PDE-OBS: CONTROLLED EVALUATION ACROSS OBSERVATION PATTERNS

Anonymous authors
Paper under double-blind review

ABSTRACT

Physical-field reconstruction and forecasting depend on both measurement density and spatial layout, yet evaluation under a single observation pattern does not characterize performance when that pattern changes. We introduce PDE-OBS, an integrated benchmarking platform spanning numerical data generation, model training, and inference and evaluation under varying observation conditions. It combines 560,000 fields and trajectories from seven partial differential equation families with configurable observation operators and seven adapted baseline methods for stationary reconstruction and short-horizon forecasting. Separating observation construction from physical records allows users to specify parameterized patterns and deterministic mixtures for training and testing while preserving prediction targets and data splits. The evaluation protocol uses references trained for each test pattern to compare models on identical test observations and targets, alongside equal-count groups for spatial-layout comparisons. On a 14,000-record subset, we evaluate 441 trained models under nine test patterns, yielding 3,969 evaluations. Mean cross-pattern error exceeds mean matched-pattern error in all 49 PDE–method pairs, and this finding persists in a configuration-matched subset of 117 models. Denser test observations do not consistently reduce error for a fixed model. Mixed-pattern training on five completed pairs reduces large single-pattern transfer errors, although destination-trained references usually remain more accurate. Together, the benchmark and findings support systematic evaluation of observation-pattern sensitivity and provide a reusable workflow for developing methods under changing measurement conditions. Code: Anonymous code repository.

1 INTRODUCTION

Reconstructing physical fields and predicting their evolution from partial observations (Evensen, 1994; Carrassi et al., 2018; Raissi et al., 2018; Brajard et al., 2020; Williams et al., 2024) are central problems in scientific machine learning (Karniadakis et al., 2021; Brunton et al., 2020; Willard et al., 2022; Kovachki et al., 2023; Azizzadenesheli et al., 2024). The information available to a model depends on both the number and the locations of its measurements (Manohar et al., 2018; Yildirim et al., 2009; Krause et al., 2008). Scattered, clustered, and line-based layouts can contain the same number of measurements yet reveal different information about the underlying field (Joshi & Boyd, 2009; Clark et al., 2019; Chaturantabut & Sorensen, 2010; Drmač & Gugercin, 2016; Peherstorfer et al., 2020). A model's accuracy under its training observation pattern therefore needs to be considered alongside its performance when measurement density or spatial layout changes. This distinction connects prediction from partial observations to the broader challenge of evaluation under distribution shift (Koh et al., 2021; Gulrajani & Lopez-Paz, 2021) and motivates treating observation conditions as an explicit dimension of a scientific benchmark.

Shared benchmarks support systematic comparisons by providing datasets, baseline implementations, and common evaluation interfaces (Otness et al., 2021; Bonnet et al., 2022; Luo et al., 2023; Herde et al., 2024). PDEBench combines configurable numerical generation with baseline models (Takamoto et al., 2022), while The Well unifies training and evaluation across diverse physical simulations (Ohana et al., 2024). Learning from partial observations is also an established setting (Raissi, 2018; Raissi et al., 2019; 2020; Jin et al., 2021; Sun & Wang, 2020), including LANO and its POBench-PDE benchmark (Hou et al., 2026). Building on these efforts, we address a complementary evaluation

