RF-LIN: LINEAR PDE REPRESENTATIONS VIA INVERTIBLE RANDOM-FEATURE COORDINATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear representations of nonlinear PDEs must preserve state information while capturing dynamics. We introduce RF-LIN, which learns invertible random-feature coordinates and a finite-dimensional linear evolution operator. Frozen nonlinear features make each appended additive-coupling readout a conditional least-squares problem. For fixed coordinates, a projected singular-value solve fits the operator with a stationarity certificate in the retained subspace. Exact reconstruction and multistep composition separate coordinate invertibility from dynamical accuracy. We derive a finite-horizon error bound linking latent residuals, matrix-power growth, and decoder sensitivity, and show that a disconnected complete fixed-point set precludes global homeomorphic linearization. In three-seed Burgers experiments at two viscosities, RF-LIN achieves mean spectral errors of 0.927% and 4.415%, reductions of 28.3% and 17.1% relative to RF-EDMD, the strongest tested spectral baseline. Reconstruction errors remain below 2.8×10^{-14}, while RF-EDMD yields lower mean rollout error. Allen–Cahn tests expose a spurious near-fixed bridge between attractors. A single Kuramoto–Sivashinsky validation window exhibits strong transient amplification and poor rollout accuracy despite a spectral radius below one. These results demonstrate spectral gains from learned invertible coordinates and identify geometric and finite-time limits on their predictive use.

1 INTRODUCTION

Linear representations are attractive for nonlinear PDEs because matrix operations can replace repeated evaluation of a nonlinear evolution model. They also expose decay rates, invariant directions, and the sensitivity of forward and inverse calculations. Koopman methods pursue this representation through observables: DMD fits a linear operator to state snapshots, and EDMD fits one in a chosen dictionary (Tu et al., 2014; Williams et al., 2015). Learned dictionaries and autoencoders adapt the representation to the data (Lusch et al., 2018; Azencot et al., 2020).

Two errors arise in this construction. The representation may discard state information, and the fitted linear dynamics may fail to predict the retained information. An accurate encoder–decoder addresses the first error but leaves the second unresolved. Invertible coordinate maps eliminate reconstruction loss by construction. Existing examples include FlowDMD, EDMD-IDL, and Koopmanizing Flows (Meng et al., 2024; Jin et al., 2024; Bevanda et al., 2022). Their use raises a further question: how can the coordinate map be fitted while keeping the resulting operator solve explicit and its numerical error measurable?

RF-LIN places frozen random features inside invertible coupling layers and learns their linear readouts. This choice preserves nonlinear coordinates while making each appended readout a linear least-squares problem when the preceding layers and the latent operator are fixed. The latent matrix is then refitted by projected least squares. Unlike a free decoder, the analytic inverse introduces no separate reconstruction objective. Unlike joint coordinate–operator optimization alone, the final refit has a known coordinate-conditional solution. The method can use only these alternating solves or add gradient refinement of the readouts when selected by validation.

Our analysis connects these computational choices to prediction. It separates the operator-fit residual, growth of matrix powers, and decoder sensitivity in a physical-space error bound. It also identifies a geometric obstruction: a global homeomorphic conjugacy to a linear map must have a connected

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.

fixed set. This gives a concrete way to test where a linear latent model imposes the wrong dynamics, even when reconstruction is accurate.

The contributions are threefold. First, we give an invertible random-feature architecture whose coordinate readouts and latent operator can be fitted through alternating linear problems. Second, we derive the projected fitting certificate, composition identity, rollout bound, and fixed-point diagnostic for this model. Third, we evaluate spectral recovery and physical prediction separately on PDE examples with analytic or geometric structure. Burgers provides known heat-flow decay factors, Allen–Cahn exposes the interaction between two attractors, and a KS validation trajectory tests finite-time forecast behavior. These experiments show why a useful spectral representation and an accurate forecaster are related but distinct goals.

2 METHOD

RF-LIN encodes the physical state, advances it with a fitted linear operator, and recovers the forecast through the analytic inverse of the coordinate map. An N-step endpoint forecast applies A^N between one encoding and one decoding (Proposition 1).

2.1 INVERTIBLE RANDOM-FEATURE COORDINATES

Let $S_{\Delta t} : \mathbb{R}^d \to \mathbb{R}^d$ be a time-step map of a spatially discretized PDE. Given paired states (u_i, v_i) with $v_i = S_{\Delta t}(u_i)$, we learn a coordinate map g_θ and a matrix $A \in \mathbb{R}^{d\times d}$ satisfying

$$g_\theta(v_i) \approx A g_\theta(u_i), \qquad \widehat{S}(u) = g_\theta^{-1}(A g_\theta(u)). \tag{1}$$

The latent dimension equals the state dimension. Each additive coupling splits a mixed input $z = (z_1, z_2)$ into unchanged and updated coordinates:

$$z_1' = z_1, \qquad z_2' = z_2 + W\phi(\Omega z_1 + b), \qquad \phi = m^{-1/2}\tanh. \tag{2}$$

The feature weights Ω and biases b are sampled once and held fixed; the readout W is trained. Inversion subtracts the same conditioner from z_2'. Seeded orthogonal mixing between couplings allows successive blocks to act on different directions. The implementation also supports invertible coordinate scaling through actnorm. The inverse of the complete map applies the layer inverses in reverse order.

For an unchanged input of dimension p and feature width m, the implementation samples $\Omega_{ij} \sim \mathcal{N}(0, \gamma^2/p)$ and $b_i \sim \mathcal{N}(0, 1)$ independently, where γ is the feature bandwidth. Division by $\sqrt{p}$ controls the scale of the preactivations; division by $\sqrt{m}$ in ϕ normalizes the aggregate feature scale as width changes. The random seed fixes these features for the entire fit. Bandwidth and width therefore control the available nonlinear coordinates, while the readout solve determines their contribution to each state update.

This parameterization uses random features to fit coordinate changes rather than a direct state predictor or a fixed EDMD dictionary. Since the features depend only on the unchanged coordinates of a block, they are independent of its readout. The updated latent snapshots are therefore affine functions of W, which is the property needed for the blockwise solve below. The nonlinear feature map need not itself be invertible.

The unchanged coordinates carry the information needed to evaluate the inverse. The feature width can therefore be chosen independently of the state dimension: features parameterize a shift, while the state itself passes through the coupling. With K blocks, each updating q coordinates using m features, the additive readouts contain Kqm coefficients. The dense latent operator contributes another d^2 coefficients. Frozen feature weights are stored but are not optimized. Optional actnorm layers add $2d$ scale and bias parameters per block. These layers are initialized to the identity and can be updated during gradient refinement. The representation preserves the resolved state rather than compressing it into fewer latent variables.

For gradient refinement, an affine coupling may replace the additive update:

$$z_2' = z_2 \odot e^{\ell(z_1)} + t(z_1), \qquad \ell(z_1) = s_{\max}\tanh\big(s(z_1)/s_{\max}\big), \quad s_{\max} = 2. \tag{3}$$

Here s and t have frozen nonlinear features and trainable readouts. The inverse subtracts t and multiplies by $e^{-\ell}$. A zero scale readout initializes this block as its additive counterpart. Bounding ℓ controls pointwise scaling; derivatives of the conditioner still affect the condition number of the full map (Appendix C.1).

2.2 Projected fitting of the latent operator

Let $X, Y \in \mathbb{R}^{n\times d}$ contain the encoded inputs and targets as rows. We use $B = A^\top$, so the fitted relation is $Y \approx XB$. If $X = U\Sigma V^\top$ and V_r contains the retained right-singular directions, the operator problem is

$$\min_{C\in\mathbb{R}^{r\times d}} \|Y - XV_rC\|_F^2 + \mu\|C\|_F^2, \qquad B = V_rC, \quad \mu \geq 0. \tag{4}$$

Truncation limits amplification along poorly resolved input directions; the optional ridge parameter shrinks the retained coefficients.

Theorem 1 (Projected least-squares solution)**.** *For retained positive singular values in Σ_r, the unique minimizer of equation equation 4 is*

$$C_\star = (\Sigma_r^2 + \mu I)^{-1}\Sigma_r U_r^\top Y, \qquad A_\star = C_\star^\top V_r^\top. \tag{5}$$

It satisfies $V_r^\top X^\top (Y - XV_rC_\star) - \mu C_\star = 0$.

The resulting numerical certificate is the Frobenius norm of this projected normal residual divided by $\|Y\|_F$, with nonzero target norm assumed. We report it with the fit residual, retained rank, condition number, and discarded-target contribution. If nonzero singular directions are removed, the certificate applies to the constrained problem in equation 4; the full normal residual also contains the effect of truncation. The unregularized solve becomes ordinary minimum-norm least squares when every nonzero direction is retained. Appendix C.1 gives a two-dimensional example of this bias–sensitivity tradeoff.

2.3 Blockwise readout fitting

Suppose a prefix of the coordinate map is fixed and a new additive block is appended. After its fixed mixing, write the input and target states as $X_0, Y_0 \in \mathbb{R}^{n\times d}$. Let $H_x, H_y \in \mathbb{R}^{n\times m}$ be the frozen feature matrices, and let $E \in \mathbb{R}^{d\times q}$ select the q updated coordinates. For $W \in \mathbb{R}^{q\times m}$,

$$X(W) = X_0 + H_xW^\top E^\top, \qquad Y(W) = Y_0 + H_yW^\top E^\top. \tag{6}$$

With B fixed, set $R_0 = Y_0 - X_0B$ and define

$$\mathcal{L}_B(W) = H_yW^\top E^\top - H_xW^\top E^\top B, \qquad \min_W \|R_0 + \mathcal{L}_B(W)\|_F^2. \tag{7}$$

This is linear least squares in qm unknowns. A matrix-free solver evaluates $\mathcal{L}_B$ and its adjoint without storing the $nd \times qm$ vectorized design matrix. The adjoint and numerical checks are given in Appendix C.1. For the next operator update, the coordinates are fixed and equation equation 4 is solved again.

The training procedure starts with the identity map and an operator fit. For each proposed block, it alternates readout and operator updates. A joint readout/operator candidate is accepted when the relative training latent loss does not increase, up to numerical tolerance; otherwise the readout step is damped or rejected. A completed block is retained when the relative validation latent loss improves by the specified margin. The readout features can be formed on a deterministic subset for efficiency, while acceptance uses the full training split. The final model always receives an operator refit on the full training split.

Freezing the prefix is important for this construction. Once a preceding readout changes, the inputs to later feature maps also change, and their dependence on the earlier readout is nonlinear. The cascade isolates a block for which H_x and H_y are fixed during the solve. It thereby retains a linear update without requiring the complete coordinate map to be linear. The two update types also act on different objects: a readout changes the state representation, whereas the operator update selects the best allowed linear evolution in that representation. Their separate residuals make it possible to distinguish a poorly solved conditional problem from an inadequate set of coordinates.

