WHEN REVERSIBLE NORMALIZATION CHANGES WHAT EQUIVARIANT FORECASTERS CAN EXPRESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Finite-sample calibration can change which architecture a forecasting experiment recommends, even with the physical task and supervised data fixed. In Lorenz–96, position inputs (POS) and a parameter-matched equivariant-feature control (SHAM) receive the same three-rate validation search. On 128 test trajectories generated after locking selection, POS lowers horizon-32 error from 0.796 to 0.330 with four site-wise calibration trajectories. Increasing only the calibration pool to 256 reverses the mean ranking. With global normalization, the same architecture search favors SHAM and attains 0.084 using four calibration trajectories. Metric certificates preserve all four mean rankings under raw and global-scale errors. A separate five-block centering study repeats the calibration-dependent mean ranking across a sixteen-fold supervision range at fixed training budgets. An equal-dimensional construction distinguishes symmetry compatibility from statistical model choice, although its predicted supervision effect has the opposite sign in the jointly trained CNN. Gray–Scott FNO supplies a second small-pool POS benefit; CNN and long-horizon results delimit transfer. The practical consequence is that a position-input gain within one preprocessing protocol does not establish a need for position-dependent physics. Calibration choices should be controlled when interpreting architecture ablations.

1 INTRODUCTION

An architecture ablation measures a benefit conditional on preprocessing. Position inputs can improve a forecaster even when the physical rule is identical at every site. For the periodic one-site shift, $x_j^+ = x_{j-1}$, one circular convolution is exact. Site-wise centering, $z_j = x_j - \mu_j$, instead requires

$$z_j^+ = z_{j-1} + \mu_{j-1} - \mu_j. \tag{1}$$

The shared physical shift now requires a site-dependent correction. Section 3 gives a four-site input at which every strictly shift-equivariant map fails, while identity dynamics remain exactly representable. The coordinates change which physical maps the architecture admits without removing information. This pointwise obstruction is not a forecast-risk lower bound.

With translation-invariant population moments, finite sampling can create this conflict. Feature-wise scaling estimates each site's mean and scale separately unless space is pooled (scikit-learn developers, 2026); frozen fluctuations then enter the represented translation action. Shared global statistics commute with translations at every sample size, even when estimated inaccurately. To assess the consequence, we vary the statistics pool without adding labels and test matched architecture interventions.

Our central experiment compares position inputs (POS) with a parameter-matched equivariant-feature control (SHAM) on Lorenz–96. Both receive the same three-rate validation search before evaluation on 128 newly generated test trajectories. With four calibration trajectories, POS wins under site-wise normalization (P), reducing error from 0.796 to 0.330. Yet global normalization (G), estimated from the same pool, favors SHAM and reaches 0.084 (Table 1). Enlarging the P pool also reverses the ranking. An ablation that holds P fixed recommends position inputs; comparing calibration protocols recommends a pipeline without them.

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

A validation replay selects G+SHAM in every crossed cell, with twice the candidates of a fixed-normalizer search. Metric certificates and all completed rate pairings support the rankings within the searched grid.

In a separate five-block centering-only study, site-wise centering (Mc) favors POS at calibration sizes 4 and 32 and favors SHAM in mean at 256. This calibration-dependent mean ranking persists across three supervision levels, from 7 to 112 trajectories, at fixed optimization budgets.

Our deterministic calculation separates compatibility from model choice. Ordinary-shift incompatibility persists at every finite pool size, yet the preferred fitted class reverses, even with one parameter per class and equal least-squares curvature. The fitted direction determines both the calibration error corrected and the supervised estimation cost. This fixed-backbone construction gives a possibility result for model choice. We test calibration and supervision separately to distinguish that result from the behavior of a jointly trained network.

Allen–Cahn and Gray–Scott test transfer, with Kuramoto–Sivashinsky as a boundary check. FNO gaps decrease between pools of 4 and 256 in all three primary systems. Gray–Scott FNO repeats the small-pool POS advantage, but not the large-pool reverse ranking; CNN and frozen-backbone outcomes limit transfer. Our contribution is a controlled demonstration that a position-input benefit can recommend a feature absent from the preferred alternative calibration pipeline. This motivates a calibration control for architecture evaluation; the prevalence of changed decisions across public benchmarks remains open.

2 Related work

Equivariance and operator learning. Group-equivariant networks specify how features transform (Cohen & Welling, 2016; Kondor & Trivedi, 2018; Weiler & Cesa, 2019). Fourier neural operators parameterize maps between function spaces (Li et al., 2021), and group-equivariant Fourier operators encode additional symmetries (Helwig et al., 2023). Shared convolutions and Fourier multipliers commute with periodic grid translations when their surrounding operations do. We examine the complete physical predictor, including the input and output coordinate maps.

Symmetry and normalization in forecasting. Compatibility between normalization and equivariance has been studied in several settings. Meng et al. (2024) propose geometric normalization for $\mathrm{SE}(3)$-equivariant particle simulation; Scanvic et al. (2025) characterize layer compatibility with discrete and continuous translations. Spherical FNO applies instance normalization globally across space to retain equivariance (Bonev et al., 2023), and Herbreteau et al. (2023) study equivariance to intensity offset and scale. We examine how finite calibration pools affect measured architecture benefits, using label-fixed forecasting interventions, matched architecture controls, and explicit failure cases. The conjugation identity supplies the analysis, not a new symmetry construction.

Lie-symmetry augmentation transforms differential-equation training examples (Brandstetter et al., 2022); the coordinates in which transformations act must remain consistent. RevIN addresses distribution shift with input-window-dependent statistics (Kim et al., 2022). Such statistics can transform with the window, unlike the fixed maps tested here. Equivariance then depends on this joint transformation and the learned affine parameters. We do not evaluate RevIN or transfer our fixed-map conclusions to it.

Group averaging and statistical risk. Group averaging is established in analyses of data augmentation (Chen et al., 2020), with frame averaging offering related constructions (Puny et al., 2022). We use it to distinguish physical orbit loss, distributional risk, and recursive error.

3 Normalization changes the representation of symmetry

Let $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ be a physical one-step evolution map. A finite group Γ acts through orthogonal permutation matrices T_g, and physical equivariance means $\Phi T_g = T_g \Phi$ for every $g \in \Gamma$. Our experiments use cyclic translations on a periodic lattice. Fix an invertible affine normalizer

$$z = N(x) = S^{-1}(x - \mu), \qquad S = \mathrm{diag}(\sigma_1, \ldots, \sigma_d), \quad \sigma_j > 0. \tag{2}$$

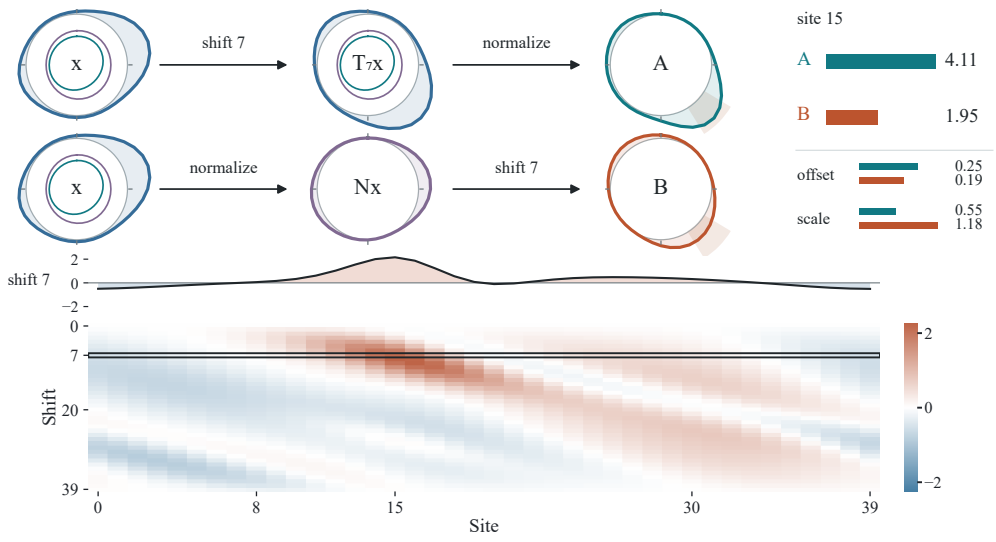


Figure 1: A fixed site-wise normalizer changes the represented translation. The upper route shifts the physical state by seven sites before normalization; the lower route reverses these operations. Blue contours show physical states; nested contours show fixed offsets and scales. At the highlighted site, different calibrations of the same incoming value produce 4.11 versus 1.95. The seven-shift profile and 40×40 discrepancy matrix share a site axis; the outlined row gives the profile. This synthetic illustration does not measure forecast error.

A normalized predictor f gives the physical predictor $\widehat{\Phi} = N^{-1} f N$; the normalized target is $\Phi_N = N \Phi N^{-1}$. If $\mathcal{H}_{\text{shift}}$ is a class constrained to ordinary shift equivariance in normalized coordinates, its physical hypothesis class is

$$\widehat{\mathcal{H}}_N = \{N^{-1} f N : f \in \mathcal{H}_{\text{shift}}\}. \tag{3}$$

Invertibility preserves the state information; it does not make $\widehat{\mathcal{H}}_N$ independent of N.

3.1 A FOUR-SITE OBSTRUCTION AND A NEGATIVE CONTROL

Let the physical dynamics be a one-site shift, $\Phi = T_1$, and first isolate centering by setting $S = I$. In normalized coordinates,

$$[\Phi_N(z)]_j = z_{j-1} + \mu_{j-1} - \mu_j. \tag{4}$$

For $d = 4$, $\mu = (0, 1, 0, -1)^\top$, and $z = 0$, the target is $\Phi_N(0) = (-1, -1, 1, 1)^\top$. Any ordinarily shift-equivariant f obeys $f(0) = T_1 f(0)$ because $T_1 0 = 0$; hence $f(0)$ must be spatially constant and cannot equal this target. Indeed, $\min_c \|c\mathbf{1} - \Phi_N(0)\|_2^2 = 4$. This is an exact expression obstruction at one constructed input, not an in-distribution forecast-risk bound. It also predicts a negative control: if $\Phi = I$, then $N \Phi N^{-1} = I$ for every invertible N, so the identity dynamics remain compatible with ordinary equivariance. The conflict depends jointly on the coordinate map and the dynamics.

3.2 THE INDUCED ACTION AND THE IMPLEMENTED ACTION

Proposition 1 (Affine transport of the action)**.** *The physical action T_g is represented in normalized coordinates by*

$$T_g^N = N T_g N^{-1}, \qquad T_g^N(z) = S^{-1} T_g S z + S^{-1}(T_g \mu - \mu). \tag{5}$$

For a fixed g, $T_g^N = T_g$ on all of $\mathbb{R}^d$ if and only if $T_g \mu = \mu$ and $T_g S = S T_g$. Moreover, $\widehat{\Phi}$ is physically equivariant if and only if $f T_g^N = T_g^N f$ for every g.

For the full cyclic group, the conditions in Proposition 1 require constant means and scales across positions. Global statistics satisfy them for every finite pool. Position-wise statistics need not. The transformed action in Equation (5) can have both a position-dependent linear part and an offset; using an ordinary array shift discards both corrections.

For the cyclic convention $(T_g z)_j = z_{j-g}$, with indices taken modulo d, Equation (5) is the explicit componentwise rule

$$(T_g^N z)_j = \frac{\sigma_{j-g}}{\sigma_j}\, z_{j-g} + \frac{\mu_{j-g} - \mu_j}{\sigma_j}. \tag{6}$$

Thus the transported action has a scale ratio and an offset at every site. Under the trajectory-level assumptions in Appendix A.4, the scale ratios converge to 1 and the offsets converge to 0 as the number of independent trajectories grows; this statement does not assert a finite-sample forecasting law.

The distinction can be measured directly. If $x = N^{-1}z$, then

$$\widehat{\Phi}(T_g x) - T_g \widehat{\Phi}(x) = S\big[f(T_g^N z) - T_g^N f(z)\big]. \tag{7}$$

Thus, the relevant commutator in normalized coordinates uses T_g^N. An ordinary shift-equivariant model instead satisfies $fT_g = T_g f$ and induces equivariance under the generally different physical action $N^{-1}T_g N$. The four-site example shows that this difference can obstruct expression; the identity control shows that it need not.

A target commutator bounds approximation error on two related inputs, but not test risk without a condition on their probability mass (Appendix A). Indeed, the Gaussian example in Appendix F has decreasing optimal equivariant risk despite a growing pointwise obstruction. No commutator-only law orders prediction risk.

3.3 Calibration and supervision can favor opposite architectures

Let $X \sim \mathcal{N}(0, s_x^2 I_d)$ and $\Phi(x) = Bx + r(x)$ be deterministic and shift-equivariant. The fixed circulant backbone satisfies $B\mathbf{1} = \mathbf{1}$, with $\mathbb{E}r(X) = 0$ and covariance Σ. Estimate $\hat{\mu}$ from n calibration states, independently of m supervised states. G and P here use *centering only*: $h_G = \Pi\hat{\mu}$ and $h_P = \hat{\mu}$, where $\Pi = \mathbf{1}\mathbf{1}^\top/d$.

Proposition 2 (Risk of fitting a fixed correction subspace)**.** *Fit $Bz + U\theta$ by least squares to centered targets, where $U^\top U = I_k$ and $Q = UU^\top$ are fixed independently of the samples. Let $\Omega = s_x^2(B-I)(B-I)^\top$ and $R_0 = \mathrm{tr}(\Sigma)/d$. Expected physical per-site squared risks satisfy*

$$\begin{aligned} R_Q^G &= R_0 + \frac{\mathrm{tr}(Q\Sigma)}{md}, \\ R_Q^P &= R_0 + \frac{\mathrm{tr}[(I-Q)\Omega]}{nd} + \frac{\mathrm{tr}(Q\Sigma)}{md}. \end{aligned} \tag{8}$$

The fitted directions determine both the calibration error left uncorrected and the supervised estimation cost. Their number alone does not determine this tradeoff.

For a homogeneous example, take $d = 8$, $s_x = 1$, $B = T_1$, and

$$r(x) = (I - \tfrac{1}{4}T_1)(x \odot x - \mathbf{1}). \tag{9}$$

Compare one coefficient along $u = \mathbf{1}/\sqrt{8}$ with one along the fixed position vector $v = (1,-1,1,-1,1,-1,1,-1)^\top/\sqrt{8}$. Both fits have one parameter and least-squares Hessian $2/d = 1/4$. Here $\Sigma = \frac{9}{8}I + \frac{1}{2}\Omega$, $u^\top\Omega u = 0$, and $v^\top\Omega v = 4$. Writing $D = R_u - R_v$ gives

$$D_G = -\frac{1}{4m}, \qquad D_P = \frac{1}{2n} - \frac{1}{4m}, \qquad n_\star = 2m. \tag{10}$$

Positive values favor the position direction. Increasing calibration lowers D_P by reducing the correction to be learned; increasing supervision raises it by reducing the sampling variance of the fitted coefficient. With $m = 16$, P favors the position direction at $n = 4$ and the shared direction at $n = 256$, while G favors the shared direction at both sizes.

Corollary 3 (Compatibility does not select the fitted class)**.** *In this example, the P centering map fails to commute with T_1 almost surely at every finite n, yet the expected ranking reverses at $n = 2m$.*

Appendix H.1 proves both results. This matches fitted dimension and quadratic curvature with exact labels. The functional classes still differ, and the backbone is fixed. The calculation concerns centering and iid states; it does not predict a neural threshold or allocate neural error between expression and optimization. Section 5.2 tests the directional comparison with a jointly trained backbone and finds support for the calibration direction but the opposite supervision direction.

3.4 What finite samples do and do not imply

Attributing nonuniform statistics to sampling requires an invariant population, beyond physical equivariance alone. Invariant initial states propagated by equivariant dynamics supply one. Independent trajectories yield consistent moments under finite second moments and positive variances, despite within-trajectory correlation. The represented action then converges without implying a forecasting rate or monotonic improvement. We nest trajectory pools and fix supervised transitions (Appendix A.4).

3.5 Complementary controls

Physical group averaging separates one-step orbit risk into averaged-map risk and non-equivariant variance. Coordinate transport instead represents a frozen G predictor in P coordinates without changing its physical function. Appendix B derives both controls. Neither orders recursive errors or assigns a causal share of the POS/SHAM difference to symmetry.

4 Controlled forecasting protocol

Periodic systems. Lorenz–96 (L96) is our densely replicated reference system (Lorenz, 1996):

$$\dot{x}_j = (x_{j+1} - x_{j-2})x_{j-1} - x_j + 8, \qquad j \in \mathbb{Z}/40\mathbb{Z}. \tag{11}$$

We add two 64×64 periodic reaction–diffusion systems. Allen–Cahn uses

$$u_t = 0.0025\nabla^2 u + u - u^3, \tag{12}$$

and Gray–Scott uses

$$\begin{aligned} u_t &= 2\times 10^{-5}\nabla^2 u - uv^2 + 0.0367(1-u), \\ v_t &= 10^{-5}\nabla^2 v + uv^2 - (0.0367 + 0.0649)v, \end{aligned} \tag{13}$$

Solver refinement, grid, and translation checks precede training. Each two-dimensional system has 16 training, 16 validation, 256 external-pool, and 64 confirmation trajectories from disjoint seed ranges. KS is a boundary check; NS2D is excluded because its archived `pool_seed` did not alter the generated pool (Appendix C).

Paired normalization intervention. For each two-dimensional system, architecture, training seed a, pool seed b, and $n \in \{4, 16, 64, 256\}$, G and P use the same supervised data, initial parameter tensors, sampled starts, batches, optimizer, and physical loss. Only the normalizer differs. Three training seeds and three pool seeds are fully crossed, so a design cell contains nine matched G/P pairs. The two-dimensional base runs receive 20,000 Adam updates at 10^{-3}; endpoint checkpoints at $n = 4, 256$ continue to 100,000 updates without restarting. This yields 216 formal records per system. Both the circular CNN and FNO have periodic boundaries, receive no coordinates, and start as the identity through a zero output layer. All planned outcomes, including divergent rollouts, remain in the audit.

The L96 fixed-rate extension contains 30 CNN pairs at each of seven pool sizes and 15 FNO40 pairs at each of four sizes (Appendix C.2). Frozen controls, the G/M/S/P factorial, and KS remain separate protocols.

Metrics and statistical units. For confirmation trajectory i, physical error is

$$e_{i,h} = \frac{\|D_{\mathrm{ref}}^{-1}(\widehat{x}_{i,h} - x_{i,h})\|_2}{\sqrt{d}}, \qquad E_h = \frac{1}{M}\sum_{i=1}^{M} e_{i,h}, \qquad \Delta_n = E_{32}(\mathrm{P}_n) - E_{32}(\mathrm{G}_n). \tag{14}$$

The diagonal D_{ref} is fixed across matched arms using training standard deviations: one global scalar for primary L96 and Allen–Cahn, per-channel scales for Gray–Scott, and per-site scales for the L96 architecture interaction and adapter control. Positive Δ_n means larger P error. Nonfinite outcomes remain failures. The paired unit fixes system, architecture, budget, n, training seed, and pool seed; shared confirmation trajectories and horizons do not add model replicates.

For complete 3×3 cells, a bootstrap replicate samples the three training seed labels and three pool seed labels independently with replacement, retains their Cartesian product, and recomputes the paired mean, following crossed row/column resampling (Owen, 2007). We report this crossed-factor interval in the main tables. A second audit jointly resamples the shared confirmation trajectories with the same indices in every arm. Both are descriptive conditional intervals: three levels per factor do not support an asymptotic population claim or a test of monotonicity.

Frozen one-step diagnostics. For every two-dimensional G/P checkpoint at $n = 4,256$, we evaluate all 64 translations in $H = \{(8a, 8b) : a, b \in \mathbb{Z}_8\}$ on 32 shared confirmation trajectories. Equation (35) then separates orbit risk into the risk of the exactly averaged predictor and its non-equivariant variance. The identity is checked record by record. A separate target-map audit evaluates $N\Phi N^{-1}$ under eight fixed ordinary shifts. If $N^{-1}T_g N(x)$ makes the reference solver nonfinite, the shift is retained as an evaluation failure. Solver failure alone does not establish that the input lies outside the data support.

Architecture interactions. POS appends fixed sine/cosine channels: two in L96 and four in Gray–Scott. SHAM supplies equally many translated input-state channels, matching parameter counts. Pairs copy the state branch and zero-initialize added inputs, sharing the initial function, starts, batches, and data. Effective capacity and subsequent optimization can differ (Appendices E.12 and G).

The L96 equal-search comparison crosses five training and five pool seeds under G/P at $n = 4,256$. Each POS/SHAM arm selects among $\{3\times10^{-4}, 10^{-3}, 3\times10^{-3}\}$ after 100,000 updates by validation E_{32}, breaking ties toward the lower rate. We exclude the first 28 trajectories of the original validation and test files, which duplicate training trajectories, leaving 100 per set. After inspecting the original test results, we locked all 200 selections before generating 128 new test trajectories, with no initial-state overlap with the 15 earlier files (Appendix E.9). M remains fixed-rate. Gray–Scott crosses three training and three pool seeds at rate 10^{-3}, with 20k runs continued to 100k (Appendix G).

For matched POS/SHAM comparisons, define

$$\begin{aligned} D_{N,n} &= E_{32}(\mathrm{SHAM}; N, n) - E_{32}(\mathrm{POS}; N, n), \\ K_n &= D_{\mathrm{P},n} - D_{\mathrm{G},n}, \qquad J_{\mathrm{SHAM}} = K_4 - K_{256}. \end{aligned} \tag{15}$$

Positive D favors POS; K_n compares that advantage under P and G, and J_{SHAM} measures its calibration sensitivity. We report D separately because positive interactions need not imply a P benefit. Fresh L96 summaries reanalyze locked evaluations using 50,000 crossed-factor bootstrap draws and shared indices, conditional on selected rates and test trajectories (Appendix E.10). They do not assign a causal error fraction to symmetry.

5 Results

5.1 Preprocessing changes the observed architecture comparison

The original EQ/POS comparisons show a selective small-pool benefit in L96 and Gray–Scott FNO, but not uniformly in Gray–Scott CNN (Appendix Tables 19 and 20). We prioritize parameter-matched POS/SHAM contrasts.

The L96 three-rate search yielded 200 selected checkpoints from 600 candidates at 100,000 updates (Appendix E.8). Choices were locked before generating 128 new test trajectories; an exact initial-state audit found no overlap with training, pool, or earlier confirmation files (Appendix E.9).

Table 1: More calibration data reverses the validation-based architecture decision without adding supervised examples. Fresh-test errors use 128 new trajectories and 200 locked checkpoints after equal three-rate searches. Means span 25 crossed seed cells. Positive $D = E_{32}(\mathrm{SHAM}) - E_{32}(\mathrm{POS})$ favors POS; intervals resample both seed factors. The last column replays arm selection from original validation errors, with the number of matching choices. This retrospective analysis does not select on test error (Appendix E.11).