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

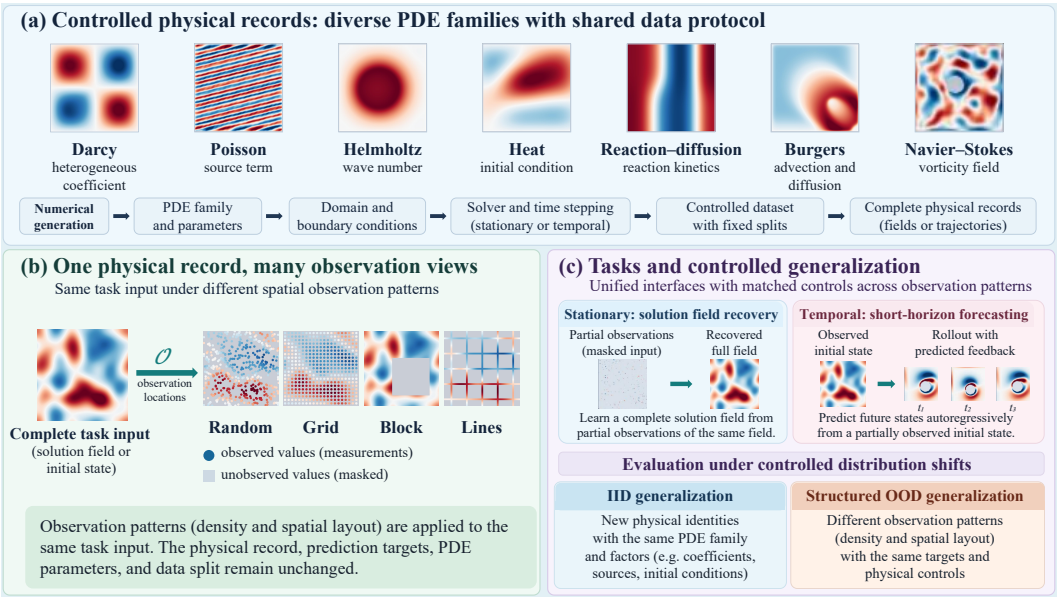

Figure 1: **Overview of PDE-OBS.** The benchmark connects numerical data generation, observation construction, model training, and inference across observation conditions. **(a)** Numerical solvers generate complete fields and trajectories with fixed train/test splits. **(b)** Different observation patterns are applied to the same physical record while prediction targets remain unchanged. The observed field is the solution for reconstruction and the initial state for forecasting. **(c)** Models perform stationary reconstruction or autoregressive forecasting and are evaluated under matched and shifted observation patterns. For each test pattern, a reference model is trained on the training split using that pattern and evaluated on identical test observations and targets. Field maps are schematic; Section 3 specifies the benchmark tasks and observation protocols.

question: *how does prediction performance change when observation conditions at inference differ from those used in training?* Studying this question across physical systems and methods calls for a reusable workflow that connects numerical data generation, model training, and inference while allowing training and test observation patterns to be specified separately.

We introduce PDE-OBS, an integrated benchmarking platform for stationary field reconstruction and short-horizon forecasting under partial observations (Figure 1). The platform connects numerical data generation, observation construction, model training, checkpoint-based inference, and evaluation through shared definitions of physical records, model inputs, and prediction targets. Stationary reconstruction recovers a complete solution field from measurements of that same field; forecasting predicts future states from a partially observed initial state using autoregressive prediction. Both tasks use complete target fields for training supervision, while inference receives observed values, observation masks, and domain geometry. Users can generate or load physical records, train models under specified observation patterns or deterministic mixtures, and evaluate the resulting checkpoints under alternative observation conditions. Throughout these comparisons, the underlying physical records, prediction targets, and train/test splits remain fixed (Section 3).

Using a 14,000-record subset, we evaluate 441 trained models spanning seven PDE families, seven adapted baseline methods, and nine training observation patterns. Testing each model under all nine patterns produces 3,969 evaluations and 49 complete training-pattern-by-test-pattern matrices. Mean cross-pattern error exceeds mean matched-pattern error in every PDE–method pair, and the same finding holds in a configuration-matched subset of 117 models. Comparisons at equal measurement counts reveal performance differences across spatial layouts, while fixed-model density comparisons show that denser test observations do not consistently reduce error. Additional mixed-pattern training on five completed pairs reduces large transfer errors, but does not uniformly outperform destination-trained references (Section 4, Appendix G). These findings motivate reporting observation-pattern

sensitivity alongside matched accuracy and illustrate how the integrated workflow supports systematic investigation of training and inference conditions.

Our contributions are fourfold.

- **An integrated benchmark workflow.** We connect numerical data generation, observation construction, model training, inference, and evaluation within a reusable platform. Seven adapted baseline methods share explicit task and input definitions, supporting method development and evaluation under configurable observation conditions.
- **A configurable physical data resource.** We provide 560,000 fields and trajectories across seven PDE families, four boundary protocols, and ten condition constructions. Numerical solvers and condition-field generators support selectable physical regimes, resolutions, and data tiers, with record-level metadata and reproducible data splits.
- **Systematic evaluation across observation patterns.** Nine reference patterns, including two equal-count layout groups, define matched and cross-pattern comparisons on shared physical records. For each test pattern, a destination-trained reference is evaluated on identical observations and targets. Parameterized operators and deterministic mixtures extend these comparisons to additional training and test conditions.
- **A comprehensive empirical result resource.** We provide complete cross-pattern results, full-precision per-record errors, and training configurations for the 441-model evaluation. Analyses of spatial layout, test density, a configuration-matched subset, and mixed-pattern training characterize complementary aspects of observation-pattern sensitivity (Section 4).