Gradient refinement optimizes the readouts and any actnorm parameters with Adam and periodic operator refreshes. In the Burgers runs, its checkpoint is selected by relative validation latent loss. It can include multistep or isometry terms. These terms belong to the refinement objective; the linear readout solve in equation 7 uses the unnormalized one-step latent residual. Keeping these objectives separate preserves the linear structure of the blockwise fit. The reported Burgers configurations use refinement at $\nu = 0.03$ and readout-only fitting at $\nu = 0.10$.

2.4 Prediction and computational form

After fitting, prediction uses the same pair (g, A) at every step. The coupling inverse gives the following identities.

Proposition 1 (Reconstruction and multistep composition)**.** *For $\widehat{S} = g^{-1} \circ A \circ g$, exact arithmetic gives*

$$g^{-1}(g(u)) = u, \qquad \widehat{S}^N(u) = g^{-1}(A^N g(u)), \quad N \geq 0. \tag{8}$$

If A is nonsingular, the same construction gives $\widehat{S}^{-N} = g^{-1} \circ A^{-N} \circ g$.

An endpoint forecast thus needs one encoding and one decoding. The latent state can be advanced by repeated matrix–vector products or by a matrix power computed through repeated squaring. For a dense operator these cost $O(Nd^2)$ and $O(d^3 \log N)$, respectively, before encoding and decoding. The preferable implementation depends on horizon, dimension, and whether powers are reused across initial states. A full physical trajectory still requires one decoding per output time. Appendix C.2 details the work counts and floating-point checks. Regularized backward recovery and sparse-observation objectives follow from the same representation (Appendix C.2).

3 Theory

The analysis addresses optimization, finite-time prediction, and the domain on which linearization can hold. Complete proofs are in Appendix A.

3.1 Accepted updates and coordinate scale

Theorem 2 (Monotonicity under explicit loss checks)**.** *For a fixed prefix, let $F(W, A) = \|Y(W) - X(W)A^\top\|_F^2$. If the readout and operator candidates satisfy*

$$F(W_{j+1}, A_j) \leq F(W_j, A_j), \qquad F(W_{j+1}, A_{j+1}) \leq F(W_{j+1}, A_j), \tag{9}$$

the accepted objective sequence is nonincreasing and converges to a finite limit.

This result gives sufficient acceptance conditions for iterative readout solves and operator truncation. The implementation checks a joint update using relative latent loss, whose normalization changes with the coordinates; this check alone does not establish the two inequalities above for F. Even under those inequalities, objective convergence does not establish a joint global optimum. Validation latent loss determines which blocks to retain.

Coordinate scale also matters. Rescaling g by a small positive constant reduces an absolute latent residual without improving the decoded prediction. For this reason we evaluate relative latent residuals, physical errors, and inverse conditioning together. Proposition 3 formalizes this scaling example. The approximation result in Proposition 2 assumes that a useful conjugacy exists and that the coupling and feature classes approximate it on the relevant compact set; the domain restriction below is therefore essential.

3.2 From a latent residual to a physical forecast error

Theorem 3 (Finite-horizon error)**.** *Let K be forward invariant under $S = S_{\Delta t}$ and suppose $\|g(S(u)) - Ag(u)\| \leq \varepsilon$ on K. If g^{-1} is L-Lipschitz on a set containing the exact and predicted latent points up to horizon N, then*

$$\|S^N(u) - g^{-1}(A^N g(u))\| \leq L\varepsilon \sum_{j=0}^{N-1} \|A^j\|. \tag{10}$$

The three factors have different roles: ε measures the local representation error, matrix powers transport that error through time, and L converts latent error into state error. A small spectral radius alone does not bound transient amplification. For example, a matrix with a nonzero upper off-diagonal entry can have both eigenvalues strictly inside the unit circle while its early powers grow. Appendix C.2 works out this example and the exact residual convolution. Sampled Jacobian singular values estimate local decoder sensitivity; a uniform bound requires control throughout the latent region visited by both trajectories.

3.3 Fixed points restrict global linearization

Theorem 4 (Global fixed-point obstruction)**.** *If a homeomorphism $g : \mathbb{R}^d \to \mathbb{R}^d$ satisfies $g \circ S = A \circ g$ globally, then*

$$\mathrm{Fix}(S) = g^{-1}(\ker(A - I)). \tag{11}$$

Consequently, a map with a disconnected fixed set cannot have such a global linearization.

The kernel is a connected linear subspace, and a homeomorphism preserves connectedness. On a restricted invariant domain Ω, the corresponding condition concerns $g(\Omega) \cap \ker(A - I)$, which is connected when $g(\Omega)$ is convex. Thus a useful basin-local model can exist even when a global model is excluded.

Theorem 5 (Approximate fixed-point distance)**.** *Let $p = S(p)$, $M = A - I$, and $\|g(S(p)) - Ag(p)\| \le \varepsilon$. For $M \neq 0$,*

$$\mathrm{dist}(g(p), \ker M) \le \frac{\varepsilon}{\sigma^{+}_{\min}(M)}, \tag{12}$$

where $\sigma^{+}_{\min}$ is the smallest positive singular value. The distance is zero when $M = 0$.

This diagnostic relates an endpoint defect to the near-fixed latent subspace. It also motivates the Allen–Cahn path experiment: if a linear operator nearly fixes two encoded equilibria, it nearly fixes their connecting latent segment. Decoding the segment can then produce a continuum of nearly stationary predictions at states where the PDE actually moves. The relevant test is the displacement along this path, in addition to ordinary basin-local forecasts.

4 Experimental design

Data and evaluation roles. The Burgers comparison and Allen–Cahn control use three matched seeds. Burgers has 256 periodic grid points, sampling interval $\Delta t = 0.025$, viscosities $\nu \in \{0.03, 0.10\}$, and independent 512/64/64 training, validation, and test trajectories. Allen–Cahn uses matched two-basin and positive-basin splits. A separate KS diagnostic uses one saved model and the first 100-step window of its validation split; its results are reported separately from the three-seed tests. The KS data use 256 points on a periodic interval of length 22 and $\Delta t = 0.25$. Appendix D specifies the training configurations and sample selection.

Baselines. We compare RF-LIN with exact DMD, degree-3 polynomial EDMD, and random-feature EDMD. DMD tests a linear model in state coordinates; the EDMD baselines test nonlinear fixed dictionaries with fitted decoders. This comparison separates the benefit of learned invertible coordinates from increasing observable expressiveness alone. The Burgers test numbers use the validation-selected configurations, including gradient refinement at the lower viscosity. Solver convergence and split details are given in Appendix D.

Spectral and coordinate diagnostics. For mean-zero periodic Burgers flow, the Cole–Hopf representation leads to heat multipliers $\exp(-\nu k^2 \Delta t)$. We compare the 20 largest learned eigenvalue magnitudes with the prescribed ordered heat factors, including their multiplicities. Each run reports the median relative discrepancy; the table reports the mean and sample standard deviation of these medians across seeds. A held-out linear fit from discretized Cole–Hopf coordinates to learned coordinates measures alignment. Its commutation defect measures compatibility with the two coordinate evolutions. Derivation and normalization details appear in Appendix B.1.

Physical metrics. One-step prediction starts from each reference state; autonomous rollout is initialized once and uses no subsequent reference-state resets. Mean rollout error averages the relative state L_2 error over predicted times and then trajectories, excluding the exact initial state. Reconstruction measures $\|g^{-1}(g(u)) - u\|_\infty$. For the fixed-point test, we compare learned and reference one-step displacement along the decoded segment between the encoded equilibria. All reported mean–SD pairs summarize seeds, rather than treating correlated time samples as independent observations. The rankings are descriptive comparisons of those means.

5 Results

5.1 Burgers: spectral recovery and prediction

Table 1 compares the four methods. RF-LIN has the lowest mean spectral error at both viscosities: 0.927% and 4.415%, respectively. These are 28.3% and 17.1% below RF-EDMD, the strongest spectral baseline in these experiments. All six RF-LIN runs reconstruct with maximum error below 2.8×10^{-14}; RF-EDMD nevertheless has lower mean rollout error at both viscosities.

Table 1: RF-LIN has the lowest spectral error at both viscosities, while RF-EDMD has the lowest rollout error. Values are mean ± sample SD over three seeds. Spectrum is the within-run median relative error of the 20 leading eigenvalue magnitudes; rollout is the time-mean relative state error. Bold marks the lowest observed mean within each viscosity and metric.

ν	Method	Spectrum error	One-step error	Rollout error	Latent residual
0.03	DMD	1.576e-02 ± 0.00e+00	1.008e-02 ± 0.00e+00	6.835e-01 ± 0.00e+00	1.371e-02 ± 0.00e+00
0.03	Polynomial EDMD	1.680e-02 ± 1.82e-05	1.297e-02 ± 2.03e-04	5.218e-01 ± 1.86e-04	3.959e-02 ± 1.58e-04
0.03	RF-EDMD	1.293e-02 ± 1.41e-04	2.577e-02 ± 8.77e-04	**3.726e-01 ± 5.63e-03**	**5.488e-03 ± 7.25e-05**
0.03	RF-LIN	**9.269e-03 ± 4.88e-04**	**8.776e-03 ± 2.09e-04**	4.631e-01 ± 3.06e-03	9.759e-03 ± 3.26e-04
0.10	DMD	5.891e-02 ± 0.00e+00	6.856e-03 ± 0.00e+00	6.348e-01 ± 0.00e+00	9.532e-03 ± 0.00e+00
0.10	Polynomial EDMD	5.813e-02 ± 1.40e-05	**3.762e-04 ± 3.73e-07**	3.993e-01 ± 2.24e-04	2.760e-02 ± 9.06e-05
0.10	RF-EDMD	5.328e-02 ± 1.46e-04	6.518e-03 ± 2.90e-04	**2.318e-01 ± 7.90e-03**	**1.842e-03 ± 1.86e-05**
0.10	RF-LIN	**4.415e-02 ± 2.56e-04**	1.554e-03 ± 7.94e-06	4.089e-01 ± 4.25e-03	2.361e-03 ± 2.71e-05

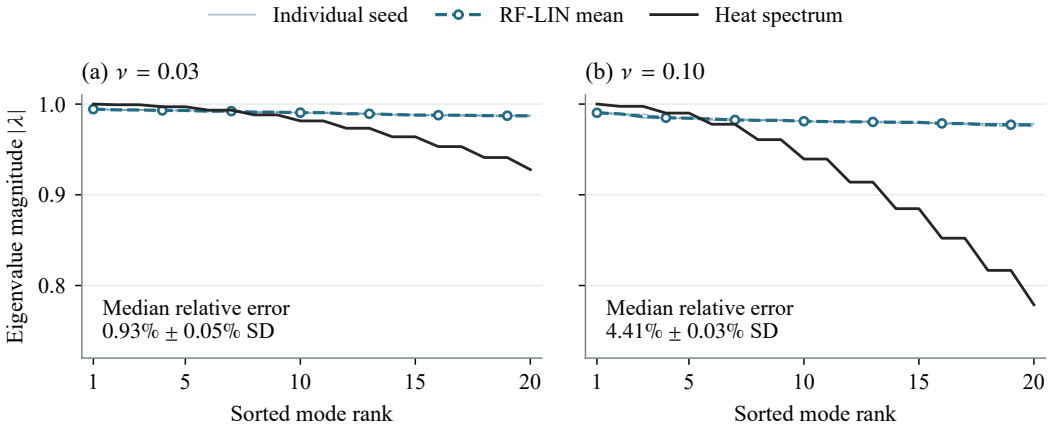


Figure 1: RF-LIN eigenvalue magnitudes and the analytic heat-flow reference. Thin curves show individual seeds and the emphasized curve shows their mean. Sorted rank indexes the leading 20 magnitudes. The mismatch at larger ranks is more pronounced at the higher viscosity.