Statistics	Pool n	POS E_{32}	SHAM E_{32}	D [95% interval]	Validation choice
Global	4	0.164	0.084	-0.080 $[-0.087, -0.073]$	SHAM (25/25)
Global	256	0.163	0.083	-0.080 $[-0.086, -0.074]$	SHAM (25/25)
Site-wise	4	0.330	0.796	$+0.467$ $[+0.436, +0.496]$	POS (25/25)
Site-wise	256	0.248	0.208	-0.040 $[-0.049, -0.030]$	SHAM (25/25)

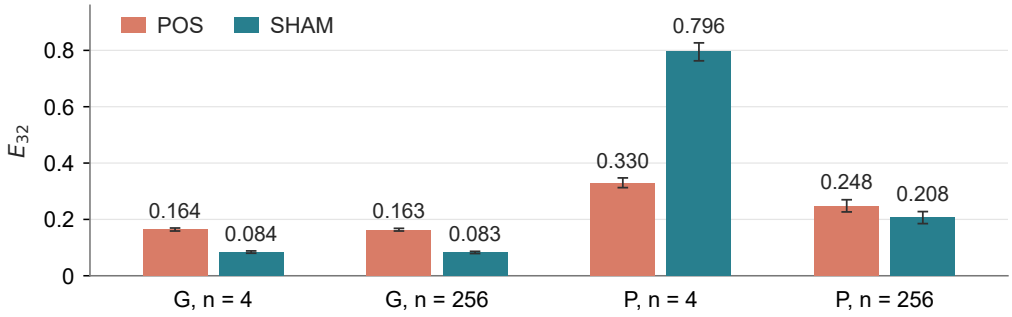


Figure 2: The architecture comparison depends on calibration. Lower E_{32} is better. Bars show means over 25 crossed training/pool-seed cells after equal three-rate validation search; whiskers show 95% crossed-factor bootstrap intervals, conditional on selected rates and the shared fresh 128-trajectory test set. POS wins only under small-pool P; G+SHAM has the lowest error at either pool size. Only $n = 4, 256$ were measured for this matched comparison, so the location of a ranking transition is unknown.

On the fresh test set, POS cuts mean error from 0.796 to 0.330 under site-wise normalization at $n = 4$: $D_{\mathrm{P},4} = 0.467$ $[0.436, 0.496]$, positive in all 25 cells (Table 1). At $n = 256$, the ranking reverses: $D_{\mathrm{P},256} = -0.040$ $[-0.049, -0.030]$, negative in 24 of 25 cells. Under global normalization, SHAM wins by about 0.08 at both sizes. Calibration therefore changes the architecture ranking with the physical task and supervision fixed.

The change in D between pools is 0.507 $[0.477, 0.536]$ under P and 0.00051 $[-0.00410, 0.00467]$ under G, giving $J_{\mathrm{SHAM}} = 0.506$ $[0.476, 0.536]$, positive in all 25 crossed cells (Appendix E.10). Representation and optimization contributions remain unresolved. The earlier test split agrees; M was not searched, and no sustained horizon-128 gain appears (Appendix E.9).

Robustness to metrics and rate choices. A norm-bound certificate preserves all four mean rankings under raw physical RMSE and RMSE divided by a global training standard deviation. The P SHAM-minus-POS contrast has deterministic global-scale enclosures $[0.435, 0.496]$ at $n = 4$ and $[-0.053, -0.028]$ at $n = 256$, using the same fresh evaluations and five hash-verified reference arrays (Appendix E.11). On the earlier 100-trajectory holdout, all nine completed rate pairings favor POS at $n = 4$ in all 25 cells, and SHAM in mean at $n = 256$. The reversal survives in-grid rate choices; boundary selections leave rates beyond the grid and fully optimized performance untested.

The preferred architecture depends on which calibration choices are allowed. Replaying validation decisions with P fixed chooses POS in all 25 small-pool cells, with fresh-test error 0.330. Fixing G chooses SHAM throughout, reaching 0.084 (Figure 2). Joint selection chooses G+SHAM at both pool sizes, reducing small-pool error by 0.246 $[0.227, 0.265]$ relative to P-only selection without more calibration trajectories or labels, but with twelve candidates per cell instead of six. This

retrospective readout establishes a decision consequence within the controlled workflow, not a public benchmark result (Appendix E.11).

In the matched Gray–Scott control, FNO's small-pool P advantage persists: $D_{\mathrm{P},4} = 0.2321$ $[0.1822, 0.2983]$ at 20k and 0.3214 $[0.2765, 0.3678]$ after continuation to 100k, positive in all nine cells at each budget. The endpoint interactions are $J_{\mathrm{SHAM}} = 0.2398$ $[0.1819, 0.3109]$ and 0.3313 $[0.2843, 0.3852]$; both large-pool POS–SHAM contrast intervals overlap zero. CNN's 20k interaction spans zero; at 100k a CNN $n = 256$ SHAM/P rollout has $E_{32} = 9.38 \times 10^4$ and dominates its group mean. Figure 12 retains this instability (Appendix G). Matching input-weight counts does not exclude effective capacity or optimization differences.

A 72-fit frozen-backbone L96 control fails to reproduce selective P repair. After 5,000 updates, matched 640-parameter POS/SHAM adapters at $n = 4$ have P errors 0.793/0.790 and G errors 0.170/0.135. SHAM's larger G improvement drives the interaction, leaving effective capacity and joint optimization unresolved (Appendix Table 21).

5.2 Calibration changes the ranking across supervision levels

A separate L96 study crosses calibration size $n \in \{4, 32, 256\}$ with $m \in \{7, 28, 112\}$ supervised trajectories. Gc uses a shared mean; Mc uses site-wise means, both with one external scalar scale. Across five independent blocks, 540 candidates trained for 100,000 updates yield 180 validation-locked choices with finite test metrics. The primary risk R is physical one-step MSE on 128 test trajectories per block; E_{32} is secondary. This centering-only design differs from the earlier G/P study (Appendix J).

For $D_{c,n,m,b} = R_{\mathrm{SHAM}} - R_{\mathrm{POS}}$ in block b, the prespecified contrasts are

$$C_{n,c,b} = \tfrac{1}{3}\sum_{m}(D_{c,4,m,b} - D_{c,256,m,b}), \qquad C_{m,c,b} = \tfrac{1}{3}\sum_{n}(D_{c,n,112,b} - D_{c,n,7,b}). \tag{16}$$

The restricted calculation motivated positive Mc contrasts for both changes. The calibration contrast is 0.00510 $[0.00330, 0.00690]$, positive in all five blocks. At each supervision size, Mc favors POS in mean at $n = 4, 32$ and SHAM at $n = 256$. POS wins all 30 small/intermediate-pool block/grid comparisons; SHAM wins 14/15 at $n = 256$ and all 45 under Gc. These paired counts are not independent repetitions (Appendix Figure 14).

The supervision contrast is -0.000217 $[-0.000288, -0.000146]$, negative in all five blocks and opposite to the prespecified prediction. Both primary Gc intervals span zero. Intervals are descriptive t_4 intervals over five blocks, conditional on the shared scale and protocol. Secondary E_{32} retains a positive Mc calibration contrast, 0.519 $[0.412, 0.627]$, while its supervision interval spans zero. Thus calibration-dependent ranking persists across the supervision grid, but the fixed-backbone supervision prediction does not describe the jointly trained CNN.

5.3 The L96 gap is an early-horizon, pool-dependent effect

Expanded fixed-rate L96 studies find smaller G/P gaps at $n = 256$ than at $n = 4$ for CNN and FNO40, including 100,000-update continuations. Intermediate pools support the trend (Figure 5; Appendix D). Both arms approach the decorrelation scale by horizon 128; finite rollouts through horizon 512 do not establish accurate invariant statistics.

5.4 Cross-system transfer and failure modes

Each two-dimensional audit contains 216 formal records. In the base G/P study, all 213 metric-complete cells across 23 complete groups have $E_{32}(\mathrm{P}) > E_{32}(\mathrm{G})$. These counts share crossed seeds and confirmation trajectories and are not independent replications.

Across both systems and budgets, FNO's G/P gap decreases by 74.8–87.6% between $n = 4$ and 256 (Table 2), using Equation (14). CNN does not follow the same pattern.

Gray–Scott CNN gaps at 20k are nonmonotone: 0.345, 0.559, 0.544, and 0.215 for $n = 4, 16, 64, 256$. At 100k, endpoints 0.428/0.431 show no recovery. Allen–Cahn CNN is unstable: 17 of 36 P runs at 20k have nonfinite errors before horizon 128, and all nine $n = 256$ P continuations at 100k fail;

Table 2: Crossed two-dimensional replication. Each complete Δ_n averages nine matched cells; brackets are descriptive crossed-factor 95% intervals. Positive means P is worse. The last column gives nonfinite P runs at $n = 4; n = 256$. Allen–Cahn CNN at 100k/$n = 256$ is incomplete because 3/9 E_{32} values fail; no available-case mean is substituted. Reference scales are global for Allen–Cahn and channel-wise for Gray–Scott (Equation (14)).

System	Architecture	Updates	Δ_4 [95% interval]	Δ_{256} [95% interval]	nonfinite P runs
Allen–Cahn	CNN	20k	1.144 [0.613, 2.012]	153.976 [1.302, 563.251]	0/9; 2/9
Allen–Cahn	CNN	100k	1.460 [0.953, 1.871]	incomplete	0/9; 9/9
Allen–Cahn	FNO	20k	0.331 [0.324, 0.337]	0.071 [0.066, 0.076]	0/9; 0/9
Allen–Cahn	FNO	100k	0.331 [0.324, 0.337]	0.083 [0.078, 0.089]	0/9; 0/9
Gray–Scott	CNN	20k	0.345 [0.293, 0.404]	0.215 [0.195, 0.234]	0/9; 0/9
Gray–Scott	CNN	100k	0.428 [0.320, 0.533]	0.431 [0.202, 0.910]	0/9; 0/9
Gray–Scott	FNO	20k	0.787 [0.664, 0.943]	0.104 [0.086, 0.120]	0/9; 0/9
Gray–Scott	FNO	100k	0.855 [0.741, 0.988]	0.106 [0.088, 0.124]	0/9; 0/9

three are already nonfinite at horizon 32. We therefore report the latter cell as an incomplete nine-pair estimand, rather than averaging its six finite values.

In KS, the FNO64 gap decreases from 0.149 to 0.040, whereas the three-seed CNN mean changes from -0.340 to 0.035; one CNN G seed dominates the reversal, and both seed ranges span zero. The result rules out a universal architecture-independent ordering but does not distinguish capacity from optimization (Appendix Table 11).

5.5 The represented-action defect is real but does not determine rollout cost

Allen–Cahn's P target defect decreases from 0.3155 to 0.00646 as the pool grows; G is at numerical zero. Gray–Scott has incomplete small-pool target audits because six of eight prescribed transformed inputs give nonfinite reference evolutions. The frozen one-step orbit decomposition holds in all 288 records (maximum residual 1.71×10^{-13}), but its removed variance does not order the CNN/FNO rollout responses. Failed target evaluations and complete orbit summaries remain in Appendix E.4. Neither diagnostic identifies the causal share of the architecture ranking attributable to expression constraints.

6 Discussion and limitations

In L96, small-pool site-wise normalization favors position inputs, whereas allowing global normalization selects a pipeline without them, with unchanged physics, supervision, and calibration count. Raw and global-scale certificates preserve the rankings. The architecture conclusion must therefore specify the calibration protocol.

For tasks with the stated symmetry, compare shared and site-wise statistics with matched pools, architectures, rates, and budgets. Report the preferred architecture within each protocol and across protocols, including the added search cost. Contrasts $D_{N,n}$, K_n, and J_{SHAM} measure this sensitivity. A gain that survives the control supports calibration robustness; it does not by itself establish heterogeneous physics. Forcing and boundaries determine the applicable symmetry (Appendix I).

The centering study repeats the mean ranking change at all three supervision sizes. Its opposing supervision contrast limits the fixed-backbone explanation of this effect. Increasing available trajectories at fixed update counts leaves expression, effective capacity, and optimization intertwined. The mathematical construction separates compatibility from model choice under explicit assumptions; it does not identify their contributions in the trained networks.

Gray–Scott FNO repeats the small-pool benefit without a full reversal; CNN and frozen L96 results limit transfer. The centering study leaves the full G/P transition and unsearched rates unresolved. Limited seed levels constrain population inference, and public-benchmark prevalence remains open. For architecture evaluation, the supported control is to test whether the preference survives a change in calibration.

AI USE STATEMENT

Generative AI tools were used to help with and polish the writing of this paper, including language editing, rewriting for clarity, and translation. They were also used for search and discovery, such as locating related work. Every cited source was checked against its original publication, and the authors reviewed and edited every AI-assisted change, verified the final text, and take responsibility for the claims and supporting artifacts.

REPRODUCIBILITY STATEMENT

The protocol specifies normalization estimands, matched training factors, validation-based rate selection, a newly generated confirmation set, failure handling, and crossed-seed uncertainty. The appendix records data generation, architectures, optimization schedules, seed structure, intervention definitions, and robustness scope. The anonymous supplement includes the 600-candidate L96 search and its 200 fresh confirmations, plus all 540 centering-study candidates and 180 locked choices, per-trajectory test readouts, frozen protocols, source, hashes, and aggregation scripts. The new study's checkpoint and raw-array identities were verified on the experiment host; their tensors are not included in the compact supplement.

REFERENCES

Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical Fourier neural operators: Learning stable dynamics on the sphere. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2806–2823. PMLR, 2023. URL https://proceedings.mlr.press/v202/bonev23a.html.

Johannes Brandstetter, Max Welling, and Daniel E. Worrall. Lie point symmetry data augmentation for neural pde solvers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2241–2256. PMLR, 2022. URL https://proceedings.mlr.press/v162/brandstetter22a.html.

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020. URL https://www.jmlr.org/papers/v21/20-163.html.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999. PMLR, 2016. URL https://proceedings.mlr.press/v48/cohenc16.html.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 2959–2969. PMLR, 2021. URL https://proceedings.mlr.press/v139/elesedy21a.html.

Jacob Helwig, Xuan Zhang, Cong Fu, Jerry Kurtin, Stephan Wojtowytsch, and Shuiwang Ji. Group equivariant fourier neural operators for partial differential equations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12907–12930. PMLR, 2023. URL https://proceedings.mlr.press/v202/helwig23a.html.

Sébastien Herbreteau, Emmanuel Moebel, and Charles Kervrann. Normalization-equivariant neural networks with application to image denoising. In *Advances in Neural Information Processing Systems*, volume 36, pp. 5706–5728. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/12143893d9d37c3569dda800b95cabd9-Paper-Conference.pdf.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=cGDAkQo1C0p.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2747–2755. PMLR, 2018. URL https://proceedings.mlr.press/v80/kondor18a.html.

Zongyi Li, Nikola 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.

Edward N. Lorenz. Predictability: A problem partly solved. In *Proceedings of the Seminar on Predictability*, pp. 1–18. European Centre for Medium-Range Weather Forecasts, 1996. URL https://www.ecmwf.int/en/elibrary/75462-predictability-problem-partly-solved.

Ziqiao Meng, Liang Zeng, Zixing Song, Tingyang Xu, Peilin Zhao, and Irwin King. Towards geometric normalization techniques in SE(3) equivariant graph neural networks for physical dynamics simulations. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5981–5989, 2024. doi: 10.24963/ijcai.2024/661. URL https://www.ijcai.org/proceedings/2024/0661.pdf.

Art B. Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411, 2007. doi: 10.1214/07-AOAS122. URL https://arxiv.org/abs/0712.1111.

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J. Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.03336.

Jérémy Scanvic, Quentin Barthélemy, and Julián Tachella. Translation-equivariance of normalization layers and aliasing in convolutional neural networks. *arXiv preprint arXiv:2505.19805*, 2025. URL https://arxiv.org/abs/2505.19805.

scikit-learn developers. StandardScaler: Feature standardization. scikit-learn documentation, 2026. URL https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. Accessed September 26, 2026.

Maurice Weiler and Gabriele Cesa. General E(2)-equivariant steerable CNNs. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper/9580-general-e2-equivariant-steerable-cnns.

A COORDINATE IDENTITIES AND THEIR SCOPE

A.1 PROOF OF PROPOSITION 1

Since $N^{-1}(z) = Sz + \mu$, direct substitution gives

$$NT_gN^{-1}(z) = S^{-1}\big(T_g(Sz+\mu) - \mu\big) \tag{17}$$

$$= S^{-1}T_gSz + S^{-1}(T_g\mu - \mu). \tag{18}$$

Equality with the linear map T_g for every z requires equality of both the linear and constant parts. The constant part vanishes precisely when $T_g\mu = \mu$, and the linear part equals T_g precisely when $T_gS = ST_g$. Conversely, these two conditions imply equality. For a permutation action, the second condition means that scales are constant along the permutation's cycles. Requiring it for the cyclic group on d positions makes all scales equal; the corresponding statement holds for means.

Writing $x = N^{-1}z$, we have

$$\widehat{\Phi}T_g(x) = N^{-1}fT_g^N(z), \tag{19}$$

$$T_g\widehat{\Phi}(x) = N^{-1}T_g^N f(z). \tag{20}$$

Invertibility of N proves equivalence of the two equivariance conditions. Subtracting the physical expressions yields Equation (7), because the translation part of N^{-1} cancels. Finally, if $fT_g = T_gf$, conjugating this equality by N^{-1} proves equivariance of $\widehat{\Phi}$ under $N^{-1}T_gN$. None of these statements requires a statistical assumption or a learned model. □

A.2 ACTION MISMATCH GIVES AN UPPER BOUND ON THE PHYSICAL DEFECT

Let $\delta_g(z) = T_g^N z - T_g z$. Equation (5) implies

$$\delta_g(z) = S^{-1}(T_gS - ST_g)z + S^{-1}(T_g\mu - \mu). \tag{21}$$

If $s_{\min} = \min_j \sigma_j$, then

$$\|\delta_g(z)\|_2 \le \frac{\|T_gS - ST_g\|_2 \, \|z\|_2 + \|T_g\mu - \mu\|_2}{s_{\min}}. \tag{22}$$

Suppose f commutes with the ordinary action and is L-Lipschitz on the relevant pairs of inputs. Adding and subtracting $f(T_gz) = T_gf(z)$ gives

$$f(T_g^N z) - T_g^N f(z) = f(T_gz + \delta_g(z)) - f(T_gz) - \delta_g(f(z)), \tag{23}$$

$$\left\|\widehat{\Phi}(T_gx) - T_g\widehat{\Phi}(x)\right\|_2 \le \|S\|_2 \left(L\,\|\delta_g(z)\|_2 + \|\delta_g(f(z))\|_2\right). \tag{24}$$

If the Lipschitz domain also includes 0 and z, then $\|f(z)\|_2 \le \|f(0)\|_2 + L\,\|z\|_2$. With $a_g = \left\|S^{-1}(T_gS - ST_g)\right\|_2$ and $b_g = \left\|S^{-1}(T_g\mu - \mu)\right\|_2$, this closes the bound:

$$\left\|\widehat{\Phi}(T_gx) - T_g\widehat{\Phi}(x)\right\|_2 \le \|S\|_2 \left[2La_g\,\|z\|_2 + a_g\,\|f(0)\|_2 + (L+1)b_g\right]. \tag{25}$$

This is an upper bound, not evidence that the defect is nonzero. The terms cancel for the identity predictor.

A.3 PROOF AND INTERPRETATION OF THE TWO-POINT BOUND

If the true normalized target has commutator $D_g(z) = \Phi_N(T_gz) - T_g\Phi_N(z)$, every ordinary shift-equivariant f satisfies

$$\frac{\|f(z) - \Phi_N(z)\|_2^2 + \|f(T_gz) - \Phi_N(T_gz)\|_2^2}{2} \ge \frac{\|D_g(z)\|_2^2}{4}. \tag{26}$$

Let $e(z) = f(z) - \Phi_N(z)$. Ordinary equivariance of f yields

$$D_g(z) = \Phi_N(T_gz) - T_g\Phi_N(z) = -e(T_gz) + T_ge(z). \tag{27}$$

Since T_g is orthogonal, the squared triangle inequality gives

$$\|D_g(z)\|_2^2 \le 2\,\|e(T_g z)\|_2^2 + 2\,\|e(z)\|_2^2\,, \tag{28}$$

which proves Equation (26). In physical units, each pointwise squared error is at least $s_{\min}^2$ times its normalized-coordinate counterpart. A distributional lower bound would additionally require justified mass on the relevant pairs, or an invariant input measure for the *ordinary normalized action*. The physical pair corresponding to $z, T_g z$ is $x, N^{-1}T_g N(x)$, which is generally not the physical orbit pair $x, T_g x$. We make no claim that the confirmation set satisfies this extra support condition. □

A linear example. Let $\Phi(x) = Ax$ with A circulant, so that $AT_g = T_g A$, and let $S = sI$. Then

$$\Phi_N(z) = Az + \frac{A\mu - \mu}{s}. \tag{29}$$

Although A commutes with ordinary shifts, this affine map does so only when its offset $b = (A\mu - \mu)/s$ is shift invariant. A nonuniform mean can therefore introduce a normalized target offset incompatible with ordinary shift equivariance. Conjugated equivariance still holds exactly. The case $A = I$ has $b = 0$ for every μ, demonstrating why nonuniform statistics alone cannot imply an approximation obstruction.

Closed-form four-site control. Table 3 completes the minimal comparison suggested by Equation (4). The entries are per-site squared residuals at the zero-code probe for the shift task; the identity row uses an arbitrary probe. G uses a shared mean, P uses $\mu = (0, 1, 0, -1)$, and "transported" uses the exact action NT_1N^{-1}. No fitting or random seed is involved.

Table 3: Exact four-site sanity check. The P/ordinary residual is an expression obstruction at the stated probe, not a distributional test error.

Physical task	Coordinates	Enforced action	Probe MSE
one-site shift	G	ordinary shift	0
one-site shift	P	ordinary shift	1
one-site shift	P	transported shift	0
identity	P	ordinary shift	0

The released verification script constructs the four matrices and asserts all four entries. This control checks the coordinate algebra; it is not a proposed training method.

A.4 Consistency under an invariant trajectory law

Corollary 4 (Trajectory-level consistency)**.** *Let $X_i = (X_{i,1}, \ldots, X_{i,L})$, $i = 1, 2, \ldots$, be independent identically distributed trajectories in $(\mathbb{R}^d)^L$, with fixed finite L. Assume that each time marginal is invariant under the transitive permutation group Γ, that each coordinate has finite second moment, and that the variance of the mixture over saved times is positive. Form position-wise empirical means and standard deviations from all nL states, adding the same fixed $\epsilon \ge 0$ to each standard deviation. Then the means and scales converge almost surely to spatial constants. Consequently, for each g, $N_n T_g N_n^{-1} \to T_g$ uniformly on compact sets.*

Proof. For each position j, define trajectory averages

$$A_{i,j} = L^{-1}\sum_{t=1}^{L} X_{i,t,j}, \qquad B_{i,j} = L^{-1}\sum_{t=1}^{L} X_{i,t,j}^2. \tag{30}$$

The finite second-moment assumption makes both variables integrable. The strong law applied across trajectories yields $n^{-1}\sum_i A_{i,j} \to m_j$ and $n^{-1}\sum_i B_{i,j} \to q_j$ almost surely. Group invariance and transitivity imply $m_j = m$ and $q_j = q$ for all j. Thus the empirical variance converges to $q - m^2 > 0$ and the stabilized scale to $s = \sqrt{q - m^2 + \epsilon} > 0$. Since d is finite, these convergences hold jointly over coordinates. The affine coefficients in Equation (21) then converge to zero, proving uniform convergence on each compact set. □

No independence across saved times was used. The statement concerns increasing independent trajectories at a fixed sampling protocol. An invariant initial distribution followed by an equivariant deterministic evolution supplies the required invariant marginals, even after a fixed burn-in. Conversely, a noninvariant initial distribution or spatially inhomogeneous sampling can produce nonconstant population statistics despite equivariant dynamics. The corollary does not apply in that case. Nor does it give a convergence rate for learning or a monotonic relation between n and rollout accuracy.

