $p = 3$ | 900007 | 0.1532 | 36.9597 | -0.1357 | 36.9772 |
| $p = 3$ | 900008 | 0.2473 | 28.9932 | 0.2839 | 29.5245 |
| $p = 3$ | 900009 | 0.4123 | 45.1543 | 0.5245 | 46.0911 |
| $p = 3$ | 900010 | 0.2253 | 35.8760 | 0.3223 | 36.4236 |
| $p = 5$ | 900001 | 0.1270 | 34.0815 | 0.1813 | 34.3899 |
| $p = 5$ | 900002 | 0.2007 | 34.4120 | 0.6315 | 35.2441 |
| $p = 5$ | 900003 | 0.4130 | 46.8472 | 0.8428 | 48.1030 |
| $p = 5$ | 900004 | 0.2192 | 50.0841 | 0.6541 | 50.9574 |
| $p = 5$ | 900005 | 0.2421 | 48.0873 | 0.6115 | 48.9409 |
| $p = 5$ | 900006 | 0.5248 | 40.7574 | 1.0145 | 42.2966 |
| $p = 5$ | 900007 | 0.1827 | 33.4938 | 0.2627 | 33.9392 |
| $p = 5$ | 900008 | 0.3215 | 54.8827 | 0.4058 | 55.6100 |
| $p = 5$ | 900009 | 0.5743 | 57.7225 | 0.8687 | 59.1654 |
| $p = 5$ | 900010 | 0.2957 | 38.5476 | 0.5056 | 39.3489 |
| $H$ | 900001 | 0.0078 | 1.3107 | 0.0204 | 1.3389 |
| $H$ | 900002 | 0.0131 | 1.4603 | 0.0235 | 1.4969 |
| $H$ | 900003 | 0.0236 | 2.1242 | 0.0165 | 2.1642 |
| $H$ | 900004 | 0.0136 | 2.9644 | 0.0317 | 3.0097 |
| $H$ | 900005 | 0.0193 | 1.9145 | 0.0411 | 1.9748 |
| $H$ | 900006 | 0.0359 | 2.1338 | 0.0126 | 2.1823 |
| $H$ | 900007 | 0.0136 | 1.4547 | 0.0157 | 1.4839 |
| $H$ | 900008 | 0.0288 | 2.2290 | 0.0215 | 2.2792 |
| $H$ | 900009 | 0.0363 | 2.7080 | 0.0223 | 2.7666 |
| $H$ | 900010 | 0.0186 | 1.4824 | 0.0186 | 1.5195 |

# L JOINT RECONSTRUCTION OF AUTONOMOUS PREDICTION ERRORS

## L.1 FIXED MODELS, PAIRED WINDOWS, AND SEPARATE HORIZONS

The reference-forced quantity in Eq. 6 combines the model's newly injected defects with its response to preceding deviations. We measure how accurately this joint quantity, and one correction around its resulting path, reconstruct the actual neural trajectory. This differs from measuring response to an initial donor direction alone (Appendix K.2). The analysis uses existing numerical data and fixed models, with no additional training or numerical simulation. Its identities, windows, amplitudes, contrasts, and endpoints were frozen before the new reconstruction outcomes.

The primary panel contains seven final checkpoints: U-Net trained on $p = 1, p = 5, H$ with BP4 continuation, U-Net trained on $H$ with TF4 continuation, and FNO trained on $p = 1, p = 5, H$ with BP4 continuation. These are the fixed-final models of Appendices H and J, trained with the shared protocol in Appendix B.3.

The primary cohort is the same ten explored identities, 900001–900010. At zero-based start $s \in \{0, 8, 16, 32, 48\}$, every path receives reference frames $s : s + 5$ and predicts the next eight frames; the four-step score is its prefix. Thus the primary comparison comprises 350 eight-step windows, clustered within ten identities. A separate continuous 56-step calculation starts only at $s = 0$. Short reference-started windows are never concatenated to represent that autonomous endpoint.

Table 28: Frozen reconstruction design. Counts are model–identity–start–amplitude combinations, not independent statistical samples. All paths within a cell use the listed precision. The interface check is excluded from primary summaries.

| Analysis | Models | IDs | Starts | Amplitudes | Steps / precision | Cells |
|---|---|---|---|---|---|---|
| Interface check | 7 | 2 | 2 | 1 | 8 / FP32 | 28 |
| Primary | 7 | 10 | 5 | 1 | 8 / FP32 | 350 |
| Continuous endpoint | 7 | 10 | 1 | 1 | 56 / FP32 | 70 |
| Defect amplitude | 7 | 2 | 2 | 6 | 8 / FP32 | 168 |
| Precision subset | 7 | 1 | 2 | 6 | 8 / FP64 | 84 |
| $H$ replication | 4 | 20 | 5 | 1 | 8 / FP32 | 400 |
| $H$ continuous replication | 4 | 20 | 1 | 1 | 56 / FP32 | 80 |

The amplitude subset uses identities 900001/900002 and starts 0/32; FP64 repeats exactly the identity-900001 subset. The replication panel uses the existing U-Net $H$-TF4/$H$-BP4 models on identities 910001–910020. Those identities were fresh for the earlier performance confirmation, but are reused for this subsequently specified diagnostic.

## L.2 THREE TRAJECTORIES AT THE SAME DEFECT AMPLITUDE

Use the normalized complete five-frame, seven-channel histories of Appendix D. The next-frame map includes the residual connection, and every Jacobian–vector product includes all recurrent channels. For each window, let $x_t^\star$ be its reference history, $b_t = G_\theta(x_t^\star) - x_{t+1}^\star$, and $A_t = DG_\theta(x_t^\star)$. Compute

$$\xi_0 = 0, \qquad \xi_{t+1} = A_t\xi_t + b_t, \qquad q_t^\epsilon = x_t^\star + \epsilon\xi_t, \tag{77}$$

$$G_t^\epsilon(x) = G_\theta(x) - (1-\epsilon)b_t, \qquad x_0^\epsilon = x_0^\star, \qquad x_{t+1}^\epsilon = G_t^\epsilon(x_t^\epsilon), \tag{78}$$

$$c_t^\epsilon = G_t^\epsilon(q_t^\epsilon) - q_{t+1}^\epsilon, \qquad d_0^\epsilon = 0, \qquad d_{t+1}^\epsilon = DG_\theta(q_t^\epsilon)d_t^\epsilon + c_t^\epsilon. \tag{79}$$

The corrected path is $p_t^\epsilon = q_t^\epsilon + d_t^\epsilon$. At $\epsilon = 1$, $x^\epsilon$ is the original autonomous neural trajectory. At other amplitudes, the modified map scales the continuously injected reference defects; it does not merely perturb the initial history. The correction differentiates at $q^\epsilon$ and is applied once. It is never repeatedly recentered at $p^\epsilon$, and the corrected output is not fed back as a new autonomous predictor. Both proxies are constructed from the fixed model and reference path alone; the realized trajectory is used afterward to assess reconstruction and measure $\delta_j$. All three paths share their initial history, frozen weights, normalization, and arithmetic precision.