The complete spectral curves in Figure 1 show where agreement is strongest. The leading magnitudes are close to one, while the learned tail decays more slowly than the heat reference. The mean within-run maximum spectral errors are 6.390% and 25.453%. Thus the median ranking reflects improved agreement over the selected spectrum, with a remaining tail error that matters for repeated application of the operator. A slightly overestimated decay factor can retain latent components for too long as the horizon increases.

For an isolated scalar mode, this effect has a simple expression. If a nonzero true multiplier λ is estimated as $\widehat{\lambda} = \lambda(1+\delta)$, then

$$\frac{\widehat{\lambda}^N}{\lambda^N} = (1+\delta)^N. \tag{13}$$

Thus a persistent relative bias in one-step decay compounds across time. This observation concerns the mode multiplier; assigning it to a physical error additionally requires its amplitude and decoded spatial shape. It helps explain why the full spectral curve is more informative than its median alone, and why we pair spectral measurements with autonomous state-space forecasts.

The physical-error ordering differs. At $\nu = 0.03$, RF-LIN has the best one-step error, 12.9% below DMD, but RF-EDMD has the lowest mean rollout error. At $\nu = 0.10$, polynomial EDMD leads one-step prediction and RF-EDMD again leads rollout. RF-LIN's mean rollout errors are 46.31% and 40.89%, compared with 37.26% and 23.18% for RF-EDMD. The corresponding RF-LIN rollout ranks are second and third. The errors show that recovering decay factors does not by itself recover their physical amplitudes, spatial structure, or long-horizon interaction.

Coordinate alignment gives a further distinction. Held-out Cole–Hopf alignment reaches $R^2 = 0.721 \pm 0.006$ and 0.774 ± 0.006, below the specified 0.95 target. Relative commutation defects are 0.107 ± 0.003 and 0.366 ± 0.453. The learned coordinates therefore capture part of the analytic representation but do not recover it up to the tested linear alignment. Appendix B.1 defines these diagnostics; Appendix B.1 shows the reference solution profiles that underlie the competition between nonlinear steepening and viscous decay.

5.2 Allen–Cahn: A near-fixed bridge between attractors

The two-basin experiment probes the path predicted by the fixed-point analysis. Figure 2 compares displacement along the decoded straight segment between the encoded constant equilibria. The reference dynamics move the interior states strongly, while RF-LIN nearly fixes them. The within-seed median model/reference displacement ratios have mean 1.94×10^{-4} and sample SD 1.37×10^{-4} across three seeds.

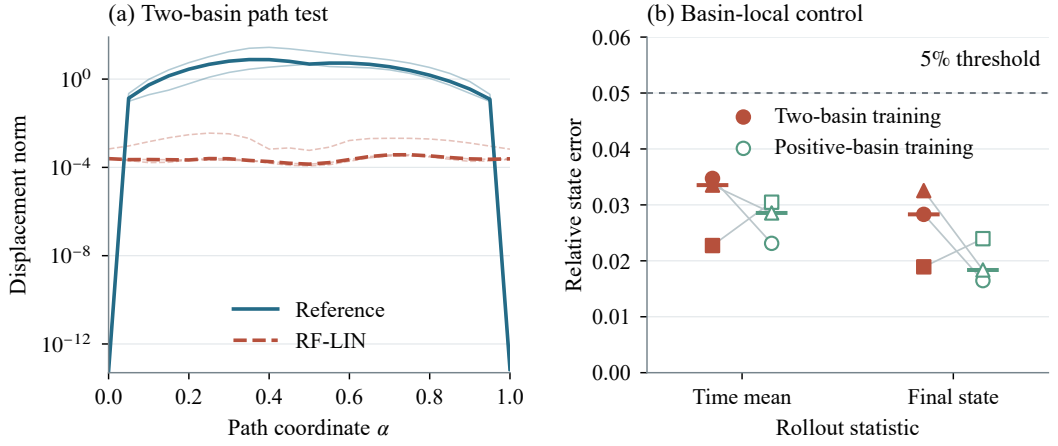


Figure 2: RF-LIN nearly fixes the decoded bridge while the reference dynamics move its interior states. Left: learned and reference displacement along the segment; thin curves show seeds and thick curves their pointwise medians. Right: paired seed forecast errors for both-basin and positive-basin training, with median bars.

Table 2: Allen–Cahn forecast errors, mean ± sample SD over three seeds.

Training domain	One-step error	Rollout mean	Rollout final
Both basins	$3.975\text{e}{-4} \pm 7.11\text{e}{-5}$	$3.033\text{e}{-2} \pm 6.61\text{e}{-3}$	$2.659\text{e}{-2} \pm 6.97\text{e}{-3}$
Positive basin	$2.997\text{e}{-4} \pm 3.12\text{e}{-5}$	$2.739\text{e}{-2} \pm 3.81\text{e}{-3}$	$1.959\text{e}{-2} \pm 3.90\text{e}{-3}$

Restricting training to the positive basin reduces mean rollout error by 9.7% (Table 2), with machine-precision reconstruction in both regimes. This control separates local approximation from global geometry. A model can forecast trajectories near an attractor while assigning incorrect dynamics to the connecting region. The path is a targeted geometric probe rather than an in-domain random test trajectory. Applying the global theorem requires a disconnected complete fixed set; the experiment tests its near-fixed-bridge mechanism without classifying every equilibrium of the discretized PDE.

5.3 KS: RECONSTRUCTION, FITTING, AND TRANSIENT AMPLIFICATION

The KS example tests whether a well-fitted latent operator remains predictive on a demanding trajectory. We use a validation-selected readout-only model with six accepted blocks, 512 features per conditioner, bandwidth 2, and seed 20260730. The forecast starts at index 80,000, the first validation state, and covers 100 steps, or 25 physical time units.

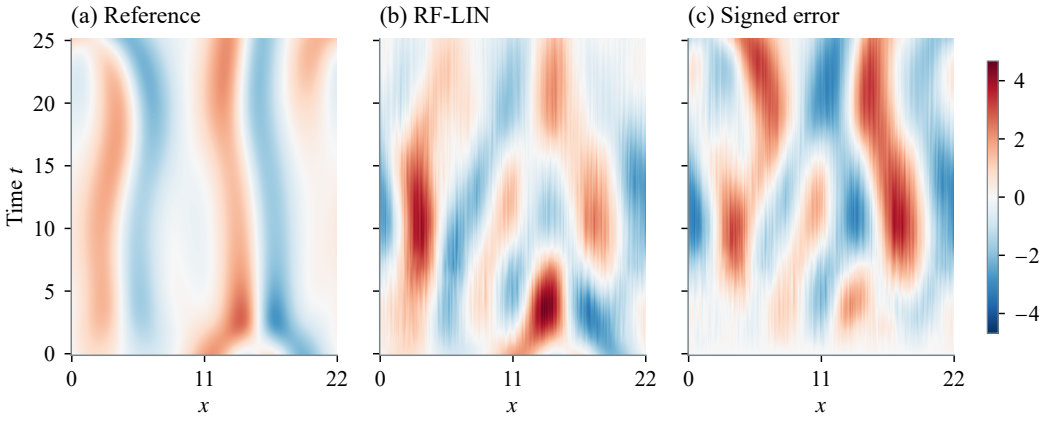


Figure 3: The KS autonomous prediction separates from the reference on this validation window. Panels show reference, prediction, and signed error with one shared colorbar and symmetric limits about zero.

The prediction initially follows the reference before its spatial structures and amplitudes separate (Figure 3). On this window, the mean one-step error is 1.829%, the mean autonomous error is 130.32%, and the final error is 125.72%. The reconstruction error is 1.38×10^{-14}. The one-step/rollout difference is consistent with accumulation of dynamical error despite accurate coordinate inversion.

The final operator fit retains all 256 singular directions with zero ridge. It has relative training residual 0.01404 and normal-equation certificate 3.74×10^{-10}. Its spectral radius is 0.993958, yet $\|A^{20}\|_2 \approx 3.77 \times 10^5$ in the fitted coordinates. These quantities illustrate the separate terms in Theorem 3: least-squares stationarity and asymptotically decaying eigenvalues can coexist with large finite-time amplification. The operator norm depends on the coordinate system, so it should be interpreted together with decoder sensitivity. Appendix B.2 contains the error, energy, and operator-power curves. This is a single-window validation diagnosis; a three-seed KS baseline comparison and component ablations remain open.

Implications for model selection. The three examples expose different requirements on the representation. Burgers tests whether a learned operator matches known decay factors; Allen–Cahn tests

whether its near-fixed states have the correct geometry; KS tests whether finite-time propagation remains useful after local errors are repeatedly transported. A validation objective based only on one-step latent fit can miss all three distinctions. The readout-only blocks and Burgers refinement checkpoints are selected by relative latent validation loss. Spectrum, reconstruction, power norms, and decoded rollout error characterize the selected model together. This links the fit to the scientific use of the learned coordinates.

6 Related work

Fixed and learned observables. DMD and EDMD provide explicit finite-dimensional operator estimates (Tu et al., 2014; Williams et al., 2015). Neural embeddings and consistent autoencoders learn observables together with reconstruction or dynamical objectives (Lusch et al., 2018; Azencot et al., 2020). Online Fourier-feature methods adapt Koopman representations using streamed data (Salam et al., 2023). RF-LIN instead puts frozen nonlinear features inside a learned invertible map. The resulting readout is linear in each block update, while the complete coordinates remain nonlinear.

Invertible representations. FlowDMD and EDMD-IDL use invertible neural coordinates to recover physical states (Meng et al., 2024; Jin et al., 2024). Koopmanizing Flows learn diffeomorphic coordinates for stable continuous-time models (Bevanda et al., 2022), and invertible Koopman neural operators apply invertible representations to PDE modeling (Jin et al., 2025). RF-LIN's distinction is the readout-based fitting structure and the coordinate-conditional operator solve. Invertibility is shared with these approaches; the specific contribution is how the coordinates are parameterized, optimized, and connected to measurable prediction errors.