2 Related Work

PDE benchmarks and autoregressive forecasting. Shared benchmarks support comparisons of physical prediction methods through common datasets and evaluation protocols (Rasp et al., 2020; 2024; Chung et al., 2023; Johnson et al., 2023). PDEBench combines configurable numerical data generation with baseline implementations (Takamoto et al., 2022), while The Well provides standardized model interfaces across diverse physical simulations (Ohana et al., 2024). APEBench focuses on autoregressive PDE emulation and examines how training procedures affect successive-step prediction accuracy (Koehler et al., 2024). PDEArena supports neural surrogate modeling (Gupta & Brandstetter, 2023), and AL4PDE studies data-efficient learning for neural PDE solvers (Musekamp et al., 2025). Complementary resources address inverse problems through PDEInvBench (Goel et al., 2026) and physical prediction from real measurements through RealPDEBench (Hu et al., 2026). PDE-OBS contributes an integrated benchmark workflow spanning numerical data generation, model training, and inference and evaluation under varying observation conditions. By specifying training and test observation patterns separately on fixed physical records, targets, and data splits, it makes observation-pattern sensitivity a systematic dimension of benchmark evaluation.

Reconstruction from sparse sensors. Recovering physical fields from limited measurements connects classical sparse reconstruction and inverse problems (Donoho, 2006; Candes et al., 2006; Stuart, 2010; Callaham et al., 2019) with data-driven field estimation. Gappy proper orthogonal decomposition reconstructs fields by fitting the coefficients of a low-rank basis to observed entries (Everson & Sirovich, 1995; Willcox, 2006; Gunes et al., 2006; Rowley & Dawson, 2017). Learned approaches map reduced measurements to complete fields (Erichson et al., 2020; Fukami et al., 2019; Dubois et al., 2022; Fukami et al., 2023). Senseiver supports reconstruction from sparse, irregular sensors (Santos et al., 2023), and Voronoi-based reconstruction accommodates varying sensor numbers and locations (Fukami et al., 2021). These approaches motivate examining reconstruction accuracy across measurement densities and spatial layouts. PDE-OBS provides a common evaluation protocol for this purpose: each trained model is tested under multiple observation patterns on shared physical records, references trained for each test pattern are evaluated on identical observations and targets, and equal-count groups compare layouts at fixed measurement counts.

Learning from partial observations. Training supervision is another important distinction among partial-observation prediction settings. LANO introduces POBench-PDE for temporal prediction with point-wise and patch-wise masking, using partially observed inputs and outputs during training and complete target fields for evaluation (Hou et al., 2026). PDE-OBS uses complete training targets

for stationary reconstruction and short-horizon forecasting, while specifying the input observation patterns separately for training and inference. This setting supports comparisons of single-pattern and mixed-pattern training, together with matched and shifted test observations, under complete-target supervision.

DiffusionPDE and FunDPS use diffusion models to predict physical fields conditioned on partial observations (Huang et al., 2024; Yao et al., 2025), while CORAL and DINo employ neural-field or implicit representations for operator learning and forecasting (Serrano et al., 2023; Yin et al., 2023). These methods offer complementary approaches to prediction and field representation. PDE-OBS provides reusable data, observation operators, task definitions, and evaluation interfaces through which additional methods can be adapted and compared under shared measurement conditions and prediction targets.

3 PDE-OBS: A BENCHMARK FOR LEARNING FROM PARTIAL OBSERVATIONS

Following the modular benchmark design of PDEBench (Takamoto et al., 2022), PDE-OBS integrates numerical data generation, observation construction, model training, and inference and evaluation under varying observation conditions. These stages share physical records, task definitions, and data splits, allowing training and test observation patterns to be specified separately. We first define the prediction tasks, then describe the data resource, observation protocols, evaluation metrics, baseline models, and end-to-end workflow. Detailed benchmark specifications and usage instructions are provided in Appendix A and Appendix B.

3.1 GENERAL PROBLEM DEFINITION

A physical record i contains a complete condition field c_i, such as a coefficient, source, or initial state; a complete solution or trajectory; a geometry field G_i marking walls or obstacles; and metadata. For a given task, let x_i denote the input field and y_i the prediction target. We use v and w to denote the training and test observation protocols, respectively. Each protocol specifies how a record-specific observation mask is constructed. Given a binary mask $M_{i,w}$ with ones at observed locations, the model input and prediction are

$$z_{i,w} = M_{i,w} \odot x_i, \qquad \widehat{y}_i^{v,w} = f_{\theta_v}\big(z_{i,w}, M_{i,w}, G_i\big), \tag{1}$$

where $\odot$ denotes elementwise multiplication and θ_v denotes parameters learned under protocol v. Unobserved entries are zero-filled, and the mask distinguishes missing entries from measured zeros. The geometry field describes the physical domain separately from observation availability. Changing w changes the observations supplied to a fixed model while preserving the underlying record and prediction target.