Table 29: **Where joint reconstruction separates the recorded means.** Entries count passing contrasts / total contrasts. The mean-gap condition uses the measured vector discrepancy, not the theorem's independent bound.

| Comparison group | Steps | Original $q$ | Corrected $p$ |
|---|---|---|---|
| Both architectures | 4 | 7/7 | 7/7 |
| Both architectures | 8 | 7/7 | 7/7 |
| U-Net | 56 | 4/4 | 4/4 |
| FNO | 56 | 0/3 | 0/3 |
| U-Net $H$, seeds 41/42 | 56 | 0/2 | 2/2 |

In exact arithmetic, the original proxy's first step and the corrected proxy's first two steps equal the corresponding actual trajectory. We retain per-step results, but assess reconstruction at four/eight steps and separately at 56 steps.

### L.3 Vector discrepancies, signed terms, and paired contrasts

For a window of length $T$, write $y = \operatorname{stack}_{t=1}^{T} O(x_t^\epsilon - x_t^\star)$, $y^{(0)} = \operatorname{stack} O(q_t^\epsilon - x_t^\star)$, and $y^{(1)} = \operatorname{stack} O(p_t^\epsilon - x_t^\star)$. With the window's original raw six-channel reference energy $Q$, report

$$E = \frac{\|y\|}{\sqrt{Q}}, \qquad \widehat{E}_j = \frac{\|y^{(j)}\|}{\sqrt{Q}}, \qquad \delta_j = \frac{\|y - y^{(j)}\|}{\sqrt{Q}}, \quad j \in \{0, 1\}. \tag{80}$$

We also report the relative vector discrepancy $\delta_j / E$ and cosine with $y$. Corresponding measurements retain all seven newest-frame channels in raw and standardized coordinates, and all five frames of the seven-channel history after each update. Overlap between successive histories belongs to that explicitly defined norm; it does not create additional independent samples.

Subtraction is performed after casting each stored normalized path to float64, then channel standard deviations restore physical units. $Q$ is computed directly from the original raw reference, not from its normalization roundtrip. We separately rescore reconstructed physical outputs against the original raw targets, retaining the small roundtrip displacement in both total error and local defect. That representation difference is not a compulsory lower floor for $\delta_j$, because the common reference cancels between actual and approximate error vectors.

Let $r = \operatorname{stack}_{t=1}^{T} O(\epsilon b_{t-1})$, taken from the directly retained injection array. For each $\widetilde{y} \in \{y, y^{(0)}, y^{(1)}\}$, compute

$$R = \frac{\|r\|^2}{Q}, \qquad D = \frac{\|\widetilde{y} - r\|^2}{Q}, \qquad C = \frac{2\langle r, \widetilde{y} - r\rangle}{Q}, \qquad \frac{\|\widetilde{y}\|^2}{Q} = R + D + C. \tag{81}$$

$R$ is common to the three paths; $D$ and the signed $C$ need not be. These equalities concern each window's squared error, and averaging them does not give the square of the mean $E$.

For models $A, B$, vector discrepancies give the empirical comparison

$$\left|(E_A - E_B) - (\widehat{E}_{j,A} - \widehat{E}_{j,B})\right| \le \delta_{j,A} + \delta_{j,B}. \tag{82}$$

We report the actual gap, proxy gap, discrepancy budget, and margin $M_j = |\widehat{E}_{j,A} - \widehat{E}_{j,B}| - (\delta_{j,A} + \delta_{j,B})$, checking both margin and agreement of signs. Unlike $\Delta$ in Theorem 2, these $\delta_j$ use the actual neural rollout. A separate arithmetic sensitivity check requires the margin to exceed 64 times float64 machine precision times the largest participating score or budget.

Every primary contrast is retained: $p = 5 - p = 1$, $H - p = 5$, and $H - p = 1$ within each architecture, plus U-Net $H$-BP4 minus $H$-TF4. We average the five prespecified starts within each identity before averaging identities. The mean-gap condition uses the corresponding mean $\delta$ budget. Bootstrap intervals use 5,000 paired identity resamples, preserving all starts and models within an identity. They are marginal, unadjusted, conditional on the fixed checkpoints, and exploratory on reused cohorts. Window-level relative discrepancies are averaged in this same order; they are not ratios of aggregate means.

### L.4 Reconstruction at the actual amplitude

All 350 primary windows are available. Table 30 reports the four/eight-step results, and Figure 8 compares observed and full-history vector discrepancies. At eight steps, the original proxy's mean relative vector discrepancy ranges from $1.14\%$ to $10.27\%$ across models; after one correction, it ranges from $0.017\%$ to $1.05\%$. In the complete standardized five-frame, seven-channel history, the corresponding ranges are $0.79\%$–$11.36\%$ and $0.013\%$–$1.53\%$. Thus, agreement is not confined to the six scored channels. The worst eight-step complete-history discrepancy for degree-five U-Net is $67.4\%$ before correction and $21.4\%$ after it; for degree-five FNO it is $36.4\%$ and $7.32\%$.