Spectral analysis and approximation. Residual DMD studies verified spectral approximation (Colbrook et al., 2023). Its spectral-verification question differs from the stationarity of a finite projected regression problem addressed here. Coupling-flow approximation theory (Teshima et al., 2020) motivates expressive invertible coordinates, while the conditional statement in Appendix A identifies the approximation assumptions used by RF-LIN. Our PDE tests combine a known analytic spectral reference with a fixed-point geometry test, allowing the representation to be assessed beyond its one-step regression residual.

7 Limitations and broader impact

The current experiments establish a spectral advantage over three Burgers baselines, not a general forecasting advantage. The spectral tail and coordinate-alignment errors remain substantial, and the KS example shows large autonomous errors. A matched comparison with neural predictors, a complete component ablation, and same-hardware timing are needed to assess accuracy and computational cost more broadly.

On the mathematical side, the approximation statement assumes a useful conjugacy and gives no finite-width or sample-complexity rate. The rollout bound requires control of the decoder on the visited latent region, and the fixed-point theorem excludes some global domains even when local models are accurate. These limits suggest two concrete directions: conditioning-aware coordinate learning and separate local representations for different basins. For scientific surrogate use, the physical error and conditioning diagnostics provide checks that complement spectral interpretation and exact reconstruction.

8 Conclusion

RF-LIN learns nonlinear PDE coordinates through invertible random-feature couplings with linear readouts and an explicit projected operator fit. This construction gives tractable coordinate updates, exact multistep composition, and a clear separation between fitting error, transient amplification, and decoder sensitivity. Burgers experiments show improved spectral recovery, while the Allen–Cahn and KS results identify geometric and finite-time limitations. Model selection should therefore check spectral agreement together with decoded rollout error and fixed-point geometry.

REPRODUCIBILITY STATEMENT

The operator and readout formulations are specified in Section 2; Appendix A supplies the proofs. Appendix D records configurations, splits, data sources, and the numerical precision. Reference-solver checks and diagnostic definitions accompany the supplementary solution plots. Result tables are derived from the recorded seed aggregates.

AI USE STATEMENT

OpenAI Codex was used for coding assistance and language editing. The authors reviewed all AI-assisted content and take full responsibility for the manuscript, implementation, and reported results.

REFERENCES

Omid Azencot, N. Benjamin Erichson, Vanessa Lin, and Michael W. Mahoney. Forecasting sequential data using consistent koopman autoencoders. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 475–485, 2020.

Petar Bevanda, Max Beier, Sebastian Kerz, Armin Lederer, Stefan Sosnowski, and Sandra Hirche. Diffeomorphically learning stable koopman operators. *arXiv preprint arXiv:2112.04085*, 2022. URL https://arxiv.org/abs/2112.04085.

Matthew J. Colbrook, Lorna J. Ayton, and Máté Szőke. Residual dynamic mode decomposition: Robust and verified koopmanism. *Journal of Fluid Mechanics*, 955:A21, 2023. doi: 10.1017/jfm.2022.1052.

Peng Jin, Shuai Cong, Zhi Hou, Yizhou Gao, Yingjie Ge, Yiming Zhu, Jin Feng, and Bin Li. Invertible koopman neural operator for data-driven modeling of partial differential equations. *arXiv preprint arXiv:2503.19717*, 2025.

Yuhong Jin, Lei Hou, and Shun Zhong. Extended dynamic mode decomposition with invertible dictionary learning. *Neural Networks*, pp. 106177, 2024. doi: 10.1016/j.neunet.2024.106177.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9:4950, 2018. doi: 10.1038/s41467-018-07210-0.

Yuhuang Meng, Jianguo Huang, and Yue Qiu. Koopman operator learning using invertible neural networks. *Journal of Computational Physics*, 501:112795, 2024. doi: 10.1016/j.jcp.2024.112795.

Tahiya Salam, Alice Kate Li, and M. Ani Hsieh. Online estimation of the koopman operator using fourier features. In *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*, volume 211 of *Proceedings of Machine Learning Research*, pp. 1271–1283, 2023. URL https://proceedings.mlr.press/v211/salam23a.html.

Takeshi Teshima, Isao Ishikawa, Koichi Tojo, Shohei Oono, Masahiro Ikeda, and Masashi Sugiyama. Coupling-based invertible neural networks are universal diffeomorphism approximators. In *Advances in Neural Information Processing Systems*, 2020.

Jonathan H. Tu, Clarence W. Rowley, Dirk M. Luchtenburg, Steven L. Brunton, and J. Nathan Kutz. On dynamic mode decomposition: Theory and applications. *Journal of Computational Dynamics*, 1:391–421, 2014. doi: 10.3934/jcd.2014.1.391.

Matthew O. Williams, Ioannis G. Kevrekidis, and Clarence W. Rowley. A data-driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25:1307–1346, 2015. doi: 10.1007/s00332-015-9258-5.

A Proofs and auxiliary results

A.1 Fitting, reconstruction, and prediction

Projected least squares.

Proof of Theorem 1. Since $XV_r = U_r\Sigma_r$, differentiating equation 4 gives

$$\nabla_C F = 2\{(\Sigma_r^2 + \mu I)C - \Sigma_r U_r^\top Y\}. \tag{14}$$

The retained singular values are positive, so $\Sigma_r^2 + \mu I$ is positive definite. Setting the gradient to zero gives the unique solution in equation 5 and its projected normal equation. The map $C \mapsto V_r C$ is injective, so it determines a unique B in the retained subspace. For rank zero the only admissible coefficient is $B = 0$. □

Corollary 1 (Untruncated least squares)**.** *When $\mu = 0$ and all nonzero singular directions are retained, $B_\star = X^\dagger Y$ is the minimum-norm minimizer of $\min_B \|Y - XB\|_F^2$ and satisfies $X^\top(Y - XB_\star) = 0$.*

Proof. Under these assumptions, equation equation 5 is the pseudoinverse formula. Components of B in $\ker X$ leave the fit unchanged and increase its norm unless zero. □

Reconstruction and composition.

Proof of Proposition 1. An additive block is inverted by subtracting its conditioner; an affine block is inverted by subtracting its shift and dividing by its nonzero scale. Orthogonal mixing and invertible actnorm also have explicit inverses. Applying these inverses in reverse order proves $g^{-1} \circ g = I$. Adjacent $g \circ g^{-1}$ factors cancel in the N-fold composition of $g^{-1} \circ A \circ g$, giving equation 8 by induction. When A is nonsingular, applying the same argument to A^{-1} proves the backward identity. □

Accepted updates.

Proof of Theorem 2. Combining the two inequalities in equation 9 yields $F(W_{j+1}, A_{j+1}) \le F(W_j, A_j)$. These values form a nonincreasing sequence bounded below by zero and therefore converge. This argument uses the acceptance tests, so it also covers inexact candidates from an iterative readout solve or a truncated operator fit. □

Finite-horizon prediction.

Proof of Theorem 3. Write $z_n = g(S^n(u))$ and $r_n = z_{n+1} - Az_n$. Forward invariance ensures $\|r_n\| \le \varepsilon$ at every required exact state. Repeated substitution gives

$$z_N - A^N z_0 = \sum_{n=0}^{N-1} A^{N-1-n} r_n. \tag{15}$$

The triangle inequality bounds its norm by $\varepsilon \sum_{j=0}^{N-1} \|A^j\|$. Applying the assumed Lipschitz inequality for g^{-1} proves equation 10. For $N = 0$ both trajectories equal the initial state. □

If $\max_{0 \le j < N} \|A^j\| \le C$, the bound is at most $LCN\varepsilon$. A nonuniform estimate instead retains each residual norm in equation 15, identifying the steps that contribute most strongly to the final error.

A.2 Equilibria and coordinate representations

Fixed points.

Proof of Theorem 4. Conjugacy and injectivity of g imply that $S(u) = u$ if and only if $Ag(u) = g(u)$. Surjectivity identifies the complete fixed set as in equation 11. The kernel of $A - I$ is a connected linear subspace, and its image under the continuous map g^{-1} is connected. □

Corollary 2 (Restricted domain)**.** *Let Ω be invariant and $g : \Omega \to g(\Omega)$ be a homeomorphism satisfying the conjugacy there. If $g(\Omega) \cap \ker(A - I)$ is connected, then $\mathrm{Fix}(S) \cap \Omega$ is connected. The intersection condition holds when $g(\Omega)$ is convex.*

Proof. Use the same fixed-set identity with the kernel restricted to $g(\Omega)$. The inverse map preserves connectedness. An intersection of a convex set and a linear subspace is convex; an empty intersection gives no fixed points. □

Proof of Theorem 5. Since $S(p) = p$, the assumption gives $\|Mg(p)\| \le \varepsilon$. The orthogonal projection onto $(\ker M)^\perp$ is $M^\dagger M$. Hence

$$\operatorname{dist}(g(p), \ker M) = \|M^\dagger M g(p)\| \le \|M^\dagger\|\varepsilon. \tag{16}$$

For $M \neq 0$, $\|M^\dagger\| = 1/\sigma^+_{\min}(M)$. For $M = 0$, the kernel is the entire space and the distance is zero. □

Conditional coordinate approximation.

Proposition 2 (Approximation of an existing conjugacy)**.** *Let K be compact. Suppose a C^1 diffeomorphism h on a neighborhood of $K \cup S(K)$ satisfies*

$$\sup_{u\in K} \|h(S(u)) - Ah(u)\| \le \eta. \tag{17}$$

Assume coupling maps approximate h uniformly on this set and their conditioners can be approximated uniformly by the specified random-feature class as width increases. For every $\delta > 0$, a sufficiently expressive RF coupling map g_m then satisfies

$$\sup_{u\in K} \|g_m(S(u)) - Ag_m(u)\| \le \eta + (1 + \|A\|)\delta. \tag{18}$$

Proof. Choose g_m with $\sup_{v\in K\cup S(K)} \|g_m(v) - h(v)\| \le \delta$ under the stated approximation assumptions. Add and subtract $h(S(u))$ and $Ah(u)$ inside the residual. The approximation terms are bounded by δ and $\|A\|\delta$, respectively; equation equation 17 bounds the remaining term. □

The density assumptions specify the approximation topology needed here. A finite-width learning rate would additionally require quantitative feature approximation and sampling estimates.

Coordinate scaling.

Proposition 3 (An arbitrarily small absolute latent residual)**.** *For finite pairs (x_i, y_i) and any $\varepsilon > 0$, an invertible affine map g and matrix A can satisfy $\max_i \|g(y_i) - Ag(x_i)\| < \varepsilon$.*

Proof. If $M = \max_i \|y_i\| > 0$, take $g(u) = \delta u$, $A = 0$, and $0 < \delta < \varepsilon/M$. If $M = 0$, use $g = I$ and $A = 0$. □

Uniform rescaling $g \mapsto \delta g$ leaves a normalized latent residual unchanged when its denominator is nonzero, while the inverse Lipschitz factor scales as δ^{-1}. This explains the different physical interpretation of absolute and normalized latent errors.