Implemented moments and sampling variance. With the trajectory averages $A_{i,j}$ and $B_{i,j}$ defined in the proof, the implementation pools all saved frames through

$$\widehat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} A_{i,j}, \qquad \widehat{v}_j = \frac{1}{n}\sum_{i=1}^{n} B_{i,j} - \widehat{\mu}_j^2, \qquad \widehat{\sigma}_j = \sqrt{\widehat{v}_j} + \epsilon. \tag{31}$$

Thus $\widehat{v}_j$ estimates the pooled-frame second central moment $v_j^* = \mathbb{E}[B_{1,j}] - \mu_j^{*2}$, not $\mathrm{Var}(A_{1,j})$. If the trajectories are iid and have finite fourth moments, then

$$\mathbb{E}[(\widehat{\mu}_j - \mu_j^*)^2] = \frac{\mathrm{Var}(A_{1,j})}{n} = \frac{1}{nL^2}\sum_{t=1}^{L}\sum_{u=1}^{L}\mathrm{Cov}(X_{1,t,j}, X_{1,u,j}), \tag{32}$$

$$\widehat{v}_j - v_j^* = O_{\mathbb{P}}(n^{-1/2}). \tag{33}$$

The covariance sum makes explicit why saved frames are not counted as independent replicates. The stabilizer ϵ is added outside the square root, as in the released implementation.

B Group-average projection and coordinate control

B.1 Definitions and scope of the controls

For a fixed physical predictor F, define its finite-group average

$$(\mathcal{R}F)(x) = \frac{1}{|\Gamma|}\sum_{g\in\Gamma} T_g^{-1} F(T_g x). \tag{34}$$

Our R arm applies this operation to the frozen P predictor at every forecast step. All shifts and inverse shifts in Equation (34) act in physical coordinates.

Proposition 5 (Group averaging and squared error)**.** *For a finite group acting through orthogonal transformations, $\mathcal{R}F$ is equivariant and $\mathcal{R}^2F = \mathcal{R}F$. For each x, let $a_g = T_g^{-1}F(T_g x)$ and $\bar{a} = (\mathcal{R}F)(x)$. If Φ is equivariant, then*

$$\frac{1}{|\Gamma|}\sum_g \|F(T_g x) - \Phi(T_g x)\|_2^2 = \|\bar{a} - \Phi(x)\|_2^2 + \frac{1}{|\Gamma|}\sum_g \|a_g - \bar{a}\|_2^2 . \tag{35}$$

If, additionally, the input distribution ρ is group invariant and $F, \Phi \in L^2(\rho)$, $\mathcal{R}$ is an orthogonal projection and

$$\mathbb{E}_\rho \|F - \Phi\|_2^2 = \mathbb{E}_\rho \|\mathcal{R}F - \Phi\|_2^2 + \mathbb{E}_\rho \|F - \mathcal{R}F\|_2^2 . \tag{36}$$

The variance identity for a_g and any fixed target is algebraic; orthogonality and target equivariance identify its left side with the physical orbit loss in Equation (35). This does not compare projected and unprojected error at one unshifted input. Equation (36) does so in expectation under the additional distributional assumption. Neither equation orders autoregressive errors: $(\mathcal{R}F)^h$ generally differs from $\mathcal{R}(F^h)$.

The C control answers a different question. Let N_G use global statistics, N_P use position-wise statistics, and f_G be a frozen global-normalization model. Define its representation in position-wise coordinates by

$$c = N_P N_G^{-1} f_G N_G N_P^{-1}. \tag{37}$$

Then $N_P^{-1} c N_P = N_G^{-1} f_G N_G$ exactly. If f_G is ordinary shift-equivariant, c is equivariant under $N_P T_g N_P^{-1}$. This construction changes coordinates around the G model; it neither retrains a model nor corrects the separately trained P model.

B.2 Proof of Proposition 5

For $h \in \Gamma$, make the change of index $k = gh$ in Equation (34). Using $T_{kh^{-1}}^{-1} = T_h T_k^{-1}$ gives

$$(\mathcal{R}F)(T_h x) = \frac{1}{|\Gamma|} \sum_g T_g^{-1} F(T_{gh} x) \tag{38}$$

$$= T_h \frac{1}{|\Gamma|} \sum_k T_k^{-1} F(T_k x) = T_h (\mathcal{R}F)(x). \tag{39}$$

If F is already equivariant, every summand equals $F(x)$; hence $\mathcal{R}F = F$. Together these facts prove idempotence.

For the pointwise decomposition, expand $a_g - y = (a_g - \bar{a}) + (\bar{a} - y)$ with $y = \Phi(x)$. The cross term vanishes after summation because $\sum_g (a_g - \bar{a}) = 0$. This proves Equation (35). Physical equivariance of Φ and orthogonality of T_g identify its left-hand side with

$$\frac{1}{|\Gamma|} \sum_g \|F(T_g x) - \Phi(T_g x)\|_2^2 . \tag{40}$$

The algebraic variance decomposition itself holds for any fixed y; equivariance is needed for this interpretation as prediction loss on the transformed physical examples.

For the population result, equip $L^2(\rho; \mathbb{R}^d)$ with inner product $\langle F, H \rangle = \mathbb{E}_\rho[F(x)^\top H(x)]$. Define $U_g F(x) = T_g^{-1} F(T_g x)$. Invariance of ρ and orthogonality of T_g make U_g unitary, with $U_g^* = U_{g^{-1}}$. Therefore

$$\mathcal{R}^* = \frac{1}{|\Gamma|} \sum_g U_g^* = \frac{1}{|\Gamma|} \sum_g U_{g^{-1}} = \mathcal{R}. \tag{41}$$

Self-adjointness and idempotence make $\mathcal{R}$ an orthogonal projection onto the equivariant subspace. Since $\mathcal{R}\Phi = \Phi$, the decomposition

$$F - \Phi = (\mathcal{R}F - \Phi) + (F - \mathcal{R}F) \tag{42}$$

is orthogonal, proving Equation (36). The same argument holds for the weighted norm $\|v\|_D^2 = d^{-1}\|D^{-1}v\|_2^2$ when a fixed positive diagonal D commutes with every T_g: translations remain isometries in this inner product. Global and spatially constant per-channel scales satisfy this condition, including the Gray–Scott reference scales. Arbitrary site-wise scales need not; we do not apply this decomposition to the site-weighted L96 adapter errors. □

B.3 Why the projection theorem does not order rollouts

For a general nonlinear map, averaging and composition do not commute. For example, $(\mathcal{R}F)^2(x)$ evaluates the next averaged step at the *averaged* first prediction; $\mathcal{R}(F^2)(x)$ instead propagates each transformed prediction with F before averaging. Proposition 5 concerns the one-step map, or a separately projected h-step map, and does not identify these constructions. Furthermore, an invariant data distribution need not be preserved by the learned autoregressive predictor. Applying the population statement at successive predicted states would require additional assumptions about their distribution. We evaluate the actual composition $(\mathcal{R}F)^h$ empirically.

B.4 The coordinate wrapper is the same physical map

Let $Q = N_G N_P^{-1}$. Then $c = Q^{-1} f_G Q$, and cancellation gives

$$N_P^{-1} c N_P = N_G^{-1} f_G N_G. \tag{43}$$

Since the global normalizer commutes with physical translations,

$$Q(N_P T_g N_P^{-1})Q^{-1} = N_G T_g N_G^{-1} = T_g. \tag{44}$$

Consequently, ordinary equivariance of f_G implies conjugated equivariance of c. With parameter-independent normalizers and the same physical loss, equality of the physical output as a function of parameters also implies equality of losses and gradients wherever the model is differentiable. These are exact-arithmetic statements. Finite-precision discrepancies can accumulate under repeated composition, particularly in chaotic dynamics. The available one-step checks do not substitute for a separately executed finite-precision rollout comparison.

Table 4: Frozen physical-orbit diagnostic on 128 confirmation states and all 40 cyclic shifts. MSE columns are multiplied by 10^4; $L_{\mathrm{orb}} = L_R + V_A$ up to a maximum absolute residual of 3.1×10^{-18}. The checkpoints, normalizers, source hash, and per-sample arrays are archived in `orbit_results/orbit_manifest.json`.

Seed	Arm	direct	L_{orb}	L_R	V_A
0	G	0.730	0.730	0.730	0.000
0	P	2.240	2.247	0.972	1.275
1	G	0.538	0.538	0.538	0.000
1	P	2.262	2.251	0.706	1.545
2	G	0.423	0.423	0.423	0.000
2	P	1.991	1.900	0.586	1.314

Table 5: Baseline Lorenz–96/CNN family, with normalizers estimated from 16 training trajectories. Each cell is the mean of three seed-level means over 128 shared confirmation trajectories. E_h uses the fixed physical reference scale; VPT is in saved steps. R averages the frozen P checkpoint over 40 translations at every step. All baseline rollouts remained finite through 512 steps.

Arm	$E_8 \downarrow$	$E_{32} \downarrow$	$E_{128} \downarrow$	$E_{512} \downarrow$	VPT↑
G	0.0247	0.2576	1.3806	1.4035	45.26
P	0.0822	0.5910	1.3892	1.4129	31.57
R	0.0342	0.3678	1.3880	1.4195	39.97

B.5 Empirical risk scales are not interchangeable

The group-average identity in Proposition 5 is a one-step or orbit-averaged statement. In the audited Lorenz–96 records, the mean physical one-step MSE difference between the frozen position-wise predictor and the global predictor is 0.000147 on 2,048 correlated state–target pairs per seed. The same records give a mean horizon-32 rollout difference of 0.2831, defined as the average across the three seeds of $E_{32}(\mathrm{P})^2 - E_{32}(\mathrm{G})^2$, where each E_{32} is the mean per-trajectory RMSE. These quantities use the common physical scale but measure different objects: the first is one-step/orbit risk, whereas the second is the recursively composed rollout. Since $(\mathcal{R}F)^h$ generally differs from $\mathcal{R}(F^h)$, the one-step decomposition cannot be reported as a rollout decomposition and is not the risk in Equation (36). The archived records do not contain the per-state rollout tensors needed to estimate a rollout-level projection residual. These values are descriptive audit quantities, not a causal partition of forecast error.

As a separate frozen-checkpoint audit, we evaluated all 40 physical shifts on 128 confirmation states for each of the three baseline seeds. Table 4 reports the direct one-step MSE, the orbit-averaged loss L_{orb}, the group-averaged prediction loss L_R, and the removed orbit variance V_A in units of 10^{-4} (all on the physical reference scale). The identity residual is below 3.1×10^{-18} in every row. For the position-wise predictor, L_R is lower than the direct loss in every seed; the global predictor has zero orbit variance up to floating-point precision. This diagnostic is one-step evidence and does not replace the measured autoregressive rollout comparison.

C Experimental details and evidence definitions

C.1 Normalization, initialization, and model

For a statistics pool $X \in \mathbb{R}^{n \times 521 \times 40}$, the position-wise normalizer uses

$$\widehat{\mu}_j = \frac{1}{521n} \sum_{i,t} X_{i,t,j}, \qquad \widehat{\sigma}_j = \sqrt{\frac{1}{521n} \sum_{i,t} (X_{i,t,j} - \widehat{\mu}_j)^2 + 10^{-6}}. \tag{45}$$

The global mean and standard deviation use all three axes, with the same additive 10^{-6} stabilization. The baseline estimates statistics from its 16 training trajectories. The pool intervention estimates them from the first n trajectories of an independent 256-trajectory pool, with reductions computed

Table 6: Label-fixed external statistics-pool intervention. Each entry averages three paired training/pool realizations evaluated on the same 128 confirmation trajectories. n counts trajectories used to estimate the normalizer; the four-trajectory pool is nested within the 256-trajectory pool per realization. Δ_n is $E_{32}(\mathrm{P}_n) - E_{32}(\mathrm{G}_n)$, computed before rounding.

	$E_{32} \downarrow$			VPT↑ (steps)	
n	G_n	P_n	Δ_n	G_n	P_n
4	0.2546	0.8099	0.5553	45.28	23.97
256	0.2399	0.3108	0.0709	46.20	42.05

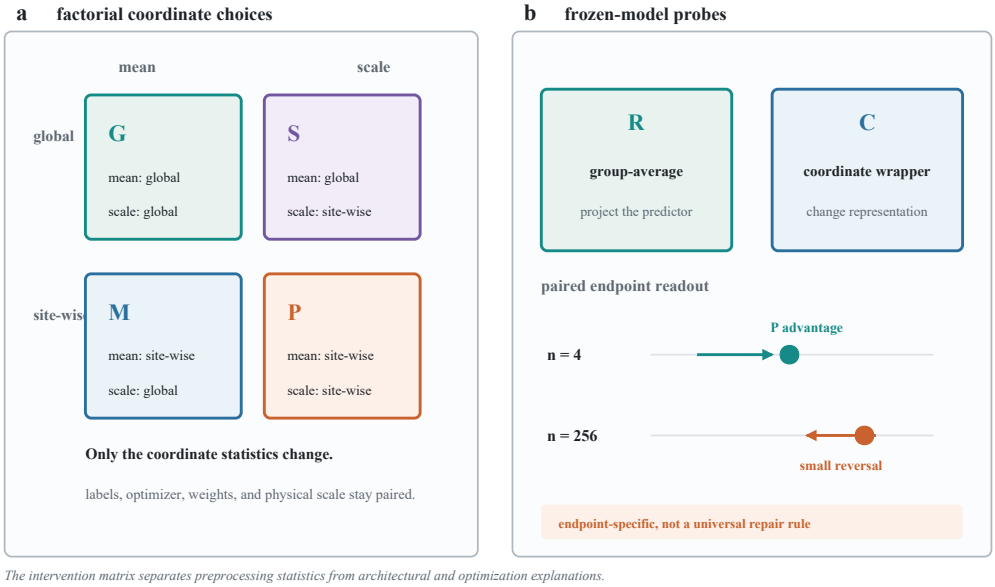


Figure 3: Schematic of the L96 interventions. Panel (a) classifies G/M/S/P by whether centering and scaling use shared or site-wise statistics. Panel (b) distinguishes physical averaging of a frozen P predictor (R) from expressing a frozen G predictor in P coordinates (C). Its lower arrows indicate only the direction of the POS–SHAM ranking under P at the two pool sizes; they encode neither error magnitude nor uncertainty. The measured ranking is in Table 1, the G/P pool values in Table 6, and frozen-control values in Tables 4 and 15.

in float64 and stored statistics used in float32. Its loss and evaluation reference scale remains the baseline training-data global scale, even when G_n uses a different scale internally.

The circular CNN forms the five-position neighborhood at offsets $-2, -1, 0, 1, 2$, applies a shared affine map to 64 channels and SiLU, repeats a five-position shared map to 64 channels and SiLU, and applies a shared pointwise map to one channel. Its output is $f_\theta(z) = z + 0.05\, h_\theta(z)$. The second neighborhood expands the receptive field but does not introduce position-dependent parameters. All training arms within a realization start from the same parameter tensors. Initial functions in *physical* coordinates need not be identical across normalizers, because the coordinate maps differ. The paired design controls the initialization procedure rather than asserting equality of those physical functions.

Each learning-rate candidate in the original families is trained for the full 10,000-update budget. Baseline validation averages the trajectory RMSEs over the four validation trajectories and horizons 8 and 32. Pool-intervention validation instead pools squared error across validation trajectories at each horizon, then averages the two horizon-wise RMSEs. Neither procedure selects a time point from a training curve. The baseline implementation repeats training at the selected rate from the same initialization and batch sequence before saving its final checkpoint; the pool implementation saves the trained candidates and evaluates the selected saved candidate. Thus the learning procedure

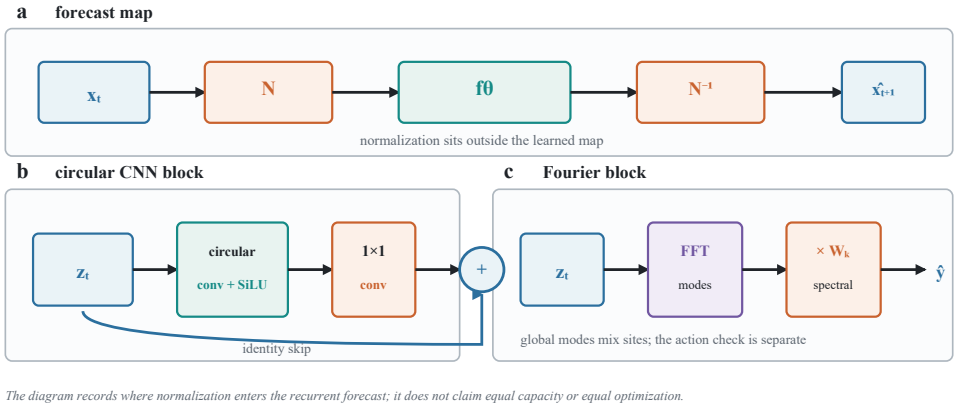


Figure 4: Implemented forecasting map. A fixed affine normalizer maps a 40-site physical state to the CNN input. Two shared five-site circular maps, SiLU activations, and a pointwise map produce $f_\theta(z) = z + 0.05h_\theta(z)$; N^{-1} returns to physical units. Shared weights and periodic boundaries imply ordinary normalized-array shift equivariance. Figure 1 explains why this action can differ from the physical one.

is specified separately for each family. The reported inference controls operate on frozen selected checkpoints.

The separate fixed-rate factorial uses fixed Adam learning rate 10^{-3} without validation selection. The same external pool file supplies the first 4 or 256 trajectories; the saved supervised training data, 4,096 starts, 10,000 batches, and initial parameter content match across pool sizes within seed and architecture. Normalizer means and scales are calculated in float64 and stored in float32, using the same 10^{-6} stabilizer. The four-arm CNN training reinitializes the same weights before each arm; the two-arm FNO40 training likewise pairs weights within that architecture. FNO40 uses a pointwise lift, four width-32 spectral/pointwise blocks retaining all 21 real Fourier modes on 40 sites, GELU, and a pointwise head with a 0.05 residual step. Neither model receives spatial coordinates. The FNO and CNN parameter initializations are distinct.

The final recovery-audited extension is summarized in `evidence/records/dynamics/overnight_20260915/196_summary.json`. Its primary crossed CNN study has seven pool sizes, ten training/data seeds, and three pool seeds (30 paired cells per size); the crossed FNO40 study has four pool sizes, five training/data seeds, and the same three pool seeds (15 paired cells per size). The summary resolves historical retries by `retry_of`, records 354 complete logical pair jobs and 708 training arms, and reports zero missing, invalid, nonfinite, or unresolved selected records. Budget, learning-rate, augmentation, and width groups are retained as robustness factors rather than pooled into the primary curve.

C.2 Optimization coordinates and initialization

For one sample, write $r_j = f_\theta(N_a x)_j$ and $y'_j = (y_j - \mu_{a,j})/\sigma_{a,j}$. Its loss and output derivative are

$$\ell = \frac{1}{d}\sum_j \frac{\sigma_{a,j}^2}{s_{\mathrm{ref}}^2}(r_j - y'_j)^2, \qquad \frac{\partial \ell}{\partial r_j} = \frac{2\sigma_{a,j}^2}{d s_{\mathrm{ref}}^2}(r_j - y'_j). \tag{46}$$

Thus a common physical metric does not equalize curvature in normalized output coordinates. These factors pass through the network Jacobian; they are not independent learning rates for its shared parameters. The residual initialization also gives

$$\widehat{\Phi}_{a,\theta_0}(x) = x + 0.05 S_a h_{\theta_0}\left(S_a^{-1}(x - \mu_a)\right). \tag{47}$$

Identical parameter tensors therefore control initialization procedure, not the initial physical function across normalizers. Fixed learning rates remove arm-specific selection but do not resolve these

effects. The executed EQ/POS interaction in Section 5 tests architectural specificity with an identical initial function within each normalizer. Because POS acquires trainable position dependence after initialization, it still changes the hypothesis class and cannot alone assign a causal fraction to symmetry mismatch.

At 40 sites FNO40 retains every real Fourier mode. Its spectral linear block is a full-domain circular convolution parameterized in frequency space, without low-pass truncation. This is a fixed-grid architecture comparison, not a test of resolution transfer. Whether full bandwidth increases sensitivity to spatially varying normalization would require a controlled mode-truncation experiment.

C.3 Units of replication and provenance

Training data and initialization are indexed by seeds $0, 1, 2$. The external pool seeds are respectively $9132701, 9132702, 9132703$. The shared confirmation set uses seed 9132601 and shape $128 \times 521 \times 40$. The first 512 saved transition starts are eligible for training; using 521 states makes the required targets available. One-step confirmation diagnostics use time indices $0, 32, \ldots, 480$ and the immediately following targets. Rollout evaluations start at each trajectory's first saved state.

The three seed summaries are the units displayed for the original families. In the final crossed extension, training/data seeds and pool seeds are crossed as described above; they jointly vary training data, initialization, and pool draw. Confirmation trajectories are paired across arms and reused across seeds and experiment families. Pool sizes are nested within a realization. These dependencies preclude interpreting all arm–trajectory–seed combinations as independent repetitions. Reported mean changes are descriptive; we do not attach significance claims or independent-confirmation counts to them.

C.4 Bootstrap aggregation for the crossed extension

For each evaluated pool size, the primary crossed summary first forms one paired difference $D_{a,b,n} = E_{32,a,b,n}(\mathrm{P}) - E_{32,a,b,n}(\mathrm{G})$ for every training/data seed a and pool seed b. A bootstrap replicate samples the training/data-seed labels and pool-seed labels independently with replacement, takes their Cartesian product, and recomputes the mean of the resulting paired differences. The G and P values remain paired, and training/data variation is kept separate from pool-draw variation. The executed aggregator also draws 128 confirmation-trajectory indices with replacement and uses the same indices in every cell of that replicate. These are pointwise intervals for each pool size, not a jointly resampled confidence band across the nested pools. The reported interval is the percentile interval of these replicate means. The inspectable L96 summary stores the resulting point estimates, intervals, resampling scope, and the bootstrap settings ($2{,}000$ replicates with RNG seed 8001409). This crossed-factor interval is descriptive and is not a claim of independent confirmation or a test of a monotonic law. Only three external-pool seeds are observed: drawing three labels gives ten possible unordered multiplicity patterns, with all three draws equal in $1/9$ of replicates. Additional bootstrap repetitions do not increase this support; the nominal 95% label should not be read as established finite-sample coverage.

The accompanying supplementary files are stored in the `evidence/legacy_sources` directory next to the manuscript source. Family paths below are relative to that directory; paths starting with `evidence/` are relative to the bundle root. The file `FILE_MANIFEST.json` records the included files and their SHA-256 hashes. Numerical records support inspection of the reported summaries; reproducing training also requires the specified generators, configurations, and software environment. For each listed path, `seed*` or `formal_seed*` denotes precisely seeds $0, 1, 2$.

- Baseline training and selected checkpoints: `normalization_causal_v1/seed*/metrics.json`. The physical CNN and simulator are defined in `normalization_causal_v1/transport_sequence_linear_v1.py`. The file hash of `normalization_causal_v1/normalization_causal_v1.py` matches the executed audit-script hash recorded in all three training manifests. The later `v2` and `v3` variants have different hashes and are not treated as the executed baseline training source.
- Baseline and physical-average confirmation results: `normalization_reynolds_v1/seed*/reynolds_eval_corrected_audited.json`. Table 5 averages the

horizon_rmse values in its models.G, models.P, and models.R_P trajectory rows, respectively.

