WHEN ARE PHYSICS GRADIENTS CORRECTIVE? UPDATE-SPACE DEPENDENCE IN AUTOREGRESSIVE NEURAL PDE OPERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive neural PDE predictors can remain locally accurate under a physical-coefficient shift while their long rollouts deteriorate. Governing-equation residuals offer label-free test-time objectives, but a small residual does not establish that its gradient will correct predictive error, nor does a corrective direction guarantee that an optimizer can use it. We separate these questions in controlled viscous Burgers and forced Navier–Stokes systems. Our central experiment holds real update-space dimension fixed. In two Fourier-operator training regimes, architecture-defined pointwise ($d = 648$) and output-head ($d = 1{,}249$) spaces have larger mean horizon-100 corrections than ten frozen random spaces at the same size. A two-way model-seed/initial-condition bootstrap supports the pointwise effect under both one- and five-step protocols; the head effects remain positive in mean but statistically inconclusive. A structured $d = 4{,}096$ counterexample rules out a blanket preference for structure. Complementary interventions separate scalar residual value from gradient direction: pooled residual–error association is near zero but heterogeneous across seeds, while equal-norm physics updates rank above matched random motion and reversal in aggregate. Rank-four adapters can nevertheless have high alignment and almost no rollout gain. A well-fit non-spectral DilResNet also exhibits positive alignment, directional specificity, and a 12.10% mean physics-only adaptation gain. Correctability is therefore a joint property of the objective and accessible update space, not residual magnitude, cosine, or dimension alone.

1 INTRODUCTION

Neural PDE surrogates amortize expensive simulation across initial states and coefficients, but deployment often requires feeding predictions back into the learned time-stepper. Error under this closed-loop use need not track teacher-forced one-step error. Physical-parameter shifts are particularly revealing: in our Burgers and controlled forced Navier–Stokes settings, locally accurate predictions coexist with a markedly larger long-horizon error outside the training coefficient range (Table 1). This is an empirical observation in the studied systems, not a claim of universal instability.

Known equations provide self-supervision without future solution labels. Physics-informed neural operators already use residuals for training and instance-wise refinement (Li et al., 2021), and recent work adapts PDE foundation models using physics-informed full or low-rank updates (Medvedev et al., 2026; Zhu et al., 2025; Song et al., 2026). The open question here is different: when is the *parameter gradient* of a physical objective actually corrective for a deployed autoregressive model? A residual's scalar value may indicate equation violation without reliably ranking future predictive error. Even a gradient that locally agrees with an unavailable supervised direction may fail when only a restricted parameter set can move.

We distinguish three quantities: residual magnitude, directional agreement with an analysis-only oracle gradient, and realized error reduction under a frozen update protocol. Our central test compares architecture-defined and random update spaces at exactly matched real-coordinate dimension. It asks whether dimension alone explains correction, using both a matched-norm single step and five-

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.

step Adam. Complementary analyses inspect pooled and group-specific residual–error association and intervene with equal-norm random, orthogonal, and sign-reversed directions.

Our empirical contributions are: (i) exact-dimension comparisons showing a replicated pointwise-space advantage over random spaces, alongside a head-space result weakened by clustered uncertainty and a structured counterexample; (ii) a value-versus-direction analysis across Burgers and controlled Navier–Stokes; (iii) equal-norm interventional tests showing why cosine alignment alone is insufficient; and (iv) one well-fit non-spectral convolutional replication. The study does not introduce a new test-time adaptation algorithm. Its novelty is a controlled mechanism study of update-space dependence.

2 Related work

PDE surrogates and rollout robustness. Fourier neural operators (FNOs) learn solution maps through spectral kernels (Li et al., 2020); DeepONet is a distinct branch–trunk construction (Lu et al., 2019). PDEBench and PDEArena provide broader evaluation and architecture context (Takamoto et al., 2022; Gupta & Brandstetter, 2022). DRIFT-Net addresses drift with coupled spectral and image branches (Li & Salim, 2025); HERO uses history-enriched rollout training (Zhang et al., 2026). Our HERO-compatible model follows the latter training idea but is not an exact official reproduction. We study post-training gradient correctivity rather than proposing a more stable training architecture.

Physics at inference. PINO combines operator learning and PDE constraints, including test-instance optimization (Li et al., 2021). Recent foundation-model work performs physics-informed fine-tuning with residual and boundary losses (Medvedev et al., 2026; Zhu et al., 2025); unsupervised adaptation also combines residual objectives with low-rank updates (Song et al., 2026). Operator splitting instead composes pretrained dynamics at test time without modifying weights (Serrano et al., 2026); residual-informed diffusion modulation avoids iterative guidance (Sawhney et al., 2025). These papers preclude novelty claims for physics-only refinement, low-rank adaptation, or test-time physical computation themselves. Our question is mechanistic and interventional: whether a residual gradient agrees with an unavailable supervised correction and whether an equally sized update space exposes that correction.

Update subspaces and optimizer coordinates. Random-subspace training measures how many degrees of freedom a task needs (Li et al., 2018), while intrinsic-dimension analyses and LoRA show that low-dimensional or low-rank parameterizations can support effective adaptation (Aghajanyan et al., 2021; Hu et al., 2022). Those results establish that small spaces can work; they do not imply that two equal-dimensional spaces transfer a particular physics gradient equally well. Coordinate choice also matters for adaptive optimization: Adam is not rotation invariant and can behave differently after a basis change (Zhang et al., 2025). This motivates reporting both a matched-norm one-step comparison, which removes Adam, and a separate multi-step Adam outcome. Our fixed-size result concerns location and orientation relative to a physics correction, not an intrinsic-dimension estimate or a universal preference for architectural structure.

Gradient interactions. Physics-informed training can suffer gradient pathologies (Wang et al., 2020). Such work motivates inspecting directions, but training-loss balancing is distinct from the deployment-time comparison between an equation-derived gradient and held-out rollout labels. We use those labels only to study the mechanism, never to select a physics-adaptation step.

