Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Measuring and Mitigating Forgetting when Extending a PDE Foundation Model to Simulated Flow Regimes

wasif masood

wasif.masood@empirischtech.at

Empirisch Tech GmbH https://orcid.org/0000-0003-3231-8104

Shoaib Noor

Empirisch Tech GmbH

Research Article

Keywords: continual learning, catastrophic forgetting, PDE foundation models, neural operators, experience replay, fine-tuning, parameter-efficient adaptation, LoRA, transfer learning, scientific machine learning, flow prediction, out-of-distribution generalization

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11745898/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Measuring and Mitigating Forgetting when Extending a PDE Foundation Model to Simulated Flow Regimes

Wasif Masood
Empirisch Tech GmbH
wasif.masood@empirischtech.at

Shoaib Noor
Empirisch Tech GmbH
shoaib.noor@empirischtech.at

Abstract

Foundation models for partial differential equations adapt efficiently to new tasks, but their effect on the original pretraining distribution is rarely measured. We study this retention problem on Poseidon, a neural operator pretrained on periodic-domain compressible Euler and incompressible Navier–Stokes families. We adapt it to two regimes absent from pretraining, a compressible Kelvin–Helmholtz flow and wall-bounded flow past a circular cylinder. Fine-tuning adapts efficiently but degrades the pretraining families. LoRA still forgets substantially, and freezing the whole trunk reduces forgetting without preventing it. We show that low-weight replay and joint training resolve this trade-off, extending the model to the new regimes while keeping the pretraining families near their original accuracy. These results suggest that lightweight replay can extend a PDE foundation model while substantially preserving its existing capabilities.[1]

1 Introduction

Foundation models for partial differential equations (PDEs) are pretrained across multiple equation families and subsequently adapted to downstream tasks. The reported benefit is predominantly *forward transfer*, whereby adaptation to a new task requires fewer samples and less training than learning the task from scratch. Several studies evaluate this form of transfer. Subramanian et al. [2023] assess generalisation to unseen coefficients and equations, Hao et al. [2024] and McCabe et al. [2023] fine-tune large multi-family pretrained models on new downstream systems, and Ye et al. [2025] evaluate zero- and few-shot adaptation to unseen PDEs. In each case, however, evaluation is confined to the downstream task, leaving the effect of adaptation on previously learned regimes unmeasured. This effect is critical to determining whether a foundation model can be *extended* or must instead be *forked*. If each new adaptation degrades performance on existing regimes, the result is effectively a collection of single-purpose surrogates rather than a single model that accumulates capability. We therefore evaluate retention, defined as performance on the pretraining distribution after adaptation, as a measure of whether new capability can be added without sacrificing existing capability.

Forgetting under fine-tuning is well documented in language and vision, where replay and regularization-based methods are commonly used to mitigate it [McCloskey and Cohen, 1989, Kirkpatrick et al., 2017, Lopez-Paz and Ranzato, 2017, Chaudhry et al., 2019]. In operator learning, the limited work addressing this problem has primarily focused on designing forgetting-resistant architectures [Tripura and Chakraborty, 2025], rather than quantifying the retention of a standard pretrained model under conventional fine-tuning. This latter setting more closely reflects how a practitioner would extend an existing PDE foundation model.

[1]The jointly trained model is available at `huggingface.co/empirischtech/poseidon-b-joint-cyl-kh` and the simulation data at `huggingface.co/datasets/empirischtech/pyfr-cylinder-kh` (Appendix E).

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Representations for the Physical Sciences Workshop.

We study this question on Poseidon [Herde et al., 2024], a PDE foundation model pretrained on six two-dimensional families spanning compressible Euler and incompressible Navier–Stokes physics, all posed on periodic domains without internal boundaries. Two properties suit it to a retention study. First, its pretraining distribution is well-characterised and bounded (six families, all periodic and boundary-free), so a wall-bounded target is unambiguously out of distribution and retention is measured against a known distribution rather than a broad corpus in which out-of-distribution status is ill-defined. Second, its state is represented in physical variables including a density channel, which lets us encode a solid body directly in that channel (§2). The new regime is thus admitted without adding inputs or altering the architecture, so any forgetting is attributable to fine-tuning rather than to architectural change. We adapt it to two regimes selected to differ substantially from the pretraining distribution, and ask what this adaptation costs the pretraining families and whether that cost can be mitigated.

To answer these questions, we proceed in three steps. We first measure how much fine-tuning to a new flow regime degrades the pretraining families. We then ask whether parameter-efficient adaptation avoids this, and find that neither LoRA nor a frozen trunk does. Finally, we show that low-weight replay reverses the degradation, extending the model to both new regimes while retaining, and often improving, the pretraining families.

2 Method

Model and pretraining families. The base model is Poseidon-B (scOT) [Herde et al., 2024], a lead-time-conditioned neural operator acting on a four-channel state (ρ, u, v, p) at 128^2 resolution. It was pretrained on six periodic, boundary-free families, and we track four of them for retention, namely CE-Gauss, CE-RP, CE-CRP, and CE-KH, with 10 trajectories $\times$ 21 frames each. We then adapt the model to the two new regimes described below.