- Corrected, unshifted one-step comparison: normalization_reynolds_v1/audit_run/audited_one_step.json, specifically mse_common_global_sd2. The older one_step_P_MSE field in the Reynolds records contains the projected error; it is not used as the unprojected P error.
- Pool training: normalization_sample_intervention_v1/formal_seed*/metrics.json. Confirmation rows are in normalization_sample_intervention_v1/eval_seed0_v2/results.json, eval_seed1/results.json, and eval_seed2/results.json under the same family. The aggregate is recorded in normalization_sample_intervention_v1/aggregate_audit.json.
- True target-action diagnostic: evidence/target_commutator.json, regenerated by evidence/target_commutator.py from the archived confirmation states, normalizers, and Lorenz–96 update formula. The numerical update was checked against the immediately saved next states (maximum absolute difference zero in the archived float32 data). Table 8 averages its three seed rows at each pool size; the file retains all seed-level values.
- Fixed-rate factorial and architecture replication: paper_experiments_20260914/snapshot_20260914/runs/cnn_v3_s*_n*/metrics.json and fno_s*_n*/metrics.json under that family. Each arm's confirmation record is in the corresponding *_eval/results.json, keys mean_per_trajectory, mean_vpt, and nonfinite_count. Executed training and evaluation scripts are paper_experiments_20260914/l96_tasks.py and paper_experiments_20260914/evaluate.py; the CNN and FNO40 class definitions are archived separately. Table 9 and Figure 6 use only status=ok evaluations of checkpoints with completed_updates=10000 and failure=null. All 36 requested arm records meet these conditions.
- Final crossed-pool extension and recovery-aware audit: overnight_20260915/l96_summary.json, overnight_20260915/l96_recovery_audit.json, and overnight_20260915/RESULTS_SUMMARY_20260915.md. Table 7 uses the selected primary attempts and paired horizon-32 gaps from this summary; historical deployment failures are excluded by the recorded recovery policy.
- Coordinate control: normalization_causal_v1/eval_seed*/standalone_eval.json. The reported loss/gradient checks use controls; the trained output check uses trained_val64_C_vs_G_max_abs; the ordinary and conjugated commutators use models.C_from_G.normalized_shift7. This family contains no evaluated C rollout rows.
- Frozen physical-orbit diagnostic: orbit_results/orbit_manifest.json and the six seed directories under orbit_results/. The audit evaluates direct one-step risk, L_{orb}, L_R, V_A, identity residual, and physical equivariance on 128 confirmation states and all 40 shifts; these values are not used as rollout or causal decompositions.

The confirmation data hash recorded across these families is

fe94fe4aa089a665208d84775b2733c123a1eb3defd9cfc158221b12dc1888c5.

The baseline source hash is

3c9b112165285ac2335bfe458b46300b82f7b515b844e514cab5533712b38a37.

Per-run records additionally identify the normalizer, checkpoint, sampled starts, and batch sequence. These identifiers distinguish recorded runs; a recorded hash alone does not establish that an implementation is correct. The result families above are kept separate because they differ in training or inference protocol.

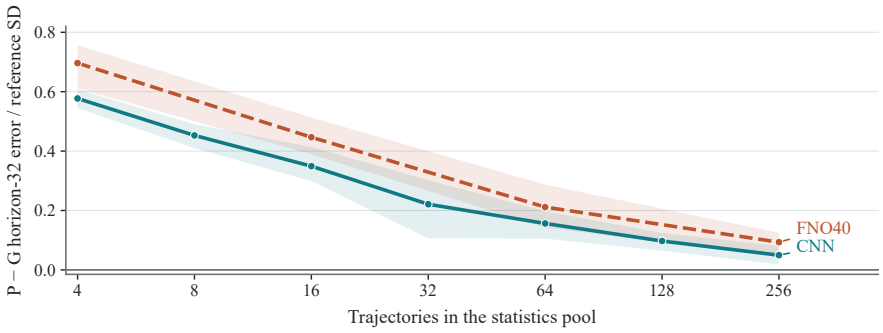


Figure 5: L96 paired horizon-32 error difference $E_{32}(P) - E_{32}(G)$ across the measured statistics-pool sizes. Points are crossed-cell means and light bands span the observed 10th–90th percentiles, not confidence intervals. CNN has seven measured pool sizes and FNO40 has four; lines connect only measured conditions. Errors use the shared physical reference scale. This base G/P study does not measure the intermediate-pool POS/SHAM ranking.

Table 7: Recovery-audited crossed-pool result. Entries are paired horizon-32 gaps $E_{32}(\mathrm{P}) - E_{32}(\mathrm{G})$; brackets give the crossed-factor 95% percentile interval. CNN uses 30 paired data/pool-seed cells per n; FNO40 uses 15. All selected attempts passed the recovery-aware audit.

Pool n	CNN (30 pairs)	FNO40 (15 pairs)
4	0.577 [0.537, 0.618]	0.696 [0.642, 0.751]
8	0.453 [0.409, 0.499]	—
16	0.349 [0.300, 0.404]	0.447 [0.394, 0.499]
32	0.221 [0.139, 0.294]	—
64	0.156 [0.118, 0.197]	0.212 [0.159, 0.275]
128	0.097 [0.072, 0.122]	—
256	0.050 [0.028, 0.071]	0.093 [0.055, 0.141]

D RESULTS FOR INDIVIDUAL REALIZATIONS

The KS experiment solves $u_t + uu_x + u_{xx} + u_{xxxx} = 0$ on the periodic domain $[0, 22)$, using 64 sites ($\Delta x = 22/64$), ETDRK4 step $1/32$, and a two-thirds Fourier cutoff at every nonlinear product. Initial fields are $\sum_{q=1}^{4} q^{-2}[a_q \sin(2\pi qx/22) + b_q \cos(2\pi qx/22)]$ with iid standard-normal coefficients, centered and normalized to unit spatial RMS. We discard 100 time units and save 257 consecutive states. The CNN has two width-64, five-site circular convolutions with GELU. The KS model archived as `fno40` has width 48 and retains all 33 real modes at 64 sites; both use a 0.1 residual step. These differ from the Lorenz–96 implementations. AdamW uses weight decay 0.01 and unscaled physical MSE. KS VPT uses raw RMSE threshold 2, so Lorenz–96's normalized threshold 0.5 does not diagnose failure here.

Table 12 retains all 48 formal arm records, including the M/S controls. The CNN G mean is strongly affected by seed 1 (raw $E_{32} = 1.528$ at $n = 4$); seeds 0 and 2 give 0.140 and 0.059. A nine-site receptive field, capacity, or optimization could explain its poor and variable fit, but no controlled KS architecture ablation identifies the cause. In particular, the negative average P–G gap is not evidence that P improves every realization. The source audit reports no nonfinite predictions through 256 steps; finite output is distinct from accurate forecasting. Solver, data, and evaluation hashes are retained in `overnight_20260915/ks_formal_audit.json`.

The group-average evaluator also compares a batched implementation against serial evaluation of the same 40 physical transformations, checks that averaging the already equivariant G predictor leaves its output unchanged up to finite precision, and tests the physical shift commutator of R. The orbit-loss decomposition and the independently evaluated unshifted one-step errors are distinct diagnostics. Tables 13–15 describe the recorded finite-sample runs; they do not establish a universal error ordering.

Table 8: Direct target-action diagnostic on the independent confirmation states, averaged over three realizations. Commutators are RMS in physical coordinates divided by s_{ref}. M varies only centering across sites; S varies only scale. These target-map calculations are distinct from trained-model forecasting in Table 9. The last column is the paired $\mathrm{P}_n - \mathrm{G}_n$ horizon-32 forecast-error difference from the original pool family. Four shifts and 2,048 correlated states are used per realization.

Pool n	G target	M target	S target	P target	$E_{32}(\mathrm{P}_n) - E_{32}(\mathrm{G}_n)$
4	5.74×10^{-8}	0.0268	0.0221	0.0316	0.555
256	6.09×10^{-8}	0.00372	0.00270	0.00417	0.071

Table 9: Summary of the separate fixed-10^{-3} protocol. Entries are mean E_{32} over three paired seeds on the shared physical-scale confirmation set; M/S vary site-wise mean/scale alone and FNO40 has only G/P. The records share the archived training data and initialization procedure with the corresponding factorial runs and are not an independent confirmation set.

Model	Pool n	G	M	S	P
CNN	4	0.236	0.737	0.623	0.795
CNN	256	0.240	0.306	0.273	0.311
FNO40	4	0.244	—	—	0.903
FNO40	256	0.254	—	—	0.357

E TWO-DIMENSIONAL SYSTEMS AND INTERACTION DETAILS

E.1 DATA GENERATION AND TRAINING

Allen–Cahn initial fields are stationary Gaussian Fourier polynomials with $|k_x|, |k_y| \leq 4$, ensemble RMS 0.5, and a stochastic constant mode. The reference solver alternates the exact heat and reaction subflows. Gray–Scott initial conditions contain a random periodic spot plus low-amplitude Fourier perturbations in both channels. Its reference solver uses exact Fourier half-steps for diffusion and RK4 for the reaction. Both reference solvers run in float64; arrays are stored in float32 only after time-refinement, grid, and translation checks pass. No simulated state is clipped or repaired.

Both domains are $[0, 1)^2$ with a 64×64 periodic grid. Allen–Cahn uses internal step 0.005 and saves every 0.05 time units; Gray–Scott uses internal step 0.2 and saves every 2 time units. Each saved interval therefore contains ten split steps. Trajectories start from the specified initial ensemble, without an additional burn-in. Allen–Cahn training, validation, and confirmation trajectories contain 129 saved states, and its calibration trajectories contain 32; the corresponding Gray–Scott counts are 33 and 9. Confirmation evaluates 128 steps for Allen–Cahn and 32 for Gray–Scott. All adjacent training-state pairs are available for one-step training, with batches of 32.

The base CNN uses four width-32 circular 3×3 hidden convolutions with GELU, followed by a 3×3 output convolution. The FNO uses width 24 and four spectral-plus-pointwise blocks with GELU, retaining 12×12 modes in each positive/negative frequency block, with 1×1 lifting and output layers. The models have one physical channel for Allen–Cahn and two for Gray–Scott. Their residual increments are multiplied by 0.05 and 1, respectively; zero-initialized output layers give the identity initial map. These are the base architectures; the separate POS/SHAM comparison adds the features specified in Appendix G.

Each system uses independently seeded sets of 16 training, 16 validation, 256 pool, and 64 confirmation trajectories. The three training seeds and three pool seeds are crossed, not collapsed into nine independent draws. Pool sizes are nested prefixes within a pool seed. G and P at a fixed cell share initialization, data, sample indices, and 100,000 pre-generated batches. The 100k endpoint jobs continue the 20k optimizer state and checkpoint. Thus the 216 records comprise 144 separately initialized base arms and 72 continuations. The all-outcomes rule was frozen before P results were inspected.

The complete audit stores one row per formal record, one row per matched pair, crossed and crossed-plus-trajectory bootstrap summaries, direction counts, failure rates, first failure steps, target-action results, and frozen physical subgroup results. Allen–Cahn and Gray–Scott each have 216/216 formal

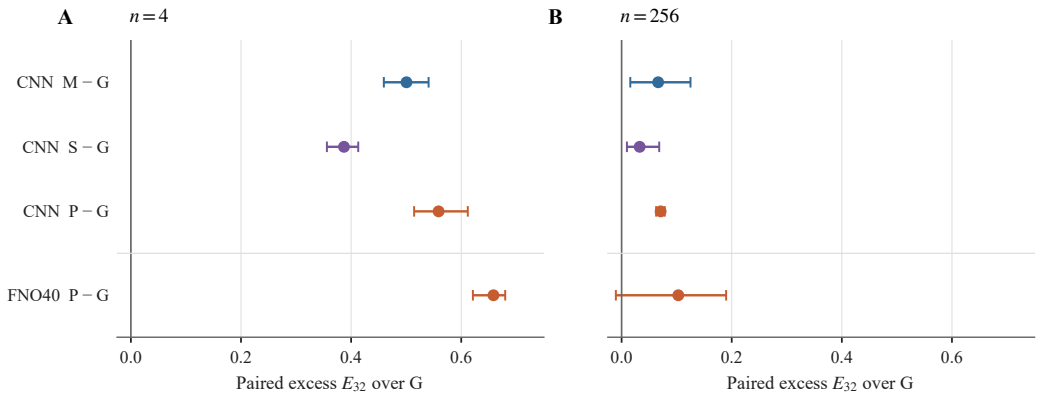


Figure 6: Separate fixed-rate protocol on the shared Lorenz–96 confirmation set. Panels show $n = 4$ and $n = 256$ on a shared horizontal scale. Each marker is the mean paired E_{32} difference from G across three training/pool seeds; its interval is the observed seed range, not a confidence interval. CNN M and S change only position-dependent centering and scaling, respectively. FNO40 is a second equivariant architecture on the same system. Both panels use the same physical error scale; the records share the archived data and initialization procedure with the corresponding factorial runs, so no independent confirmation count is implied.

Table 10: Individual horizon-32 errors in the separate fixed-rate family; all rows use 128 shared confirmation trajectories. The four CNN arms are paired within each seed and pool; FNO40 has only G/P arms. A dash means an arm was not trained, not a missing evaluation.

Model	Seed	Pool n	G	M	S	P
CNN	0	4	0.22218	0.76288	0.61429	0.83411
CNN	1	4	0.23691	0.69632	0.64983	0.75132
CNN	2	4	0.24911	0.75066	0.60513	0.79939
CNN	0	256	0.21546	0.27374	0.23588	0.29379
CNN	1	256	0.23674	0.36187	0.24643	0.30829
CNN	2	256	0.26755	0.28333	0.33587	0.33026
FNO40	0	4	0.23879	—	—	0.91335
FNO40	1	4	0.23004	—	—	0.85122
FNO40	2	4	0.26402	—	—	0.94387
FNO40	0	256	0.19093	—	—	0.32038
FNO40	1	256	0.22864	—	—	0.41849
FNO40	2	256	0.34128	—	—	0.33087

records and no G/P pairing-metadata error. The exact subgroup audit has 288/288 records; the refreshed target-action audit has 48/48. Source summaries and regenerated tables are cross-checked in the analysis manifest.

For Allen–Cahn, the reference standard deviation pools all training entries. For Gray–Scott, it pools trajectories, times, and spatial sites separately for each physical channel. Thus D_{ref} repeats one scalar in Allen–Cahn and two channel-specific scalars across the Gray–Scott grid. These fixed reference scales are shared by G/P; they are distinct from the intervened normalizer statistics. We first compute each trajectory's scaled RMS across channels and sites, then average trajectories; we do not pool squared errors across trajectories before taking the square root.

E.2 Failures and finite estimands

Allen–Cahn has 26 P-CNN formal runs with a nonfinite error before the full horizon: 17 among the 20k four-pool family and all nine $n = 256$ 100k continuations. At 20k, all nine $n = 16$ P runs, six $n = 64$ runs, and two $n = 256$ runs fail before horizon 128. Their E_{32} values are nevertheless finite and are retained in Table 16. At 100k and $n = 256$, three E_{32} values are nonfinite, so neither a

Table 11: KS formal boundary check. Each row averages three paired seeds at the indicated statistics pool. E_{32} is raw physical mean per-trajectory RMSE, without the Lorenz–96 reference-scale division; Δ is $E_{32}(\mathrm{P}) - E_{32}(\mathrm{G})$. Brackets give the observed minimum and maximum of the three paired seed differences, not a population confidence interval. For three observations the exact 95% percentile bootstrap equals this range, so resampling does not add information. The source audit records zero failed or nonfinite trajectories in every cell; first-failure step is therefore not applicable.

Architecture	n	$E_{32}(\mathrm{G})$	$E_{32}(\mathrm{P})$	Δ	seed range	failures
FNO64	4	0.0021	0.1506	0.1486	[0.1303, 0.1813]	0/3
FNO64	256	0.0020	0.0416	0.0397	[0.0341, 0.0466]	0/3
CNN	4	0.5758	0.2362	−0.3396	[−1.2959, 0.1524]	0/3
CNN	256	0.5803	0.6150	0.0347	[−0.6761, 0.4749]	0/3

Table 12: All KS seed-level E_{32} values in raw physical units, including position-wise mean-only (M) and scale-only (S) controls. FNO64 denotes the 64-site spectral model stored as `fno40`; it is not the 40-site Lorenz–96 model. Differences are computed before rounding.

Model	n	seed	G	M	S	P	Δ
CNN	4	0	0.14032	1.80201	0.21078	0.26489	0.12456
CNN	4	1	1.52760	0.18563	0.21716	0.23174	-1.29586
CNN	4	2	0.05943	0.40126	0.29638	0.21186	0.15243
CNN	256	0	0.16262	0.15243	0.32851	0.46784	0.30521
CNN	256	1	1.51711	0.73305	0.69258	0.84103	-0.67609
CNN	256	2	0.06121	0.06381	0.85763	0.53615	0.47494
FNO64	4	0	0.00225	0.13477	0.10848	0.13253	0.13028
FNO64	4	1	0.00194	0.14867	0.10333	0.13599	0.13405
FNO64	4	2	0.00209	0.12388	0.15303	0.18342	0.18133
FNO64	256	0	0.00171	0.04496	0.02305	0.04005	0.03834
FNO64	256	1	0.00203	0.05942	0.01957	0.04862	0.04659
FNO64	256	2	0.00214	0.05178	0.02046	0.03625	0.03411

nine-cell mean nor a crossed interval exists. Predictor-state nonfiniteness and earlier squared-error overflow are tracked separately. Gray–Scott has no nonfinite formal rollout. A finite rollout is an evaluability condition, not evidence of accurate forecasting.

E.3 Target action and frozen physical subgroup

For the target audit, G has normalized RMS defect below 2.2×10^{-15} in Allen–Cahn and below 9.8×10^{-15} in Gray–Scott. Table 17 therefore reports P. Gray–Scott small-pool values are conditional on the two finite shifts; each of the other six prescribed shifts yields a nonfinite reference evolution from the transformed physical input. The recorded failures are part of the result and preclude treating those rows as full eight-shift means. The archived status `partial_off_support` names this numerical outcome; it is not a support-membership test. Distinguishing solver instability from failure of the physical evolution would require a separate refinement check on these transformed inputs.

For each frozen model, the physical subgroup evaluator computes direct one-step risk, orbit-averaged squared risk, group-averaged risk, non-equivariant variance, and physical commutator norms. All prediction fractions equal one and the largest identity residual is 1.7053×10^{-13}, below the prespecified 10^{-10} audit tolerance. Small negative floating-point variance values in G are retained; none exceeds that tolerance in magnitude. Paired correlations between these diagnostics and E_{32} are exploratory: the nine cells share seed levels and a confirmation set, and pooling design cells would confound system, architecture, budget, and pool size.

E.4 Target defects and frozen orbit risk: detailed results

The target-map audit directly tests the coordinate claim. Allen–Cahn G is at numerical zero, while the P normalized-target RMS defect decreases from 0.3155 at $n = 4$ to 0.00646 at $n = 256$. Gray–Scott G is likewise at numerical zero. For Gray–Scott P at $n = 4, 16, 64$, however, only two of eight

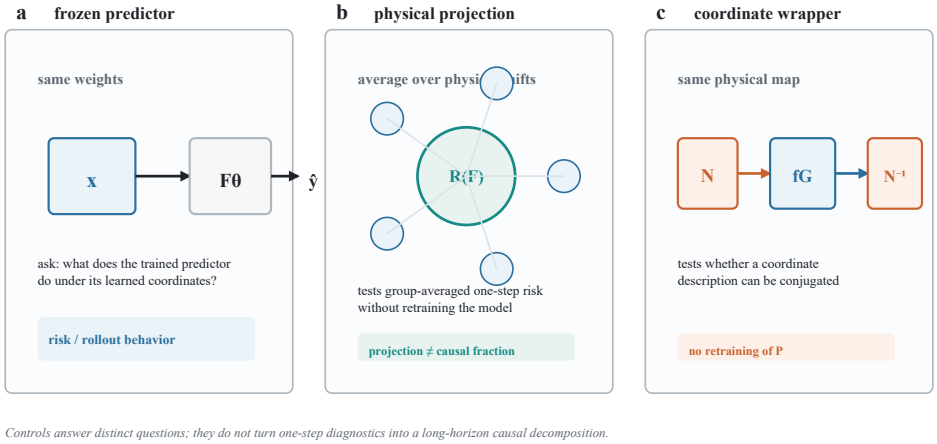


Figure 7: Conceptual map of three frozen-model questions. Panel (a) asks how the learned predictor behaves without changing its weights; panel (b) averages its physical predictions over translations to test one-step projection; panel (c) wraps the same physical map in another coordinate system to test action conjugation. These panels contain no numerical results. The measured one-step and orbit risks appear in Table 4, and the coordinate checks in Table 15; neither is a causal decomposition of rollout error.

Table 13: Baseline CNN confirmation results, separated by training/data seed. Each row summarizes the same 128 confirmation trajectories. R uses the corresponding frozen P checkpoint. VPT is measured in saved steps; all rows have zero nonfinite trajectories through 512 steps.

Seed	Arm	E_8	E_{32}	E_{128}	E_{512}	VPT
0	G	0.02715	0.27765	1.38061	1.39746	44.23
0	P	0.08078	0.56064	1.39388	1.40366	32.02
0	R	0.03574	0.35583	1.38497	1.41949	40.60
1	G	0.02391	0.22703	1.35518	1.40034	46.85
1	P	0.08829	0.62380	1.38385	1.40233	30.87
1	R	0.03513	0.37166	1.36380	1.42434	39.42
2	G	0.02317	0.26823	1.40591	1.41278	44.70
2	P	0.07761	0.58852	1.38980	1.43283	31.83
2	R	0.03161	0.37599	1.41531	1.41462	39.89

prescribed shifts produce finite reference evolutions from the transformed physical inputs; all six failed shifts are retained. At $n = 256$, all shifts are finite and the normalized defect is 0.115. The partial small-pool means therefore cannot be compared as complete eight-shift estimands. The finite evaluations measure action mismatch; the failures limit this diagnostic, without proving support exclusion or a forecast-risk bound.

The frozen predictor audit verifies the exact identity in all 288 records (maximum residual 1.71×10^{-13}). Relative to G, P has larger non-equivariant variance, group-averaged risk, orbit risk, and commutator RMS in all 144 paired cells. Group averaging removes 91.5%/85.2% of Gray–Scott FNO orbit risk at $n = 4/256$, versus 23.1%/10.2% for Allen–Cahn FNO. Yet Gray–Scott CNN has shares of 83.1%/85.1% while its 100k rollout gap does not recover. Equation (35) thus describes one-step predictor components, not a causal partition of autoregressive E_{32} (Appendix E).

E.5 Architecture Interactions

The original EQ/POS comparison uses

$$I_n = \Delta_{\mathrm{EQ},n} - \Delta_{\mathrm{POS},n}, \qquad J = I_4 - I_{256}. \tag{48}$$

Table 14: Individual paired statistics-pool realizations. Each seed identifies both a training realization and its associated external pool. Both G_n and P_n are retrained with the corresponding normalizer. The displayed gap is computed before rounding.

Seed	n	$E_{32}(\mathrm{G}_n)$	$E_{32}(\mathrm{P}_n)$	Δ_n	VPT(G_n)	VPT(P_n)
0	4	0.27788	0.83805	0.56017	44.20	23.11
0	256	0.21546	0.29379	0.07833	47.59	43.39
1	4	0.23691	0.79235	0.55545	46.18	24.44
1	256	0.23674	0.30829	0.07155	46.25	42.69
2	4	0.24911	0.79939	0.55028	45.47	24.36
2	256	0.26755	0.33026	0.06272	44.74	40.08

Table 15: Coordinate checks from the standalone evaluation family. All quantities are maximum absolute differences. Output checks use 64 validation states; the two normalized-coordinate commutators use shift 7 on the eight original test initial states. Their units are respectively physical and position-wise normalized state units.

Seed	Trained physical C vs. G	Ordinary commutator	Conjugated commutator
0	2.38×10^{-6}	0.07949	4.77×10^{-7}
1	1.91×10^{-6}	0.07560	4.77×10^{-7}
2	2.38×10^{-6}	0.06041	5.36×10^{-7}

The matched SHAM counterpart is defined in Equation (15).