3 Problem setup

For state u_t and physical coefficient λ, a learned time-stepper is recursively applied as

$$\hat{u}_{t+1} = G_\theta(\hat{u}_t; \lambda), \qquad \hat{u}_0 = u_0. \tag{1}$$

Evaluation uses relative L_2 error of the resulting trajectory or its stated horizon. We study coefficients outside the training interval with held-out, paired initial conditions. The deployment-time objective $L_{\mathrm{phys}}(\theta)$ is computed from the known PDE on a short predicted rollout, without future

Table 1: **Studied systems and shift.** Relative L_2 means are not directly comparable across PDEs. Burgers local error is teacher-forced at the final snapshot; Navier–Stokes local error is at the first step. FNO/HERO share a spectral architecture; DilResNet is a non-spectral local implementation.

PDE and model	Training viscosity	Test condition	Local → long error
Burgers, FNO	$[.001, .004]$	$\nu = .010$	$.0026 \to .0556$ $(h = 50)$
NS, bounded FNO	$[.002, .006]$	ID	$.00958 \to .546$ $(h = 100)$
NS, bounded FNO	$[.002, .006]$	high OOD	$.01096 \to .786$ $(h = 100)$
NS, DilResNet	$[.002, .006]$	ID; OOD mean	$.00130 \to .10761; .25928$ $(h = 100)$

labels. An analysis-only objective $L_{\text{true}}(\theta)$ measures that rollout against held-out truth. The latter is never available to physics test-time adaptation.

Let $g_{\text{phys}} = \nabla_\theta L_{\text{phys}}$ and $g_{\text{true}} = \nabla_\theta L_{\text{true}}$. Their cosine is $A = g_{\text{phys}}^\top g_{\text{true}} / (\|g_{\text{phys}}\| \|g_{\text{true}}\|)$. For an allowed parameter region S, we measure its restricted gradients $g_{\text{phys}}^S, g_{\text{true}}^S$, cosine $A(S)$, oracle norm ratio $C(S) = \|g_{\text{true}}^S\| / \|g_{\text{true}}\|$, and directional-magnitude product $Q(S) = (g_{\text{phys}}^S)^\top g_{\text{true}}^S$. For an augmented adapter, the reported norm ratio compares differently parameterized gradients and is *not* a literal orthogonal projection. Both C and Q require labels; neither is a deployable criterion nor a sufficient predictor. The oracle objective matches the short physical-loss window (10 steps in the controlled Navier–Stokes analyses and 11 in Burgers); the principal outcome is much longer-horizon error. Thus alignment to this oracle is a mechanistic diagnostic, not an identity with the evaluation metric.

For a small explicit step $\Delta\theta = -\eta g_{\text{phys}}$, Taylor expansion gives $L_{\text{true}}(\theta + \Delta\theta) \approx L_{\text{true}}(\theta) - \eta g_{\text{true}}^\top g_{\text{phys}}$. Within a coordinate space S, the corresponding first-order term is $-\eta (g_{\text{true}}^S)^\top g_{\text{phys}}^S$. This motivates separate study of direction and accessible magnitude. It is a *local interpretation*, not a theorem for five-step Adam updates, where curvature, coordinates, anchors, and nonlinear rollouts intervene.

Our directional test applies a single explicit displacement along $+g_{\text{phys}}$, its reversal, and independent isotropic or orthogonal directions, with norms matched within each model state. The reported physics percentile is the fraction of sampled random directions whose error reduction is smaller. Multi-step physics TTA is evaluated separately under interpolation-calibrated settings; absolute gains from these two protocols are not interchangeable.

4 Experimental setup

We use periodic 1D viscous Burgers and a controlled periodic 2D forced Navier–Stokes vorticity system. The latter has continuous-viscosity training on 64 independent trajectories; evaluation uses three model seeds and ten held-out paired initial conditions at each of one in-distribution and six out-of-distribution viscosities. Its 128^2 numerical truth is downsampled to a 64^2 learned grid. Low-viscosity conclusions inherit a measured resolution discrepancy, discussed below. Burgers supplies an independent PDE setting and a direct gradient comparison; most intervention and subspace tests use Navier–Stokes. Appendix B fixes PDEs, splits, numerical methods, and controls.

The Navier–Stokes models are a bounded-coefficient FNO, an FNO under a *HERO-compatible* rollout-training regime (not the exact official implementation), and a locally implemented periodic DilResNet following the PDEArena core (Gupta & Brandstetter, 2022). The DilResNet is not an official-package reproduction. Its fit threshold was frozen and passed on interpolation data before out-of-distribution mechanism analysis. All physics step sizes and multi-step adaptation settings were selected on interpolation validation only. Oracle gradients and oracle-labelled quantities are computed afterward for analysis; no out-of-distribution future labels enter physics adaptation.

In the controlled FNO, the ID one-step/horizon-100 errors are $0.00958/0.546$; at the strong-low and strong-high coefficient shifts, one-step error changes only to $0.01038/0.01096$ while horizon-100 error rises to $0.732/0.786$ (ratios $1.34/1.44$). The local metric and closed-loop metric are not interchangeable: the former starts from reference states, whereas the latter repeatedly consumes the model's own predictions. The HERO-compatible training regime and the non-spectral candidate preserve a qualitative local/long-horizon separation, but their absolute error scales differ. This base-