Metric. We report the per-channel normalised root-mean-square error (NRMSE), the root-mean-square error divided by the channel's global value range over the evaluation set, computed with per-family normalisation. Per-family normalisation is essential because family statistics differ sharply, and under shared CE statistics the RP pressure channel is displaced by approximately 5.4σ, a shift that alone dominates the resulting error. Persistence (no-change baseline, $u(t+\Delta t) = u(t)$) serves as a baseline throughout, and because it is non-monotonic in lead time on periodic flows, it is compared per horizon.

Data. We adapt Poseidon to two regimes chosen to lie far from its periodic, boundary-free pretraining distribution. The first regime is KHC, our custom Kelvin–Helmholtz family, distinct from the pretraining CE-KH family. It is a compressible Navier–Stokes double shear layer with $\mu = 0.01$ and $\mathrm{Pr} = 0.72$, sweeping Mach 0.20–0.32 for training and 0.32–0.35 for holdout at amplitudes 0.03–0.10, with $50 + 10$ trajectories $\times$ 101 frames. The second regime is Cylinder, wall-bounded flow past a circular cylinder from PyFR [Witherden et al., 2014], using artificial-compressibility Navier–Stokes at order 3 on 3,427 elements, and it spans Re 62–180 stratified into 150 trajectories, with a 180–190 holdout of 30. Both regimes were run with PyFR on NumericalAI (numericalai.net), a cloud simulation platform developed by the authors' company. Fields are interpolated to 128^2, and the body is written into the density channel with $\rho = 1$ in fluid and 0 in solid, an encoding that is slot-compatible with pretraining and doubles as the geometry mask. The regime therefore enters without new inputs or architectural change, so any forgetting is attributable to fine-tuning. Checkpoint selection uses validation error on the 30 held-out cylinder trajectories (Appendix C). All other reported evaluations, namely the pretraining families, the extended band (Re 190–240), the KHC holdout, and the rollout correlations, enter no training or selection decision. Cylinder NRMSE uses training-set channel statistics, and each pretraining family uses its own pretraining statistics. Two independent physics checks validate the sweep (Appendix C). A growth-law fit recovers the vortex-shedding onset Reynolds number, giving $\mathrm{Re}_c \approx 45.9$ against 46.2 reported by Jackson [1987], and Strouhal numbers match the parallel-shedding correlation of Williamson [1988] to $+0.04\%$.

3 Results

Following the three steps of §1, we first measure how well the model transfers to the cylinder regime (§3.1), then how much this adaptation degrades the pretraining families (§3.2), and finally show

that low-weight replay reverses the degradation while fitting both new regimes and producing stable rollouts (§3.3).

Table 1: Extrapolation beyond the training band ($\mathrm{Re} \leq 180$), lead-1 NRMSE over three seeds, with the in-band holdout for reference.

sub-band	Re	fine-tuned	from scratch	persistence	ratio
holdout	180–190	0.00528 ± 0.0001	0.0437 ± 0.012	0.0393	$8.3\times$
near	190–206	0.00567 ± 0.0002	0.0440 ± 0.011	0.0398	$7.8\times$
mid	206–223	0.00627 ± 0.0002	0.0445 ± 0.011	0.0404	$7.1\times$
far	223–240	0.00711 ± 0.0003	0.0450 ± 0.011	0.0411	$6.3\times$

3.1 Transfer to the new regime

Fine-tuning the pretrained model on the cylinder regime reaches an NRMSE of 0.00528 ± 0.00011 at lead 1 (three seeds, $N = 150$, Table 1). This is an $8.3\times$ improvement over training from scratch (0.0437 ± 0.0116). Fine-tuning also matches the from-scratch model's final accuracy in roughly $25\times$ fewer updates, by epoch 2. The gain is acquired during fine-tuning rather than inherited. The zero-shot pretrained model is in fact worse than the no-change (persistence) baseline by $1.39\times$.

Pretraining also stabilises optimisation. The seed-to-seed spread is 2.1% for fine-tuning versus 26.6% from scratch, and the two seed distributions do not overlap. The worst fine-tuned seed still beats the best from-scratch seed by $5.7\times$. The advantage is likewise insensitive to training-set size. The fine-tuned error is flat across a $15\times$ range ($N = 10$ to 150, Fig. 1), while the from-scratch baseline stays an order of magnitude more seed-variable. This is because the target family is intrinsically low-dimensional.[2] Six modes capture 99% of its variance [Berkooz et al., 1993] (Fig. 4, Appendix C), so extra trajectories are largely redundant. This intrinsic rank is a cheap diagnostic of how much data a family needs.