Table 30: Actual-amplitude reconstruction on ten explored identities. $E_z$ is the actual six-channel error; $\delta_q, \delta_p$ are vector discrepancies from it. $h_q, h_p$ average relative discrepancies in the full standardized five-frame, seven-channel history. All means first average the five windows within identity. In particular $h$ is a mean of window-level ratios. $F/N$ count physical/nonfinite failures for actual/original/corrected paths, each out of 50 correlated windows.

| $T$ | Model | $E_z$ | $\delta_q$ | $\delta_p$ | $h_q$ | $h_p$ | $F_{z/q/p}$ | $N_{z/q/p}$ |
|---|---|---|---|---|---|---|---|---|
| 4 | U-Net $p=1$ | 0.0126 | $2.9600\times10^{-4}$ | $2.1600\times10^{-7}$ | 0.0090 | $2.5400\times10^{-5}$ | 0/0/0 | 0/0/0 |
| 4 | U-Net $p=5$ | 0.0134 | $3.2600\times10^{-4}$ | $3.5700\times10^{-6}$ | 0.0323 | $6.4100\times10^{-4}$ | 2/2/2 | 0/0/0 |
| 4 | U-Net $H$ BP4 | 0.0021 | $1.3400\times10^{-5}$ | $5.6100\times10^{-8}$ | 0.0023 | $3.6300\times10^{-5}$ | 0/0/0 | 0/0/0 |
| 4 | U-Net $H$ TF4 | 0.0022 | $1.5100\times10^{-5}$ | $5.6900\times10^{-8}$ | 0.0026 | $3.2700\times10^{-5}$ | 0/0/0 | 0/0/0 |
| 4 | FNO $p=1$ | 0.0164 | 0.0021 | $1.3000\times10^{-5}$ | 0.0450 | $2.2400\times10^{-4}$ | 0/0/0 | 0/0/0 |
| 4 | FNO $p=5$ | 0.0272 | 0.0020 | $1.4700\times10^{-4}$ | 0.0364 | 0.0012 | 1/4/1 | 0/0/0 |
| 4 | FNO $H$ | 0.0067 | $1.8300\times10^{-4}$ | $9.2200\times10^{-6}$ | 0.0130 | $4.0500\times10^{-4}$ | 0/0/0 | 0/0/0 |
| 8 | U-Net $p=1$ | 0.0147 | $7.7900\times10^{-4}$ | $5.0200\times10^{-6}$ | 0.0239 | $1.3200\times10^{-4}$ | 0/0/0 | 0/0/0 |
| 8 | U-Net $p=5$ | 0.0237 | 0.0024 | $2.1900\times10^{-4}$ | 0.1140 | 0.0153 | 7/8/7 | 0/0/0 |
| 8 | U-Net $H$ BP4 | 0.0033 | $7.7800\times10^{-5}$ | $1.6700\times10^{-6}$ | 0.0079 | $1.2500\times10^{-4}$ | 0/0/0 | 0/0/0 |
| 8 | U-Net $H$ TF4 | 0.0037 | $9.3600\times10^{-5}$ | $3.0000\times10^{-6}$ | 0.0096 | $1.9400\times10^{-4}$ | 0/0/0 | 0/0/0 |
| 8 | FNO $p=1$ | 0.0192 | 0.0033 | $1.4600\times10^{-4}$ | 0.0800 | 0.0024 | 0/0/0 | 0/0/0 |
| 8 | FNO $p=5$ | 0.0453 | 0.0073 | $9.4600\times10^{-4}$ | 0.1050 | 0.0101 | 8/12/8 | 0/0/0 |
| 8 | FNO $H$ | 0.0113 | $5.3700\times10^{-4}$ | $4.9800\times10^{-5}$ | 0.0247 | 0.0015 | 0/0/0 | 0/0/0 |

For the eight-step $p=5-p=1$ comparisons, original-proxy mean gaps are $0.0097$ for U-Net and $0.0288$ for FNO, against measured budgets $0.0031$ and $0.0106$. For U-Net $H$-BP4 minus $H$-TF4, the gap is $-4.383\times10^{-4}$ against budget $1.714\times10^{-4}$. All seven prespecified contrasts pass the mean-gap condition and numerical guard, with the correct sign, at four and eight steps (Table 31). This does not imply that every paired window passes. For example, the original eight-step U-Net $p=5-p=1$ comparison passes in 44/50 windows, and the within-$H$ comparison in 47/50.

The signed terms are reconstructed as well as the scalar scores (Table 32). At each of the four/eight-step endpoints, both proxies reproduce the actual $C$ sign in all 50 windows for each model. Degree-one cancellation is not universal at the window level: at eight steps, actual $C<0$ in 28/50 U-Net and 27/50 FNO windows, whereas its mean over the five starts is negative on every identity. Degree-five $C$ is positive in every sampled window. These local results do not impose the earlier 56-step sign pattern on every reset window.

All constructed initial reference histories are physically valid. An independent audit checks all 61 frames of the 30 reference identities, retaining all seven channels in the pressure calculation: none has a nonfinite or physically invalid cell. This property is not imposed on predictions. There is no physical clipping, rescaling, rejection, or truncation at first failure. At eight steps, degree-five U-Net has 7/8/7 physically failing actual/original/corrected windows out of 50; FNO has 8/12/8. All other primary eight-step branches have no physical failures, and none of the 350 primary windows is numerically nonfinite. Counts concern correlated windows, not independent identities.

### L.5 Continuous 56-step reconstruction

The continuous endpoint initializes each model once from the first five reference frames, then computes $x, q, p$ without resets for all 56 steps. Its reference denominator, accumulated signed terms,