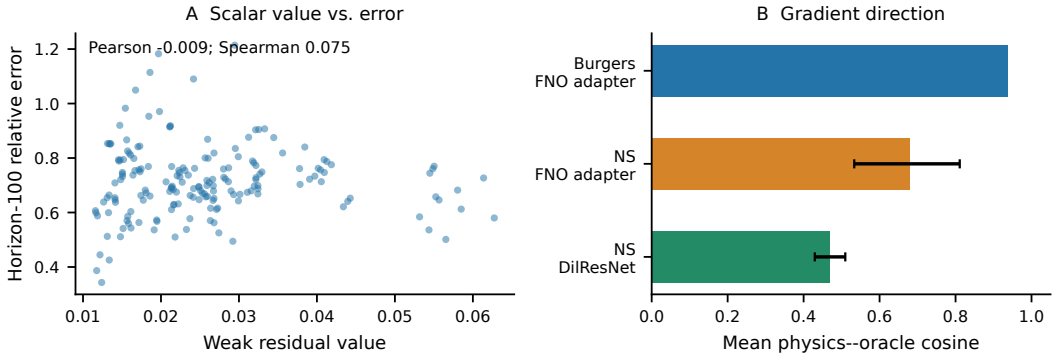

Figure 1: **A residual value is not its gradient.** Left: pooled state-level weak-residual value versus horizon-100 relative error for the controlled Navier–Stokes FNO; grouped correlations and clustered uncertainty are in Appendix H. Right: positive mean physics–oracle cosine in separately measured Burgers and Navier–Stokes diagnostics. Oracle labels are analysis-only; bars use distinct parameterizations and horizons.

line comparison motivates an unlabeled correction objective; it does not itself show that a physics residual supplies the right update.

We report mean relative error, gradient cosine, random-direction percentile, sign-reversal frequency, and trajectory-wise relative gain $(E_{\mathrm{pre}}-E_{\mathrm{post}})/E_{\mathrm{pre}}$. The primary uncertainty audit uses a two-way crossed cluster bootstrap: model seeds and initial conditions are independently resampled, while all viscosities in each selected cell are retained (Owen & Eckles, 2012). Archived row-bootstrap intervals are retained for traceability but are not treated as independent-population intervals. With only three trained seeds, even the clustered intervals remain descriptive. One-step tests use 32 random directions per FNO state and 16 per DilResNet state; the fixed-dimension experiment uses five random-coordinate and five implicit dense-orthonormal definitions per size. Appendix H gives grouped estimates and both uncertainty interpretations.

5 Residual value and corrective direction

Figure 1 separates a scalar diagnostic from an optimization direction. For the controlled Navier–Stokes FNO, the weak-form residual across 180 out-of-distribution trajectories has pooled Pearson -0.009 and Spearman 0.075 association with horizon-100 error. Dependence-aware intervals are wide (Pearson $[-0.437, 0.520]$, Spearman $[-0.436, 0.567]$), and seed-wise Spearman values are -0.023, -0.165, and 0.500; viscosity-wise values range from -0.031 to 0.152. The pooled near-zero number therefore does *not* establish universal decorrelation. It shows instead that this scalar is not a stable cross-model ranking rule. A separately evaluated rank-four parameterization has mean short-horizon physics–oracle gradient cosine 0.681 at horizon 10, positive in 86.1% of measured states; its isotropic-vector control has mean cosine 0.025. In Burgers, the matching horizon-11 cosine is 0.939. Residual magnitude does not by itself establish whether its gradient will reduce a particular model's rollout error; the scalar and gradient analyses also use related but nonidentical parameterizations.

Direction under intervention. Gradient cosine is still correlational. We therefore apply one conservative, explicit, norm-matched update per state and compare $+g_{\mathrm{phys}}$ with $-g_{\mathrm{phys}}$, isotropic random motion, and orthogonal random motion. In the FNO and HERO-compatible regime, $+g_{\mathrm{phys}}$ lies at mean percentiles 0.779 and 0.761 among 32 matched random directions and beats its reversal in 78.9% and 78.3% of states. Mean one-step gains are 0.454% and 1.155%, but two-way clustered intervals $[-0.127, 1.000]\%$ and $[-0.903, 3.357]\%$ include zero; these are aggregate directional effects, not population-level guarantees. Reversal means are negative and physics-orthogonal controls are near zero. Structured rollout-derived random objectives sometimes helped and are not treated as isotropic nulls.

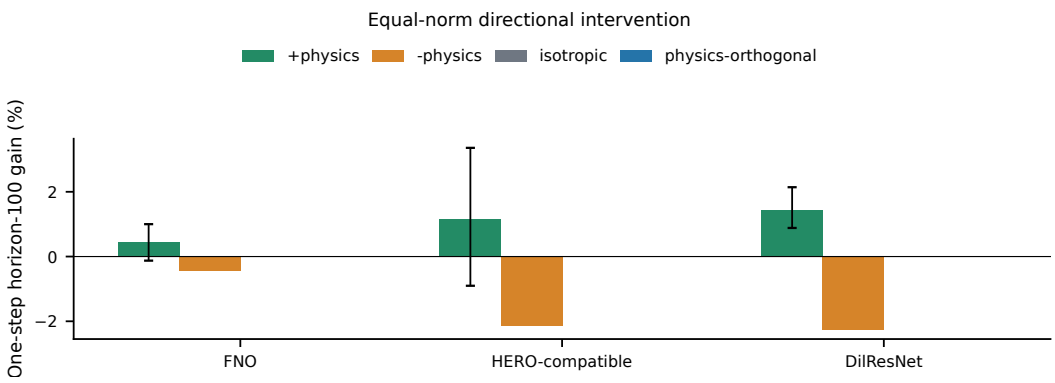

Figure 2: **Equal-norm intervention tests directional specificity.** Mean horizon-100 relative-error gain from a single explicit displacement along physics, its reversal, and matched random controls. Error bars are two-way model-seed/initial-condition cluster-bootstrap 95% intervals for the physics mean. FNO/HERO use 32 random directions per state; DilResNet uses 16. Step magnitudes differ by model.

Table 2: **Cosine is insufficient.** Five-step horizon-100 gain in the same controlled Navier–Stokes evaluation. C is an analysis-only oracle-gradient norm ratio; for the augmented rank-four adapter it is not a literal parent-space projection. The full-space counterexample shows C is not sufficient either.