Beyond the training band. We next push the regime parameter (the Reynolds number Re) up to 33% past the training maximum, over 30 held-out cases (Table 1). The fine-tuned error grows smoothly with distance while the from-scratch error stays flat. The transfer advantage therefore decays gracefully from $8.3\times$ in-band to $6.3\times$ at the farthest band rather than collapsing. The fine-tuned model beats both baselines throughout, and its lower seed variance persists ($\leq 4\%$ versus $\geq 24\%$). The from-scratch baseline is flat because it never learns the parameter dependence. The fine-tuned model's degradation therefore isolates the cost of extrapolating a *learned* dependence. A 5.5% shift in the flow's velocity statistics already erodes about a quarter of the advantage.

3.2 Forgetting under adaptation

The same checkpoint that §3.1 evaluated on the cylinder, we now evaluate on the four pretraining families, where it degrades every one of them (Table 2). CE worsens by a factor of 1.83, RP by 1.99, and CRP by 1.52. The resulting errors on CE and RP even exceed the persistence (no-change) baseline. The forgetting is not spread evenly across the four output channels. It concentrates in the transverse-velocity channel, which is the one the cylinder regime drives most strongly, so adaptation rewrites that channel and the pretraining families lose it.

Parameter-efficient adaptation. A natural hypothesis is that freezing the transformer trunk prevents forgetting. We do not find this to be the case (Table 2). LoRA [Hu et al., 2022] with $r = 16$ fits the cylinder comparably to full fine-tuning (0.008 versus 0.005, Table 2), so it is a genuine parameter-efficient alternative on the target task. It does not, however, retain. It still trains 4.0% of the parameters yet degrades CE by a factor of 1.76, comparable to the 1.83 under full fine-tuning. For LoRA, the parameter saving buys efficiency, not retention. We attribute this to the shared input embedding, output head, and adapter-induced residual-stream changes, which every family's forward pass traverses. Forgetting is not confined to the trunk weights. Training only the head and embeddings forgets least among the parameter-efficient variants, but it still degrades CE and RP by about $1.5\times$ and fits the cylinder less tightly than full fine-tuning (0.012 versus 0.005). We compare it with replay in §3.3.

[2] At these Reynolds numbers the wake settles into periodic vortex shedding, which is a limit cycle. Sweeping Re then traces a one-parameter family of such cycles, which is why the dynamics collapse onto so few modes.

Table 2: Lead-2 NRMSE by family, with relative-to-pretrained factors in parentheses. $\star$ marks families held out of every replay buffer. Every adapted row is a mean$_{\pm\sigma}$ over three seeds. CE-KH (a pretraining family) is shown as its density channel. Table 4 (Appendix D) gives all four channels. KHC is reported separately in Appendix D. The pretrained row is the unchanged base model.

model	trainable	CE	RP	CRP*	CE-KH*	CYL
pretrained	—	0.055	0.046	0.040	0.379	0.098
cylinder-only	100%	$0.101_{\pm.003}$ (1.83×)	$0.092_{\pm.001}$ (1.99×)	$0.061_{\pm.001}$ (1.52×)	$0.408_{\pm.001}$ (1.08×)	$0.005_{\pm.000}$
LoRA r=16	4.0%	$0.097_{\pm.002}$ (1.76×)	$0.076_{\pm.001}$ (1.65×)	$0.054_{\pm.000}$ (1.35×)	$0.350_{\pm.005}$ (0.92×)	$0.008_{\pm.000}$
head + embed	<1%	$0.084_{\pm.001}$ (1.53×)	$0.070_{\pm.001}$ (1.52×)	$0.048_{\pm.000}$ (1.20×)	$\mathbf{0.249}_{\pm.001}$ (**0.66**×)	$0.012_{\pm.000}$
replay W=0.25	100%	$\mathbf{0.024}_{\pm.002}$ (**0.43**×)	$\mathbf{0.026}_{\pm.000}$ (**0.55**×)	$0.046_{\pm.000}$ (1.15×)	$0.314_{\pm.001}$ (0.83×)	$0.011_{\pm.001}$
joint (cyl + KHC)	100%	$0.035_{\pm.002}$ (0.63×)	$0.032_{\pm.005}$ (0.69×)	$0.047_{\pm.001}$ (1.16×)	$0.292_{\pm.007}$ (0.77×)	$0.018_{\pm.001}$

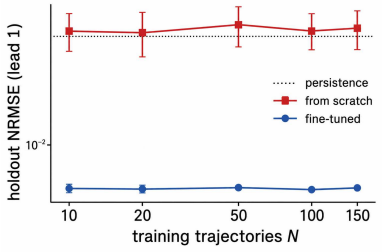


Figure 1: Scaling at fixed compute. Fine-tuned flat across the 15× range while from-scratch an order of magnitude more seed-variable.