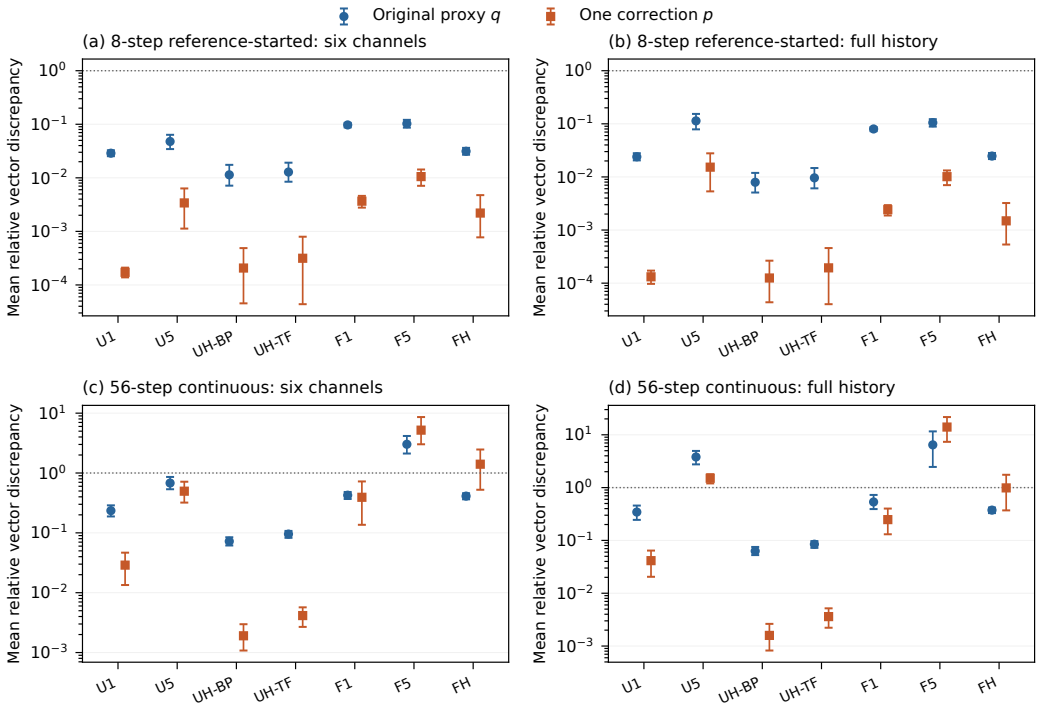



Figure 8: **Joint reconstruction has a horizon-dependent scope.** Original and corrected vector discrepancies relative to actual neural errors, in the six physical output channels and the full standardized five-frame, seven-channel history. Eight-step reference-started windows and the separately computed 56-step autonomous paths are shown separately. Each statistic averages window-level ratios within identity, then identities; it is not a ratio of mean errors. Error bars are marginal, unadjusted 95% paired identity-bootstrap intervals, conditional on the fixed models and reused cohort. The dashed line denotes discrepancy equal to the actual error. U/F denote U-Net/FNO; H-BP/H-TF are the two $H$ continuation objectives.

Table 31: All frozen short-window contrasts at actual amplitude. The four numerical columns and margin intervals are multiplied by $10^3$. Budget is $\overline{\delta}_A + \overline{\delta}_B$; margin is the absolute mean proxy gap minus that budget. A yes entry requires a positive mean margin beyond the arithmetic guard and agreement with the actual sign; it does not require the interval to exclude zero. Intervals are unadjusted paired identity bootstrap intervals, conditional on fixed weights and exploratory on reused identities.

| $T$ | $A-B$ | Path | Actual | Proxy | Budget | Margin [95% CI] | Mean pass |
|---|---|---|---|---|---|---|---|
| 4 | U: $5-1$ | $q$ | 0.870 | 0.992 | 0.622 | 0.370 [-0.177, 0.879] | Yes |
| 4 | U: $5-1$ | $p$ | 0.870 | 0.870 | 0.004 | 0.866 [0.249, 1.463] | Yes |
| 4 | F: $5-1$ | $q$ | 10.750 | 11.429 | 4.062 | 7.367 [6.193, 8.579] | Yes |
| 4 | F: $5-1$ | $p$ | 10.750 | 10.760 | 0.160 | 10.600 [8.991, 12.301] | Yes |
| 4 | U: $H-5$ | $q$ | -11.365 | -11.409 | 0.339 | 11.070 [9.675, 12.613] | Yes |
| 4 | U: $H-5$ | $p$ | -11.365 | -11.365 | 0.004 | 11.362 [9.852, 13.039] | Yes |
| 4 | F: $H-5$ | $q$ | -20.468 | -20.859 | 2.164 | 18.695 [16.610, 20.952] | Yes |
| 4 | F: $H-5$ | $p$ | -20.468 | -20.478 | 0.156 | 20.322 [17.849, 23.001] | Yes |
| 4 | U: $H-1$ | $q$ | -10.496 | -10.417 | 0.310 | 10.107 [8.945, 11.349] | Yes |
| 4 | U: $H-1$ | $p$ | -10.496 | -10.496 | $2.730 \times 10^{-4}$ | 10.496 [9.262, 11.811] | Yes |
| 4 | F: $H-1$ | $q$ | -9.718 | -9.430 | 2.264 | 7.165 [6.384, 7.943] | Yes |
| 4 | F: $H-1$ | $p$ | -9.718 | -9.718 | 0.022 | 9.696 [8.569, 10.843] | Yes |
| 4 | U: $H_{\mathrm{BP}}-H_{\mathrm{TF}}$ | $q$ | -0.160 | -0.160 | 0.029 | 0.132 [0.121, 0.144] | Yes |
| 4 | U: $H_{\mathrm{BP}}-H_{\mathrm{TF}}$ | $p$ | -0.160 | -0.160 | $1.130 \times 10^{-4}$ | 0.160 [0.147, 0.175] | Yes |
| 8 | U: $5-1$ | $q$ | 9.066 | 9.703 | 3.138 | 6.565 [5.582, 7.496] | Yes |
| 8 | U: $5-1$ | $p$ | 9.066 | 9.084 | 0.224 | 8.860 [7.332, 10.475] | Yes |
| 8 | F: $5-1$ | $q$ | 26.089 | 28.772 | 10.595 | 18.177 [16.407, 20.020] | Yes |
| 8 | F: $5-1$ | $p$ | 26.089 | 26.173 | 1.093 | 25.081 [22.345, 28.026] | Yes |
| 8 | U: $H-5$ | $q$ | -20.415 | -20.873 | 2.437 | 18.437 [16.329, 20.696] | Yes |
| 8 | U: $H-5$ | $p$ | -20.415 | -20.432 | 0.221 | 20.211 [17.463, 23.248] | Yes |
| 8 | F: $H-5$ | $q$ | -33.944 | -36.352 | 7.800 | 28.552 [25.940, 31.332] | Yes |
| 8 | F: $H-5$ | $p$ | -33.944 | -34.025 | 0.996 | 33.029 [29.504, 36.802] | Yes |
| 8 | U: $H-1$ | $q$ | -11.348 | -11.171 | 0.857 | 10.314 [9.034, 11.691] | Yes |
| 8 | U: $H-1$ | $p$ | -11.348 | -11.348 | 0.007 | 11.341 [9.866, 12.945] | Yes |
| 8 | F: $H-1$ | $q$ | -7.855 | -7.580 | 3.869 | 3.711 [3.117, 4.309] | Yes |
| 8 | F: $H-1$ | $p$ | -7.855 | -7.852 | 0.196 | 7.656 [6.632, 8.717] | Yes |
| 8 | U: $H_{\mathrm{BP}}-H_{\mathrm{TF}}$ | $q$ | -0.439 | -0.438 | 0.171 | 0.267 [0.168, 0.341] | Yes |
| 8 | U: $H_{\mathrm{BP}}-H_{\mathrm{TF}}$ | $p$ | -0.439 | -0.439 | 0.005 | 0.434 [0.392, 0.477] | Yes |