Model	Update parameters	Dimension	Cosine	C	Gain
FNO	rank-4 adapter	192	.695	.029	0.118%
FNO	last block	166,536	.689	.345	8.43%
FNO	full	667,513	.711	1.000	3.74%
HERO-compatible	rank-4 adapter	192	.752	.031	0.018%
HERO-compatible	last block	166,536	.626	.362	7.18%
HERO-compatible	full	667,513	.694	1.000	2.06%

The non-spectral model supplies a further interventional check: its mean physics–oracle cosine is 0.470 (clustered interval $[0.241, 0.627]$), and its mean one-step physics gain is 1.431% ($[0.883, 2.142]\%$). Physics is at percentile 0.933 among equal-norm random directions and beats reversal in 96.1% of states. Table 3 keeps these measurements beside, but distinct from, five-step Adam adaptation. The first-order argument in Section 3 is local and does not guarantee any multi-step result.

6 Correction depends on the update space

Alignment is not sufficient. Under the same Navier–Stokes evaluation, the rank-four adapter has mean physics–oracle cosine 0.695 for FNO and 0.752 for HERO-compatible FNO, yet its five-step horizon-100 gains are only 0.118% and 0.018%. Updating the last block yields lower cosine (0.689/0.626) but larger gains (8.43%/7.18%). The adapter's analysis-only oracle-gradient norm ratios are 0.029/0.031, versus 0.345/0.362 in the last block. It is thus possible for the accessible directions to agree while the available correction is too small or the transfer protocol too ineffective. This contrast motivates the update-space study; the adapter's augmented coordinates prevent interpreting its ratio as a literal projection.

Coverage alone is no law: the whole network has ratio one by construction but gains only 3.74%/2.06% under the same five-step protocol, less than selected blocks. The output head also has substantial oracle coverage without universally strong transferred gain. Labels make these analyses explanatory, not available at deployment. Five-step Adam furthermore changes effective steps by coordinate, so we directly control dimension and inspect a matched-norm single step below.

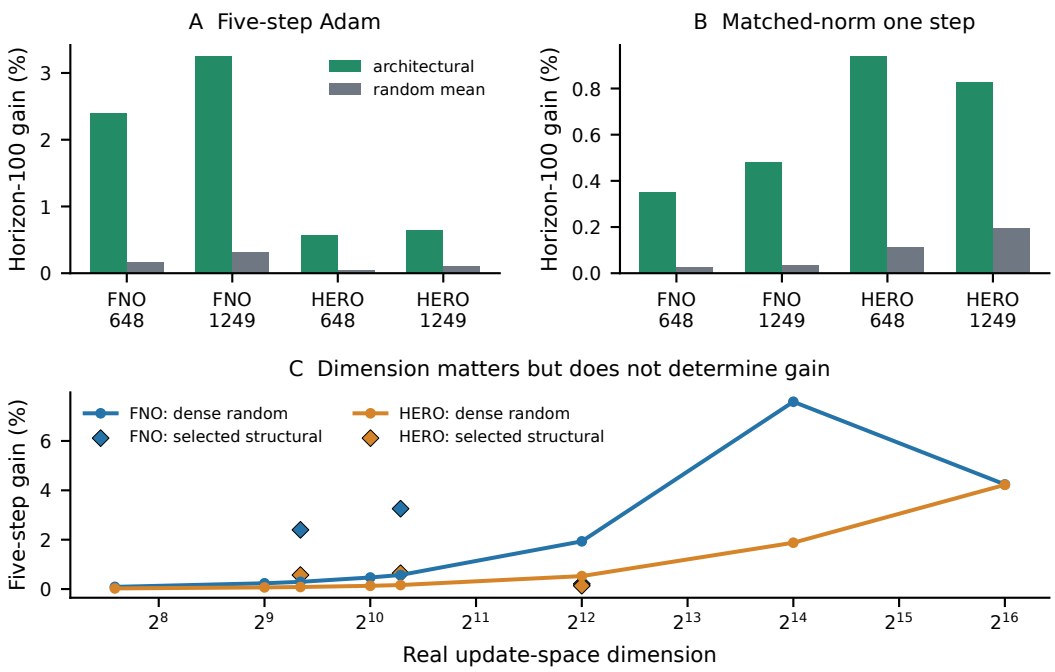

Figure 3: **Equal dimension does not imply equal correction.** Top: exact-size architecture-defined spaces versus the mean of ten random definitions; panel A shows five-step Adam and panel B the matched-norm one-step protocol. Bottom: selected spaces across dimension, including the poor contiguous $d = 4{,}096$ architectural counterexample. Bars and curves are descriptive means; clustered intervals are in Appendix H.

Exact-dimension controls. Within the same 333,673-real-coordinate last-block-plus-head parent, we compare architecture-defined regions against five random-coordinate and five implicit dense-orthonormal spaces at each *identical real dimension*. At $d = 648$, the final pointwise region's five-step gain is 2.398% versus a 0.159% mean over ten random definitions for FNO; HERO is 0.568% versus 0.050%. At $d = 1{,}249$, the output head reaches 3.256% versus 0.311% for FNO and 0.642% versus 0.105% for HERO. Each architecture-defined aggregate mean exceeds all ten corresponding random definitions. Under two-way model-seed/initial-condition clustering, structured-minus-random intervals exclude zero for both models and both protocols at $d = 648$; all four $d = 1{,}249$ means remain positive, but their intervals include zero. The archived row-resampled intervals exclude zero in all eight cases and are less conservative. The one-step comparison fixes the update norm, removing the simplest step-magnitude explanation; Appendix H reports every clustered interval.

Dimension plainly matters: larger random spaces generally expose more useful directions. It is not the sole determinant. Nor is "structured" a sufficient descriptor: a contiguous $d = 4{,}096$ region sits at only the 30th percentile of matched random spaces for FNO. Location, orientation, gradient concentration, curvature, and optimizer coordinates can all contribute; the experiment establishes a fixed-size difference, not which of these factors uniquely causes it.