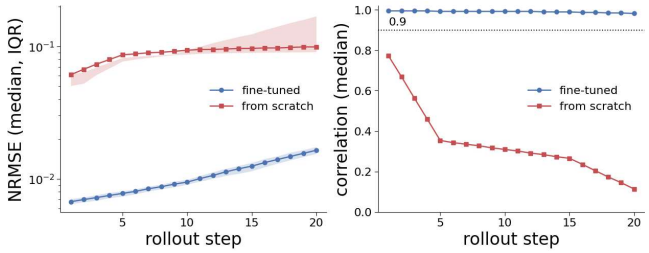


Figure 2: Rollout stability over 270 rollouts. Left, median NRMSE (log scale). Right, median correlation. Fine-tuned stays above 0.9 through 20 steps, while from-scratch falls below immediately.

3.3 Mitigation and rollout stability

Including the CE and RP families in the training mixture at weight $W = 0.25$ largely reverses the degradation of §3.2. Across three seeds, CE and RP improve to $0.43\times$ and $0.55\times$ their pretrained errors, respectively, while held-out CE-KH improves to $0.83\times$ on density. We track CE-KH through density because it carries the structure unique to this family. The per-channel breakdown (Table 4, Appendix D) shows replay improving the ρ, u and p channels below the pretrained model while roughly preserving v. The remaining held-out family, CRP, recovers from $1.52\times$ to 1.15 ± 0.005. Most of the loss is undone and the result stays below persistence despite a 15% residual degradation. The replay buffer contains none of this family's data, so the recovery is not memorisation. We also sweep the replay weight over $W \in \{0.25, 0.5, 1.0\}$ (Table 3). Increasing W from 0.25 to 0.5 improves CE and RP at a small cylinder cost. At $W = 1.0$, one seed degrades substantially, inflating the CE spread by an order of magnitude without reliable gain. Held-out CE-KH remains insensitive to W, with errors below the pretrained. Thus, replay weight mainly controls adaptation stability and target-task cost, motivating the conservative choice $W = 0.25$ elsewhere in the paper.

Frozen-trunk baseline. Training only the patch embedding and output head gives the closest competitor to replay in Table 2. Without any replay, it fits the cylinder about as well (0.012 versus 0.011) and retains the held-out families comparably ($1.20\times$ versus $1.15\times$ on CRP and $0.66\times$ versus $0.83\times$ on CE-KH density). It does not, however, protect the replayed families. It degrades CE and RP by about $1.5\times$, whereas replay brings them to $0.43\times$ and $0.55\times$. On CE-KH, its advantage is confined to density. Replay is better on the other three channels, most clearly transverse velocity, where the baseline's error rises to $1.8\times$ pretrained against $1.06\times$ for replay (Table 4, Appendix D). Averaged over channels, replay is marginally ahead on both held-out families, with $0.86\times$ versus $0.89\times$ on CE-KH and $1.15\times$ versus $1.20\times$ on CRP. Replay also fits the cylinder less tightly than full fine-tuning (0.011 versus 0.005), so part of its held-out retention may reflect less aggressive adaptation rather than replay itself. A comparison at matched target error would separate the two.

Joint training. Replay adds one new regime at a time. We also test whether both new regimes can be added at once, training from the base model with cylinder and KHC in the mixture (both at weight 1.0) alongside CE and RP replay (weight 0.25). Across three seeds this learns both, with a cylinder error of 0.018 (Table 2) and a KHC error of 0.0377 (Appendix D), each well below its persistence baseline. All four tracked pretraining families lie between 0.63 and 1.16 times their pretrained error.

The joint model reduces the held-out CE-KH density to 0.292 ± 0.007 ($0.77\times$ pretrained) while fitting both new regimes, even though CE-KH appears in no training mixture.

Rollout stability. We roll the model out autoregressively for 20 steps on all 30 held-out trajectories (3 start phases $\times$ 3 seeds, 270 rollouts, Fig. 2). The fine-tuned model reproduces the position, spacing, and number of vortices in the wake. Median NRMSE grows only from 0.007 to 0.017 with a narrow interquartile band, and correlation with the reference exceeds 0.9 at every step of every rollout (median 0.985 at step 20). The from-scratch model, by contrast, drops below 0.9 after the first step and reaches 0.11 by step 20. Over 40 steps the error grows more slowly than the $\sqrt{40} \approx 6.3$ expected if per-step errors were independent (5.5 ± 0.2), which indicates the model integrates the dynamics rather than retrieving stored states. The rollout is biased toward large scales, keeping the high-energy structure while smoothing fine detail (it retains 46% of the fine-scale energy, cf. the over-smoothing in Fig. 5).