Table 32: Signed decomposition at eight steps, averaged over starts within identity and then identities. Every displayed energy is multiplied by $10^3$. The directly retained local-defect energy $R$ is common to actual, original-proxy and corrected paths; $\overline{R} + \overline{D} + \overline{C} = \overline{E^2}$ for each path, not $(\overline{E})^2$. Both positive and negative interactions are retained.

| Model | $R$ | $D_z$ | $C_z$ | $D_q$ | $C_q$ | $D_p$ | $C_p$ |
|---|---|---|---|---|---|---|---|
| U-Net $p=1$ | 0.325 | 0.300 | -0.215 | 0.292 | -0.217 | 0.300 | -0.215 |
| U-Net $p=5$ | 0.092 | 0.634 | 0.174 | 0.686 | 0.188 | 0.636 | 0.174 |
| U-Net $H$ BP4 | 0.003 | 0.011 | 0.001 | 0.011 | 0.001 | 0.011 | 0.001 |
| U-Net $H$ TF4 | 0.003 | 0.015 | 0.002 | 0.015 | 0.002 | 0.015 | 0.002 |
| FNO $p=1$ | 0.586 | 0.681 | -0.679 | 0.644 | -0.659 | 0.681 | -0.679 |
| FNO $p=5$ | 0.295 | 1.771 | 0.534 | 2.209 | 0.584 | 1.788 | 0.534 |
| FNO $H$ | 0.021 | 0.118 | 0.019 | 0.119 | 0.019 | 0.118 | 0.019 |

and full-history comparisons are evaluated over that complete path. Table 33 retains all seven models and all ten identities; no result is inferred by assembling the five short windows.

All 70 continuous paths are available, with no nonfinite approximation. Both proxies pass the mean-gap condition for all four prespecified U-Net contrasts, but neither passes for any of the three FNO contrasts. For FNO $H - p = 1$, the corrected mean gap even has the opposite sign to the actual gap.

Ordinal separation also does not guarantee accurate error-vector reconstruction. Degree-five U-Net's mean six-channel relative discrepancy decreases from $68\%$ to $49.7\%$ after correction, while its complete-history discrepancy decreases from $380.3\%$ to $150.4\%$. Its large source gap remains separable despite these substantial errors. Degree-five FNO worsens: the six-channel discrepancy rises from $302.1\%$ to $520.1\%$, and the complete-history discrepancy from $644.6\%$ to $1401.2\%$; correction increases the discrepancy on nine of ten identities. FNO-$H$ also worsens from $41.1\%$ to $140.3\%$ in the scored vector. These finite failures remain in every summary.

The U-Net $H$ comparison is much better reconstructed. Corrected mean relative discrepancies are $0.191\%$ for BP4 and $0.415\%$ for TF4 in the six-channel vector, and $0.159\%$/$0.361\%$ in the full history. For the mean BP4–TF4 gap, the measured budget relative to the absolute actual gap falls from $84.8\%$ to $3.3\%$; the corrected comparison passes on all ten identities.

At 56 steps, the original proxy reproduces each model's $C$ sign on all ten identities. The corrected proxy misidentifies the sign on one degree-five U-Net identity and one FNO-$H$ identity, while agreeing on the other model–identity pairs. Degree-five FNO retains the correct sign despite its large vector error, illustrating why sign agreement alone is insufficient. Physical failure counts for all actual and approximate paths remain in Table 33.

Table 33: Continuous 56-step paths from one shared initial history per identity. Top: six-channel errors and discrepancies, with physical/nonfinite failure counts for actual/original/corrected paths, each out of ten identities. Bottom: all frozen mean-gap comparisons in the convention of Table 31, including every failed criterion; only the bottom panel is multiplied by $10^3$. No short-window concatenation is used.

| Model | $E_z$ | $E_q$ | $E_p$ | $\delta_q$ | $\delta_p$ | $F_{z/q/p}$ | $N_{z/q/p}$ |
|---|---|---|---|---|---|---|---|
| U-Net $p = 1$ | 0.0257 | 0.0246 | 0.0256 | 0.0064 | $8.5400 \times 10^{-4}$ | 0/0/0 | 0/0/0 |
| U-Net $p = 5$ | 0.0731 | 0.0890 | 0.0841 | 0.0522 | 0.0395 | 10/10/10 | 0/0/0 |
| U-Net $H$ BP4 | 0.0099 | 0.0100 | 0.0099 | $7.4800 \times 10^{-4}$ | $2.1200 \times 10^{-5}$ | 0/1/0 | 0/0/0 |
| U-Net $H$ TF4 | 0.0122 | 0.0123 | 0.0122 | 0.0012 | $5.4700 \times 10^{-5}$ | 1/1/1 | 0/0/0 |
| FNO $p = 1$ | 0.0450 | 0.0503 | 0.0527 | 0.0191 | 0.0187 | 3/10/8 | 0/0/0 |
| FNO $p = 5$ | 0.2107 | 0.6890 | 1.1500 | 0.6450 | 1.1100 | 10/10/10 | 0/0/0 |
| FNO $H$ | 0.0446 | 0.0479 | 0.0850 | 0.0185 | 0.0609 | 5/10/9 | 0/0/0 |