7 A NON-SPECTRAL REPLICATION

The periodic DilResNet is a local adaptation of the PDEArena dilated residual core, not the complete official package. It has 206,801 parameters and uses exactly the same 64 training trajectories as the matched-data FNO. Its three-seed interpolation horizon-1 and horizon-100 errors are 0.148 and 0.272 times the matched FNO errors, passing a predeclared at-most-two-times fit gate *before* OOD analysis. Test ID errors are 0.00130 at one step and 0.10761 at horizon 100; pooled OOD horizon-100 error is 0.25928.

Table 3: **Directional specificity and transferred correction.** Cosine/percentile/reversal are measured in each model's stated intervention space. TTA gains use separately calibrated five-step Adam spaces, so the last column is not the outcome of the one-step intervention. Entries are trajectory-wise means (180 OOD states/model).

Model	Cosine	Random pct.	$+ > -$	TTA gain / space
FNO	.725	.779	78.9%	10.57% / block+head
HERO-compatible FNO	.719	.761	78.3%	7.41% / block+head
PDEArena-style DilResNet	.470	.933	96.1%	12.10% / full

Within its natural last-block-plus-head region, 95.0% of physics–oracle cosines are positive, and Figure 2 shows the physics direction's matched-random and reversal advantages. Under a separately interpolation-selected five-step *full-model* protocol, mean horizon-100 error falls from 0.25928 to 0.22695: a 12.10% mean trajectory-wise gain (two-way clustered interval $[8.10, 16.00]\%$), with 88.9% of states improved. This is an existence result beyond spectral/FNO parameterization on one controlled forced Navier–Stokes task, not architecture-general evidence.

8 Limitations

The evidence spans two PDE families and only one well-fit non-FNO architecture. Our DilResNet follows a published core but is not an exact official-package reproduction; the HERO-compatible training regime likewise is not an official reproduction. The exact-size study isolates raw dimension but not location, curvature, or optimizer-coordinate effects, especially for five-step Adam in rotated dense spaces. The matched-norm step removes Adam but not subspace orientation. Only three model seeds are available: two-way clustering widens uncertainty, leaves the $d = 648$ result supported, and makes the $d = 1{,}249$ head result inconclusive. Low-viscosity truth also retains a measured $N = 64$ versus $N = 128$ endpoint discrepancy. Finally, tiny adapters nearly fail on Navier–Stokes, while underfit and external-checkpoint cases show no general TTA guarantee. Appendix G reports these boundaries.

Deployment implication. The experiments support treating update-space choice and step size as architecture-specific hyperparameters calibrated on in-range validation trajectories. They do *not* support selecting an update per state from residual magnitude, oracle cosine, or oracle coverage: the latter two require labels, and a frozen label-free diagnostic reached ROC-AUC 0.641 but gain Spearman 0.009, performing worse than always adapting. Nor should practitioners assume that LoRA, the largest space, or any structured region is automatically best. Without a validated label-free gate, the safe conclusion is diagnostic—compare candidate spaces before deployment—rather than a promise of blind OOD improvement.

9 Conclusion

In the evaluated autoregressive PDE predictors, residual value, gradient direction, and realized correction are distinct. The central fixed-dimensional comparison shows that a 648-coordinate architecture-defined pointwise space transfers more correction than matched random spaces across two training regimes and both one- and five-step protocols; the weaker 1,249-coordinate evidence and a 4,096-coordinate counterexample prevent a general structure claim. Equal-norm interventions and the DilResNet replication further show that direction can matter, while clustered uncertainty limits population-wide conclusions. Practitioners can validation-select an update space, but cannot infer per-state correctability from residual magnitude or oracle-only diagnostics. Predicting it without future labels remains open.

Reproducibility statement

Training, interpolation, and OOD splits were fixed; viscosity tests use paired held-out initial conditions and three model seeds. Physics hyperparameters were calibrated on interpolation states, with

oracle labels reserved for analysis. Canonical raw outcomes, experiment configurations, seeded checkpoints, and a registry are supplied; included scripts regenerate figures, grouped correlations, and the two-way clustered uncertainty audit. Appendix I records solver, architecture, optimization, statistical, and compute details.

AI USE STATEMENT

Generative AI tools assisted hypothesis refinement, experimental design and critique, implementation and debugging, organization and interpretation of results, literature discovery, and drafting and editing this manuscript. The paper retains only claims traced to saved experimental outputs or verified primary literature. The human authors remain responsible for independently reviewing AI-assisted code and prose, checking quantitative claims and citations, and approving the submitted work.

REFERENCES

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of ACL-IJCNLP*, 2021. URL https://arxiv.org/abs/2012.13255.

Jayesh K. Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized pde modeling. *arXiv preprint arXiv:2209.15616*, 2022. URL https://arxiv.org/abs/2209.15616.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1804.08838.

Jiayi Li and Flora D. Salim. Drift-net: A spectral–coupled neural operator for pdes learning. *arXiv preprint arXiv:2509.24868*, 2025. URL https://arxiv.org/abs/2509.24868.

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

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *arXiv preprint arXiv:2111.03794*, 2021. URL https://arxiv.org/abs/2111.03794.

Lu Lu, Pengzhan Jin, and George Em Karniadakis. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv preprint arXiv:1910.03193*, 2019. URL https://arxiv.org/abs/1910.03193.

Vlad Medvedev, Leon Armbruster, Christopher Straub, Georg Kruse, and Andreas Rosskopf. Physics-informed fine-tuning of foundation models for partial differential equations. *arXiv preprint arXiv:2603.15431*, 2026. URL https://arxiv.org/abs/2603.15431.

Art B. Owen and Dean Eckles. Bootstrapping data arrays of arbitrary order. *The Annals of Applied Statistics*, 6(3):895–927, 2012. URL https://arxiv.org/abs/1106.2125.

Medha Sawhney, Abhilash Neog, Mridul Khurana, and Anuj Karpatne. Beyond loss guidance: Using pde residuals as spectral attention in diffusion neural operators. *arXiv preprint arXiv:2512.01370*, 2025. URL https://arxiv.org/abs/2512.01370.