Table 3: Replay-weight sweep (lead-2 NRMSE, mean $\pm\, \sigma$ over three seeds). Moderate replay ($W = 0.5$) improves the replayed families (CE, RP) over $W = 0.25$ at a small CYL cost. At $W = 1.0$ one of three seeds degrades, inflating the CE spread by an order of magnitude, while held-out CRP and CE-KH stay insensitive to W within seed noise.

family	$W = 0.25$	$W = 0.5$	$W = 1.0$
CE (replayed)	0.024 ± 0.002	0.019 ± 0.000	0.026 ± 0.016
RP (replayed)	0.025 ± 0.000	0.022 ± 0.001	0.025 ± 0.009
CRP*	0.046 ± 0.000	0.048 ± 0.001	0.047 ± 0.003
CE-KH*	0.314 ± 0.001	0.298 ± 0.047	0.298 ± 0.037
CYL (new regime)	0.011 ± 0.001	0.013 ± 0.001	0.019 ± 0.004

Acknowledgments and Disclosure of Funding

The authors would like to thank the Austrian Federal Ministry of Finance and the Austrian Research Promotion Agency (FFG) for funding this work through the LUMEIK-5G project under the grant agreement n. FO999923407. LUMEIK-5G is funded under the research and technology development of gigabit applications as part of the light-house projects of Breitband Austria 2030: GigaApp, which was initiated by the Austrian Federal Ministry of Agriculture, Regions, and Tourism.

References

B. F. Armaly, F. Durst, J. C. F. Pereira, and B. Schönung. Experimental and theoretical investigation of backward-facing step flow. *Journal of Fluid Mechanics*, 127:473–496, 1983.

D. Barkley and R. D. Henderson. Three-dimensional Floquet stability analysis of the wake of a circular cylinder. *Journal of Fluid Mechanics*, 322:215–241, 1996.

G. Berkooz, P. Holmes, and J. L. Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Review of Fluid Mechanics*, 25:539–575, 1993.

A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations*, 2019.

Z. Hao, C. Su, S. Liu, J. Berner, C. Ying, H. Su, A. Anandkumar, J. Song, and J. Zhu. DPOT: Auto-regressive denoising operator transformer for large-scale PDE pre-training. In *International Conference on Machine Learning*, 2024.

M. Herde, B. Raonić, T. Rohner, R. Käppeli, R. Molinaro, E. de Bézenac, and S. Mishra. Poseidon: Efficient foundation models for PDEs. In *Advances in Neural Information Processing Systems*, 2024.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

C. P. Jackson. A finite-element study of the onset of vortex shedding in flow past variously shaped bodies. *Journal of Fluid Mechanics*, 182:23–45, 1987.

G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

D. Lopez-Paz and M. Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, 2017.

M. McCabe, B. Régaldo-Saint Blancard, L. H. Parker, R. Ohana, M. Cranmer, A. Bietti, M. Eickenberg, S. Golkar, G. Krawezik, F. Lanusse, et al. Multiple physics pretraining for physical surrogate models. *arXiv preprint arXiv:2310.02994*, 2023.

M. McCabe, P. Mukhopadhyay, T. Marwah, B. Régaldo-Saint Blancard, F. Rozet, C. Diaconu, L. Meyer, et al. Walrus: A cross-domain foundation model for continuum dynamics. *arXiv preprint arXiv:2511.15684*, 2025.

M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

R. Ohana, M. McCabe, L. Meyer, R. Morel, F. Agocs, M. Beneitez, M. Berger, B. Burkhart, S. Dalziel, D. Fielding, et al. The Well: A large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems*, 2024.

A. Roshko. On the development of turbulent wakes from vortex streets. Technical Report NACA Report 1191, National Advisory Committee for Aeronautics, 1954.

S. Subramanian, P. Harrington, K. Keutzer, W. Bhimji, D. Morozov, M. W. Mahoney, and A. Gholami. Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior. In *Advances in Neural Information Processing Systems*, 2023.

M. Takamoto, T. Praditia, R. Leiteritz, D. MacKinlay, F. Alesiani, D. Pflüger, and M. Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022.

Tapas Tripura and Souvik Chakraborty. Neural combinatorial wavelet neural operator for catastrophic forgetting free in-context operator learning of multiple partial differential equations. *Computer Physics Communications*, 2025.

C. H. K. Williamson. Defining a universal and continuous Strouhal–Reynolds number relationship for the laminar vortex shedding of a circular cylinder. *Physics of Fluids*, 31(10):2742–2744, 1988.

C. H. K. Williamson. Vortex dynamics in the cylinder wake. *Annual Review of Fluid Mechanics*, 28:477–539, 1996.

F. D. Witherden, A. M. Farrington, and P. E. Vincent. PyFR: An open source framework for solving advection–diffusion type problems on streaming architectures using the flux reconstruction approach. *Computer Physics Communications*, 185(11):3028–3040, 2014.

Z. Ye, Z. Liu, B. Wu, H. Jiang, L. Chen, M. Zhang, X. Huang, Q. Meng, J. Zou, H. Liu, and B. Dong. PDEformer-2: A versatile foundation model for two-dimensional partial differential equations. *arXiv preprint arXiv:2507.15409*, 2025.

H. Zhang, C. Kang, Y. Wang, and D. Zou. F-Adapter: Frequency-adaptive parameter-efficient fine-tuning in scientific machine learning. In *Advances in Neural Information Processing Systems*, 2025.