L96 CNN. For this archived intervention, the training loss and reported errors use $s_j = \mathrm{std}_{i,t}(x^{\mathrm{train}}_{i,t,j}) + 10^{-6}$ fixed across G/P and EQ/POS, with $e_{i,h} = [d^{-1}\sum_j((\widehat{x}_{i,h,j} - x_{i,h,j})/s_j)^2]^{1/2}$. This is the per-site-reference case of Equation (14); interaction magnitudes are compared only within this matched protocol.

The corrected interaction contains 72/72 arm records and 18 complete four-arm units, evaluated on the same 128 confirmation trajectories. EQ/POS initial outputs agree within each normalizer, EQ commutators are at numerical precision, and no rollout failure is reported. An earlier 14-record partial launch used unmatched POS state-branch initialization; it was invalidated before aggregation and is not mixed with the corrected matrix. The corrected aggregate retains nine matched J units. Its intervals independently resample the three training-seed and three pool-seed levels, keeping a full crossed submatrix and reusing the same draws across contrasts.

The positive I_4 is not caused by a degraded G/POS arm: mean G is 0.1733 under EQ and 0.1747 under POS, while mean P falls from 0.7997 to 0.3977. At $n = 256$, G is 0.1713/0.1736 and P is 0.2290/0.2350 under EQ/POS. Position channels thus selectively help P only at the small pool in this L96/CNN experiment.

Gray–Scott archived EQ/POS comparison (CNN and FNO). This archived comparison contains 144/144 POS records and the corresponding 144 archived EQ records from `gray_scott_architecture_interaction_v1`, giving 72/72 complete four-arm cells; it is distinct from the new matched SHAM study in Appendix G. Each architecture, budget, and pool endpoint contains a crossed 3×3 grid. The POS models add four fixed periodic channels, $\sin(2\pi x)$, $\cos(2\pi x)$, $\sin(2\pi y)$, and $\cos(2\pi y)$. Their state branches copy EQ and their new incoming weights start at zero. All initialization checks, G/P and EQ/POS batch prefix checks, checkpoint hashes, and 72 POS continuation links pass; each 100k run continues the corresponding 20k checkpoint without a loss discontinuity.

All 288 EQ/POS records used here have finite E_{32} and zero rollout failures. The POS records are archived under `evidence_sync/20260918/gray_scott_architecture_interaction_v1/results/gray_scott_2d`; the matched EQ records are under `evidence_sync/20260918/backup_suite_v1/results/gray_scott_2d`. An independent reconstruction from their raw records reproduced all 72 paired cells, means, intervals, and direction counts exactly. Intervals independently resample the three training-seed and three pool-seed

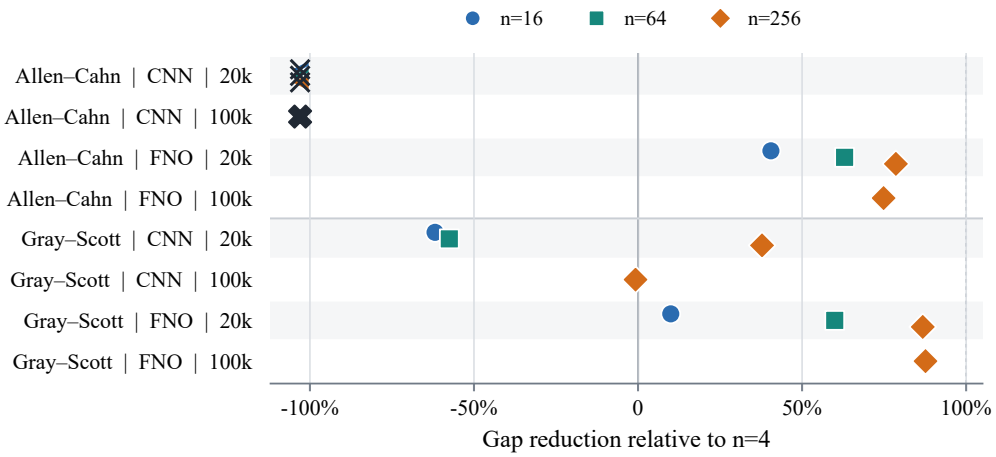


Figure 8: Pool response across the two-dimensional audits. Markers show the fraction of the matched $n = 4$ gap removed. Left-clipped triangles fall below -100%; thin crosses retain nonfinite P rollouts, and the filled cross marks the incomplete Allen–Cahn CNN 100k endpoint (3/9 nonfinite E_{32} values). Absolute gaps and crossed intervals are in Table 2.

Table 16: Complete 20k two-dimensional pool response. Δ_n is the mean paired $E_{32}(\mathrm{P}) - E_{32}(\mathrm{G})$ over nine crossed cells. The final column gives P runs with finite error through the system's full rollout; a finite E_{32} can therefore precede a horizon-128 failure. Large finite Allen–Cahn CNN values are retained; they are not evidence of a stable forecast. Reference standard deviations are global for Allen–Cahn and channel-wise for Gray–Scott, as specified in Equation (14).

System	Architecture	n	Δ_n	finite P runs
Allen–Cahn	CNN	4	1.144	9/9
		16	4.29×10^{9}	0/9
		64	2.35×10^{7}	3/9
		256	153.976	7/9
Allen–Cahn	FNO	4	0.331	9/9
		16	0.197	9/9
		64	0.123	9/9
		256	0.071	9/9
Gray–Scott	CNN	4	0.345	9/9
		16	0.559	9/9
		64	0.544	9/9
		256	0.215	9/9
Gray–Scott	FNO	4	0.787	9/9
		16	0.708	9/9
		64	0.315	9/9
		256	0.104	9/9

labels, retain the crossed grid, and use 20,000 draws with seed 20260918. They are descriptive conditional intervals for these fixed seed levels and shared confirmation set, not population confidence intervals or equivalence tests. The complete result supports a selective FNO effect at both budgets; CNN is uncertain at 20k and non-selective at 100k.

E.6 Frozen-backbone position and equivariant adapters

We reuse the 36 corrected EQ checkpoints at $n = 4, 256$ and cross three training seeds with three pool seeds, retaining their 128-trajectory confirmation set. Each G/P parent receives two zero-initialized adapters. POS supplies $\sqrt{2}\sin(2\pi j/40)$ and $\sqrt{2}\cos(2\pi j/40)$. SHAM supplies two relative shifts of the normalized input, divided by its training-only global RMS. SHAM therefore commutes with

Table 17: Normalized-target action defect for P. RMS values average three pool seeds on 32 shared confirmation states. "Finite shifts" is per seed out of eight prescribed ordinary normalized shifts. Gray–Scott values marked † condition on the two finite shifts and are not comparable eight-shift estimands.

System	n	normalized RMS defect	finite shifts
Allen–Cahn	4	0.31550	8/8
	16	0.03401	8/8
	64	0.01414	8/8
	256	0.00646	8/8
Gray–Scott	4	1.23563†	2/8
	16	0.65387†	2/8
	64	0.18337†	2/8
	256	0.11544	8/8

Table 18: Frozen one-step physical subgroup diagnostic at 20k updates. The variance share is V_H/L_{orb} in Equation (35); commutator values are physical RMS. Each entry averages nine matched trained models and uses all 64 translations on 32 shared trajectories.

		variance share		commutator RMS	
System	Architecture	$n = 4$	$n = 256$	$n = 4$	$n = 256$
Allen–Cahn	CNN	0.310	0.266	0.0182	0.00462
Allen–Cahn	FNO	0.231	0.102	0.0181	0.00409
Gray–Scott	CNN	0.831	0.851	0.0509	0.0124
Gray–Scott	FNO	0.915	0.852	0.116	0.00971

Table 19: L96/CNN normalization–architecture interaction. Each contrast uses nine matched training-seed × pool-seed cells. Intervals independently resample the two seed factors while reusing the same draws across contrasts. $I_n = \Delta_{\mathrm{EQ},n} - \Delta_{\mathrm{POS},n}$ and $J = I_4 - I_{256}$.

Contrast	Mean	crossed 95% interval	positive cells
$\Delta_{\mathrm{EQ},4}$	0.6264	[0.5781, 0.6684]	9/9
$\Delta_{\mathrm{POS},4}$	0.2230	[0.1605, 0.2740]	9/9
I_4	0.4034	[0.3071, 0.4759]	9/9
$\Delta_{\mathrm{EQ},256}$	0.0577	[0.0489, 0.0649]	9/9
$\Delta_{\mathrm{POS},256}$	0.0614	[0.0524, 0.0703]	9/9
I_{256}	−0.0037	[−0.0143, 0.0043]	4/9
J	0.4071	[0.3201, 0.4744]	9/9

ordinary cyclic shifts. Each adapter trains exactly 640 additional first-layer weights; all parent weights and biases stay fixed. Adam uses 10^{-3}, batches of 64, and 5,000 updates without validation selection. All four arms share sample indices within a crossed cell. Preflight verifies parent and data hashes, adapter-to-parent identity, frozen gradients, and normalized SHAM equivariance; postflight verifies unchanged parent tensors. All 72 fits and all rollouts through horizon 128 are finite. Four single-thread CPU workers run this small control separately from the GPU experiments.

Define $K_n = \Delta_{\mathrm{SHAM},n} - \Delta_{\mathrm{POS},n}$. Shared crossed resampling gives $K_4 = 0.0319$ with interval $[0.0095, 0.0542]$ and $K_{256} = 0.0071$ with $[-0.0026, 0.0162]$; the endpoint contrast is 0.0248 with $[0.0104, 0.0419]$ (5,000 bootstrap draws). These positive contrasts do not establish selective P repair. At $n = 4$, the absolute P/SHAM-minus-P/POS contrast is -0.0030 with $[-0.0183, 0.0107]$, whereas G/SHAM-minus-G/POS is -0.0348 with $[-0.0483, -0.0238]$. Relative to its parent, P/POS improves by only 0.0062 with $[-0.0047, 0.0181]$. At $n = 256$, SHAM improves both arms more than POS. The frozen criterion requiring absolute P improvement is therefore not met. A scalar-reference sensitivity gives $K_4 = 0.0317$ and the same interpretation. These are local adaptation outcomes; they do not negate the full-model intervention or isolate its optimization path.

Table 20: Gray–Scott normalization–architecture interaction. Entries are means with descriptive conditional crossed-factor 95% intervals. Each row uses nine matched training-seed × pool-seed cells. Positive arm improvement means lower error under POS; $I_n = \Delta_{\mathrm{EQ},n} - \Delta_{\mathrm{POS},n}$.

Updates	Arch.	n	Δ_{EQ}	Δ_{POS}	I_n	$I_n > 0$
20k	CNN	4	0.3453 [0.2935, 0.4036]	0.3255 [0.2734, 0.3717]	0.0198 [−0.0350, 0.0631]	5/9
20k	CNN	256	0.2148 [0.1948, 0.2342]	0.2136 [0.1930, 0.2284]	0.0012 [−0.0220, 0.0323]	5/9
20k	FNO	4	0.7869 [0.6637, 0.9434]	0.5514 [0.4599, 0.6782]	0.2355 [0.1892, 0.3050]	9/9
20k	FNO	256	0.1041 [0.0861, 0.1200]	0.1105 [0.0944, 0.1236]	−0.0065 [−0.0121, 0.0001]	2/9
100k	CNN	4	0.4280 [0.3202, 0.5330]	0.3022 [0.2664, 0.3498]	0.1257 [0.0335, 0.2395]	8/9
100k	CNN	256	0.4312 [0.2017, 0.9098]	0.2171 [0.1883, 0.2819]	0.2142 [0.0047, 0.6834]	7/9
100k	FNO	4	0.8548 [0.7526, 0.9834]	0.5204 [0.4409, 0.6451]	0.3344 [0.2746, 0.3887]	9/9
100k	FNO	256	0.1058 [0.0880, 0.1242]	0.1144 [0.0960, 0.1295]	−0.0086 [−0.0162, −0.0014]	1/9

Updates	Arch.	n	P improvement	G improvement
20k	CNN	4	0.0178 [−0.0367, 0.0633]	−0.0020 [−0.0035, 0.0003]
20k	CNN	256	−0.0020 [−0.0269, 0.0276]	−0.0032 [−0.0049, −0.0006]
20k	FNO	4	0.2343 [0.1880, 0.3038]	−0.0012 [−0.0017, −0.0003]
20k	FNO	256	−0.0074 [−0.0131, −0.0005]	−0.0009 [−0.0016, 0.0003]
100k	CNN	4	0.1262 [0.0241, 0.2457]	0.0004 [−0.0060, 0.0074]
100k	CNN	256	0.2139 [0.0013, 0.6852]	−0.0003 [−0.0064, 0.0078]
100k	FNO	4	0.3337 [0.2732, 0.3886]	−0.0007 [−0.0014, 0.0000]
100k	FNO	256	−0.0087 [−0.0165, −0.0011]	−0.0002 [−0.0009, 0.0007]

Updates	Arch.	$J = I_4 - I_{256}$	$J > 0$
20k	CNN	0.0186 [−0.0503, 0.0743]	5/9
20k	FNO	0.2420 [0.1913, 0.3142]	9/9
100k	CNN	−0.0884 [−0.6392, 0.2125]	5/9
100k	FNO	0.3430 [0.2809, 0.3958]	9/9

Table 21: Frozen-backbone L96 repair control: 72 fits, 640 trainable parameters per fit, and 5,000 fixed updates. Entries average nine crossed cells with fixed per-site reference scales from the parent protocol. POS adds absolute position; SHAM adds equally sized shift-covariant input features. $K = \Delta_{\mathrm{SHAM}} - \Delta_{\mathrm{POS}}$; brackets are descriptive crossed-factor 95% intervals. Lower arm errors are better; positive K alone does not establish P repair.

n	G/SHAM	P/SHAM	G/POS	P/POS	K [95% interval]
4	0.135	0.790	0.170	0.793	0.032 [0.010, 0.054]
256	0.136	0.202	0.169	0.229	0.007 [-0.003, 0.016]

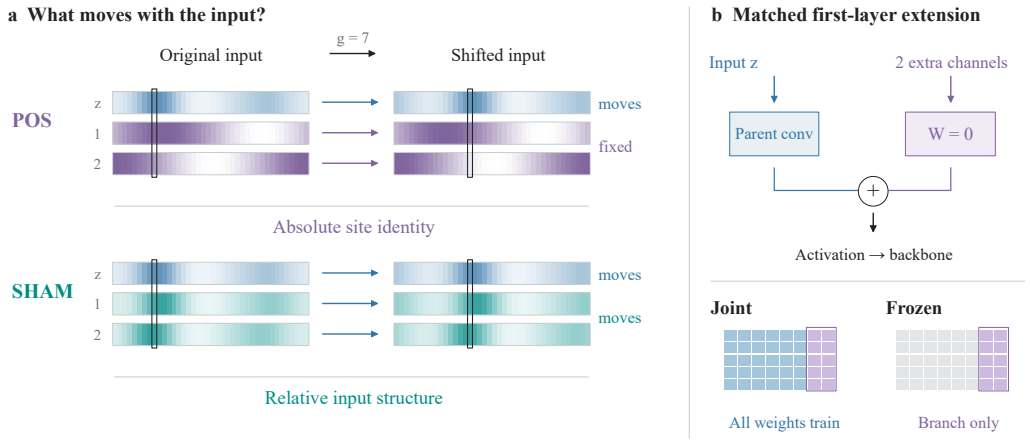


Figure 9: L96 position and equivariant control channels. (a) POS adds fixed sinusoidal site channels; SHAM adds relative shifts of the input that move with it. Intensities illustrate transformations, not performance. (b) A zero-initialized extra convolution preserves the parent predictor at initialization. Joint training updates all weights; frozen training updates only the added branch. Both L96 arms add 640 weights; the tiles indicate trainable blocks, not literal parameter counts. Matching counts and initial functions does not match subsequent optimization trajectories. The two-channel diagram describes L96, not the four-channel two-dimensional controls.

E.7 New L96 fixed-rate architecture confirmation

The new queue contains $5 \times 5 \times 2 \times 3 \times 2 = 300$ fits: training and pool seeds 61001–61005, statistics pools of 4 and 256 trajectories, normalizers G/M/P, and the POS/SHAM feature branches. Every fit reaches 100,000 Adam updates at learning rate 10^{-3} with batch size 64; all backbones change from initialization and all stored rollouts remain finite through horizon 128. Within each six-arm seed/pool/size cell, the source data, initial trainable parameters, sampled starts, and complete batch schedule match. The $n = 4$ pool is the prefix of the $n = 256$ pool for the same pool seed. G is re-estimated at each size. These fits contain no arm-specific validation choice or learning-rate search.

An audit of the generated arrays found an exact train/evaluation overlap: `train_61001.npy` has 28 trajectories, and each equals the first 513 time states of the same-index trajectory in `confirmation_61001.npy`. Their SHA-256 hashes match the job manifest. We checked initial-state hashes for every trajectory in all five training and five statistics-pool files against the original 128 evaluation trajectories and compared full common time prefixes for every hash match. Only these 28 training matches were found; no pool trajectory matched. Accordingly, every reported arm and seed uses only evaluation indices 28–127, leaving 100 common trajectories with no exact match to any training or pool trajectory. The 128-row summaries originally emitted by the runner are not used for the new manuscript estimates. The source arrays and audit are retained in the accompanying evidence; the discovery audit remains a separate experiment.

We first average each model's 100 held-out per-trajectory errors, then form within-cell SHAM-minus-POS differences. Each of 50,000 bootstrap replicates draws five training-seed labels and five pool-seed labels independently with replacement and averages their Cartesian product while preserving all arm pairings. The 95% and 98.75% percentile intervals condition on this shared holdout. Five levels per factor limit distributional inference, and the 25 crossed cells are not 25 independent training blocks. Table 22 gives all three reported horizons. The sign of $D_{\mathrm{P},4}$ is positive in 25/25 cells, while $D_{\mathrm{P},256}$ is negative in 22/25. The 98.75% intervals for $D_{\mathrm{G},4}$, $D_{\mathrm{P},4}$, and $D_{\mathrm{P},256}$ are $[-0.106, -0.051]$, $[0.345, 0.453]$, and $[-0.055, -0.008]$, respectively. The last upper endpoint does not clear the prespecified -0.01 practical threshold under this fixed-rate analysis. The matched validation-search results appear separately in Appendix E.8.

Table 22: New L96 fixed-rate confirmation after removing the 28 duplicated evaluation trajectories. Each value is the mean of 25 model-level means over the same 100 held-out trajectories. Columns retain the direct absolute errors at horizons 8, 32, and 128; the full 300 per-model records and 50,000-draw aggregation script are archived. Lower error is better.

	E_8		E_{32}		E_{128}	
Normalizer, n	POS	SHAM	POS	SHAM	POS	SHAM
G, 4	0.015	0.008	0.190	0.107	1.394	1.376
M, 4	0.027	0.126	0.303	0.738	1.398	1.405
P, 4	0.036	0.142	0.390	0.797	1.400	1.409
G, 256	0.015	0.009	0.191	0.113	1.392	1.373
M, 256	0.021	0.019	0.244	0.206	1.395	1.394
P, 256	0.023	0.021	0.260	0.228	1.397	1.393

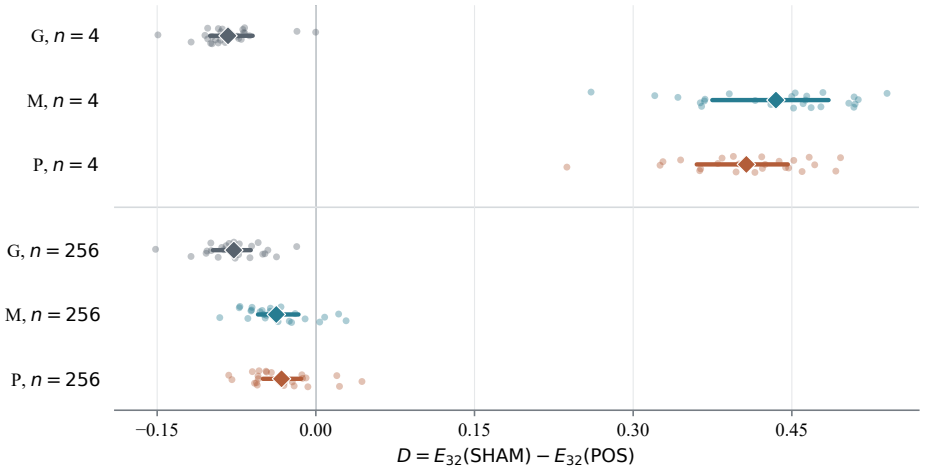


Figure 10: All 150 L96 SHAM-minus-POS paired E_{32} contrasts in the new fixed-rate confirmation. Each dot is one training-seed/pool-seed cell after averaging the same 100 held-out trajectories; there are 25 dots per row. Diamonds are crossed-cell means, bars are conditional 95% crossed-factor intervals, and the vertical line is zero. Positive values favor POS. The six rows share one horizontal scale; no trajectory is treated as an independent training replicate.

E.8 L96 EQUAL VALIDATION-SEARCH FOLLOW-UP

We evaluated the same 100,000-update EQ CNN protocol used by the fixed-rate confirmation. For each of five training seeds, five independently drawn pool seeds, G/P normalization, pool sizes $n \in \{4, 256\}$, and feature branches POS/SHAM, we compared Adam rates 3×10^{-4}, 10^{-3}, and 3×10^{-3}. The 10^{-3} candidate reuses the completed fixed-rate checkpoint; the other two rates are new fits. Thus the follow-up adds 400 fits, or 40 million optimizer updates, and reevaluates 200 existing checkpoints. All 600 result records are terminal, and every candidate reached 100,000 updates. The paired POS/SHAM fits share initial trainable tensors, sampled starts, batch indices, and input data. Both arms see the same three candidate rates and the same validation trajectories.

For each seed/pool/normalizer/size/arm cell, we selected the finite candidate with the lowest mean E_{32} on trajectories 28–127 of `confirmation_61002.npy`; ties favor the lower rate. The first 28 validation trajectories exactly repeat `train_61002.npy` and are excluded. We then evaluated the selected checkpoint on trajectories 28–127 of `confirmation_61001.npy`. Its first 28 trajectories repeat `train_61001.npy`; we excluded them as well. Initial-state checks found no match between the retained validation and test trajectories or with the calibration pools. The original fixed-rate results on this test set had already been inspected. This set was not used for the present learning-rate selection, but it was not a previously unseen benchmark.

Table 23: Locked POS/SHAM ranking on the previously inspected test split and the newly generated 128-trajectory set. Entries are paired $D = E_{32}(\mathrm{SHAM}) - E_{32}(\mathrm{POS})$ with 95% crossed-factor intervals; positive favors POS. The models and selected rates are identical across columns.

Norm., n	Earlier 100 trajectories	New 128 trajectories
G, 4	-0.088 $[-0.096, -0.079]$	-0.080 $[-0.087, -0.073]$
G, 256	-0.086 $[-0.093, -0.080]$	-0.080 $[-0.086, -0.074]$
P, 4	$+0.456$ $[+0.419, +0.493]$	$+0.467$ $[+0.436, +0.496]$
P, 256	-0.042 $[-0.051, -0.032]$	-0.040 $[-0.049, -0.030]$

The selected learning rates vary by condition. POS chose 3×10^{-3} in all 25 small-pool P cells, while SHAM chose 3×10^{-4} in 19 of them. Under G at $n = 4$, both branches most often selected 3×10^{-4} (20/25 POS and 16/25 SHAM). On the previously inspected 100-trajectory test split, the P contrast changes by 0.498 between pool sizes, with a 95% crossed-factor interval $[0.456, 0.542]$. POS wins in all 25 small-pool P cells and loses in 24 of 25 large-pool P cells. Table 23 compares these values with the later, independent-trajectory confirmation. Crossed-factor intervals condition on the selected rates and the common test trajectories; they resample five training and five pool seed levels while preserving the full Cartesian pairing. They do not treat 25 cells as independent models or quantify variation over new evaluation sets. The complete 600 per-model records, selected-cell table, validation array, analysis source, frozen manifest, and hashes accompany the submission source.