Louis Serrano, Jiequn Han, Edouard Oyallon, Shirley Ho, and Rudy Morel. Test-time generalization for physics through neural operator splitting. *arXiv preprint arXiv:2602.00884*, 2026. URL https://arxiv.org/abs/2602.00884.

Ziye Song, Zhao Wei, Xin Yu, Ivor Tsang, and Yueming Lyu. Unsupervised adaptation of pde foundation models. *arXiv preprint arXiv:2608.07053*, 2026. URL https://arxiv.org/abs/2608.07053.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Dan MacKinlay, Francesco Alesiani, Dirk Pflueger, and Mathias Niepert. Pdebench: An extensive benchmark for scientific machine learning. *arXiv preprint arXiv:2210.07182*, 2022. URL https://arxiv.org/abs/2210.07182.

Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient pathologies in physics-informed neural networks. *arXiv preprint arXiv:2001.04536*, 2020. URL https://arxiv.org/abs/2001.04536.

Jiaquan Zhang, Shuxu Chen, Haifan Meng, Yi Lu, Zhihan Lyu, Fan Mo, Wei Dong, Yang Yang, and Chaoning Zhang. Hero: History-enriched rollout training for long-horizon autoregressive neural operators. *arXiv preprint arXiv:2607.29135*, 2026. URL https://arxiv.org/abs/2607.29135.

Tianyue H. Zhang, Lucas Maes, Alan Milligan, Alexia Jolicoeur-Martineau, Ioannis Mitliagkas, Damien Scieur, Simon Lacoste-Julien, and Charles Guille-Escuret. Understanding adam requires better rotation dependent assumptions. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2410.19964.

Min Zhu, Jingmin Sun, Zecheng Zhang, Hayden Schaeffer, and Lu Lu. Pi-mfm: Physics-informed multimodal foundation model for solving partial differential equations. *arXiv preprint arXiv:2512.23056*, 2025. URL https://arxiv.org/abs/2512.23056.

A PDE DEFINITIONS

The Burgers study uses periodic 1D viscous Burgers dynamics, $u_t + uu_x = \nu u_{xx}$, with physical coefficient ν. Initial conditions are sampled from a six-mode Fourier family; the training set has four viscosities $\{.001, .002, .003, .004\}$ and 128 initial conditions each. A separate continuous-viscosity replication uses 64 coefficient values over $[.001, .004]$, eight initial conditions each, matching the 512-trajectory count. The main-text Burgers number refers to the original discrete-viscosity baseline, not to that replication.

The controlled Navier–Stokes system is periodic on $[0, 2\pi)^2$ and written in vorticity ω: $\partial_t \omega + \mathbf{v} \cdot \nabla\omega = \nu\Delta\omega + 0.1\cos(4y)$, with incompressible velocity obtained from the vorticity/streamfunction relation. The physical shift is in ν alone. We do not claim that this forcing, domain, or shift represents arbitrary fluids.

B DATA, SPLITS, AND NUMERICAL REFERENCE

For Burgers, an ETDRK4 reference solver uses 1,024 points, time step 5×10^{-5}, and snapshots every 0.02 through $T = 1$; learned states use 128 points. The original evaluation uses 32 paired initial conditions at each held-out viscosity $\{.005, .006, .008, .010\}$. The continuous-viscosity replication does not replace that evaluation pool.

The Navier–Stokes reference is a 128^2 pseudo-spectral RK4 integration with two-thirds de-aliasing, time step 0.0025, snapshots every 0.05, and terminal time $T = 5$ (100 learned transitions). Reference snapshots are spectrally downsampled to 64^2. Training uses 64 independent initial conditions, each with one log-uniform viscosity in $[.002, .006]$. Interpolation validation uses eight disjoint initial conditions and $\nu \in \{.0025, .0035, .005\}$. Ten further paired initial conditions are evaluated at ID $\nu = .004$, low OOD $\{.0015, .001, .00075\}$, and high OOD $\{.008, .010, .012\}$; no test state is used for tuning.

Solver calibration included an analytic steady-state check, with relative right-hand-side residual 2.94×10^{-14} and integrated drift 3.67×10^{-16}. Trajectories at both grid resolutions were finite through $T = 5$; maximum reported CFL proxy was 0.103. At the strongest low-viscosity shift, $N = 64$ versus $N = 128$ endpoint disagreement reaches 15.1% (time mean 6.7%). All learned-model errors are against downsampled $N = 128$ truth, but this discrepancy limits the low-viscosity physical-fidelity interpretation.

C MODEL ARCHITECTURES

The Burgers base FNO has four Fourier blocks, width 48, 16 retained modes, and 161,809 parameters. Its rank-four residual adapter has 436 parameters. Continuous-viscosity conditioning variants and a larger FNO are retained in the experiment registry; the larger-model effect differs between discrete and continuous training and is not used as a paper claim about scaling.

For controlled Navier–Stokes, the main FNO and HERO-compatible FNO share the same spectral architecture (667,513 model parameters in the restricted-subspace analysis). The compatible training regime uses lagged/current rollout comparisons inspired by Zhang et al. (2026); no claim of exact official implementation is made. The final spectral block has 166,536 real trainable parameters, its pointwise component 648, and the output head 1,249. The rank-four adapter is an augmented 192-parameter parameterization, not a coordinate subset of the parent network. The last-block-plus-head parent contains 333,673 *real* coordinates when complex spectral weights are expanded into real and imaginary components.

The non-spectral candidate is a 206,801-parameter, locally implemented periodic DilResNet following the PDEArena core (Gupta & Brandstetter, 2022). It has two residual blocks, each with seven circularly padded convolutions at dilations $[1, 2, 4, 8, 4, 2, 1]$, GroupNorm, GELU, bounded viscosity conditioning, and a normalized residual prediction head. Its natural last-block-plus-head mechanism region has 103,321 parameters. It is not the exact official PDEArena package.