B Supplementary PDE solutions and diagnostics

B.1 Burgers solutions, solver checks, and the Cole–Hopf anchor

Controlled viscosity comparison. The reference solutions below compare viscosity at a common initial state and periodic domain:

$$u_t + uu_x = \nu u_{xx}, \quad x \in [0, 2\pi), \qquad u(x, 0) = \sin x + \tfrac{1}{4}\sin(2x). \tag{19}$$

The common initial state isolates the effect of viscosity. Figure 4 shows the entire interval $0 \le t \le 5$ and four predetermined spatial profiles.

The nonlinear term transports amplitude at a state-dependent speed, creating a steeper transition near the center of the domain. Diffusion opposes this steepening. The lower-viscosity solution retains a sharper transition and larger amplitude for longer; the higher-viscosity solution smooths and decays more rapidly. A surrogate therefore has to represent both spatial deformation and dissipation. Matching the decay rates of selected modes addresses only part of this task.

The solver uses 256 Fourier grid points, quadratic dealiasing, and ETDRK4 with internal step 0.00125; every 20 internal steps produces one displayed state. Halving the internal step changes the

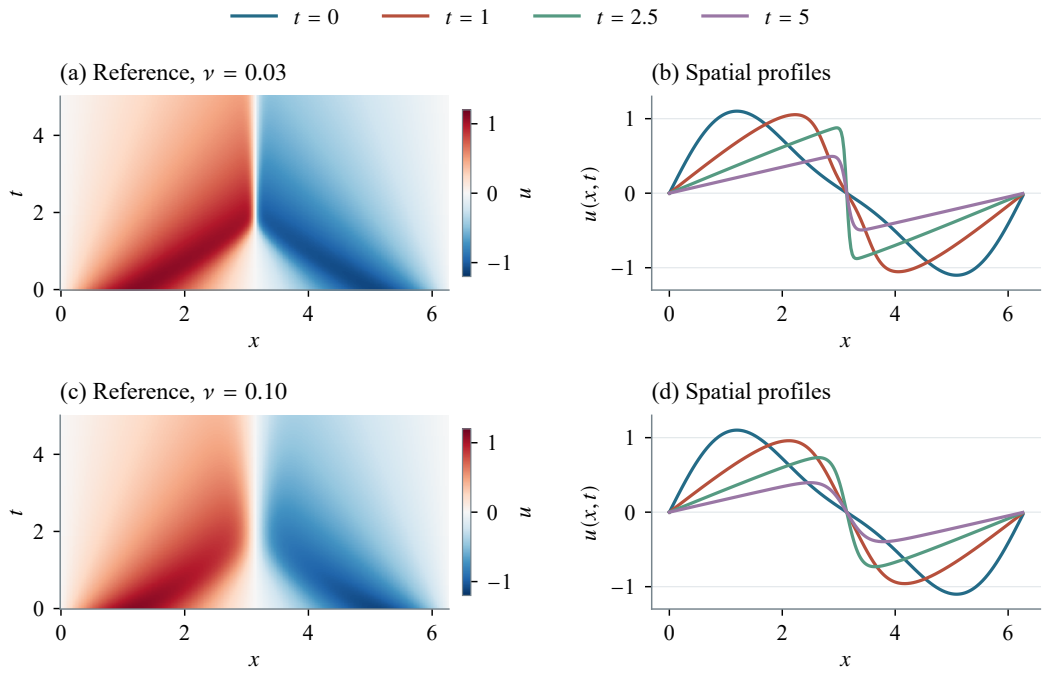


Figure 4: Burgers reference solutions with identical initial data. Left: space–time fields on common color limits. Right: profiles at $t = 0, 1, 2.5, 5$. All panels show reference-solver output.

complete stored solution by relative L_2 norms 1.32×10^{-11} and 3.97×10^{-13} for $\nu = 0.03$ and 0.10, respectively. The comparison measures temporal sensitivity at fixed spatial resolution.

Figure 5 gives a geometric view of the same solutions. The surfaces expose the changing profile shape, while the gradient maps localize steep transitions without hiding them behind a three-dimensional view. Both viscosities use the same amplitude and gradient color scales.

Energy, gradients, and the spatial spectrum. For a smooth periodic solution, multiplying equation equation 19 by u, averaging in space, and integrating by parts gives

$$E(t) = \tfrac{1}{2}\langle u^2 \rangle, \qquad \frac{dE}{dt} = -\nu \langle u_x^2 \rangle, \qquad \langle f \rangle = \frac{1}{2\pi} \int_0^{2\pi} f(x)\, dx. \tag{20}$$

The transport contribution vanishes because it is the integral of $\partial_x(u^3/3)$. The dissipation term is nonnegative, so the exact energy is nonincreasing. This identity gives a physical diagnostic independent of a coordinate map, a decoder, or a fitted linear operator.

Both stored trajectories have strictly decreasing energy at every output step. From the common initial energy $E(0) = 0.265625$, the values at $t = 5$ are 0.04526 and 0.03312 for $\nu = 0.03$ and 0.10, respectively. The maximum resolved gradient reaches 16.55 at $t = 2.025$ for the lower viscosity and 3.684 at $t = 2.05$ for the higher viscosity. The spatial Fourier coefficients are $\widehat{u}_k = 256^{-1} \sum_j u_j e^{-2\pi i j k/256}$. At lower viscosity, steeper spatial gradients populate more high-frequency modes before they decay. The final smooth profiles alone would conceal this intermediate transfer of energy to shorter spatial scales. Values near the floating-point floor should not be interpreted as resolved physical content.

The spatial Fourier spectrum $|\widehat{u}_k|^2$ describes the distribution of state energy at a given time, whereas the operator spectrum describes latent evolution. Energy balance and gradient resolution therefore provide physical checks beyond the spectral comparison in the main text.

Reference discretization and independent solver checks. The reference integrator separates a diagonal linear Fourier operator from the nonlinear product. With angular wave numbers $k = 2\pi j/L$, its linear symbols are $-\nu k^2$ for Burgers, $k^2 - k^4$ for KS, and $1 - \epsilon^2 k^2$ for Allen–Cahn. The nonlinear

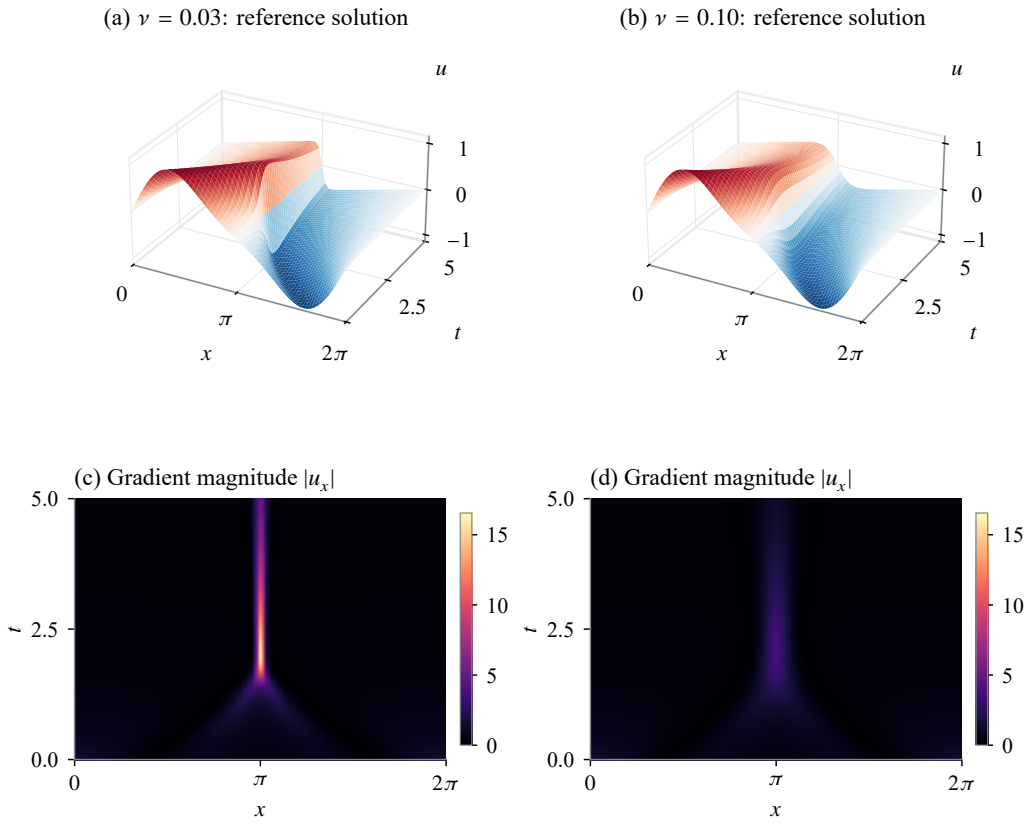


Figure 5: Geometry of the Burgers reference solutions. Top: space–time surfaces for $\nu = 0.03$ and 0.10, with color encoding u on a common scale. Bottom: $|u_x|$ computed by Fourier differentiation on the full stored grid, with a shared color scale. Surface meshes are thinned only for rendering; the underlying solutions are unchanged. These are reference-solver solutions, not learned-model predictions.

Fourier terms are respectively $-ik\widehat{u^2}/2$, $-ik\widehat{u^2}/2$, and $-\widehat{u^3}$. Products are evaluated in physical space and transformed back. The quadratic solvers retain the two-thirds band; the cubic solver retains modes up to one quarter of the grid size. A Hermitian projection maintains a real physical state after each internal step.

For a Fourier state v, write the semidiscrete equation as $\dot{v} = Lv + N(v)$. The implemented four-stage exponential update has the form

$$\begin{aligned} a &= E_2 v + QN(v), \qquad b = E_2 v + QN(a), \\ c &= E_2 a + Q\{2N(b) - N(v)\}, \\ v^+ &= Ev + f_1 N(v) + 2f_2\{N(a) + N(b)\} + f_3 N(c), \end{aligned} \tag{21}$$

where $E = e^{hL}$, $E_2 = e^{hL/2}$, h is the internal step, and Q, f_1, f_2, f_3 are diagonal coefficient functions of hL. We evaluate them with a 32-point complex contour quadrature to avoid direct subtraction of nearly equal exponentials near zero. Equation equation 21 describes the implemented update, including the factor two multiplying f_2; changing coefficient conventions without changing this factor would change the method.

The observed convergence order is computed as $\log_2(e(h)/e(h/2))$. The values approach four before roundoff dominates. The verification report also checks energy balance independently: the relative residual is 9.67×10^{-8} for Burgers and 2.06×10^{-10} for KS. For the latter equation, integration by parts gives

$$\frac{d}{dt}\tfrac{1}{2}\langle u^2 \rangle = \langle u_x^2 \rangle - \langle u_{xx}^2 \rangle. \tag{22}$$

Table 3: Temporal solver convergence against a common finer-step reference.