A Related work

Foundation models for PDEs. Large multi-family simulation corpora [Takamoto et al., 2022, Ohana et al., 2024] have enabled a rapidly growing class of models pretrained across equation families. These include multiple-physics pretraining [McCabe et al., 2023], denoising operator pretraining [Hao et al., 2024], Poseidon [Herde et al., 2024], PDEformer-2 [Ye et al., 2025], which conditions on a symbolic representation of the governing equation, and Walrus [McCabe et al., 2025], a 1.3B-parameter model spanning nineteen continuum-dynamics scenarios. Transfer studies characterise when such pretraining helps out of distribution [Subramanian et al., 2023]. As noted in the introduction, none of these re-evaluate the pretraining families after adaptation, so retention is unmeasured. It is that axis we study here.

Continual learning. Catastrophic forgetting under sequential training is long established [McCloskey and Cohen, 1989]. It is addressed either by constraining parameter updates [Kirkpatrick et al., 2017] or by replaying stored examples [Lopez-Paz and Ranzato, 2017, Chaudhry et al., 2019], and we use the latter, which requires no change to the training objective. For neural operators specifically, NCWNO [Tripura and Chakraborty, 2025] achieves forgetting-robustness architecturally (gated wavelet experts) when trained across a task sequence from scratch, and replay has been validated for physics-informed operators on Darcy-flow and elasticity problems. We differ in three ways. We start from an existing, publicly released multi-family pretrained model, so that forgetting is measured against a real pretraining distribution. We evaluate retention on families never placed in the replay buffer. And we target wall-bounded flow.

Parameter-efficient adaptation. LoRA [Hu et al., 2022] and related adapters are increasingly applied to scientific models, including frequency-adaptive variants designed for operator fine-tuning [Zhang et al., 2025]. This line of work optimises the *efficiency* of adaptation, and its effect on *retention* of the pretraining distribution is generally not evaluated. Section 3.2 provides such a measurement and finds that parameter efficiency does not by itself confer retention. Physics-informed training [Karniadakis et al., 2021] offers a complementary route to reducing data requirements that we do not explore here.

Surrogates for cylinder flow are abundant. Here the cylinder serves as a well-validated vehicle for the extension question rather than as a modelling contribution.

B Limitations

The main holdout band is adjacent to the training band, so its absolute errors reflect near-interpolation. §3.1 quantifies the consequence, and the transfer advantage is not distance-independent. Those extrapolation results span the full $\mathrm{Re} = 190$–240 sweep (30 cases) but reach only 33% beyond the training band. Behaviour at larger distances, and on flows that are not near-periodic, remains untested. Above the onset of the three-dimensional mode-A instability, placed at $\mathrm{Re} = 188.5 \pm 1.0$ by Barkley and Henderson [1996], the two-dimensional flow is no longer a faithful model of the physical wake [Williamson, 1996]. The main holdout is bounded at $\mathrm{Re} = 190$ for this reason, and the far band therefore tests extrapolation on a well-posed PDE rather than on physically realisable flow. The family is intrinsically of rank six, so the present results concern a low-dimensional target. A four-parameter wall-bounded family (a pulsatile backward-facing step, validated in the steady limit against Armaly et al. [1983]) is under construction as the corresponding higher-dimensional test. Rollouts are tested within approximately four shedding periods, and the frame-resampled joint model has a minimum usable lead of $1/20$ of the trajectory window, below which accuracy degrades sharply. Export resolution restricts all claims to wake dynamics. CE/RP gains under replay are in-distribution by construction, and the retention claim rests on the held-out families, whose validation sets are small (10 trajectories each). Held-out retention under replay is not yet separated from the effect of less aggressive adaptation (§3.3).

C Experimental details

Cylinder simulations. We simulate the artificial-compressibility Navier–Stokes equations with PyFR [Witherden et al., 2014], using flux reconstruction at order 3 on a mixed mesh of 3,427 elements. The mesh has a structured quadrilateral boundary-layer block of 28×7 elements around the cylinder and triangles elsewhere, for 35,446 solution points. The domain spans $x \in [-8, 35]$ and $y \in [-8, 8]$ diameters, with blockage 6.25%. We use dual time stepping (SDIRK33 physical, RK45 pseudo) with a converged inner solve (residual tolerance 10^{-10}, 3–40 iterations), characteristic far-field boundaries, and a no-slip cylinder. Vortex shedding is seeded by a deterministic transverse Gaussian perturbation, and per-case burn-in times follow a growth-rate law fitted from a diagnostic sweep, $\sigma \approx 0.00348\,(\mathrm{Re} - 45.9)$. Each case provides 41 frames spanning approximately 3.3 shedding periods after saturation. Cases whose harvest windows were found not fully stationary by an envelope test on in-run wake probes (32 cases at the low-Re end) were extended by a further 20 time units. After extension, $149/150$ training cases are fully stationary and the mean Strouhal residual against Williamson's correlation is $+0.04\%$ (RMS 0.0017). We checked solver-convergence sensitivity by differencing cheap and converged runs at $\mathrm{Re} = 100$. The difference concentrates in transverse velocity (up to 3.7%), which motivates the converged setting. Fig. 3 shows the Strouhal–Reynolds

validation and the per-case saturation diagnostic. Fig. 4 shows the POD spectrum, the effective rank, and the interpolation divergence of the exported tensors.