In *stationary field reconstruction*, $x_i = y_i = u_i$: the model reconstructs a complete solution field from measurements of that same field. Training supervises the complete solution, including both observed and hidden locations.

In *short-horizon forecasting*, $x_i = u_{i,0}$ is the initial state and $y_i = (u_{i,1}, u_{i,2}, u_{i,3})$ comprises the next three stored states. The predictor f_{θ_v} applies a shared one-step model autoregressively:

$$\widehat{u}_{i,1} = F_{\theta_v}\big(M_{i,w} \odot u_{i,0}, M_{i,w}, G_i\big), \qquad \widehat{u}_{i,h+1} = F_{\theta_v}\big(\widehat{u}_{i,h}, \mathbf{1}, G_i\big), \quad h = 1, 2. \tag{2}$$

The all-one mask marks the fully specified predicted field supplied at subsequent steps. Training and evaluation follow the same recurrence: predictions provide the recurrent inputs, and complete future states provide supervision during training. Gradients propagate through the three-step rollout. At inference, the baselines receive only observed values, observation masks, and domain geometry. PINO additionally uses physical coefficients, sources, or parameters in its training residual. Detailed task definitions, information access, and the supported forward and inverse extensions are specified in Appendix A.1.

3.2 OVERVIEW OF DATASETS AND OBSERVATION PROTOCOLS

Physical systems and data generation. Configurable numerical solvers generate records for seven PDE families. The stationary systems are **Darcy**, $-\nabla \cdot (a\nabla u) = f$, with positive coefficient a; **Poisson**, $-\Delta u = f$; and **Helmholtz**, $(-\Delta - k^2)u = f$, in a coercive low-wavenumber

setting. The temporal systems are **Heat**, $\partial_t u = \alpha \Delta u$; Allen–Cahn-type **reaction–diffusion**, $\partial_t u = \kappa \Delta u + \rho(u - u^3)$; scalar two-dimensional **Burgers**, $\partial_t u + u(\partial_{x_1} u + \partial_{x_2} u) = \nu \Delta u$; and **Navier–Stokes** vorticity, $\partial_t \omega + \boldsymbol{v} \cdot \nabla \omega = \nu \Delta \omega + f$. For Navier–Stokes, the state represented by u in the task definitions is the scalar vorticity ω.

The data resource contains 560,000 fields and trajectories across seven PDE families, four boundary protocols, and ten condition constructions. Each PDE–boundary–construction combination contains 2,000 records distributed across three family-specific physical regimes. Generation supports selectable spatial resolutions and nested data tiers. Generated records are assessed using family-specific discrete PDE residuals, boundary and initial-condition checks, and transition-consistency checks. Numerical methods, parameters, time grids, and validation results are documented in Appendix A.2.1; data-generation options are described in Appendix B.1.

Evaluation subset and data splits. The reported experiments use a 14,000-record subset with 2,000 records per PDE family. This subset consists of scalar 128×128 fields with smooth random-field conditions, Dirichlet boundaries for stationary systems, and periodic boundaries for temporal systems. Stationary boundary values are zero on the outer stored cell-center layer. Temporal trajectories contain 15 equally spaced states, of which frame 0 supplies the partially observed input and frames 1–3 supply the forecasting targets.

A seeded ranking within each physical regime assigns 1,800 records to training and 200 to testing per PDE family. These identity lists are shared across all methods and observation patterns. Because each identity belongs to a physical record, changing its observation pattern preserves its data split. The data design, identity conventions, and exact partitioning rules are provided in Appendix A.3.

Reference observation patterns. The benchmark constructs binary grid masks from the physical-record identity, observation protocol, and fixed observation seed. For a given record and protocol, the resulting mask is shared across methods and remains fixed across training epochs. Nine reference patterns define the main evaluation: uniformly random points at nominal densities of 50%, 65%, and 80% (R50, R65, R80); horizontal rows (H); vertical columns (V); crossed horizontal and vertical lines (LI); observations outside a missing square (BL); a perimeter band (BD); and Gaussian-clustered points (CL).

Two groups support spatial-layout comparisons at fixed measurement counts. R50/H/V/CL contain exactly 8,192 observed locations, while BL/LI/BD contain exactly 8,284. Density comparisons use the separately sampled location sets of R50, R65, and R80, thereby changing both measurement counts and observed locations. All nine reference patterns use noiseless observations and cover at least half the grid. BL and CL use non-wrapped array coordinates, and BD specifies a perimeter sampling layout on both stationary and periodic domains. Exact constructions, observation counts, and seed definitions are given in Appendix A.4.