PDE	Internal step	Relative L_2 discrepancy	Order to next step
Burgers	0.0200	8.9337×10^{-8}	3.9956
	0.0100	5.6008×10^{-9}	3.9984
	0.0050	3.5043×10^{-10}	—
KS	0.0100	4.1700×10^{-11}	3.9747
	0.0050	2.6524×10^{-12}	3.9591
	0.0025	1.7055×10^{-13}	—

The first term injects energy while the second removes it. Reproducing this balance is a different check from time-step convergence and is useful for detecting a sign or derivative convention error.

Short-horizon convergence establishes numerical consistency on the verification states. The solution examples below also have their own time-step checks because their gradients and integration intervals differ.

The Cole–Hopf anchor and spectral agreement. For a smooth mean-zero periodic Burgers field, let $P_x = u$ and define $q = \exp[-P/(2\nu)]$, up to a positive scalar factor. Choosing the additive, time-dependent gauge of P appropriately gives

$$q_t = \nu q_{xx}, \qquad u = -2\nu\, \partial_x \log q, \qquad \widehat{q}_k(t + \Delta t) = e^{-\nu k^2 \Delta t} \widehat{q}_k(t). \tag{23}$$

Indeed, substitution of the logarithmic derivative into Burgers' equation cancels the quadratic terms when q solves the heat equation. Positivity allows the logarithm, and mean-zero u makes the potential periodic. For the initial condition in equation equation 19, one can take $P(x, 0) = -\cos x - \frac{1}{8}\cos(2x)$.

The constant heat mode is unchanged; nonconstant modes decay at rates νk^2. Real sine and cosine modes have the same decay factor, so the reference ordering must retain the intended multiplicities. With $\langle q \rangle = 1$, diffusion smooths the initial profile toward the constant state $q = 1$. The strong concentration at low viscosity explains why evaluating the exponential and logarithm requires attention to numerical scaling. Subtracting the minimum initial potential before exponentiating, followed by mean normalization, avoids an unnecessarily large exponential without changing u.

Three restrictions matter when using this anchor. First, normalization removes a scalar degree of freedom: q and cq decode to the same u. Second, a finite grid and a finite retained operator need not realize the complete infinite-dimensional heat dynamics. Third, eigenvalues do not identify eigenvectors or the nonlinear coordinate map. Matrices related by similarity share a spectrum, and even matrices with the same eigenvalues need not have the same transient behavior.

The scalar gauge and mode multiplicities are fixed before comparing coordinates. This removes a normalization ambiguity from the alignment fit.

Held-out coordinate alignment. Coordinate alignment complements the spectral comparison in the main text. A linear alignment is fitted on training coordinates, its truncation tolerance is selected on validation data, and its R^2 is then evaluated on held-out coordinates. We summarize the three seeds by their mean and sample standard deviation.

Specifically, let Φ contain analytic Cole–Hopf coordinates and Z the learned coordinates as rows. The fit predicts Z from Φ:

$$\Phi_{\mathrm{train}} M^\top \approx Z_{\mathrm{train}}, \qquad R^2 = 1 - \frac{\|Z_{\mathrm{test}} - \Phi_{\mathrm{test}} M^\top\|_F^2}{\|Z_{\mathrm{test}} - \mathbf{1}\overline{z}_{\mathrm{train}}^\top\|_F^2}. \tag{24}$$

Here $\overline{z}_{\mathrm{train}}$ is the training mean, matching the implementation's denominator convention. The truncation tolerance is chosen from $\{10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}\}$ by validation R^2. Writing H for the discrete heat-step matrix, compatibility is measured as

$$e_{\mathrm{comm}} = \frac{\|MH - AM\|_F}{\|MH\|_F}. \tag{25}$$

The code uses a float64 tiny-value floor if either denominator vanishes. These definitions keep the direction of the alignment and its normalization fixed across seeds.

At $\nu = 0.03$, the held-out alignment is $R^2 = 0.721 \pm 0.006$ and the relative commutation defect is 0.107 ± 0.003. At $\nu = 0.10$, these values are 0.774 ± 0.006 and 0.366 ± 0.453, respectively. Both mean alignment scores fall below the specified target of $R^2 = 0.95$. The commutation defect also varies substantially more across seeds at the higher viscosity. Thus, agreement with the heat spectrum does not establish recovery of the Cole–Hopf coordinates or their evolution under the tested linear alignment.

B.2 KS SOLUTION GEOMETRY, FORECAST ERRORS, AND OPERATOR DIAGNOSTICS

The curves below supplement the validation trajectory in Figure 3. For its reference states u_n, teacher-forced and autonomous errors are defined by

$$e_n^{\text{one}} = \frac{\|g^{-1}(Ag(u_n)) - u_{n+1}\|_2}{\|u_{n+1}\|_2}, \qquad e_n^{\text{roll}} = \frac{\|g^{-1}(A^n g(u_0)) - u_n\|_2}{\|u_n\|_2}. \tag{26}$$

Reference norms are nonzero on this window. The error guide in Figure 6 is 0.4, and the horizontal axis uses physical time. Profiles at $t = 5$ and 25 distinguish spatial displacement from amplitude error; the energy trace summarizes amplitude over the domain.

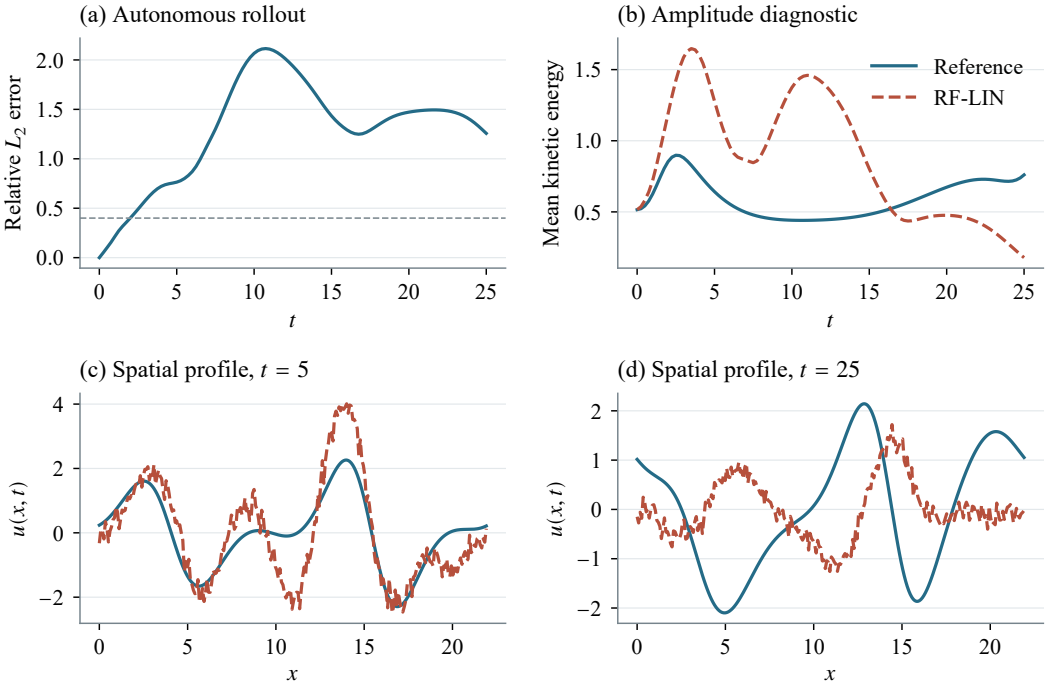


Figure 6: Physical diagnostics for the same KS validation example. The horizontal error guide is 0.4; the time axis is physical time, not Lyapunov units. Energy and profiles expose amplitude and positional errors that encode–decode reconstruction does not measure.

Figure 7 plots eigenvalues and matrix-power norms. We evaluate the powers at horizons 0, 1, 2, 5, 10, 20, 50, and 100. The amplification is already large at short horizons even though all eigenvalues lie inside the unit circle. This measurement is made in the stored coordinates. A similarity transformation would leave the eigenvalues unchanged while generally changing the Euclidean power norms.

For context, the KS equation is $u_t + uu_x + u_{xx} + u_{xxxx} = 0$. Its Fourier growth rate $k^2 - k^4$ combines growth at low wave numbers with damping at high wave numbers, while nonlinear transport couples modes. The reference energy balance is given in equation 22. This combination motivates checking spatial profiles and energy along with a single scalar forecast error.

Figure 8 displays the geometry of this same validation trajectory and its autonomous forecast. The four profile times span early, intermediate, and final portions of the stored window; they are fixed at $t = 2.5, 5, 12.5, 25$. The panels retain the forecast discrepancies and report the corresponding relative errors defined in equation 26.

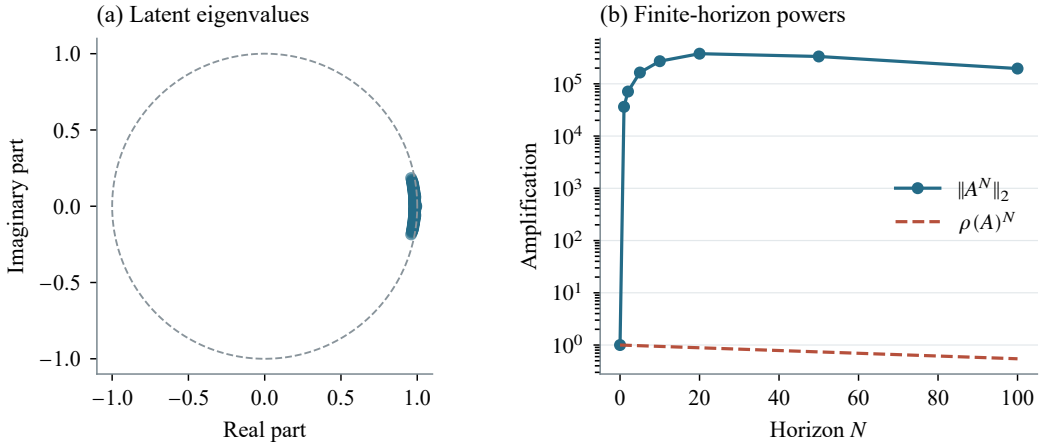


Figure 7: The latent operator of the same readout-only model. Left: eigenvalues with the unit circle as a guide. Right: measured spectral norms of selected matrix powers and the corresponding spectral-radius powers. The norms are measured in the fitted latent coordinates.