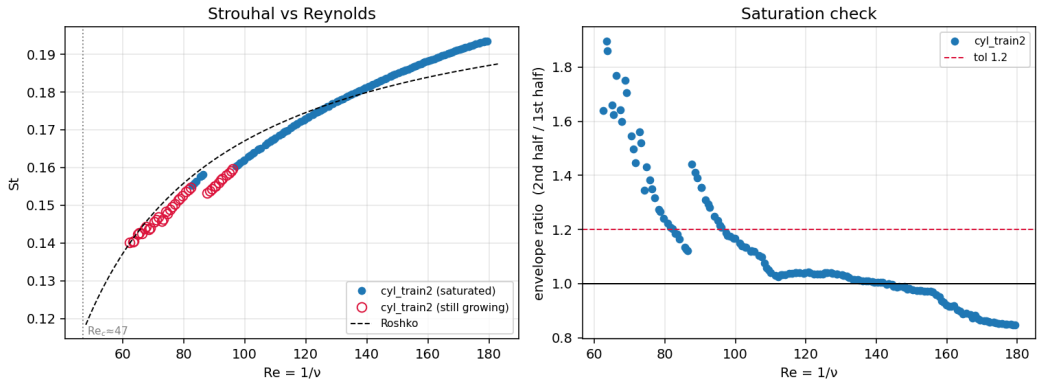


Figure 3: Physics validation of the cylinder sweep. Left, Strouhal number versus Reynolds number for the 150 training cases against the correlation of Roshko [1954] (the quantitative residuals in the text use Williamson's parallel-shedding fit, with a post-extension mean residual of $+0.04\%$). Right, amplitude-envelope ratio per case. Both panels are computed over the full simulation history including burn-in, so the "still growing" markers at low Reynolds number are an artefact of that whole-history measure. The harvest windows themselves are stationary in 149 of 150 cases.

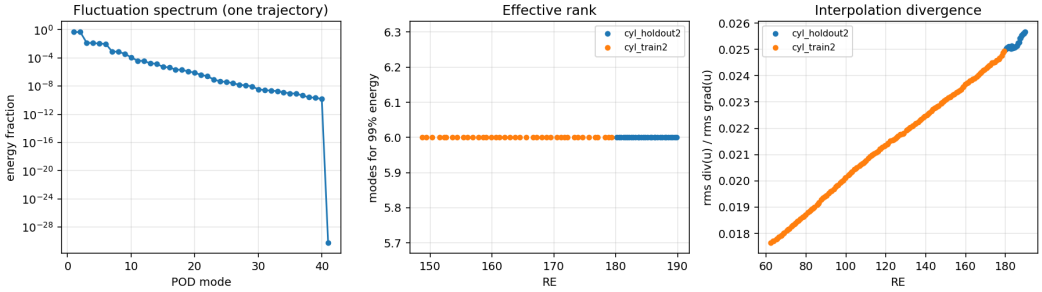


Figure 4: Tensor diagnostics. Left, the POD fluctuation-energy spectrum of one trajectory. The final mode drops to machine precision because a mean-subtracted 41-frame trajectory has rank at most 40, so its true energy is zero. Centre, the effective rank (modes for 99% of fluctuation energy) versus Reynolds number, which is exactly six in every case evaluated. Right, the interpolation divergence of the exported fields, smooth in Reynolds number across all 180 cases.

KHC simulations. The KHC family is a compressible Navier–Stokes double shear layer ($\mu = 0.01$, $\mathrm{Pr} = 0.72$, density ratio 1:2, shear-layer thickness 0.07, domain $[-10, 10]^2$) with 101 frames per trajectory over 100 time units. Mach controls both compressibility and the pressure scale ($p \propto 1/M^2$), making pressure the dominant channel in normalisation.

Training. All configurations are identical. We use AdamW (2×10^{-4}, weight decay 10^{-5}), OneCycle with 10% warm-up, L_1 loss, gradient clipping at 1.0, mixed precision, batch size 192, and 31 optimiser steps per epoch for 50 epochs. This is a fixed budget of 1,550 updates, so that dataset size and compute are never confounded. The saved checkpoint is the one with lowest validation error on the 30 held-out cylinder trajectories, which therefore serve as a validation set for checkpoint selection. There is no early stopping. For every configuration in Table 2 this selects a late-training state at the end of a monotonically improving validation curve (best epoch ≥ 41 of 50 in all runs), not an early minimum. All other reported evaluations, namely the pretraining families, the extended Reynolds band, the KHC holdout, and the rollout correlations, are computed once on the selected checkpoint and constitute untouched test sets. Replay configurations add the CE and RP families at weight W alongside the new regime(s) at weight 1.0. The CRP and CE-KH families are never included in any replay mixture and serve as held-out probes of retention. Training pairs sample a random frame gap $k \in [1, 12]$ with normalised lead time $k/(T-1)$. Multi-dataset runs sample datasets per step in proportion to

their weights and resample all trajectories to 21 frames so that a frame step corresponds to the same trajectory fraction across datasets. LoRA uses rank 16 on the attention query/key/value and dense projections (4.0% of parameters). The frozen-trunk baseline trains only the patch embedding and output head.