Configurable training and inference conditions. Beyond the reference patterns, users can specify observation densities and geometric parameters, combine patterns into deterministic mixtures, register custom operators, or supply stored masks. For example, `uniform 20` requests uniformly sampled observations at a target density of 20%, while `line 30 + random 20` defines a mixture of two component patterns. A mixture assigns one component to each physical record and retains that assignment across epochs. The same specification can be used for training and as a test observation protocol, allowing a trained checkpoint to be evaluated under alternative measurement conditions. Stored masks also support explicitly constructed nested observation sets. Observation namespaces, parameter syntax, and reference-pattern replay are detailed in Appendix B.2; mixed-pattern experiments are presented in Section 4 and Appendix G.

3.3 Overview of Metrics

The primary metric is the mean per-record relative L^2 error. For each PDE–method pair, evaluating models trained under nine reference patterns against all nine test patterns produces a 9×9 matrix:

$$E_{v,w} = \frac{1}{200} \sum_{i \in \mathcal{I}_{\text{test}}} \frac{\|\widehat{y}_i^{v,w} - y_i\|_2}{\max(\|y_i\|_2, 10^{-12})}, \qquad D = \frac{1}{9} \sum_{v} E_{v,v}, \qquad C = \frac{1}{72} \sum_{v \neq w} E_{v,w}. \quad (3)$$

Rows correspond to training patterns and columns to test patterns. The matched-pattern mean D summarizes the nine diagonal cells, while the cross-pattern mean C summarizes the 72 off-diagonal cells. The norm covers the complete target field, including observed and hidden locations. For forecasting, it is computed jointly over all three target frames.

To examine individual training-to-test directions, we compare each model with a *destination-trained reference*: a model of the same method trained on the common training split using test pattern w. The directional difference $\Delta_{v\to w} = E_{v,w} - E_{w,w}$ and ratio $R_{v\to w} = E_{v,w}/E_{w,w}$ compare predictions on identical test observations and targets. Positive differences and ratios above one indicate higher error than the reference. Averaging the 72 directional differences gives $C - D$, while individual directions and equal-count groups identify the patterns associated with larger gaps. Ratios use the positive, full-precision reference errors. Figure 2 additionally reports absolute errors on a shared logarithmic scale.

Cell-level mean $\pm$ sample standard deviation summarizes the 200 per-record errors. The standard deviations reported with D and C describe variability across their nine and 72 cell means, respectively. Forecasting results additionally include errors at each prediction horizon. The evaluator also supports stationary observed/hidden error decomposition and data-consistency diagnostics, whose definitions are provided in Appendix A.5.

Evaluation checks record identities, observation masks, array shapes, target-frame indices, and the finiteness of prediction and target entries before scoring. Each reported cell contains all 200 test records, with all finite errors included in the aggregation. Evaluation procedures, data and checkpoint bindings, and per-record result records are documented in Appendix C.

3.4 Baseline Models and Training

Seven baseline adaptations share the observation interface in Equation (1). **U-FNO** uses a two-dimensional architecture combining Fourier and U-Fourier blocks. **FNO** uses a compact mask-channel architecture, while the **CNO-inspired** encoder–decoder applies fixed low-pass filtering. **DeepONet** uses a convolutional branch to encode the masked field. **GNOT** uses a structured-grid adaptation, and **Transolver** uses structured-grid slice attention with a larger temporal preset. **PINO** augments an FNO backbone with a PDE-residual training term. The four temporal families use the autoregressive recurrence in Equation (2). Architectural configurations and parameter counts are specified in Appendix A.6, with source references, upstream revisions, licenses, and adaptation details in Appendix A.8.

Training settings. The shared data objective is the minibatch mean of per-record relative-L^2 losses over the complete target. For forecasting, the loss is computed jointly over three predicted frames, with gradients propagated through the autoregressive rollout. PINO additionally includes its physical residual. Each run follows its recorded optimizer (Kingma & Ba, 2015; Loshchilov & Hutter, 2019), learning-rate schedule (Smith & Topin, 2019), batch size, epoch budget, and training-loss stopping rule where enabled. We evaluate the final checkpoint of each run. Per-run configurations accompany the results, and Section 4 examines observation-pattern sensitivity within a configuration-matched subset. Training recipes and model presets are provided in Appendix A.6, and configurable training controls are described in Appendix B.3.