B.3 Allen–Cahn basin-local solutions and the equilibrium bridge

Reference solutions within each basin. The Allen–Cahn equation provides a complementary setting in which the geometry of equilibria matters:

$$u_t = \epsilon^2 u_{xx} + u - u^3, \quad \epsilon = 0.1, \quad x \in [0, 2\pi). \tag{27}$$

The spatially constant states $u = \pm 1$ are stable equilibria, whereas $u = 0$ is unstable to constant perturbations. We examine basin-local relaxation from $u_+(x, 0) = 1 + 0.15 \cos x$ and $u_-(x, 0) = -u_+(x, 0)$. The equation is invariant under $u \mapsto -u$, so these solutions remain exact sign reversals and have identical energy and equilibrium-distance traces.

The free energy per unit length and its dissipation identity are

$$\mathcal{E}(u) = \left\langle \tfrac{1}{2}\epsilon^2 u_x^2 + \tfrac{1}{4}(u^2 - 1)^2 \right\rangle, \qquad \frac{d\mathcal{E}}{dt} = -\langle u_t^2 \rangle. \tag{28}$$

To obtain the second equality, differentiate under the integral, integrate the gradient term by parts, and substitute equation equation 27. The decrease explains why both prescribed perturbations relax toward their nearby equilibrium. It does not imply that a single globally linearized model can represent transitions between all basins.

In the stored reference solutions, the RMS distance to the respective equilibrium decreases from 0.1061 at $t = 0$ to 4.82×10^{-6} at $t = 5$. The free energy decreases at every output step, from 1.135×10^{-2} to 2.339×10^{-11}. These are reference-solver diagnostics, not learned-model forecast errors.

These solutions use 256 Fourier points, cubic dealiasing that retains modes up to one quarter of the grid size, internal step 0.005, and output spacing 0.05. Halving the internal step gives a relative discrepancy 6.47×10^{-12} over the stored arrays. As with the Burgers atlas, this checks time-step sensitivity for the reference example. The learned-model comparison is reported in Table 2.

Displacement along a latent equilibrium bridge. The path in Figure 2 uses $z_\pm = g(p_\pm)$ and $z_\alpha = (1 - \alpha)z_- + \alpha z_+$ for $0 \le \alpha \le 1$. By linearity,

$$\|(A - I)z_\alpha\| \le (1 - \alpha)\|(A - I)z_-\| + \alpha\|(A - I)z_+\|. \tag{29}$$

If g^{-1} is L-Lipschitz between z_α and Az_α, the learned physical displacement is at most L times this endpoint bound. The reference displacement is evaluated by integrating the decoded state $g^{-1}(z_\alpha)$ for the same time interval. The experiment uses 21 equally spaced path coordinates. Endpoint defects, interior displacement, and the model/reference displacement ratio are evaluated separately.

The segment lies in the range of a global invertible map, but a model restricted to an invariant basin need not include its decoded interior. The restricted-domain condition in Corollary 2 accounts for this distinction. Thus the bridge probes the connecting region, while the positive-basin control assesses trajectories near one attractor.

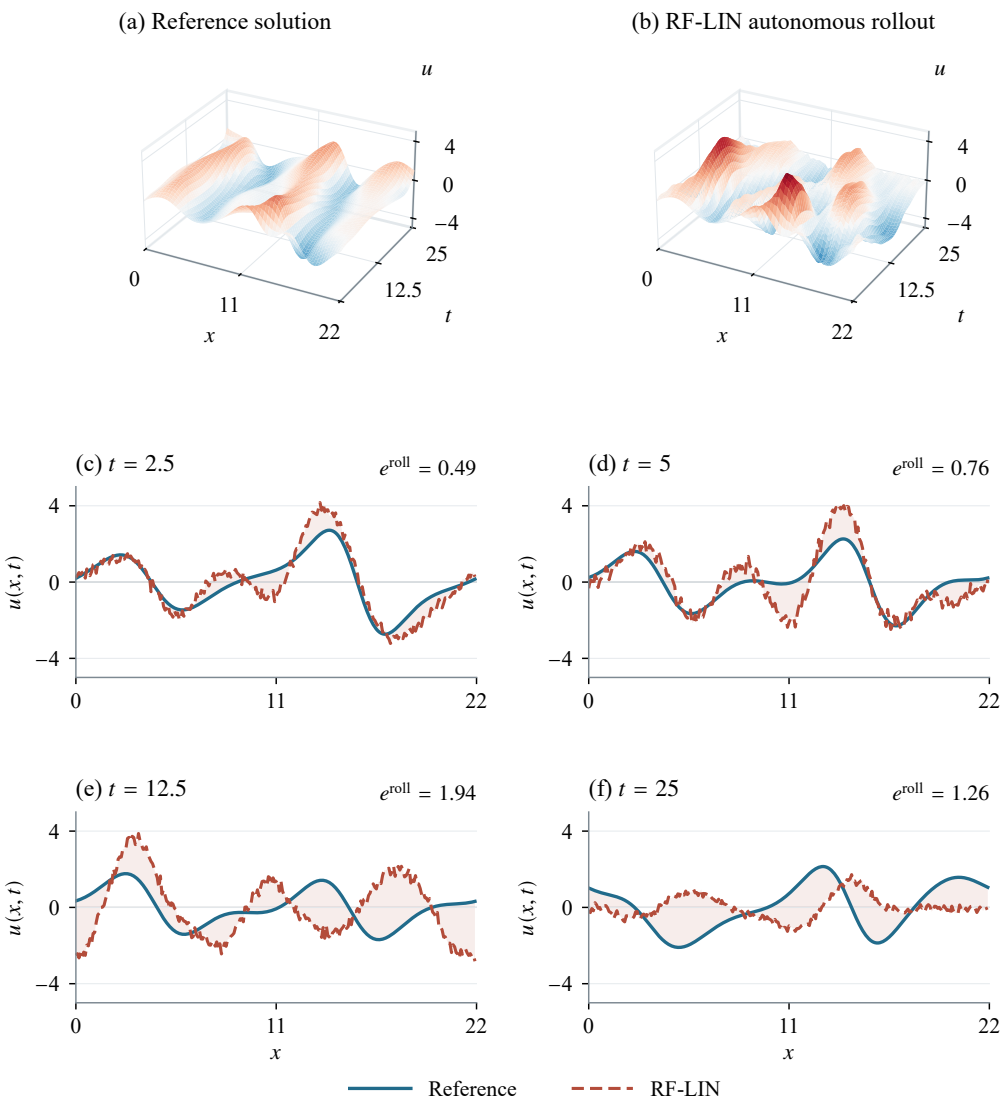


Figure 8: KS solution morphology for the saved validation example. Top: reference and RF-LIN surfaces use identical view angles, vertical limits, and amplitude color scales. Bottom: full-resolution spatial profiles at four specified times; solid blue denotes the reference and dashed red the autonomous forecast. Shading marks the gap between the curves, not uncertainty. Each profile reports its relative L_2 error. This is one saved validation window, not a multi-seed test aggregate.

C Method Details and Numerical Analysis

C.1 Invertible Coordinates and Linear Fitting

Coupling inverses, Jacobians, and conditioning. An additive coupling is invertible for every readout matrix, independently of whether the conditioner is injective. Write $f(a) = W\phi(\Omega a + b)$ and $C(a, c) = (a, c + f(a))$. Then

$$C^{-1}(a, c') = (a, c' - f(a)), \qquad DC = \begin{pmatrix} I & 0 \\ Df & I \end{pmatrix}, \qquad DC^{-1} = \begin{pmatrix} I & 0 \\ -Df & I \end{pmatrix}. \tag{30}$$

Both triangular Jacobians have determinant one. This establishes an exact algebraic inverse without solving a nonlinear system. It also shows why volume preservation is weaker than conditioning: a shear can have determinant one and very unequal singular values.

For the normalized features $\phi = m^{-1/2}\tanh$, the derivative bound gives

$$\|Df(a)\|_2 \le m^{-1/2}\|W\|_2\|\Omega\|_2, \qquad \|DC\|_2, \|DC^{-1}\|_2 \le 1 + \|Df(a)\|_2. \tag{31}$$

The second inequality follows by writing the Jacobian as $I + N$ with a single off-diagonal block and applying the triangle inequality. For a composition, operator-norm bounds multiply along the trajectory. These upper bounds can be pessimistic, but they identify the quantities that can become large. Increasing the number of invertible layers does not preserve a small Lipschitz constant automatically.

Orthogonal mixing has singular values equal to one and therefore does not amplify the Euclidean norm by itself. Its role is to change which directions enter the next conditioner. Learned actnorm scales, when present, contribute their own scale ratios. A full-map conditioning assessment must include those scales and the ordered product of coupling Jacobians, rather than examining a conditioner in isolation.

For the affine update in equation equation 3, the diagonal scale lies between $e^{-s_{\max}}$ and $e^{s_{\max}}$. Cross derivatives additionally contain $D\ell$, Dt, and the updated state, so the full-map conditioning also depends on these quantities. Floating-point cancellation can affect subtractive inverses even when their exact-arithmetic formulas are simple.

Projected operator fitting as singular-value filtering. The projected problem has two distinct controls: retaining singular directions and shrinking coefficients within the retained subspace. If $\sigma_i > 0$ is retained, its contribution to $C_\star$ is multiplied by

$$f_\mu(\sigma_i) = \frac{\sigma_i}{\sigma_i^2 + \mu}, \qquad h_\mu(\sigma_i) = \frac{\sigma_i^2}{\sigma_i^2 + \mu}. \tag{32}$$

The first factor controls coefficient sensitivity to the projected target; the second controls transfer of that target back to fitted data. Hard truncation sets both contributions to zero for discarded directions. Without regularization, the coefficient gain is $1/\sigma_i$ while the data-transfer factor is one. For $\mu > 0$, ridge bounds the coefficient gain by $1/(2\sqrt{\mu})$ and continuously attenuates poorly resolved directions; TSVD instead removes directions below its threshold.

For $\mu = 0$, the optimal projected residual decomposes as

$$Y - XB_\star = (I - U_r U_r^\top)Y, \qquad \|Y - XB_\star\|_F^2 = \|(I - U_r U_r^\top)Y\|_F^2. \tag{33}$$

It includes target components outside the retained data subspace. A small projected stationarity residual says that no allowed infinitesimal change of C improves the stated quadratic objective. It does not say that the fit residual is small, that discarded directions are irrelevant, or that the solution generalizes beyond the training pairs.

An elementary example makes the distinction visible. Let $X = \mathrm{diag}(1, \delta)$ and $Y = I_2$, with $0 < \delta < \tau$ and retain only the first direction. Then $B_\star = \mathrm{diag}(1, 0)$ has zero projected gradient, but the residual still contains the entire second target column. The unrestricted minimizer is $\mathrm{diag}(1, \delta^{-1})$ and fits exactly at the cost of high sensitivity. Thus both solutions can be correct for their respective stated problems. Reporting the retained rank, singular cutoff, ridge level, discarded-target contribution, and residual prevents these claims from being confused.