E.9 Fresh-trajectory confirmation with locked model selection

After locking the 200 selected checkpoints, we generated 128 new Lorenz–96 trajectories with RNG seed 71024, the same float32 RK4 simulator and save interval, and 100 physical time units of burn-in per trajectory. No new fit, learning-rate choice, threshold, or architecture selection used these trajectories. An exact initial-state audit checked all 15 train, pool, and confirmation files in the five-seed family and found zero matches. Each checkpoint was loaded at 100,000 updates and rolled out to horizons 8, 32, and 128 on the common new set; all 200 records are finite and complete.

On the new trajectories the P pool-size interaction is 0.507 $[0.477, 0.536]$ and positive in all 25 crossed cells. POS wins all 25 small-pool P cells; SHAM wins 24 of 25 large-pool P cells and all 25 cells under G at both sizes. Horizon-128 means lie between 1.36 and 1.40 across arms and conditions, so the ranking result remains an intermediate-horizon finding. Intervals resample five training and five pool seed levels separately with 50,000 draws and condition on the shared new trajectory set. The generating script, trajectories, initial-state audit, 200 per-checkpoint records, analysis script, and hashes accompany the source.

E.10 Calibration-specific architecture response on the fresh holdout

This is a descriptive reanalysis of the same 200 locked-checkpoint records and 128 test trajectories in Appendix E.9; it adds no fits, data, model selection, or independent replication. We form the four $D_{N,n}$ arrays over the 5×5 training/pool-seed grid, then compute $K_n = D_{\mathrm{P},n} - D_{\mathrm{G},n}$ and $J_{\mathrm{SHAM}} = K_4 - K_{256}$ within every crossed cell. All conditions use the same resampled training and pool indices in each of 50,000 bootstrap draws, with seed 20260924. Thus the intervals preserve the shared-factor pairing across normalizers, architectures, and pool sizes. They condition on the selected rates and shared test trajectories.

The residual positive K_{256} must not be interpreted as a large-pool POS win: $D_{\mathrm{P},256} < 0$ on average. It means that POS is less disadvantaged relative to SHAM under P than under G. Reporting the direct contrasts prevents an interaction alone from being described as a repair. Likewise, $J_{\mathrm{SHAM}} > 0$ does not identify how much of the response is caused by expression constraints rather than optimization. Equation (72) has that exact residual interpretation only in its specified regression model.

For completeness, the modal learning rates and counts for POS/SHAM are: 3×10^{-4} (20/16 of 25) under G at $n = 4$; 3×10^{-4} (17/19) under G at $n = 256$; $3 \times 10^{-3}/3 \times 10^{-4}$ (25/19) under P at $n = 4$; and 3×10^{-4} (17/19) under P at $n = 256$. These are the original validation selections; the reanalysis changes none. The analysis code, per-cell contrasts, source records, and SHA-256 manifest are provided in `analysis/` of this revision.

Table 24: Contrasts from the existing fresh L96 evaluation. K_n measures how the POS advantage differs between P and G; J_{SHAM} measures how this difference changes with calibration size. Positive-cell counts describe the crossed grid, not independent replications.

Contrast	Mean	95% crossed-factor interval	Positive cells
$D_{\mathrm{P},4} - D_{\mathrm{P},256}$	0.50684	$[0.47675, 0.53573]$	25/25
$D_{\mathrm{G},4} - D_{\mathrm{G},256}$	0.00051	$[-0.00410, 0.00467]$	12/25
K_4	0.54631	$[0.51341, 0.57625]$	25/25
K_{256}	0.03998	$[0.02876, 0.05256]$	25/25
J_{SHAM}	0.50633	$[0.47561, 0.53554]$	25/25

Table 25: Deterministic metric-robustness bounds for the same 200 locked models and 128 fresh L96 trajectories as Table 1. Each interval encloses the mean SHAM-minus-POS horizon-32 error under raw physical RMSE or RMSE divided by the training-seed global standard deviation. These are norm bounds, not confidence intervals or recomputed exact errors. Positive bounds favor POS. All five restored training arrays match their original recorded byte hashes.

Statistics	Pool n	Raw D bounds	Global-scale D bounds	Certified lower error
Global	4	$[-0.315, -0.266]$	$[-0.087, -0.073]$	SHAM
Global	256	$[-0.316, -0.268]$	$[-0.087, -0.073]$	SHAM
Site-wise	4	$[+1.586, +1.805]$	$[+0.435, +0.496]$	POS
Site-wise	256	$[-0.191, -0.102]$	$[-0.053, -0.028]$	SHAM

E.11 Metric, learning-rate, and architecture-decision sensitivity

These analyses use existing records after the reported outcomes were available. No models were retrained, no checkpoints or rates were reselected using test data, and no additional independent replication is claimed. We distinguish the new 128-trajectory confirmation set (200 locked models) from the earlier inspected 100-trajectory holdout (all 600 rate candidates).

A metric certificate for the fresh results. The fresh records retain per-trajectory site-scaled errors, not per-site physical prediction residuals. They cannot be converted into exact raw errors. Nevertheless, the original reference scales give rigorous bounds. For training seed t, let $r_{tj} > 0$ be its reference standard deviations, $a_t = \min_j r_{tj}$, $b_t = \max_j r_{tj}$, and $g_t > 0$ the scalar training standard deviation, with the evaluator's 10^{-6} offset. For every prediction residual e,

$$a_t \frac{\|\operatorname{diag}(r_t)^{-1} e\|_2}{\sqrt{d}} \le \frac{\|e\|_2}{\sqrt{d}} \le b_t \frac{\|\operatorname{diag}(r_t)^{-1} e\|_2}{\sqrt{d}}. \tag{49}$$

The inequality is preserved by averaging trajectories. Thus, within each paired training/pool-seed cell,

$$a_t E_{\mathrm{SHAM}} - b_t E_{\mathrm{POS}} \le D_{\mathrm{raw}} \le b_t E_{\mathrm{SHAM}} - a_t E_{\mathrm{POS}}. \tag{50}$$

Divide each cell's endpoints by g_t for globally scaled RMSE. Averaging endpoints over the 25 crossed cells gives the aggregate enclosures in Table 25. These are deterministic bounds on the observed errors, not uncertainty intervals for a population mean. Outward rounding preserves containment in the displayed table.

We restored the five original training arrays from their archived seeds and generator settings and required exact byte SHA-256 agreement with their recorded pre-evaluation hashes before using any scale. All five pass. The restored arrays contain no new random draws relative to the original data. The evidence includes the reference vectors, global scales, data and record hashes, per-cell bounds, and the restoration script. Every aggregate bound has a strict sign, including the small large-pool reversal. Each of the five training-seed means also has that sign. At the individual-cell level all G and small-pool P contrasts are certified; 24 of 25 large-pool P cells certify SHAM, while training seed 61003/pool seed 61005 is inconclusive. That cell already slightly favored POS under the original metric. The certificate therefore establishes the four mean rankings, not unanimity in the large-pool condition.

Table 26: Post hoc sensitivity within the existing three-rate grid. Entries are mean $D = E_{32}(\mathrm{SHAM}) - E_{32}(\mathrm{POS})$ on the earlier 100-trajectory holdout, paired over five training seeds crossed with five pool seeds. Every small-pool site-wise comparison is positive in all 25 cells. These reused outcomes neither establish convergence nor cover rates outside the grid.

POS rate	SHAM rate	P, $n = 4$	P, $n = 256$	$D_4 - D_{256}$
3×10^{-4}	3×10^{-4}	$+0.231$	-0.041	$+0.272$
3×10^{-4}	10^{-3}	$+0.239$	-0.020	$+0.260$
3×10^{-4}	3×10^{-3}	$+0.251$	-0.020	$+0.271$
10^{-3}	3×10^{-4}	$+0.398$	-0.053	$+0.451$
10^{-3}	10^{-3}	$+0.407$	-0.032	$+0.439$
10^{-3}	3×10^{-3}	$+0.419$	-0.032	$+0.450$
3×10^{-3}	3×10^{-4}	$+0.457$	-0.053	$+0.510$
3×10^{-3}	10^{-3}	$+0.466$	-0.032	$+0.498$
3×10^{-3}	3×10^{-3}	$+0.477$	-0.032	$+0.509$

All completed rate pairings. Table 26 crosses all three POS rates with all three SHAM rates at the common 100,000-update budget. It uses trajectories 28–127 of the earlier holdout, excluding the 28 overlapping training trajectories identically for every arm. Each entry averages the same 5×5 crossed seed grid. Under small-pool P, every pairing favors POS in every cell: the nine mean contrasts range from 0.23076 to 0.47725, and the smallest individual contrast is 0.11462. These are 225 correlated comparisons of reused fits, not 225 independent experiments. Under large-pool P, every pairing has a negative mean. Holding each arm's rate fixed across pool sizes, all nine $D_4 - D_{256}$ contrasts are positive in every cell. A conservative descriptive check also subtracts each cell's worst POS candidate from its best SHAM candidate; its mean remains positive at 0.22683. This envelope is not a test-selected deployment rule.

Within-grid robustness does not remove boundary uncertainty. At small-pool P, POS chooses 3×10^{-3} in 25/25 cells, and SHAM chooses 3×10^{-4} in 19/25. No saved candidate lies outside these rates, and equal update counts do not prove convergence. The conclusion is that the ranking change persists throughout the completed grid, not that the global optimum of either architecture has been found. Exact scalar-scale errors on the earlier disjoint holdout also preserve the four mean rankings (Table 28); those values are not substituted for the fresh-test certificate.

Replay of a validation-only architecture decision. For crossed cell c and condition (N, n), let $\ell_{A,c,N,n}$ be arm A's original validation-selected rate. Define

$$\widehat{A}_{c,N,n} = \arg\min_{A\in\{\mathrm{POS,SHAM}\}} E_{32}^{\mathrm{val}}(A, \ell_{A,c,N,n}; c, N, n). \tag{51}$$

There are no ties. The original validation values select POS for all 25 small-pool P cells and SHAM for all 25 cells in each other condition. To make the cost of carrying that decision explicit, we keep the arm label $\widehat{A}_{c,\mathrm{P},4}$ and evaluate its original condition-specific selected checkpoint at a target (N, n). We compare its stored fresh error with the fresh error of $\widehat{A}_{c,N,n}$. This carries an architecture decision, not a trained checkpoint across incompatible coordinates. Table 27 reports the resulting excess error. Fresh outcomes enter neither rate nor arm selection.

This replay was specified after the outcomes were available. It is a descriptive consequence of the controlled results, not a new prospective validation study or an independent source of statistical information. Its purpose is to identify the affected research decision: whether to retain position inputs after an ordinary validation-based ablation. Five training and five pool levels, reused test trajectories, and selected rates remain the statistical limits. The uncertainty intervals in the replay and earlier-holdout tables use 50,000 crossed-factor draws with seed 20260924, sharing resampled factor indices across compared conditions.

Allowing calibration choices in the same workflow. Equation (51) treats the normalizer as fixed. To assess the consequence of that choice, retain each arm's original validation-selected rate and define, at the same pool size,

$$(\widehat{N}_{c,n}, \widehat{A}_{c,n}) = \arg\min_{N\in\{G,P\},\, A\in\{\mathrm{POS,SHAM}\}} E_{32}^{\mathrm{val}}(A, \ell_{A,c,N,n}; c, N, n). \tag{52}$$

Table 27: Retrospective replay of validation-only architecture choices. Each arm retains its original validation-selected checkpoint. The replay chooses the arm with lower original validation error, then reuses the stored fresh 128-trajectory confirmation errors for evaluation. All 25 P, $n = 4$ decisions choose POS. Excess error compares carrying that arm label to a target condition with choosing the arm on the target validation data. This rule was analyzed after outcomes were available; the table is not a new prospective confirmation. Intervals resample the crossed seed factors.

Target	Changes	SHAM choices	Excess E_{32} [95% interval]
G, $n = 4$	25/25	25/25	$+0.080$ $[+0.073, +0.087]$
G, $n = 256$	25/25	25/25	$+0.080$ $[+0.074, +0.086]$
P, $n = 256$	25/25	25/25	$+0.040$ $[+0.030, +0.049]$

Table 28: Metric sensitivity on the earlier disjoint 100-trajectory holdout for the original validation-selected checkpoints. The scalar denominator is the training-set global standard deviation plus 10^{-6}. Both metrics exclude trajectories 0–27. Exact scalar errors are unavailable in the fresh 128-trajectory confirmation records; the rigorous certificate in Table 25 applies to that set.

Condition	Site-scaled D	Scalar-scaled D [95% interval]
G, $n = 4$	-0.088	-0.088 $[-0.096, -0.079]$
G, $n = 256$	-0.086	-0.086 $[-0.093, -0.080]$
P, $n = 4$	$+0.456$	$+0.456$ $[+0.419, +0.493]$
P, $n = 256$	-0.042	-0.042 $[-0.051, -0.032]$

This is equivalent to minimizing validation error over all twelve normalizer/architecture/rate candidates; the reconstruction checks all 600 candidate records and finds no ties. Joint selection chooses G+SHAM in every cell at both pool sizes, exactly matching G-only selection. Relative to P-only selection, the fresh-test improvement is positive in all 25 cells at each size (Table 29). The G-only and P-only searches each have six candidates, whereas their joint comparison has twelve. Both normalizers use the same calibration trajectories and supervised examples. No new training, validation, or fresh-test data enter this replay.

This comparison identifies a practical consequence beyond a reversal between two P pools: a large POS gain can recommend an architectural feature within a pipeline that loses to the alternative calibration protocol. It does not show that global statistics always win, or that the gain is entirely a correction for symmetry mismatch. The replay is retrospective, conditioned on this search grid, seed factors, and test set. Its crossed-factor intervals use the same 50,000-draw protocol as above. A new public-benchmark replication would be needed to establish the prevalence of this decision consequence outside the controlled setting.

E.12 JOINT L96 POS/SHAM CONTROL

The earlier discovery control trains the full L96 CNN from the same untrained EQ initialization for every POS/SHAM pair. POS adds the fixed sine/cosine site channels; SHAM adds two normalized relative shifts of the input. Both branches have 21,633 trainable parameters, including 640 branch parameters, and use the same 100,000-update Adam schedule, batches, starts, and loss scale. The design crosses three training seeds with three pool seeds at $n \in \{4, 256\}$, giving 72 fits, and evaluates 128 shared confirmation trajectories at horizons $1, 8, 32, 128$. No record is an evaluation-only continuation, and all 72 fits remain finite through horizon 128.

The primary horizon-32 estimand is $K_n = \Delta_{\mathrm{SHAM},n} - \Delta_{\mathrm{POS},n}$, with $J = K_4 - K_{256}$. The original frozen summary and an independent reconstruction from all 9,216 trajectory rows agree for 285 means, interval bounds, and finite-cell counts to 2.22×10^{-16}. The crossed bootstrap draws three training labels and three pool labels with replacement for each of 5,000 replicates, retaining their Cartesian product. At $n = 4$, $K_4 = 0.4928$ $[0.4217, 0.5671]$ and $P_{\mathrm{SHAM}} - P_{\mathrm{POS}} = 0.4303$ $[0.3555, 0.5061]$; at $n = 256$, $K_{256} = 0.0359$ $[0.0262, 0.0497]$ while the P contrast is -0.0337 $[-0.0423, -0.0217]$. The endpoint interaction is $J = 0.4569$ $[0.3791, 0.5317]$. All nine paired endpoint differences have the reported direction, but the nine seed pairs are not population replications. The positive

Table 29: Retrospective validation-only pipeline decisions. Every entry uses 25 crossed training/pool-seed cells and the stored fresh 128-trajectory readout. P-only validation chooses POS at $n = 4$ and SHAM at $n = 256$, each in all 25 cells. G-only and joint G/P validation both choose G+SHAM in every cell. Improvement is P-only error minus joint-choice error; intervals use paired crossed-factor resampling. Relative reductions are descriptive ratios of means. Joint search considers 12 configurations per cell versus 6 for a fixed normalizer; this replay is not a new prospective confirmation.

n	P-only: E_{32}	G-only/joint: E_{32}	Improvement [95% CI]	Reduction
4	POS: 0.330	G+SHAM: 0.084	0.246 [0.227, 0.265]	74.5%
256	SHAM: 0.208	G+SHAM: 0.083	0.125 [0.102, 0.144]	60.1%

Table 30: Existing L96 joint audit: absolute E_{32} and direct architecture comparison. Positive D favors POS. Brackets are conditional 95% crossed-seed intervals, not independent confirmation; all nine cells agree with each row's sign.

Normalization, pool	POS	SHAM	$D = \text{SHAM} - \text{POS}$ [95% interval]
G, 4	0.1756	0.1131	-0.0625 [-0.0766, -0.0486]
P, 4	0.3750	0.8053	$+0.4303$ [$+0.3555$, $+0.5061$]
G, 256	0.1742	0.1045	-0.0697 [-0.0818, -0.0502]
P, 256	0.2319	0.1982	-0.0337 [-0.0423, -0.0217]

K_{256} is compatible with POS worsening G more than P, so it is not evidence of large-pool P repair. Horizon-128 J spans zero and is treated as a secondary sensitivity.

E.13 Excluded NS2D records

The archived NS2D runner accepted a `pool_seed` argument but did not use it when generating the statistics pool. Nominal pool-seed cells therefore reuse one effective pool and cannot estimate the intended crossed factor. We exclude NS2D from every cross-system claim, table, and count. Reaggregation cannot repair this design error; admissible evidence requires a corrected rerun with distinct pool hashes.

F Distributional overlap and the four-site obstruction

The pointwise obstruction in Section 3 does not by itself quantify population risk. We examine this distinction with an exactly specified regression problem, using the usual squared-loss projection onto equivariant functions. Averaging and projection arguments are established tools in equivariant learning (Elesedy & Zaidi, 2021); the calculation here applies a density-weighted orbit projection to our centering example. It is an analytical control, not a new forecasting algorithm.

Let $x \sim \mathcal{N}(0, I_4)$ and let the physical target be Tx, where T is the one-site cyclic shift. For $\mu = a(0, 1, 0, -1)^\top$, define $z = x - \mu$, $q(z) = \mathcal{N}(z; -\mu, I_4)$, and $b = T\mu - \mu$. The normalized target is $y(z) = Tz + b$. The physical input law and physical labels are fixed as a varies; μ is a controlled calibration perturbation rather than an estimated finite-pool statistic. For the four-element cyclic group, set

$$w_g(z) = \frac{q(T_g z)}{\sum_h q(T_h z)}, \qquad a_g(z) = T_g^{-1} y(T_g z). \tag{53}$$

Every ordinarily equivariant f has $T_g^{-1} f(T_g z) = f(z)$. Changing variables by each orthogonal permutation in its population squared risk therefore gives

$$R_q(f) = \int \overline{q}(z) \sum_g w_g(z) \frac{\|f(z) - a_g(z)\|_2^2}{4}\, dz, \quad \overline{q}(z) = \frac{1}{4} \sum_g q(T_g z). \tag{54}$$

Pointwise weighted least squares yields an equivariant minimizer

$$f^*(z) = \sum_g w_g(z) a_g(z) = Tz + \sum_g w_g(z) T_g^{-1} b. \tag{55}$$

Table 31: L96 joint POS/SHAM control at horizon 32. Each entry is a mean with a 95% crossed-factor interval over three training and three pool seeds; positive $P_{\mathrm{SHAM}} - P_{\mathrm{POS}}$ favors the position-aware branch. $K = \Delta_{\mathrm{SHAM}} - \Delta_{\mathrm{POS}}$ and $J = K_4 - K_{256}$.

	$n = 4$	$n = 256$
Δ_{POS}	$0.1994\ [0.1540, 0.2411]$	$0.0577\ [0.0422, 0.0704]$
Δ_{SHAM}	$0.6922\ [0.6627, 0.7330]$	$0.0936\ [0.0694, 0.1134]$
$P_{\mathrm{SHAM}} - P_{\mathrm{POS}}$	$0.4303\ [0.3555, 0.5061]$	$-0.0337\ [-0.0423, -0.0217]$
$G_{\mathrm{SHAM}} - G_{\mathrm{POS}}$	$-0.0625\ [-0.0766, -0.0486]$	$-0.0697\ [-0.0818, -0.0502]$
K	$0.4928\ [0.4217, 0.5671]$	$0.0359\ [0.0262, 0.0497]$
$J = 0.4569\ [0.3791, 0.5317]$		

The weights relabel under a shift, which verifies equivariance of this expression. At inputs with stabilizers, summing the full group preserves the same constraint. These inputs have Gaussian probability zero here. The irreducible risk over all measurable equivariant functions is

$$\inf_f R_q(f) = \int \overline{q}(z) \sum_g w_g(z) \frac{\|a_g(z) - f^*(z)\|_2^2}{4}\, dz. \tag{56}$$

The weighted variance in this integrand is shift invariant, so its integral under q equals its integral under $\overline{q}$. Disagreement between aligned targets contributes risk when multiple orbit representatives receive appreciable probability. A commutator measures disagreement without these probability weights.

We evaluated the fixed amplitude grid $a \in \{0, .125, .25, .5, .75, 1, 1.5, 2, 3, 4, 6, 8\}$. An initial integration used 262,144 independent Gaussian draws with a fixed seed. Rare overlap events at large a motivated a documented numerical refinement: two-dimensional Gauss–Hermite quadrature of orders 64, 128, 256, and 512. The sufficient Gaussian coordinates are $(x_0 - x_2)/\sqrt{2}$ and $(x_1 - x_3)/\sqrt{2}$; the other two coordinates cancel from the residuals and relative orbit densities. The refinement changes integration accuracy, not the target or amplitude grid. Figure 11 uses the order-512 values. Numerical integration uncertainty is distinct from training-seed uncertainty; no models were trained in this control.

Risk is 0.449600 at $a = 1$, 0.274390 at $a = 2$, and 0.00157150 at $a = 4$, although the zero-input obstruction MSE grows as a^2. Gaussian supports overlap everywhere: the decrease reflects concentration of orbit likelihoods, not literal disjoint support. A flexible equivariant rule can infer an orbit representative from the input distribution. In contrast, the plain rule Tz has risk a^2. The identity target remains exact under either normalizer, and the transported shift remains exact for every a.

This example establishes a distributional boundary of the pointwise argument. It does not predict which finite CNN or FNO reaches the class optimum, nor does it identify the source of the architecture interactions in the dynamics experiments. Estimating useful orbit overlap in those high-dimensional settings remains a separate empirical question.

G GRAY–SCOTT MATCHED EQUIVARIANT FEATURE CONTROL

This control adds 72 SHAM fits: two architectures, two normalizers, two statistics pools, three training seeds, and three pool seeds. Each fit is reported at 20k and after its prespecified continuation to 100k updates. The archived EQ/POS results are reused controls, not new independent replications.

POS supplies four fixed sinusoidal coordinate channels. SHAM supplies the first normalized state channel shifted by one grid site in each of the four signed axial directions. These features translate with the input. A scalar computed from training inputs scales their RMS to $1/\sqrt{2}$, matching the POS channel RMS. Copying the parent weights and zero-initializing all added input weights preserves the parent function at initialization. Both arms use 30,082 trainable parameters for CNN and 666,170 for FNO, batch size 32, and Adam at 10^{-3}. Full networks are trained. Data, normalization, initialization, and batch-prefix hashes are checked against the archived matched controls. Equal parameter counts do not imply equal effective capacity or optimization.