3.5 Data Format, Access, and Extensibility

Complete physical records are stored as HDF5 shards with JSON metadata describing identities, PDE families, boundary protocols, condition constructions, physical regimes, and time coordinates. Observation masks are constructed when records are accessed, allowing the same physical data to support multiple training and inference conditions. The storage format and loading procedures are specified in Appendix A.3 and Appendix B.1.

The integrated workflow allows users to generate or load records, configure observations, train a model, reload its checkpoint, and perform inference under alternative observation patterns. Versioned input and prediction formats separate inference from scoring: the predictor receives observed values, masks, and geometry, while targets are supplied separately for evaluation. The same prediction

interface accepts externally supplied observations. Input preparation and checkpoint loading are detailed in Appendix B.4, with an end-to-end example in Appendix B.

Code, configurations, tests, and documentation are available in our anonymous repository under the MIT code license. The result release provides full-precision scores, per-record errors, and data and checkpoint bindings. Public registries support additional observation operators, condition constructions, prediction methods, PDE families, and metrics while preserving the nine reference-pattern definitions. Versioned formats, resolved configurations, and regression tests support maintenance and reuse. Access instructions and extension procedures are provided in Appendix B.9.

4 Results Across Observation Patterns

Evaluation setup. Using the benchmark workflow, we evaluate 441 trained models spanning seven PDE families, seven baseline methods, and nine training observation patterns. Each model is tested under all nine patterns on the same 200 held-out records for its PDE family, yielding 3,969 evaluations and 49 complete 9×9 matrices. Evaluation verifies record identities, observation masks, array shapes, finiteness, and physical-time mappings before aggregation (Appendix C). We first examine matched and cross-pattern performance, then assess the configuration-matched subset, equal-count layout changes, and fixed-model density responses. Additional mixed-pattern training experiments examine how including multiple observation patterns during training changes performance across the nine test conditions.

Does matched accuracy characterize cross-pattern performance? We first compare the matched-pattern mean D with the cross-pattern mean C within each PDE–method pair. In Figure 2, D summarizes the outlined diagonal cells and C summarizes the off-diagonal cells, with all evaluations using the same physical records and prediction targets. We find $C > D$ in 49 of 49 pairs, with median $C/D = 10.46$ (95% paired test-identity bootstrap interval [10.07, 10.71]; Appendix C.2). At the directional level, 3,281 of the 3,528 comparisons satisfy $E_{v,w} > E_{w,w}$, indicating higher error than the reference trained for the test pattern. These results reveal a widespread gap between matched and cross-pattern performance across the evaluated methods and physical systems. Figure 2 shows the absolute errors underlying these comparisons, while Table 10 summarizes matched and cross-pattern errors across the seven methods for each PDE family.

Does the finding persist under matched training configurations? We next examine the 13-pair configuration-matched subset, comprising 117 models and marked by dots in Figure 2. Within each pair, all nine models complete 500 epochs with the same architecture, optimizer, learning-rate schedule, batch size, loss, and stopping settings. Configurations differ only in the observation pattern and run-name/data-root fields. Subset membership is determined from configuration metadata and was fixed before the current prediction evaluation. We again find $C > D$ in 13 of 13 pairs, with median $C/D = 10.74$ (95% interval [10.44, 11.12], conditional on these trained models). Thus, the higher cross-pattern mean error persists within pairs whose recorded training configurations agree. The subset definition, configuration comparisons, and detailed results are provided in Appendix D.

Does spatial layout matter at fixed measurement counts? To examine layout changes at a fixed number of measured locations, we restrict training and test patterns to either R50/H/V/CL ($K = 8{,}192$) or BL/LI/BD ($K = 8{,}284$). Cross-pattern error exceeds the destination-trained reference in $545/588$ (92.69%) and $276/294$ (93.88%) directed comparisons, respectively. Table 12 separates the results for stationary reconstruction and forecasting. Performance gaps therefore remain prevalent when training and test patterns contain the same number of measurements, supporting spatial layout as an evaluation dimension alongside density. For stationary Dirichlet tasks, the measurement count includes sampled prescribed boundary values; the observation constructions and boundary coverage are specified in Appendix A.4.