D TRAINING AND ADAPTATION PROTOCOLS

All headline controlled models use seeds 0, 1, and 2. The original Burgers FNO uses Adam, cosine learning-rate decay, and 120 epochs; stronger training and larger-capacity controls are archived separately. The Navier–Stokes FNO checkpoints and HERO-compatible regime use the frozen train/validation split; configuration files and checkpoints in the anonymous supplement specify every architecture and optimizer setting. The DilResNet uses AdamW for 80 epochs, selecting checkpoints by interpolation horizon-100 error. Its fit gate compares the three-seed mean interpolation errors on the same 24 trajectories per seed against the matched-data FNO: ratios 0.148 at horizon 1 and 0.272 at horizon 100, both below the predeclared 2.0 ceiling.

The controlled Navier–Stokes weak-residual TTA protocol uses five Adam steps, learning rate 10^{-5}, and consistency-anchor weight 1 for the FNO subspace analysis. The DilResNet protocol uses five steps and learning rate 3×10^{-6} for last block, last block plus head, and full model; the full space was selected as primary on interpolation error. Interventional explicit-step magnitudes were interpolation selected: FNO 10^{-5}, HERO-compatible FNO 10^{-4}, and DilResNet 10^{-4}. For the exact-dimension experiment one magnitude is fixed *across all spaces and dimensions* within each FNO model: 1.015678×10^{-4} for FNO and 1.384953×10^{-3} for HERO-compatible FNO. No OOD labels selected a physics hyperparameter. The analysis-only oracle naturally uses held-out labels and is never used as a deployable TTA condition.

E PHYSICAL OBJECTIVES AND GRADIENT CONTROLS

The Burgers program compared a legacy central-time-difference residual, a discrete trapezoidal residual, a normalized discrete residual, and a weak-form residual. The weak form had the lowest vulnerability to a constructed random-objective collapse: legacy/discrete scalar scores could rank a collapsed trajectory as more consistent than a genuine physics-adapted one. We therefore do not extrapolate a scalar residual–error relation from one residual definition to another.

The Navier–Stokes primary objective tests the vorticity equation against 13 predeclared low-frequency real Fourier functions, including the forcing mode, on a short predicted rollout. It is label-free but deliberately insensitive to low-pass smoothing; energy and enstrophy balance terms are complementary diagnostics. Its mean value on downsampled reference truth is 4.65×10^{-8}, versus 0.273 after 10% noise, 0.148 after a y shift, and 337.5 after temporal shuffling. A prediction-consistency anchor limits drift from the frozen forecast; its weight in controlled FNO TTA is 1.

The oracle gradient differentiates a short supervised rollout loss against held-out future states. Its trajectory data are unavailable to deployment-time physics adaptation. Isotropic random vectors are drawn in the same active parameter space and norm matched to physics motion. Physics-orthogonal directions remove the physics component; oracle-orthogonal directions remove the analysis-only oracle component. Reversal uses the same magnitude and the opposite sign. A rolled-target random loss is *not* an isotropic null: it shares autoregressive structure with the model and can align with the oracle gradient. The controlled intervention figures consequently use parameter-space random/orthogonal controls, not this confounded loss.

F Additional mechanism and update-space results

In Burgers, the 11-step physics–oracle cosine has mean 0.939, median 0.992, and 99.7% positive over the archived short-horizon comparisons. Under continuous-viscosity training, raw and log conditioning preserve the local-versus-rollout gap. In the controlled Navier–Stokes FNO adapter analysis, mean physics–oracle cosine is 0.681 at horizon 10 (row-resampled 95% interval $[0.533, 0.811]$) and 0.744 at horizon 100 (interval $[0.658, 0.825]$); matched isotropic-vector means are 0.025 and -0.015. These gradients are measured in a different parameterization from the interventional last-block-plus-head experiment. The latter's mean cosines are 0.725 for FNO and 0.719 for HERO-compatible FNO. Dependence-aware intervals for paper-facing headline quantities are in Appendix H.

In the seven-space five-step analysis, last-block-plus-head mean gains are 10.57% for FNO and 7.41% for HERO-compatible FNO. The full model has maximal oracle norm ratio one yet gains 3.74% and 2.06%. The rank-four adapter gains 0.118% and 0.018%. These are means of trajectory-wise relative gains, not ratios of mean errors. The exact-dimension design evaluates dimensions $192, 512, 648, 1024, 1249, 4096, 16384, 65536$ and the full parent. Five random-coordinate and five implicit dense-orthonormal spaces are fixed at each dimension. The dense basis is an implicit real orthogonal circulant transform obtained from conjugate-symmetric unit-modulus Fourier phases and seeded coordinate selection; dense-space Adam runs in rotated coordinates. Across 360 model states, the canonical run has 31,320 finite state–space outcomes. Maximum relative one-step norm mismatch is 2.83×10^{-7}, zero-displacement functional error 4.24×10^{-6}, and basis reconstruction error 3.55×10^{-7}, all below the predeclared 5×10^{-5} tolerance.

For the non-spectral candidate, ID horizon-1/horizon-100 errors are $0.00130/0.10761$; low OOD errors are $0.00208/0.23826$ and high OOD $0.00550/0.28031$. Last-block and last-block-plus-head physics TTA also help (6.56% and 7.11%), while full-model gain is 12.10% under the interpolation-selected protocol.

G Negative results and boundary cases

Limited adapters and model fit. The small Burgers adapter was effective in its original setting, but the controlled Navier–Stokes FNO/HERO rank-four adapters produce only $0.118\%/0.018\%$ mean gain under the later seven-space comparison. A branch-feature adapter on a substantially underfit DeepONet increased Burgers error by 14.5% on average. A controlled DRIFT-Net was also underfit on the matched Navier–Stokes task (ID horizon-100 error 2.144), and its output adapter changed pooled OOD error by -0.010%. On an official DRIFT-Net NS-SL checkpoint, a 12-trajectory subset gave a 0.025% weak-update gain, essentially tied with random (0.022%) and oracle (0.027%); this native benchmark has no coefficient-OOD axis. These cases do not falsify the existence result for the well-fit DilResNet but forbid a general benefit claim.