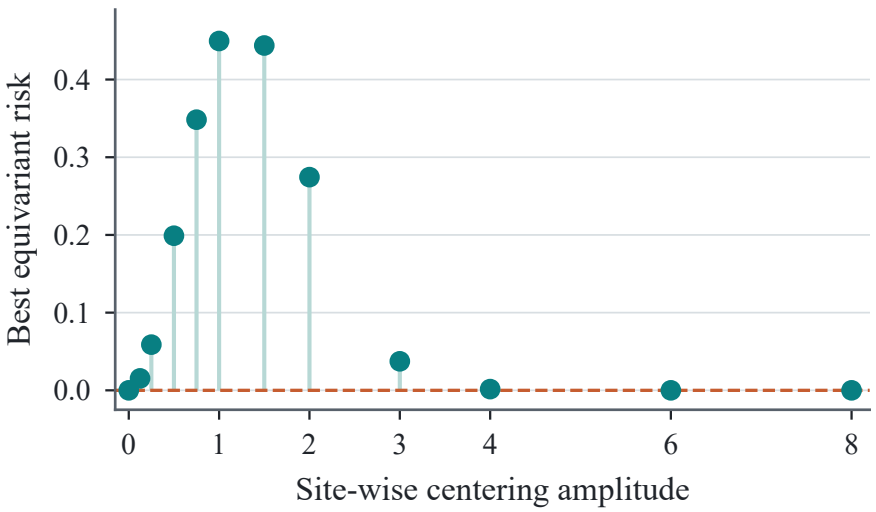


Figure 11: Population risk of the best measurable ordinarily equivariant predictor for a fixed Gaussian physical input law and a one-site shift target. Dots are the prescribed centering amplitudes and stems show their risks; no interpolation is implied. The transported shift and the identity negative control have zero risk by construction. All curves concern an analytical function-class optimum rather than fitted CNNs or FNOs.

Model	n	Budget	A_n	B_n	K_n
CNN	4	20k	0.0565 [-0.0315, 0.1569]	-0.0010 [-0.0024, 0.0001]	0.0576 [-0.0316, 0.1590]
CNN	256	20k	-0.0003 [-0.0147, 0.0136]	-0.0016 [-0.0063, 0.0017]	0.0012 [-0.0165, 0.0168]
FNO	4	20k	0.2321 [0.1822, 0.2983]	-0.0010 [-0.0016, -0.0001]	0.2331 [0.1829, 0.2998]
FNO	256	20k	-0.0078 [-0.0143, 0.0003]	-0.0011 [-0.0016, -0.0002]	-0.0068 [-0.0132, 0.0009]
CNN	4	100k	0.2648 [-0.0141, 0.7721]	0.0020 [-0.0057, 0.0084]	0.2629 [-0.0154, 0.7684]
CNN	256	100k	10423.2322 [-0.0165, 41692.8352]	0.0014 [-0.0036, 0.0057]	10423.2307 [-0.0164, 41692.8305]
FNO	4	100k	0.3214 [0.2765, 0.3678]	-0.0004 [-0.0011, 0.0004]	0.3218 [0.2761, 0.3685]
FNO	256	100k	-0.0093 [-0.0192, 0.0008]	0.0001 [-0.0006, 0.0009]	-0.0095 [-0.0192, 0.0013]

Table 32: Complete matched SHAM/POS contrasts at both budgets; every row has nine crossed cells.

Write $A_n = E_{32}(P_{\mathrm{SHAM}}) - E_{32}(P_{\mathrm{POS}})$ and $B_n = E_{32}(G_{\mathrm{SHAM}}) - E_{32}(G_{\mathrm{POS}})$. The gap interaction is

$$K_n = A_n - B_n, \qquad J_{\mathrm{SHAM}} = K_4 - K_{256}. \tag{57}$$

A positive K_n can reflect P improvement, G degradation, or both; we therefore report A_n and B_n separately. Endpoint differences are formed within each crossed cell before resampling. Intervals use 5,000 training-by-pool bootstrap draws and retain the same endpoint pairing; three levels per factor and a fixed shared confirmation set limit population inference.

All 72 SHAM fits have finite confirmation rollouts through the evaluated horizon 32 at both budgets, but one CNN $n = 256$ SHAM/P continuation has a catastrophic finite error ($E_{32} = 9.38 \times 10^4$). We retain it rather than treating it as a missing value; its group mean and endpoint interaction are therefore an instability diagnostic, not a typical effect estimate. For FNO, the small-pool POS advantage is positive in all nine cells at both budgets, while the $n = 256$ contrast is near zero and its interval overlaps zero. CNN contrasts vary with training budget and pool size. Their intervals span zero even in the group whose mean is dominated by the unstable continuation. This result strengthens the FNO architecture-specificity evidence without claiming a universal CNN response.

An early evaluator averaged finite trajectories when another trajectory failed. Before inspecting new forecast outcomes, the run was restarted after replacing this rule with full-cohort finite aggregation. Superseded artifacts remain separate. The reported records match the amended frozen source and

Model	n	Budget	G/POS	G/SHAM	P/POS	P/SHAM
CNN	4	20k	0.0075	0.0065	0.3330	0.3896
CNN	256	20k	0.0095	0.0079	0.2231	0.2228
FNO	4	20k	0.0070	0.0061	0.5585	0.7905
FNO	256	20k	0.0071	0.0060	0.1176	0.1098
CNN	4	100k	0.0086	0.0106	0.3108	0.5757
CNN	256	100k	0.0100	0.0115	0.2271	10423.4593
FNO	4	100k	0.0069	0.0064	0.5273	0.8487
FNO	256	100k	0.0068	0.0069	0.1212	0.1118

Table 33: Absolute horizon-32 errors for the same complete groups.

CNN 20k: $J_{\mathrm{SHAM}} = 0.0563$, 95% interval $[-0.0351, 0.1648]$. FNO 20k: $J_{\mathrm{SHAM}} = 0.2398$, 95% interval $[0.1819, 0.3109]$. CNN 100k: $J_{\mathrm{SHAM}} = -10422.9679$, 95% interval $[-41692.1910, 0.3013]$. FNO 100k: $J_{\mathrm{SHAM}} = 0.3313$, 95% interval $[0.2843, 0.3852]$.

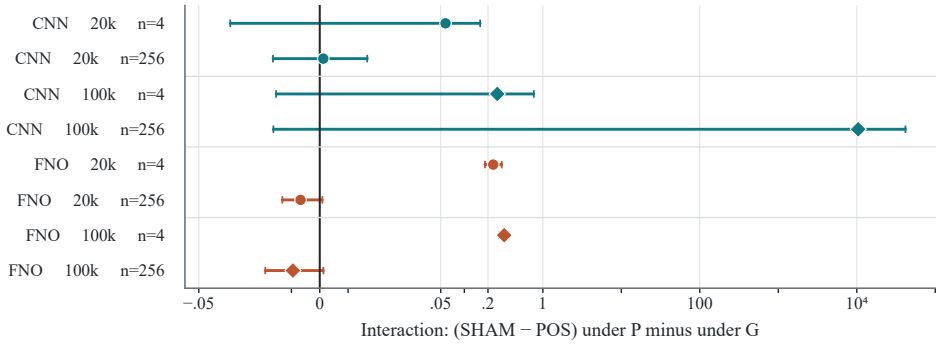


Figure 12: Gray–Scott matched SHAM/POS interaction at both training budgets: the SHAM-minus-POS horizon-32 error under P minus the same contrast under G. Points are crossed-cell means; whiskers are conditional 95% crossed-seed intervals over nine paired cells. Color distinguishes architecture and symbol distinguishes budget. The symmetric-log axis retains the extreme CNN 100k, $n = 256$ value, driven by one SHAM/P rollout with $E_{32} = 9.38 \times 10^4$; direct comparisons are in Tables 32 and 33. The 72 SHAM fits reuse archived POS controls.

manifest; nonfinite outcomes would invalidate the corresponding full-cohort mean rather than be removed from it.

H A RESTRICTED RISK COMPARISON FOR ARCHITECTURE SELECTION

H.1 PROJECTION RISK AND AN EQUAL-DIMENSIONAL CONSTRUCTION

Let $C_1, \dots, C_n$, $X_1, \dots, X_m$, and the test state X be independent draws from $\mathcal{N}(0, s_x^2 I_d)$. The target is the deterministic equivariant map $\Phi(x) = Bx + r(x)$, where the fixed circulant B satisfies $B\mathbf{1} = \mathbf{1}$, $\mathbb{E}r(X) = 0$, and $\mathrm{Cov}(r(X)) = \Sigma$. Put $\hat{\mu} = n^{-1}\sum_j C_j$, $\bar{r} = m^{-1}\sum_i r(X_i)$, $\Pi = \mathbf{1}\mathbf{1}^\top/d$, $h_G = \Pi\hat{\mu}$, and $h_P = \hat{\mu}$. For a fixed orthonormal-column matrix U, let $Q = UU^\top$ and fit

$$L_h(\theta) = \frac{1}{md}\sum_{i=1}^{m} \|B(X_i - h) + U\theta - [\Phi(X_i) - h]\|^2. \tag{58}$$

The normalized residual target is $a_h + r(X_i)$, where $a_h = (B - I)h$. Consequently,

$$\hat{\theta} = U^\top(a_h + \bar{r}), \qquad \nabla^2 L_h = \frac{2}{d} I_k, \qquad \widehat{\Phi}_{h,Q}(x) = Bx - (I - Q)a_h + Q\bar{r}. \tag{59}$$

For an independent test state, the conditional risk is

$$R_0 + \frac{1}{d}\|(I-Q)a_h\|^2 + \frac{1}{d}\|Q\bar{r}\|^2, \qquad R_0 = \frac{\operatorname{tr}\Sigma}{d}. \tag{60}$$

This follows from $\mathbb{E}r(X) = 0$ and orthogonality of the ranges of Q and $I - Q$. Under G, $a_G = 0$. Under P, $\mathrm{Cov}(a_P) = \Omega/n$, while $\mathrm{Cov}(\bar{r}) = \Sigma/m$. Taking traces proves Equation (8). It does not require Q to commute with either covariance. R_0 is fixed-backbone approximation error; it is not observation noise, and the deterministic target has zero Bayes risk.

For Equation (9), write $T = T_1$ and $q(x) = x \odot x - \mathbf{1}$. Since $q(Tx) = Tq(x)$ and $I - T/4$ commutes with T, the residual and target are equivariant under every cyclic shift. Equivalently, every site obeys the same local rule

$$\Phi_j(x) = x_{j-1} + (x_j^2 - 1) - \tfrac{1}{4}(x_{j-1}^2 - 1). \tag{61}$$

The input law is translation invariant. Gaussian moments yield $\mathbb{E}q = 0$ and $\mathrm{Cov}(q) = 2I$, so

$$\Sigma = 2(I - \tfrac{1}{4}T)(I - \tfrac{1}{4}T)^\top = \tfrac{17}{8}I - \tfrac{1}{2}(T + T^\top) = \tfrac{9}{8}I + \tfrac{1}{2}\Omega. \tag{62}$$

This is a positive-definite circulant covariance; translation invariance does not require an isotropic covariance. Since $Tu = u$ and $Tv = -v$,

	u	v
$w^\top \Omega w$	0	4
$w^\top \Sigma w$	$9/8$	$25/8$

(63)

Both projectors have rank one. Using $\operatorname{tr}\Omega = 16$ and $R_0 = 17/8$ gives

$$\begin{aligned} R_u^G &= \tfrac{17}{8} + \tfrac{9}{64m}, & R_v^G &= \tfrac{17}{8} + \tfrac{25}{64m}, \\ R_u^P &= \tfrac{17}{8} + \tfrac{2}{n} + \tfrac{9}{64m}, & R_v^P &= \tfrac{17}{8} + \tfrac{3}{2n} + \tfrac{25}{64m}. \end{aligned} \tag{64}$$

Subtracting proves Equation (10). Increasing calibration reduces D_P, whereas increasing supervision raises both D_P and D_G. If Σ were isotropic, equal ranks would give $D_G = 0$; residual anisotropy is essential for this strict reversal.

The position-dependent class is $Tz + \theta v$: for a fixed nonzero θ, its ordinary-shift defect is $2\theta v$. Commutation of its projector $vv^\top$ with T does not make the fixed intercept shift invariant. Moreover, $N_{\hat{\mu}}(Tx) - TN_{\hat{\mu}}(x) = (T-I)\hat{\mu} \neq 0$ almost surely for every finite n, since the kernel of $T - I$ consists of constant vectors. The P incompatibility verdict therefore persists across the expected-risk threshold. The example matches fitted dimension and quadratic curvature, while retaining different functional classes. It concerns iid states, a fixed backbone, and centering only; it gives no numerical threshold for correlated neural trajectories, learned scales, or autoregressive error. New CPU checks of this construction verify the matrix identities using exact rational arithmetic and compare 72 direct least-squares fits with the projected solution. A 3^8-node Gaussian cubature checks the residual moments and conditional physical risks. Twelve conditions, $n \in \{4, 16, 32, 256\}$ and $m \in \{4, 16, 64\}$, each use 100,000 independent calibration/supervised replications. Their largest discrepancy from the analytic risks and contrasts is 2.31 Monte Carlo standard errors. These are computational checks of the construction, not neural evidence.

H.2 The shared-versus-site-wise special case

The choices $Q = \Pi$ and $Q = I$ in Proposition 2 recover the shared-versus-site-wise comparison. This special case has unequal fitted dimensions and does not assume noisy labels. Let B be a fixed circulant matrix with $B\mathbf{1} = \mathbf{1}$; circulant structure also gives $\mathbf{1}^\top B = \mathbf{1}^\top$. The deterministic target $\Phi(x) = Bx + r(x)$ is equivariant under cyclic shifts, with $\mathbb{E}r(X) = 0$ and $\mathrm{Cov}(r(X)) = \Sigma$ for $X \sim \mathcal{N}(0, s_x^2 I_d)$. Calibration, supervised, and test states are independent. Write

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} U_k, \qquad \bar{r} = \frac{1}{m}\sum_{i=1}^{m} r(X_i), \qquad \Pi = \frac{\mathbf{1}\mathbf{1}^\top}{d}, \qquad a = (B - I)\hat{\mu}. \tag{65}$$

The centers are $h_G = \Pi\hat{\mu}$ and $h_P = \hat{\mu}$. Subtract the same center from input and output. Against the fixed backbone Bz, each normalized training residual is

$$[\Phi(X_i) - h] - B[X_i - h] = (B - I)h + r(X_i). \tag{66}$$

Since $(B-I)\Pi = \Pi(B-I) = 0$, least squares projects its mean onto the constant-vector subspace for a shared intercept and fits that mean directly for a site-wise intercept. The normalized fitted intercepts are

	shared	site-wise
G	$\Pi\bar{r}$	$\bar{r}$
P	$\Pi\bar{r}$	$a+\bar{r}$.

(67)

Decoding gives the physical predictors

$$\widehat{\Phi}_{G,\mathrm{shared}}(x) = Bx + \Pi\bar{r}, \qquad \widehat{\Phi}_{P,\mathrm{shared}}(x) = Bx + \Pi\bar{r} - a,$$
$$\widehat{\Phi}_{G,\mathrm{site}}(x) = \widehat{\Phi}_{P,\mathrm{site}}(x) = Bx + \bar{r}. \tag{68}$$

The two site-wise predictors coincide exactly for paired samples. This is a property of these intercept fits, not of jointly trained neural models.

For an independent test state, $\mathbb{E}\|v - r(X)\|^2/d = R_0 + \|v\|^2/d$, where $R_0 = \operatorname{tr}\Sigma/d$. Also $a \perp \mathbf{1}$, so the two terms in $\Pi\bar{r} - a$ are orthogonal for every realization. Using $\mathrm{Cov}(\bar{r}) = \Sigma/m$ and $\mathrm{Cov}(\hat{\mu}) = s_x^2 I_d/n$ gives

$$R_{G,\mathrm{shared}} = R_0 + \frac{\operatorname{tr}(\Pi\Sigma)}{md},$$
$$R_{P,\mathrm{shared}} = R_0 + \frac{\operatorname{tr}(\Pi\Sigma)}{md} + \frac{C_B}{n}, \tag{69}$$
$$R_{G,\mathrm{site}} = R_{P,\mathrm{site}} = R_0 + \frac{\operatorname{tr}\Sigma}{md}, \qquad C_B = \frac{s_x^2}{d}\|B-I\|_F^2.$$

Define

$$C_B = \frac{s_x^2}{d}\|B-I\|_F^2, \qquad V_m = \frac{\operatorname{tr}[(I-\Pi)\Sigma]}{md}. \tag{70}$$

Subtracting gives

$$R^G_{\mathrm{shared}} - R^G_{\mathrm{site}} = -V_m, \qquad R^P_{\mathrm{shared}} - R^P_{\mathrm{site}} = \frac{C_B}{n} - V_m, \tag{71}$$

and hence

$$(R^P_{\mathrm{shared}} - R^P_{\mathrm{site}}) - (R^G_{\mathrm{shared}} - R^G_{\mathrm{site}}) = \frac{C_B}{n}. \tag{72}$$

For $C_B, V_m > 0$, the threshold in this special case is

$$n_\star = C_B/V_m. \tag{73}$$

R_0 is approximation error from the omitted residual, not irreducible observation noise: the deterministic target has zero Bayes risk. Sampling different training inputs makes $\bar{r}$ fluctuate even though every label is exact.

For the explicit homogeneous example $B = T_1$, $r_j(x) = \alpha(x_{j-1}^2 - s_x^2)$, Gaussian moments give $\mathbb{E}r = 0$ and $\Sigma = 2\alpha^2 s_x^4 I_d$. Since $\operatorname{tr} T_1 = 0$ for $d \geq 2$, $C_B = 2s_x^2$. With $d = 8, m = 16, \alpha = s_x = 1$, $R_0 = 2$ and the exact risks are

Fitted class	G, either n	P, $n = 4$	P, $n = 256$
Shared intercept	2.015625	2.515625	2.0234375
Site-wise intercept	2.125000	2.125000	2.1250000

These are analytic risks against a deterministic map, not additional PDE experiments. The threshold is $n_\star = 128/7$. A reversal requires positive C_B, positive transverse residual variance, and feasible calibration sizes on opposite sides of $n_\star$. If $r = 0$, a site-wise intercept removes the centering offset exactly and never loses at finite n when $C_B > 0$. If $B = I$, calibration creates no offset for these fitted classes. Neither a universal reversal nor a universal benefit of position dependence follows.

The backbone need not equal a true linear component of Φ. If obtained by independent pretraining, the derivation holds conditional on B, provided its mean-preserving circulant constraints and $\mathbb{E}r_B(X) = 0$ hold, with $r_B = \Phi - Bx$. Its residual may include linear misspecification as well as nonlinearity. Unconditional risks average the displayed terms over B. Learning B jointly on the same supervised sample is outside this proof. The result supplies a deterministic sufficient mechanism for the calibration–capacity interaction, not a parameter-matched explanation of the observed POS/SHAM magnitudes.

Compatibility in both constructions. For $N_{\hat{\mu}}(x) = x - \hat{\mu}$,

$$N_{\hat{\mu}}(T_1 x) - T_1 N_{\hat{\mu}}(x) = (T_1 - I)\hat{\mu}. \tag{74}$$

The kernel of $T_1 - I$ is $\mathrm{span}(\mathbf{1})$. For finite n and $s_x > 0$, $\hat{\mu} \sim \mathcal{N}(0, s_x^2 I_d/n)$ has a density on $\mathbb{R}^d$ and hence lies in this one-dimensional subspace with probability zero when $d \geq 2$. P therefore fails the same binary commutation test on both sides of $n_\star$, while Equations (71) and (10) give opposite expected-risk signs in their respective constructions. The latter proves Corollary 3. G commutes for every realization. This compares expected risks, not the winner for every sample, and refers to a fixed calibration map. If the calibration mean is also translated, $N_{T_1\hat{\mu}}(T_1 x) = T_1 N_{\hat{\mu}}(x)$. The transported affine symmetry remains exact throughout. Quantitative mismatch and estimation cost determine the restricted comparison; a binary incompatibility classification alone does not.

H.3 Separate noisy-shift control

This control isolates estimation from neural-network optimization. Let T be the one-site cyclic permutation on $d \geq 2$ coordinates. Calibration states $u_1, \ldots, u_n$ are iid $\mathcal{N}(0, s_x^2 I_d)$, with $\hat{\mu} = n^{-1}\sum_k u_k$. Independent supervised states have the same law and noisy targets $y_i = Tx_i + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \tau^2 I_d)$ independently for $i = 1, \ldots, m$. Use the same fixed centering map for inputs and targets. We average physical per-site squared error against the noiseless target over calibration and supervised noise, rather than measuring error against a new noisy label.

For site-wise centering, $z = x - \hat{\mu}$ and

$$y_i - \hat{\mu} = Tz_i + a + \varepsilon_i, \qquad a = (T - I)\hat{\mu}, \qquad \mathbf{1}^\top a = 0. \tag{75}$$

The linear part T is known to both fitted models. Put $\bar{\varepsilon} = m^{-1}\sum_i \varepsilon_i$ and $\bar{\varepsilon}_{\mathrm{all}} = d^{-1}\mathbf{1}^\top \bar{\varepsilon}$. Least squares gives

$$\hat{c} = \bar{\varepsilon}_{\mathrm{all}}, \qquad \hat{b} = a + \bar{\varepsilon}. \tag{76}$$

Decoding to physical coordinates leaves errors $\hat{c}\mathbf{1} - a$ and $\hat{b} - a = \bar{\varepsilon}$. Orthogonality gives the exact decomposition

$$\frac{1}{d}\|\hat{c}\mathbf{1} - a\|^2 = \hat{c}^2 + \frac{1}{d}\|a\|^2, \qquad \frac{1}{d}\mathbb{E}\|a\|^2 = \frac{s_x^2}{nd}\,\mathrm{tr}[(T - I)^\top (T - I)] = \frac{2s_x^2}{n}. \tag{77}$$

The last equality uses $\mathrm{tr}\, T = 0$ for a nontrivial one-site cyclic shift. Since $\mathbb{E}\hat{c}^2 = \tau^2/(md)$ and $\mathbb{E}\|\bar{\varepsilon}\|^2/d = \tau^2/m$, the noisy-control risks are

$$R^G_{\mathrm{shared}} = \frac{\tau^2}{md}, \qquad R^P_{\mathrm{shared}} = \frac{2s_x^2}{n} + \frac{\tau^2}{md}, \qquad R^G_{\mathrm{site}} = R^P_{\mathrm{site}} = \frac{\tau^2}{m}. \tag{78}$$

Under any shared scalar centering, $(T - I)(\mu_G \mathbf{1}) = 0$, so the calibration term is absent even if μ_G is estimated inaccurately. For $T = I$, it is also absent under site-wise centering: at $d > 1$ and $\tau > 0$ the shared and site-wise risks still differ by their estimation variance. Only at $d = 1$ do both model classes coincide. At $\tau = 0$, this correctly specified shift control has no finite-n reversal. The deterministic residual construction above removes the need for label noise by allowing approximation error in the fixed backbone.

For $d = 8, m = 16, s_x^2 = \tau^2 = 1$, the shared risk is 0.0078125 under G, 0.5078125 under P with $n = 4$, and 0.015625 under P with $n = 256$. The site-wise risk is 0.0625 in all three cases. These are analytic values, not PDE measurements. The two classes have one and d fitted bias parameters; this calculation neither supplies a parameter-matched neural comparison nor bounds the optimal risk over all equivariant functions. It is consistent with the probability-weighted oracle in Appendix F. Its contribution is the explicit role of independently estimated preprocessing statistics in an otherwise elementary bias–variance comparison.