How does test observation density affect a fixed model? We now hold each checkpoint fixed and change its test observation pattern from R50 to R80. The ratio $E_{v,\mathrm{R80}}/E_{v,\mathrm{R50}}$ compares the resulting errors, with values below one indicating lower error under R80. Across the 49 PDE–method pairs, R80 reduces error in 13/49 R50-trained models (median ratio 1.58), 37/49 R65-trained models (0.93), and 49/49 R80-trained models (0.29). Table 12 additionally reports the responses to R65. The



Figure 2: **Performance across training and test observation patterns.** Each PDE–method panel contains a 9×9 matrix with training patterns in rows and test patterns in columns, both ordered R50, H, V, CL, BL, LI, BD, R65, R80. Color denotes mean relative-L^2 error over 200 test records on a shared absolute logarithmic scale; darker colors indicate larger errors. Outlined cells (□) identify matched training and test patterns. Dots (•) identify the 13 PDE–method pairs in the configuration-matched subset. Thin separators distinguish the two equal-count layout groups and the denser random patterns. The figure contains all 3,969 evaluations. Per-cell means, sample standard deviations, and training configurations are reported in Appendix F. CNO denotes the CNO-inspired adaptation.

benefit of denser test observations therefore depends on the training pattern. R50, R65, and R80 use separately sampled location sets, so these comparisons evaluate changes in both measurement count and sampled locations under the benchmark's random-observation protocols (Appendix A.4).

How does mixed-pattern training change performance across patterns? We examine five completed MIX models: Helmholtz/FNO, Helmholtz/DeepONet, Heat/DeepONet, Navier–Stokes/U-FNO, and Poisson/U-FNO. Each uses the same 1,800 training records as its single-pattern references, with one deterministically assigned reference pattern per record, fixed across epochs. The first four complete 500 epochs; Poisson completes 200 under the one-cycle recipe shared by its BL, BD, and CL references. All models are scored on the same 200 test records under each of nine patterns (Figure 3); actual update counts and comparison scopes are given in Appendix G.

Against destination-trained references, $R_{\mathrm{MIX}\to w} = E_{\mathrm{MIX},w}/E_{w,w}$ ranges from 0.63 to 4.07 for the four 500-epoch pairs, with respective geometric means 2.07, 2.56, 1.32, and 1.19. The three recipe-matched Poisson destinations give 3.05 (range 2.51–3.49). Navier–Stokes MIX outperforms

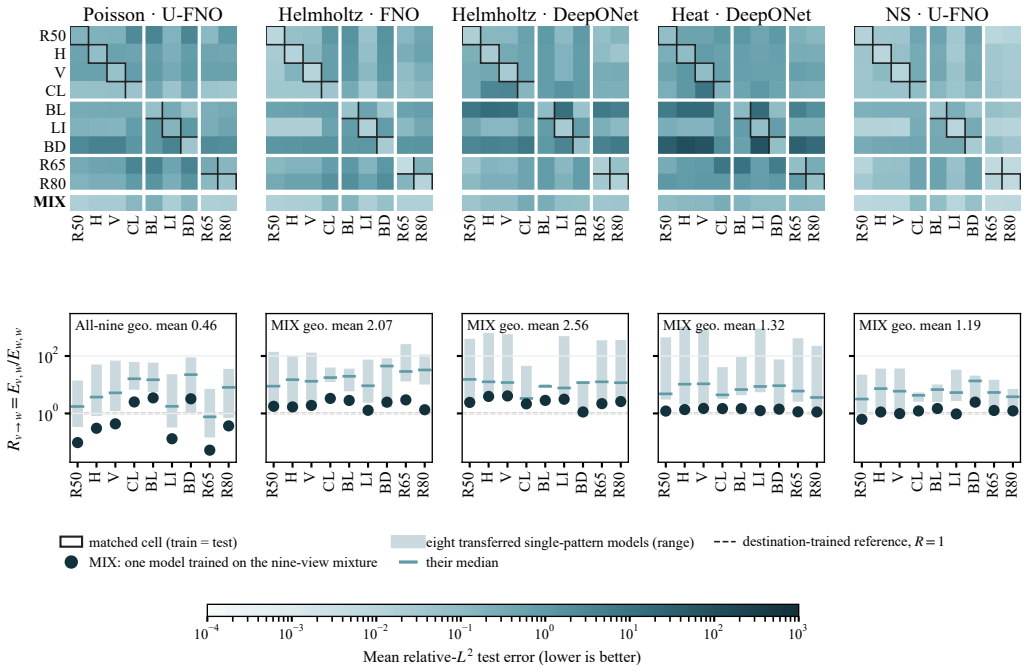

Figure 3: **Single-pattern and mixed-pattern training: five completed pairs.** Top: nine single-pattern rows and one additional MIX row, with test patterns in columns and the shared error scale of Figure 2. Outlines mark matched cells. Bottom: MIX ratios to the destination-trained reference (dots) versus the range and median of eight transferred models (bars); the dashed line is $R = 1$. Poisson's displayed geometric mean uses all nine destinations descriptively; its BL/BD/CL recipe-matched mean is 3.05. MIX rows supplement, rather than replace, the 441-model grid (Appendix G).

its destination-trained references on R50, LI, and V; thus a matched-pattern penalty is not universal. Across all 45 pair–test-pattern combinations, MIX is below the median of the eight transferred single-pattern models in 45/45 cases and below their best in 34/45; their transfer ratios reach 1,040.09. Mixed-pattern training reduces large transfer errors in these completed cases, but this descriptive comparison does not isolate the effects of training exposure from heterogeneous histories or establish a general architecture ranking.