| $T$ | $A - B$ | Path | Actual | Proxy | Budget | Margin [95% CI] | Mean pass |
|---|---|---|---|---|---|---|---|
| 56 | U: $5 - 1$ | $q$ | 47.3980 | 64.4230 | 58.6130 | 5.8100 [-1.1330, 11.8990] | Yes |
| 56 | U: $5 - 1$ | $p$ | 47.3980 | 58.4360 | 40.3940 | 18.0420 [8.8700, 26.2010] | Yes |
| 56 | F: $5 - 1$ | $q$ | 165.6910 | 638.9690 | 664.0100 | -25.0420 [-33.0860, -16.5500] | No |
| 56 | F: $5 - 1$ | $p$ | 165.6910 | 1093.4980 | 1130.8410 | -37.3430 [-71.2360, -10.3440] | No |
| 56 | U: $H - 5$ | $q$ | -63.1430 | -79.0290 | 52.9700 | 26.0600 [23.2960, 28.4360] | Yes |
| 56 | U: $H - 5$ | $p$ | -63.1430 | -74.1220 | 39.5610 | 34.5620 [27.8710, 40.3840] | Yes |
| 56 | F: $H - 5$ | $q$ | -166.1120 | -641.3100 | 663.3420 | -22.0320 [-30.0840, -13.2260] | No |
| 56 | F: $H - 5$ | $p$ | -166.1120 | -1061.1500 | 1173.1030 | -111.9530 [-199.9290, -40.1320] | No |
| 56 | U: $H - 1$ | $q$ | -15.7450 | -14.6060 | 7.1390 | 7.4670 [6.4420, 8.5340] | Yes |
| 56 | U: $H - 1$ | $p$ | -15.7450 | -15.6870 | 0.8760 | 14.8110 [12.6810, 16.8370] | Yes |
| 56 | F: $H - 1$ | $q$ | -0.4200 | -2.3410 | 37.6180 | -35.2770 [-40.1460, -30.3150] | No |
| 56 | F: $H - 1$ | $p$ | -0.4200 | 32.3480 | 79.6270 | -47.2790 [-79.7620, -21.0430] | No |
| 56 | U: $H_{\mathrm{BP}} - H_{\mathrm{TF}}$ | $q$ | -2.2960 | -2.3170 | 1.9460 | 0.3710 [-0.0570, 0.6900] | Yes |
| 56 | U: $H_{\mathrm{BP}} - H_{\mathrm{TF}}$ | $p$ | -2.2960 | -2.2940 | 0.0760 | 2.2180 [1.9320, 2.5380] | Yes |

### L.6 Defect amplitude, regularity, and precision

The amplitude experiment uses $\epsilon \in \{1/32, 1/16, 1/8, 1/4, 1/2, 1\}$ in Eq. 79. FP32 covers two fixed identities; the paired FP64 check uses identity 900001, both prescribed starts, and all seven

models, giving 84 matched cells at each precision. Every $\delta$ compares trajectories evaluated in the same precision. The FP64 calculation promotes the existing stored weights and references before arithmetic; it does not replace the numerical dataset with a more accurate simulation.

The actual U-Net uses GELU, group normalization with positive epsilon, average pooling, and bilinear interpolation; it has no ReLU or max-pooling activation boundary. The fixed-grid FNO combines smooth GELU with affine and spectral operations. These operations support the local smooth real-arithmetic interpretation in Proposition 12; they do not supply a measured Lipschitz constant or a verified neighborhood. Automatic differentiation alone would not establish those assumptions for a nonsmooth architecture.

FNO's archived wrapper casts its input to FP32 and its spectral layer allocates a complex64 buffer. The FP64 adapter removes those two casts within the loaded instance and promotes real/complex parameters to float64/complex128, preserving the original spectral algorithm and weights. Applying the same adapter to an FP32 instance leaves both its output and JVP bitwise unchanged on the fixed CPU check. All reference/actual/proxy/correction calculations within a run use its stated precision, with TF32 disabled and float64 scalar reductions.

In FP64, the original proxy's eight-step adjacent log ratios under amplitude doubling range from $1.946$ to $2.025$. For the corrected path at start 32, the small-amplitude intervals from $1/32$ through $1/4$ yield ratios $3.98$–$4.046$. The same intervals give $1.996$–$2.028$ and $3.972$–$4.117$ in the full standardized seven-channel frame, and $1.995$–$2.028$ and $3.972$–$4.111$ in the complete history. These observed ratios are consistent with the local orders in Proposition 12 where the residual is numerically resolved.

At start 0, small FP64 corrected errors for U-Net $p = 1/H$ approach $4.5 \times 10^{-16}$–$1.5 \times 10^{-15}$, making their lowest-amplitude slopes uninformative. Their eight-step ratios become close to four only at larger amplitudes. Four-step corrections are more affected by rounding: U-Net $p = 1$ at start 0 has a highest-interval ratio of only $3.657$. Conversely, FNO-$H$ at start 32 reaches a ratio $4.287$ over $1/2 \to 1$. FP32 corrected curves commonly plateau near $5 \times 10^{-8}$–$1.2 \times 10^{-7}$, while corresponding FP64 residuals can be much smaller. First-step and two-step structural residuals help diagnose arithmetic. All paired curves and adjacent ratios are retained.

### L.7 Replication and inferential scope

The seed-41/42 replication restricts attention to the frozen U-Net $H$-TF4/$H$-BP4 comparison. It repeats the five eight-step windows on all 20 previously used confirmation identities, and separately evaluates 56 steps from the initial history. Pairing, reference conventions, failure retention, and identity-first aggregation are unchanged. Different trained seeds are reported separately.

All 400 short replication windows are available, without numerical or physical failure. At eight steps, actual BP4 mean error is $11.61\%/11.5\%$ lower than TF4 for seeds 41/42. Both proxies pass the mean-gap condition and arithmetic guard at four and eight steps for each seed; all corresponding margin intervals remain above zero. The eight-step discrepancy budget relative to the absolute actual gap decreases from $55.81\%$ to $3.55\%$ after correction for seed 41, and from $58.65\%$ to $2.29\%$ for seed 42. Across these four models, mean six-channel relative discrepancies are approximately $1.38\%$–$1.48\%$ before correction and $0.034\%$–$0.0668\%$ after it; full-history discrepancies are $0.849\%$–$1.11\%$ and $0.0161\%$–$0.0398\%$.