Failed deployment gate and insufficient coverage. An interpolation-trained five-feature label-free diagnostic reached pooled OOD ROC-AUC 0.641, but score-versus-gain Spearman was 0.009

(interval $[-0.107, 0.121]$). Its frozen gate gave pooled error 0.6401, worse than adapting every state (0.6110). The proposed independent weak-basis consensus itself had pooled gain Spearman -0.088. Oracle norm ratio cannot repair deployment because it requires labels and is not sufficient even analytically: the full network's ratio one does not imply the largest five-step gain. The contiguous $d = 4096$ FNO region's 30th random percentile further rejects a universal structured-space advantage.

Alternative rollout explanation and numerical cost. Finite perturbations through 100 steps did not show systematic OOD-induced local expansion: FNO mean local gains were 0.998 low OOD, 0.992 ID, and 0.990 high OOD. Thus a simple local-expansion causal explanation was rejected; non-normal long-time growth was not measured. In 1D Burgers, 50-step adapter TTA required 1.77–2.35 seconds per trajectory, more than the 0.195-second $N = 128$ direct solver. A raw FNO forward pass was faster, so the adaptation cost matters even when accuracy improves. Unbounded z-score conditioning in a separate continuous-viscosity FNO experiment caused OOD one-step degradation and, with FiLM/hypernetwork modulation, numerical rollout divergence; bounded choices avoided catastrophic divergence without eliminating the OOD gap. These observations are specific to the tested designs.

H STATISTICAL ESTIMANDS AND DEPENDENCE

For trajectory i, relative gain is $r_i = (E_{i,\mathrm{pre}} - E_{i,\mathrm{post}})/E_{i,\mathrm{pre}}$; reported gain is the arithmetic mean of r_i, not $1 - \overline{E}_{\mathrm{post}}/\overline{E}_{\mathrm{pre}}$. A physics percentile is the within-state fraction of norm-matched random directions with lower gain than the physics direction, averaged over states. Exact-dimension random references are the within-state mean over ten frozen subspace definitions. Structured-minus-random differences are paired by model seed, viscosity, and initial condition.

The archived analyses resample the 180 trajectory rows. Because rows share three trained model seeds and ten paired initial conditions across six viscosities, the paper-facing audit additionally uses a crossed two-way cluster bootstrap (Owen & Eckles, 2012). Each of 20,000 deterministic replicates independently resamples three model seeds and ten initial-condition identifiers with replacement, then retains all six viscosities in every selected cell. This preserves the repeated-viscosity block and represents uncertainty across both trained models and initial states. Three model seeds remain too few for a population-wide probability claim; the intervals are sensitivity analyses, not asymptotic guarantees. Five random definitions per subspace family also limit tail resolution.

Table 4: Weak-residual value versus horizon-100 error by group. The pooled near-zero association is not stable across trained seeds.

Group	Value	n	Pearson	Spearman
Pooled	all	180	$-.009$	.075
Model seed	0	60	$-.183$	$-.023$
Model seed	1	60	$-.216$	$-.165$
Model seed	2	60	.331	.500
Viscosity	.00075	30	.129	.152
Viscosity	.00100	30	.115	.143
Viscosity	.00150	30	.089	.125
Viscosity	.00800	30	$-.076$	$-.031$
Viscosity	.01000	30	$-.129$	$-.031$
Viscosity	.01200	30	$-.133$	.017

Every value in Tables 4–6 is regenerated from canonical raw CSVs by the supplied clustered-statistics script. Row-bootstrap intervals remain in the archived canonical summaries for exact traceability.

Table 5: Two-way model-seed/initial-condition cluster-bootstrap intervals for headline quantities. Gains are percentage points.

Quantity	Model/protocol	Mean	95% interval
Residual–error Pearson	FNO pooled	$-.009$	$[-.437, .520]$
Residual–error Spearman	FNO pooled	$.075$	$[-.436, .567]$
One-step physics gain (pp)	FNO	$.454$	$[-.127, 1.000]$
One-step physics gain (pp)	HERO-compatible	1.155	$[-.903, 3.357]$
One-step physics gain (pp)	DilResNet	1.431	$[.883, 2.142]$
Physics–oracle cosine	DilResNet	$.470$	$[.241, .627]$
Five-step TTA gain (pp)	DilResNet full	12.096	$[8.101, 15.996]$

Table 6: Clustered structured-minus-random gain differences at exact dimension, in percentage points. All means favor the architecture-defined space; intervals exclude zero for both $d = 648$ spaces and include zero for the $d = 1{,}249$ heads.

Model	Space/protocol	d	Mean	95% interval
FNO	pointwise, one step	648	$.327$	$[.063, .578]$
FNO	pointwise, five step	648	2.239	$[.448, 3.849]$
HERO	pointwise, one step	648	$.828$	$[.062, 1.609]$
HERO	pointwise, five step	648	$.518$	$[.191, .853]$
FNO	head, one step	1,249	$.449$	$[-.034, .897]$
FNO	head, five step	1,249	2.945	$[-.138, 5.930]$
HERO	head, one step	1,249	$.634$	$[-.900, 2.251]$
HERO	head, five step	1,249	$.537$	$[-.100, 1.181]$

I COMPUTE AND REPRODUCTION

GPU experiments used one NVIDIA RTX 4090 (24 GB). The Navier–Stokes training, model, solver, and experiment configurations, seeded checkpoints, canonical raw outputs, and experiment registry are included in the anonymous supplementary package. The exact-size subspace calibration and OOD evaluations took 10.25 and 3,728.89 seconds. DilResNet training took 826.35, 910.26, and 945.31 seconds across seeds 0–2; its fit gate, frozen baseline, joint calibration, and OOD mechanism stages took 11.61, 24.74, 520.12, and 1,108.26 seconds. Re-running the plots requires the canonical result CSV/JSON files and the included plotting script, not model retraining. The supplementary code and configuration files are the source of exact software-version and optimizer settings; no local machine path or identifying repository link is embedded in the manuscript.