Evaluation. NRMSE denominators are global per-channel value ranges over each evaluation set. Each family is normalised by its own statistics. Channels that are near-constant over an evaluation set (per-channel standard deviation below 0.02) are excluded from that set's reported mean NRMSE, since their normalised error is dominated by numerical noise. Rollouts are autoregressive with the model's own output fed back. Correlation is the Pearson correlation of the predicted and reference fields. Seed spreads are over three seeds with identical data and budget. The physics validation of the cylinder sweep is summarised in Fig. 3.

D Extended results

On its custom-KH holdout the joint model attains a NRMSE of 0.0377 (self-normalised on the holdout tensors) while also fitting the cylinder regime. This value uses a different normalisation from Table 2 and is not directly comparable to it.

Table 4: Per-channel CE-KH NRMSE, held out of every replay buffer. Density (ρ) is the channel reported in Table 2, and it carries the shear-layer roll-up that distinguishes this family. Replay improves ρ, u and p and roughly preserves v. Joint training improves ρ, u and p but degrades v ($1.48\times$). The frozen-trunk baseline improves ρ, preserves p, and degrades u ($1.25\times$) and v ($1.79\times$). The last column gives the four-channel mean relative to the pretrained model. All adapted rows are three-seed means.

model	ρ (density)	u	v	p	mean
pretrained	0.379	0.066	0.060	0.093	1.00×
cylinder-only	0.408	0.099	0.125	0.082	1.19×
LoRA r=16	0.350	0.087	0.129	0.075	1.07×
head + embed	0.249	0.082	0.108	0.094	0.89×
replay W=0.25	0.314	0.053	0.064	0.083	0.86×
joint	0.292	0.051	0.089	0.080	0.86×

E Data and model availability

The joint model of §3.3, fine-tuned on the cylinder and KHC regimes with CE and RP replay, is available at `huggingface.co/empirischtech/poseidon-b-joint-cyl-kh`. It is one of the three training seeds behind the joint row of Table 2, and the repository also contains the per-family normalisation statistics used in training. The simulation data are available at `huggingface.co/datasets/empirischtech/pyfr-cylinder-kh`. They contain the 150 training and 30 holdout cylinder trajectories and the 50 training and 10 holdout KHC trajectories as raw `(T,C,128,128)` tensors with per-case metadata. The KHC tensors carry an additional total-energy channel, and the model uses only the first four channels. Data is released under the CC BY-4.0 licence. The model released under the CC BY-NC 4.0 licence following the licence of the base model, Poseidon-B [Herde et al., 2024], which is publicly available from its authors and is not redistributed here. Simulations were produced with PyFR (BSD-3-Clause) on NumericalAI (numericalai.net).

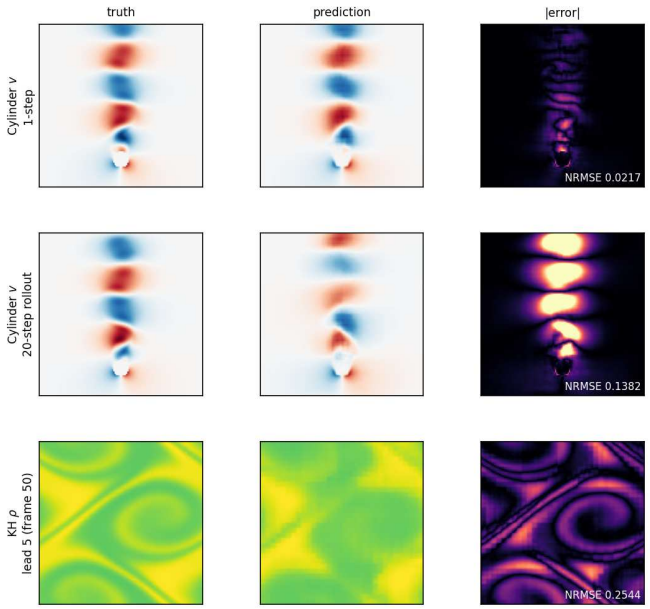


Figure 5: Held-out predictions of the joint model, showing truth, prediction and |error| for cylinder transverse velocity (minimum trained lead, 20-step autoregressive rollout) and KHC density (lead 5 from mid-trajectory). The rollout preserves vortex position and spacing, with error appearing as amplitude damping on vortex cores. KHC interfaces are over-smoothed, consistent with the $5\times$ temporal resampling of the KHC data. Per-case normalisation is used on the displayed channel, so the values are not directly comparable to Table 2.