The primary panel's perfect short-window sign reconstruction does not extend to every replication case. For seed-41 BP4 at eight steps, the original proxy reproduces the actual $C$ sign in 99/100 windows; the correction reproduces it in all 100. The other model–endpoint combinations agree in every window at four/eight steps.

All 80 continuous replication paths are also complete (Table 34). The actual 56-step BP4 benefit over TF4 is $18.18\%/18.78\%$ for seeds 41/42, but the original proxy fails the mean-gap condition for both seeds. Its measured budgets are $119.2\%/108.8\%$ of the absolute actual gaps. The once-corrected budgets fall to $10.36\%/5.83\%$, with positive margin intervals and resolution on all 20 identities for each seed. The original proxy resolves only 12/20 and 13/20 identities. Thus the corrected reconstruction reproduces this longer-horizon comparison across training seeds, while reference-linear propagation alone does not satisfy the specified adequacy criterion.

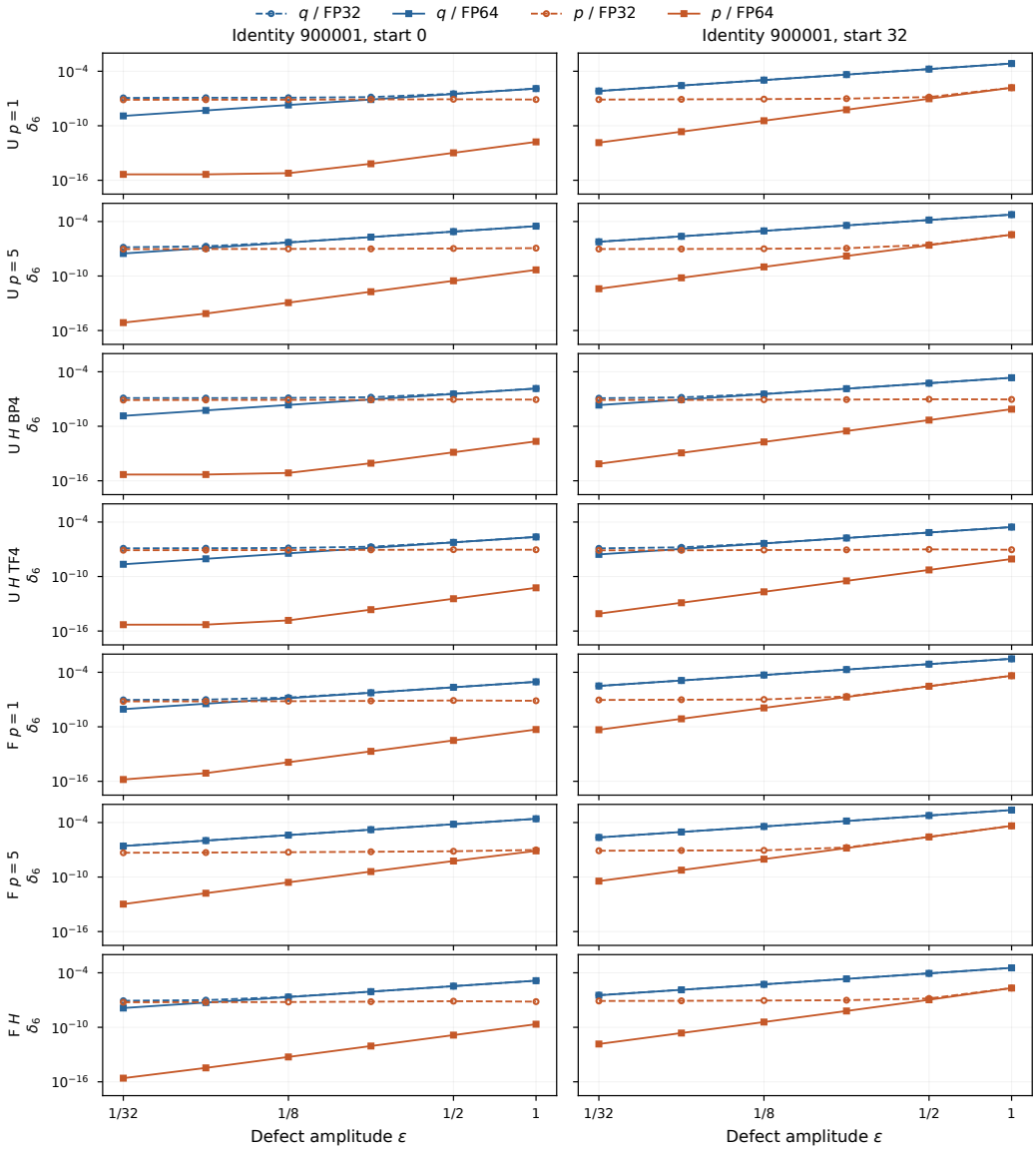



Figure 9: **Defect-amplitude behavior and arithmetic limits.** Eight-step discrepancies for identity 900001 at starts 0 and 32, with all seven models and all six frozen amplitudes. Original/corrected paths and FP32/FP64 evaluations are shown separately; each discrepancy uses the actual trajectory at its own precision. The original model occurs at $\epsilon = 1$. Flat or distorted small-amplitude curves are retained, and no slope is constrained to two or four.

Across the four replication models at 56 steps, mean six-channel relative discrepancies range from $8.9\%$ to $11.49\%$ for the original proxy and $0.296\%$ to $0.961\%$ after correction; complete-history discrepancies range from $7.39\%$ to $10.15\%$ and $0.215\%$ to $0.79\%$, respectively. Over all 80 paths, the maximum six-channel discrepancy is $26.97\%$ before correction and $4.29\%$ after it; the full-history maxima are $23.44\%$ and $3.34\%$. All actual and approximate $C$ values are positive on these paths, and none is numerically nonfinite. Physical-event reconstruction remains imperfect: the actual TF4 trajectories fail on 1/20 and 2/20 identities, whereas their corrected paths fail on 2/20 and 3/20.