The Monte Carlo implementation samples the Gaussian calibration mean directly from its exact distribution, forms $(T - I)\hat{\mu}$, draws independent supervised noise, fits both intercepts, and evaluates decoded errors. It records standard errors across independent replications after averaging coordinates within each replication. Separate identity and one-dimensional controls check that the calibration residual is zero. The twelve shift conditions and two controls use 100,000 replications each; they are computational checks of this calculation, not additional dynamical systems.

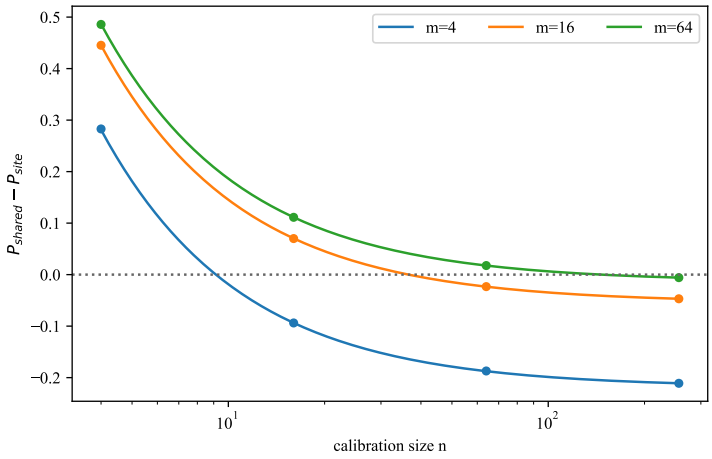


Figure 13: Restricted shift-regression risk difference, shared minus site-wise, with $d = 8$ and $s_x^2 = \tau^2 = 1$. Positive values favor site-wise biases. Curves are the analytic expression for continuous n; markers are Monte Carlo checks at $n = 4, 16, 64, 256$, with 100,000 independent replications per (n, m) cell. The displayed curves are theoretical functions, not interpolation of PDE results. Increasing calibration data can reverse the ranking without adding supervised examples.

I PUBLIC IMPLEMENTATION AUDIT AND ITS LIMITS

We inspected the public FNO fork `li-Pingan/fourier-neural-operator`, commit `f01cf052de7899c642ba738b8a322bdd488d9d38`. This is a source-level usage audit, not a reproduced benchmark or verified historical author release. In `fourier_3d.py`, lines 219–224 fit separate `UnitGaussianNormalizers` to training inputs and outputs. The class reduces over sample axis zero (`utilities3.py`, lines 73–84), retaining spatial and time coordinates: statistics have shapes $64 \times 64 \times 10$ and $64 \times 64 \times 40$. The script uses 1,000 training and 200 test examples and decodes both predictions and training targets before computing the optimized relative L^2 loss.

The same model appends fixed spatial and temporal coordinate channels. It therefore cannot be identified with the strictly spatially equivariant forecasters analyzed above. Its six-cell padding is along the temporal axis; that padding alone does not establish broken spatial translation equivariance. Dataset identity, forcing and condition transformations must be resolved before a task-level symmetry assertion or native replication. An inspected modern `neuraloperator` implementation instead accepts explicit reduction dimensions; with no dimensions specified and no mask, its `fit` computes globally shared statistics. Library names alone do not establish which convention a training pipeline uses.

The original FNO paper specifies the fixed forcing $q(x) = 0.1[\sin(2\pi(x_1 + x_2)) + \cos(2\pi(x_1 + x_2))]$ in its Navier–Stokes data-generation appendix (Li et al., 2021). For that task, a spatial shift by g preserves the forcing only when $g_1 + g_2 = 0$ modulo one. On a 64×64 grid this retains the 64 diagonal shifts $(k, -k)$, not all 4,096 spatial shifts. This subgroup calculation follows from the reported forcing; assigning it to the fork's local data files still requires verifying their provenance. A native comparison must respect the fixed condition, or transform and supply that condition explicitly.

For separate maps the physical predictor is $F = N_{\text{out}}^{-1} f N_{\text{in}}$, and the induced input/output actions must be transported separately. If reference maps are $G_{\text{in}}, G_{\text{out}}$, the representation

$$f_N = N_{\text{out}} G_{\text{out}}^{-1} h G_{\text{in}} N_{\text{in}}^{-1} \quad \Longrightarrow \quad N_{\text{out}}^{-1} f_N N_{\text{in}} = G_{\text{out}}^{-1} h G_{\text{in}} \tag{79}$$

preserves the physical function by construction. This is an implementation control, not a new repair algorithm. No public benchmark ranking reversal is claimed: the source audit establishes a concrete use of site-wise statistics, not its forecasting consequence. Pinned source URLs and file hashes accompany the analysis artifacts.

J CALIBRATION AND SUPERVISION WITH A FIXED SCALE

The centering study crosses calibration size $n \in \{4, 32, 256\}$ and supervised size $m \in \{7, 28, 112\}$ on Lorenz–96 with 40 sites and forcing 8. Each of five independent data blocks contains 112 training, 256 calibration, 128 validation, and 128 test trajectories from separate random seeds. Trajectories start from independent standard normal states, undergo a burn-in of 100 time units, and retain 513 states at intervals of 0.05; RK4 uses an internal step of 0.01. Within each block, n and m select nested prefixes of their respective pools. All 512 adjacent pairs from each of the m training trajectories are available for supervision.

Centering with a common scale. We denote the centering conditions by Gc and Mc to distinguish them from the earlier G/P normalization experiments. Gc estimates one mean over all states and sites in the n calibration trajectories; Mc estimates one mean per site from the same trajectories. Both use $N_c(x) = (x - \widehat{\mu}_c)/s$, where s is a single global standard deviation from all states and sites of 1,024 independent reference trajectories, plus 10^{-6}. The protocol fixes s before training and shares it across all blocks and conditions for input scaling, physical decoding, the training loss, and the secondary metric. Mc therefore changes centering only; it does not estimate a separate scale at each site.

Matched network training. POS and SHAM use the same circular CNN backbone and an added branch with the same parameter count and zero initial weights. POS supplies sine and cosine position channels scaled to unit spatial RMS. The backbone has two width-64 hidden layers with five-site circular stencils and SiLU activations, followed by a pointwise output layer and a residual multiplier of 0.05. Each model has 21,633 trainable parameters, including 640 added-branch weights. SHAM supplies the two neighboring normalized state values, divided by the RMS of normalized training data. The backbone and added branch are trained jointly. Within a block, candidates share parameter initialization and batch uniform random numbers; different m values map those uniforms to their available training pairs. The learning-rate grid is $\{0.0003, 0.001, 0.003\}$, giving $5 \times 3 \times 3 \times 2 \times 2 \times 3 = 540$ candidates. Each candidate has a budget of 100,000 Adam updates with batch size 64 and physical one-step MSE divided by s^2 as the loss. Training uses float32 with TF32 and cuDNN benchmarking disabled. There is no early stopping, seed replacement, or expansion of the grid; numerical failures remain in the record.

Validation and test readout. For each block, n, m, centering condition, and architecture, the protocol selects the learning rate with the lowest validation physical one-step MSE, breaking exact ties in favor of the smaller rate. A nonfinite primary metric is ineligible; a failure of the secondary metric does not change selection. After all candidates terminate, the 180 selection entries and candidate hashes are locked before a separate test readout. An entry remains missing if all three candidates fail. All successful candidates receive a test readout, but the primary comparison uses only the locked selections.

For a validation or test set of $K = 128$ trajectories, the primary metric is

$$R = \frac{1}{40K} \sum_{i=1}^{K} \|\widehat{x}_{i,1} - x_{i,1}\|_2^2, \qquad E_{32} = \frac{1}{K} \sum_{i=1}^{K} \frac{\|\widehat{x}_{i,32} - x_{i,32}\|_2}{s\sqrt{40}}. \tag{80}$$

Each forecast starts at the trajectory's initial retained state; E_{32} uses an autoregressive 32-step forecast and is secondary. The primary metric retains physical squared units.

Paired contrasts and scope. For block b and centering condition c, define

$$D_{c,n,m,b} = R_{\mathrm{SHAM},c,n,m,b} - R_{\mathrm{POS},c,n,m,b}.$$

Positive values favor POS. The prespecified contrasts are

$$\begin{aligned} C_{n,c,b} &= \frac{1}{3} \sum_{m \in \{7,28,112\}} (D_{c,4,m,b} - D_{c,256,m,b}), \\ C_{m,c,b} &= \frac{1}{3} \sum_{n \in \{4,32,256\}} (D_{c,n,112,b} - D_{c,n,7,b}). \end{aligned} \tag{81}$$

Table 34: Prespecified centering-study contrasts of primary physical one-step MSE. Means, SDs, and interval endpoints are in 10^{-3} physical squared units. Positive $D = R_{\mathrm{SHAM}} - R_{\mathrm{POS}}$ favors POS. C_n averages $D_{4,m} - D_{256,m}$ over m; C_m averages $D_{n,112} - D_{n,7}$ over n. Summaries use five independent blocks. Intervals are descriptive t intervals with four degrees of freedom, conditional on the protocol and shared reference scale. Missing blocks prevent the full mean, SD, and interval. Sign counts use available blocks; raw values and leave-one-block-out means are retained in the appendix data.

Center	Contrast	Mean	SD	95% interval	$+/-/0$	Blocks
Gc	C_n	0.00224	0.00651	$[-0.00584, 0.0103]$	2/3/0	5/5
Gc	C_m	-0.00611	0.0113	$[-0.0202, 0.00796]$	1/4/0	5/5
Mc	C_n	5.1	1.45	[3.3, 6.9]	5/0/0	5/5
Mc	C_m	-0.217	0.0574	$[-0.288, -0.146]$	0/5/0	5/5

Both contrasts have a prespecified positive direction for Mc; Gc is reported without a sign restriction. Summaries use the five independent blocks, including their raw values, mean, sample standard deviation, sign counts, and leave-one-block-out means. A complete finite set of five values also receives a two-sided descriptive 95% t interval with four degrees of freedom, conditional on the protocol and shared reference scale. A missing required pair makes that block's contrast missing and prevents the complete five-block summary. Candidates, grid cells, and trajectory frames are not independent replications. The full nine-cell grid and the corresponding secondary-metric contrasts are retained for both centerings.

The fixed-backbone calculation concerns iid states and fitted correction subspaces. The neural experiment instead trains a neural backbone jointly with its branch and uses correlated trajectory frames; n and m count trajectories. The experiment examines the prespecified directions within this grid. It cannot establish the calculation's exact $n = 2m$ threshold for neural networks or separate approximation from optimization error.

J.1 Completed readout and directional comparisons

All 540 candidates completed 100,000 updates without a recorded scientific failure. The validation rule selected all 180 entries; no entry required a missing-cell report. All 540 candidate test readouts contain 128 finite values for each metric, including the unselected rates. The primary summaries use only the locked selections.

The calibration contrast under Mc is positive in all five blocks, with mean 0.00509971 and descriptive interval $[0.00330317, 0.00689624]$. Its five leave-one-block-out means range from 0.00469494 to 0.00564361. The supervision contrast is negative in every block, with mean -0.000216801 and interval $[-0.000288035, -0.000145568]$; all five leave-one-block-out means are negative. This contradicts its prespecified positive direction. Gc's primary contrast intervals both span zero; its calibration mean changes sign when block 4 is omitted. Block identifiers in the supplied records run from 0 to 4.

The secondary Mc calibration contrast is 0.519486 $[0.412418, 0.626554]$. Its supervision contrast is 0.00553735 $[-0.0132803, 0.0243550]$, with four positive and one negative block. The secondary outcome does not overturn the primary supervision result. The full grid and contrasts below retain the observed directions; the supplement also gives all 180 absolute architecture risks and every leave-one-block-out value.

Evidence verification and reproduction. The execution host verified all 540 checkpoint hashes and all 21 data-array hashes against the frozen records before export. An independent local implementation reconstructed validation choices, both metrics from every stored per-trajectory readout, and all grid and contrast summaries, with no failed check. This is reaggregation of recorded per-trajectory errors, not independent model retraining. The compact supplement contains the complete candidate records, locked selections, test readouts, frozen code and protocol, data-generation seeds and hashes, and the server verification record under `evidence/t2_calibration_supervision`. Raw checkpoint tensors and data arrays remain outside that compact export; local audit does not claim to hash unavailable bytes.

Table 35: Primary physical one-step MSE differences under Gc. Each block value is SHAM minus POS at the indicated calibration and supervised trajectory counts. Values are in 10^{-3} physical squared units. Intervals use five independent blocks and the prespecified descriptive t rule. Positive values favor POS. A missing required block leaves the mean and interval unavailable. Full-precision values, SDs, sign counts, and leave-one-block-out means are retained in the accompanying CSV.

n	m	b_0	b_1	b_2	b_3	b_4	Mean	95% interval
4	7	−0.05	−0.0472	−0.0504	−0.0429	−0.00729	−0.0396	[−0.0623, −0.0169]
4	28	−0.0541	−0.0398	−0.0396	−0.0376	−0.0503	−0.0443	[−0.0534, −0.0351]
4	112	−0.0454	−0.0486	−0.0408	−0.0458	−0.0261	−0.0414	[−0.0525, −0.0302]
32	7	−0.0393	−0.0432	−0.0483	−0.0428	−0.00957	−0.0366	[−0.0558, −0.0174]
32	28	−0.0531	−0.0394	−0.0427	−0.0405	−0.0529	−0.0457	[−0.0541, −0.0373]
32	112	−0.0509	−0.0507	−0.044	−0.0461	−0.00754	−0.0398	[−0.0626, −0.0171]
256	7	−0.041	−0.0431	−0.0499	−0.043	−0.00747	−0.0369	[−0.0577, −0.016]
256	28	−0.051	−0.0395	−0.0425	−0.0365	−0.0543	−0.0448	[−0.0542, −0.0353]
256	112	−0.0506	−0.0507	−0.0419	−0.045	−0.0629	−0.0502	[−0.0602, −0.0403]

Table 36: Primary physical one-step MSE differences under Mc. Each block value is SHAM minus POS at the indicated calibration and supervised trajectory counts. Values are in 10^{-3} physical squared units. Intervals use five independent blocks and the prespecified descriptive t rule. Positive values favor POS. A missing required block leaves the mean and interval unavailable. Full-precision values, SDs, sign counts, and leave-one-block-out means are retained in the accompanying CSV.

n	m	b_0	b_1	b_2	b_3	b_4	Mean	95% interval
4	7	4.89	5.62	6.34	7.13	3.16	5.43	[3.54, 7.31]
4	28	4.5	5.1	5.81	6.54	2.84	4.96	[3.21, 6.71]
4	112	4.37	5.03	5.59	6.41	2.73	4.83	[3.1, 6.55]
32	7	1.18	0.248	0.644	0.614	0.569	0.651	[0.234, 1.07]
32	28	1.21	0.21	0.41	0.648	0.437	0.584	[0.106, 1.06]
32	112	1.22	0.245	0.474	0.614	0.434	0.596	[0.136, 1.06]
256	7	−0.0308	−0.0487	−0.0462	−0.0423	0.00615	−0.0324	[−0.0604, −0.00431]
256	28	−0.0132	−0.0444	−0.0377	−0.00177	−0.0422	−0.0278	[−0.0516, −0.00402]
256	112	−0.00719	−0.0486	−0.0452	−0.0283	−0.00839	−0.0275	[−0.0519, −0.00319]

Table 37: Primary physical one-step MSE contrasts in all five blocks. Values are in 10^{-3} physical squared units. For this metric, D denotes SHAM minus POS. C_n averages $D_{4,m} - D_{256,m}$ over m; C_m averages $D_{n,112} - D_{n,7}$ over n. Intervals are descriptive, conditional on the protocol and shared reference scale. The positive direction was prespecified for both Mc contrasts; all observed signs are retained.

Center	Contrast	b_0	b_1	b_2	b_3	b_4	Mean	95% interval
Gc	C_n	−0.00228	−0.000772	0.00116	−0.000589	0.0137	0.00224	[−0.00584, 0.0103]
Gc	C_m	−0.00552	−0.00551	0.00726	−0.00271	−0.0241	−0.00611	[−0.0202, 0.00796]
Mc	C_n	4.6	5.29	5.96	6.72	2.92	5.1	[3.3, 6.9]
Mc	C_m	−0.152	−0.198	−0.304	−0.236	−0.193	−0.217	[−0.288, −0.146]

Table 38: Secondary E_{32} differences under Gc. Each block value is SHAM minus POS at the indicated calibration and supervised trajectory counts. Values are dimensionless. Intervals use five independent blocks and the prespecified descriptive t rule. Positive values favor POS. A missing required block leaves the mean and interval unavailable. Full-precision values, SDs, sign counts, and leave-one-block-out means are retained in the accompanying CSV.

n	m	b_0	b_1	b_2	b_3	b_4	Mean	95% interval
4	7	0.00835	−0.0621	−0.0931	−0.0764	−0.0464	−0.0539	[−0.102, −0.00569]
4	28	−0.0396	−0.0628	−0.0221	−0.0276	−0.0559	−0.0416	[−0.0634, −0.0198]
4	112	−0.0784	−0.0567	−0.0975	−0.0707	0.0102	−0.0586	[−0.11, −0.0075]
32	7	−0.0554	−0.0628	−0.0992	−0.0766	−0.0492	−0.0687	[−0.0934, −0.0439]
32	28	−0.0485	−0.0679	−0.085	−0.0688	−0.076	−0.0693	[−0.086, −0.0525]
32	112	−0.0872	−0.0579	−0.0387	−0.0742	0.0072	−0.0502	[−0.096, −0.00442]
256	7	−0.0562	−0.0627	−0.0871	−0.0747	−0.0416	−0.0645	[−0.086, −0.0429]
256	28	−0.0361	−0.069	−0.0796	−0.0569	−0.0819	−0.0647	[−0.0881, −0.0413]
256	112	−0.0863	−0.058	−0.0643	−0.0725	−0.0496	−0.0661	[−0.0836, −0.0487]

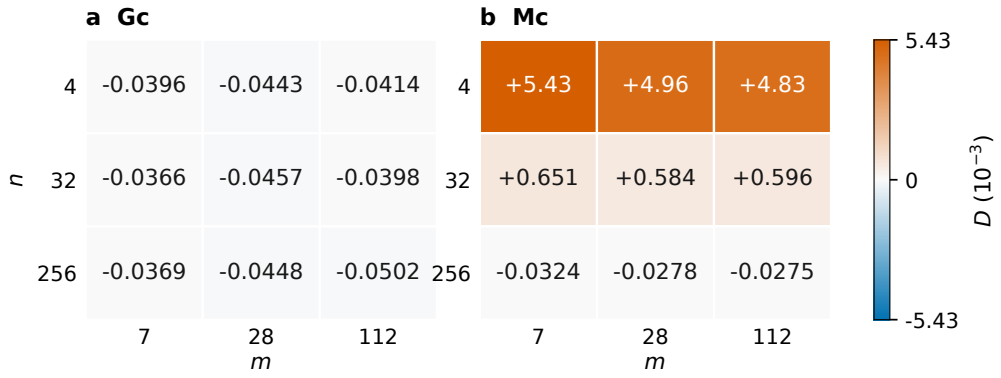


Figure 14: Calibration and supervision in the centering-only study. Cell values are five-block mean physical one-step MSE differences in units of 10^{-3}, $D = R_{\mathrm{SHAM}} - R_{\mathrm{POS}}$; positive favors POS. The two panels share one zero-centered color scale. Gc and Mc use a common external scalar scale and shared or site-wise means, respectively. Rows give calibration trajectories n; columns give supervised trajectories m. Rates are validation-locked. Under site-wise centering (Mc), calibration changes the mean winner at all three supervision levels; the prespecified Mc supervision contrast instead has the opposite sign to its prediction. Full results appear in Appendix J and the supplement.

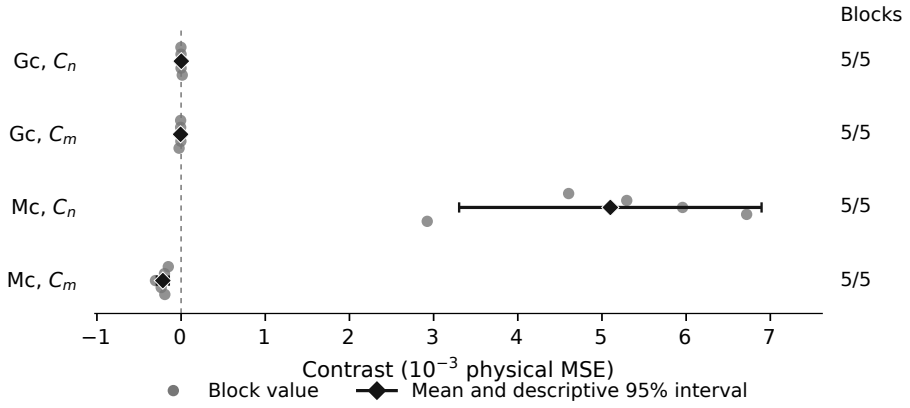


Figure 15: Prespecified one-step contrasts in 10^{-3} physical squared units. Gray points show blocks 0–4 from top to bottom within each row; diamonds and bars show the five-block mean and descriptive 95% t_4 interval. Positive C_n means the POS advantage is larger with four than with 256 calibration trajectories, averaged over supervision sizes. Positive C_m means a larger POS advantage with 112 than with seven supervised trajectories, averaged over calibration sizes. Mc supports the former direction and opposes the latter. All five blocks are present in each row.

Table 39: Secondary E_{32} differences under Mc. Each block value is SHAM minus POS at the indicated calibration and supervised trajectory counts. Values are dimensionless. Intervals use five independent blocks and the prespecified descriptive t rule. Positive values favor POS. A missing required block leaves the mean and interval unavailable. Full-precision values, SDs, sign counts, and leave-one-block-out means are retained in the accompanying CSV.

n	m	b_0	b_1	b_2	b_3	b_4	Mean	95% interval
4	7	0.468	0.506	0.507	0.557	0.356	0.479	[0.385, 0.573]
4	28	0.432	0.468	0.561	0.603	0.371	0.487	[0.37, 0.605]
4	112	0.445	0.499	0.513	0.598	0.376	0.486	[0.383, 0.588]
32	7	0.197	0.0558	0.128	0.136	0.102	0.124	[0.0599, 0.188]
32	28	0.243	0.0519	0.0834	0.147	0.155	0.136	[0.0445, 0.228]
32	112	0.199	0.0833	0.0644	0.141	0.144	0.126	[0.0598, 0.193]
256	7	−0.0452	−0.0229	−0.0563	−0.0455	−0.037	−0.0414	[−0.0568, −0.026]
256	28	−0.0351	−0.0261	0.00409	−0.0615	−0.0349	−0.0307	[−0.06, −0.00148]
256	112	−0.00851	−0.0265	−0.056	−0.0378	−0.0426	−0.0343	[−0.0565, −0.0121]

Table 40: Secondary E_{32} contrasts in all five blocks. Values are dimensionless. For this metric, D denotes SHAM minus POS. C_n averages $D_{4,m} - D_{256,m}$ over m; C_m averages $D_{n,112} - D_{n,7}$ over n. Intervals are descriptive, conditional on the protocol and shared reference scale. The positive direction was prespecified for both Mc contrasts; all observed signs are retained.

Center	Contrast	b_0	b_1	b_2	b_3	b_4	Mean	95% interval
Gc	C_n	0.023	0.00269	0.00611	0.00977	0.027	0.0137	[0.000438, 0.027]
Gc	C_m	−0.0496	0.005	0.0263	0.00336	0.035	0.00402	[−0.0368, 0.0449]
Mc	C_n	0.478	0.516	0.563	0.634	0.406	0.519	[0.412, 0.627]
Mc	C_m	0.00499	0.00557	−0.019	0.0175	0.0187	0.00554	[−0.0133, 0.0244]