Matrix-free readout adjoint and solver checks. Using the matrices in Section 2.3, the Frobenius adjoint of $\mathcal{L}_B$ is

$$\mathcal{L}_B^*(R) = E^\top R^\top H_y - E^\top B R^\top H_x, \qquad \nabla_W F = 2\mathcal{L}_B^*(R_0 + \mathcal{L}_B(W)). \tag{34}$$

For the first term, cycling the trace of $R^\top H_y W^\top E^\top$ gives the coefficient $E^\top R^\top H_y$ of W. Applying the same argument to $R^\top H_x W^\top E^\top B$ gives the second term. This proves the adjoint identity and the gradient formula without constructing a Kronecker design matrix.

LSMR uses these forward and adjoint actions to solve the readout problem. Preconditioning changes the coordinates in which the iteration is performed, so the final gradient is recomputed from the original residual. Solver termination and gradient residual assess the linear subproblem; the outer iteration separately checks the relative latent loss on the full training split, with damping or rejection of a joint update that increases this loss.

The dominant stored arrays are the feature matrices and state residuals, requiring $O(nm+nd)$ entries, together with the qm readout and d^2 operator coefficients. Explicit vectorization would instead require an $nd \times qm$ design. Matrix-free fitting avoids that storage but still depends on the number of iterative products. Readout tolerance, iteration limit, subset size, and preconditioner are therefore part of the run configuration.

C.2 Prediction stability, computational cost, and inverse problems

Error propagation beyond the spectral radius. The power sum in Theorem 3 cannot generally be replaced by a geometric series in $\rho(A)$. Consider two matrices with identical eigenvalues:

$$A_0 = aI_2, \qquad A_1 = \begin{pmatrix} a & b \\ 0 & a \end{pmatrix}, \qquad A_1^j = \begin{pmatrix} a^j & jba^{j-1} \\ 0 & a^j \end{pmatrix} \quad (j \geq 1). \tag{35}$$

Both are asymptotically stable when $|a| < 1$. Nevertheless, the off-diagonal term can initially grow. For example, with $a = 0.95$ and $b = 1$, $\|A_0^{20}\|_2 = 0.95^{20} \approx 0.358$, whereas $\|A_1^{20}\|_2 \geq 20(0.95)^{19} \approx 7.55$. Thus identical eigenvalues can produce very different finite-time amplification. This is an analytic matrix example, not a fitted PDE model.

Equation equation 15 gives the exact residual accumulation. Residual direction matters because each matrix power can amplify some directions and contract others. Replacing every $\|r_n\|$ by a common ε and then applying a global decoder bound sacrifices this directional information. The resulting theorem is conservative but explicit about what must be controlled.

If $A = V\Lambda V^{-1}$ is diagonalizable, then $\|A^j\|_2 \leq \kappa_2(V)\rho(A)^j$. The eigenvector condition number cannot be discarded. A defective matrix requires a different estimate, as equation equation 35 shows. Furthermore, the relevant amplification can be much larger than a geometric series in the spectral radius. The Jordan example isolates this effect with an explicit matrix whose powers can be calculated directly.

Collapsed prediction: algebra, work, and numerical error. The identity $\widehat{S}^N = g^{-1}A^N g$ has several computational realizations. The relevant choice depends on whether the application needs only an endpoint or an entire trajectory. If C_g denotes the work of one coordinate-map evaluation, dense matrix–vector multiplication costs order d^2, and dense matrix multiplication costs order d^3, the following leading-order counts describe different tasks.

Table 4: Indicative work for one initial state. Counts omit constants, allocation, and batching. These are operation counts.

Requested computation	Leading work
Repeated physical one-step calls	$N(2C_g + O(d^2))$
Latent recurrence, endpoint only	$2C_g + O(Nd^2)$
Repeated squaring, endpoint only	$2C_g + O(d^3 \log N + d^2)$
Latent recurrence, all outputs	$(N+1)C_g + O(Nd^2)$
New endpoint with cached A^N	$2C_g + O(d^2)$

A fair endpoint comparison should include latent recurrence. Eliminating repeated encode–decode pairs already yields that implementation without forming A^N. Repeated squaring reduces dependence on horizon but uses matrix–matrix operations, which can be more costly for short horizons and large d. Caching powers is useful when many initial states share one operator and horizon, but the setup cost and storage must be reported or amortized explicitly.

For integer N with binary digits b_j, repeated squaring uses

$$A^N = \prod_{j:b_j=1} A^{2^j}, \qquad A^{2^{j+1}} = A^{2^j}A^{2^j}. \tag{36}$$

The solution heatmaps require decoding every requested latent point, as counted in the all-outputs row of Table 4.

Floating-point evaluation also breaks literal equality between implementation paths. Repeated multiplication and repeated squaring round in different orders. An informative structural check compares

their latent results and decoded results separately, alongside encode–decode error. Disagreement can come from operator-power sensitivity, decoder amplification, or arithmetic accumulation even when the algebraic identity is correct.

Measured timings additionally depend on hardware, precision, and batch size.

Backward recovery and sparse observations. Invertibility of g removes an encoder information-loss mechanism, but it does not make a dissipative time-step operator well conditioned. Let $A^N = P\Sigma Q^\top$ be an SVD. Given a noisy encoded endpoint z_N^δ, a truncated or ridge-regularized estimate of the initial latent state is

$$z_0^{(\tau)} = \sum_{\sigma_i \geq \tau} q_i \frac{p_i^\top z_N^\delta}{\sigma_i}, \qquad z_0^{(\lambda)} = \sum_i q_i \frac{\sigma_i}{\sigma_i^2 + \lambda} p_i^\top z_N^\delta. \tag{37}$$

Decoding these estimates defines a regularized recovery procedure. It is not an exact inverse identity when singular directions are discarded or shrunk. Increasing the lookback can reduce singular values and amplify noise, even if the forward trajectory is stable.

For TSVD and encoded noise $e = z_N^\delta - A^N z_0$, the latent recovery error has an explicit bias–noise decomposition:

$$z_0^{(\tau)} - z_0 = -\sum_{\sigma_i < \tau} q_i q_i^\top z_0 + \sum_{\sigma_i \geq \tau} q_i \frac{p_i^\top e}{\sigma_i}. \tag{38}$$

The retained noise term is at most $\|e\|_2/\tau$. The first term is lost signal and does not disappear when observation noise vanishes. In physical space, a decoder Lipschitz factor multiplies the resulting latent bound. Noise originally added to the state must also pass through g, so encoded noise need not remain isotropic or have the same magnitude.

Sparse assimilation can instead fit an initial latent state directly. Let H_n select or combine observed physical coordinates at time n, let y_n be the observations, and let z_b be a background state. One explicit regularized objective is

$$J(z) = \tfrac{1}{2} \sum_{n \in \mathcal{O}} \|H_n g^{-1}(A^n z) - y_n\|^2_{R_n^{-1}} + \tfrac{\lambda}{2} \|z - z_b\|_2^2, \tag{39}$$

where R_n is positive definite and $\|v\|^2_{R_n^{-1}} = v^\top R_n^{-1} v$. Its gradient involves $(A^n)^\top Dg^{-1}(A^n z)^\top H_n^\top$ applied to the weighted observation residual. The latent dynamics are linear, but the decoder makes this objective generally nonlinear and potentially nonconvex.

A meaningful application comparison must specify the observation pattern, noise distribution, regularization-selection rule, lookback, and initialization. Recovery can be nonidentifiable if observations miss relevant directions. Good reconstruction with full clean states does not address this ambiguity. These formulas define the regularized recovery problems associated with the learned representation.

D Evaluation protocol, implementation, and data sources

Evaluation units, aggregation, and uncertainty. An evaluation unit must be specified before combining errors. If trajectory i has states $u_{i,n}$ and predictions $\widehat{u}_{i,n}$, a time-mean relative error over M trajectories and H predicted steps is

$$\overline{e} = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{H} \sum_{n=1}^{H} \frac{\|\widehat{u}_{i,n} - u_{i,n}\|_2}{\|u_{i,n}\|_2}. \tag{40}$$

The initial state is excluded because it is supplied exactly. Unlike a global ratio of Frobenius norms, this average gives equal weight to each trajectory and forecast step. The displayed KS window has nonzero reference norms throughout.

For three run-level values m_1, m_2, m_3, we report

$$\overline{m} = \tfrac{1}{3} \sum_{s=1}^{3} m_s, \qquad s_m = \sqrt{\tfrac{1}{2} \sum_{s=1}^{3} (m_s - \overline{m})^2}. \tag{41}$$

The sample standard deviation measures seed variation for the stated data and hyperparameters; correlated time points and overlapping windows remain within each run-level statistic.

For the spectral metric defined in Section 4, each m_s is a within-run median. The maximum discrepancy supplements that statistic by exposing tail mismatches. Eigenvalue phases are not part of this magnitude-based metric.

For the KS error plot, the first crossing of the 0.4 guide is measured in physical time. A trajectory that stays below the guide throughout the displayed interval gives a horizon-censored observation. No Lyapunov-time normalization or long-run distributional metric is inferred from this window.

Training configurations. The Burgers base configuration permits six additive blocks and 512 features per conditioner, with up to ten readout/operator alternations per block. The readout subset has 2,048 pairs, relative stopping tolerance 10^{-10}, and at most 5,000 iterations. Blocks require a 1% improvement in relative validation latent loss to be retained. Per-run configurations record bandwidth selection, accepted blocks, and refinement settings. The aggregate identifies readout fitting with gradient refinement at $\nu = 0.03$ and readout-only fitting at $\nu = 0.10$.

The Allen–Cahn configuration permits four blocks with 256 features, bandwidth 1, five alternations per block, and a 2,048-pair readout subset. It uses the same 1% block-acceptance margin and readout tolerance. The three evaluation seeds are matched between the two training domains.

KS diagnostic. The long trajectory contains contiguous training, validation, and test ranges of 80,000, 10,000, and 10,000 transitions. The displayed validation window starts at index 80,000 and includes the next 100 transitions. The validation-selected model is evaluated without additional fitting, using the same initial state and output times for the reference and prediction. This single-window diagnostic is separate from the Burgers and Allen–Cahn seed aggregates.

Precision and reference solutions. Final operator solves, evaluation arrays, and reference integrations use float64. The Burgers and Allen–Cahn reference solutions use the analytic initial conditions specified above. Temporal checks halve the internal step while leaving the spatial grid and output times unchanged. Spatial Fourier transforms use the stated normalization, and all displayed errors are calculated from the saved arrays.

Data sources and reproducibility. The Burgers and Allen–Cahn tables summarize recorded seed-level results. The spectral and obstruction curves correspond to those experimental runs; the coordinate-alignment statistics use the recorded means and sample standard deviations. The retained materials for these diagnostics contain numerical summaries and exported curves, but not the original training checkpoints, limiting complete reproduction from these materials alone. The KS predictions and supplementary reference solutions are retained as numerical trajectories. The geometric views in Figures 5 and 8 use these same trajectories, without retraining or additional experiments. Mesh subsampling affects only surface rendering; profile values and reported errors are evaluated at the original resolution.