Table 34: Fixed U-Net $H$ replication, with each training seed reported separately on all 20 reused confirmation identities. Top: actual error and both discrepancies; physical/nonfinite failure counts have denominator 100 correlated windows at eight steps and 20 continuous paths at 56 steps. Bottom: BP4 minus TF4 paired mean-gap comparisons, multiplied by $10^3$, in the convention of Table 31.

| $T$ | Seed | Continuation | $E_z$ | $\delta_q$ | $\delta_p$ | $F_{z/q/p}$ | $N_{z/q/p}$ |
|---|---|---|---|---|---|---|---|
| 8 | 41 | TF4 | 0.0040 | $1.3300 \times 10^{-4}$ | $9.4600 \times 10^{-6}$ | 0/0/0 | 0/0/0 |
| 8 | 41 | BP4 | 0.0035 | $1.2200 \times 10^{-4}$ | $6.7900 \times 10^{-6}$ | 0/0/0 | 0/0/0 |
| 8 | 42 | TF4 | 0.0039 | $1.3900 \times 10^{-4}$ | $6.2800 \times 10^{-6}$ | 0/0/0 | 0/0/0 |
| 8 | 42 | BP4 | 0.0035 | $1.2500 \times 10^{-4}$ | $4.0500 \times 10^{-6}$ | 0/0/0 | 0/0/0 |
| 56 | 41 | TF4 | 0.0132 | 0.0017 | $1.7400 \times 10^{-4}$ | 1/5/2 | 0/0/0 |
| 56 | 41 | BP4 | 0.0108 | 0.0011 | $7.4700 \times 10^{-5}$ | 0/2/0 | 0/0/0 |
| 56 | 42 | TF4 | 0.0134 | 0.0017 | $1.0600 \times 10^{-4}$ | 2/5/3 | 0/0/0 |
| 56 | 42 | BP4 | 0.0109 | 0.0011 | $4.0600 \times 10^{-5}$ | 1/2/1 | 0/0/0 |

| $T$ | Seed | Path | Actual | Proxy | Budget | Margin [95% CI] | Mean pass |
|---|---|---|---|---|---|---|---|
| 8 | 41 | $q$ | -0.4580 | -0.4580 | 0.2560 | 0.2020 [0.0550, 0.3210] | Yes |
| 8 | 41 | $p$ | -0.4580 | -0.4580 | 0.0160 | 0.4420 [0.3950, 0.4920] | Yes |
| 8 | 42 | $q$ | -0.4510 | -0.4520 | 0.2640 | 0.1870 [0.0190, 0.3210] | Yes |
| 8 | 42 | $p$ | -0.4510 | -0.4510 | 0.0100 | 0.4410 [0.3920, 0.4920] | Yes |
| 56 | 41 | $q$ | -2.4040 | -2.4250 | 2.8650 | -0.4390 [-1.3380, 0.2470] | No |
| 56 | 41 | $p$ | -2.4040 | -2.4060 | 0.2490 | 2.1570 [1.9110, 2.4020] | Yes |
| 56 | 42 | $q$ | -2.5150 | -2.5460 | 2.7370 | -0.1910 [-0.7400, 0.2870] | No |
| 56 | 42 | $p$ | -2.5150 | -2.5130 | 0.1470 | 2.3670 [2.1280, 2.6110] | Yes |

## M PARAMETER EXTRAPOLATION

The original extrapolation suite contains eight fixed cases per system, identities 10001–10008, disjoint from the canonical training and test populations. Each case varies one physical parameter beyond its training support while keeping the remaining parameters within support. For every case, we generate paired degree-one, degree-three, and degree-five trajectories together with an independently generated reference solution. Evaluating the 252 frozen baseline checkpoints yields 2,016 closed-loop rollouts over these extrapolation cases.

The numerical-fidelity reversal observed in-distribution remains visible at these parameter stress points. Across the eight fixed cases, the median aggregate closed-loop $p = 5/p = 1$ error ratio is $0.293$ for Allen–Cahn and $2.672$ for GLM. The direction is consistent across all eight cases in both systems: degree five improves over degree one for Allen–Cahn, whereas the GLM reversal persists outside the training parameter range.

These are fixed single-axis parameter stress tests.

**Recursive-training and capacity stress tests.** The degree-five K4 error reduction persists for both architectures and capacities in the 384-cell evaluation of all eight fixed single-axis parameter-extrapolation cases. These tests use common-reference histories, unlike the degree-matched histories of the original production OOD evaluation. Physical failures remain frequent despite the error reductions.

## N SCOPE AND LIMITATIONS

Formal references are finite numerical approximations, and changes in dissipation, dispersion, and limiting remain coupled. No source pair meets the specified $\pm 10\%$ label-error matching criterion. The source comparison therefore establishes nonmonotonic transfer rather than an equal-accuracy mechanism comparison.

The fixed-source comparisons and mechanism diagnostics use exploratory cohorts; the U-Net performance confirmation uses new identities and seeds. The adapted FNO $H$ result is conditional on seed 40, and the final $H$ route has no completed eight-point parameter-extrapolation matrix.

The DG construction supplies an exact source–architecture example. Neural reconstruction is assessed using the observed discrepancy $\delta$, rather than a certified uniform remainder bound. Physical-admissibility results remain separate from prediction-error rankings.

## O OPEN-SOURCE RELEASE

Upon publication, we will release the benchmark and the complete experimental pipeline required to reproduce the study. The release will include the paired numerical trajectories and reference solutions, frozen training/validation/test and extrapolation splits, numerical-generation configurations, training and evaluation code, model configurations, diagnostic implementations, and environment specifications.

We will also provide the processed records used to generate the reported tables and figures, together with scripts for rebuilding the manuscript figures from these records. This includes the primary degree–architecture comparisons, persistence-skill analysis, four-source experiments, source-adaptation results, response and trajectory-reconstruction diagnostics, discovery/confirmation comparisons, and the constructive DG analysis. The scalar DG construction and its full sensitivity sweep are designed to be reproducible independently on CPU.