The complete results in Appendix F include all 49 training-pattern-by-test-pattern matrices, per-cell mean $\pm$ sample standard deviation across 200 test records, pair-level D and C, error distributions, three forecasting horizons, and training epochs. Full-precision scores and per-record errors accompany the benchmark, supporting further analysis of observation-pattern sensitivity.

5 Discussion and Conclusion

PDE-OBS provides a common protocol for reconstruction and short-horizon forecasting under changing observations. All 49 PDE–method pairs have higher cross-pattern than matched mean error, including the configuration-matched subset; layout and density comparisons motivate reporting both. Five completed MIX models reduce large transfer errors, but do not uniformly beat destination-trained references. Stationary diagnostics locate error at observed points while preserving raw headline scores (Appendix E). The empirical scope is limited to selected smooth fields, boundary protocols, noiseless observations, coercive Helmholtz cases, and short horizons. Retrospective selection, one training seed, heterogeneous budgets, capacities and information access preclude causal attribution or a controlled architecture ranking. Bootstrap intervals describe test-identity uncertainty only. Formal simple baselines and forecasting ambiguity controls remain untested, and production residuals do not establish independent reference-solution accuracy.

ETHICS STATEMENT

This study uses synthetic data generated by numerical PDE simulations rather than human-subject records. The benchmark supports research on physical-field reconstruction and forecasting under partial observations. Applications to real-world sensing or control require application-specific validation and safety assessment. Data licensing and release permissions require separate verification.

AI USE STATEMENT

Generative AI tools assisted with implementation, suggestions for experimental protocols and parameter settings, analysis, figure and table preparation, literature search, and manuscript revision, including reviewer-style feedback. Reported numerical results are computed from numerical simulations, model predictions, and recorded evaluation outputs. Prediction and artifact verification procedures are described in Appendix C. The authors take full responsibility for the scientific claims, experimental results, citations, manuscript content, released artifacts, and release permissions.

REFERENCES

Kamyar Azizzadenesheli, Nikola Kovachki, Zongyi Li, Miguel Liu-Schiaffini, Jean Kossaifi, and Anima Anandkumar. Neural operators for accelerating scientific simulations and design. *Nature Reviews Physics*, 6(5):320–328, 2024. doi: 10.1038/s42254-024-00712-5.

Florent Bonnet, Jocelyn Mazari, Paola Cinnella, and Patrick Gallinari. AirfRANS: High Fidelity Computational Fluid Dynamics Dataset for Approximating Reynolds-Averaged Navier–Stokes Solutions. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23463–23478, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/94ab7b23a345f93333eac8748a66c763-Abstract-Datasets_and_Benchmarks.html.

John P. Boyd. *Chebyshev and Fourier Spectral Methods*. Dover Publications, second edition, 2001. ISBN 9780486411835.

Julien Brajard, Alberto Carrassi, Marc Bocquet, and Laurent Bertino. Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model. *Journal of Computational Science*, 44:101171, 2020. doi: 10.1016/j.jocs.2020.101171.

Steven L. Brunton, Bernd R. Noack, and Petros Koumoutsakos. Machine Learning for Fluid Mechanics. *Annual Review of Fluid Mechanics*, 52(1):477–508, 2020. doi: 10.1146/annurev-fluid-010719-060214.

Jared L. Callaham, Kazuki Maeda, and Steven L. Brunton. Robust flow reconstruction from limited measurements via sparse representation. *Physical Review Fluids*, 4(10):103907, 2019. doi: 10.1103/PhysRevFluids.4.103907.

E.J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2): 489–509, 2006. doi: 10.1109/TIT.2005.862083.

Alberto Carrassi, Marc Bocquet, Laurent Bertino, and Geir Evensen. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *WIREs Climate Change*, 9(5): e535, 2018. doi: 10.1002/wcc.535.

Saifon Chaturantabut and Danny C. Sorensen. Nonlinear Model Reduction via Discrete Empirical Interpolation. *SIAM Journal on Scientific Computing*, 32(5):2737–2764, 2010. doi: 10.1137/090766498.

Wai Tong Chung, Bassem Akoush, Pushan Sharma, Alex Tamkin, Ki Sung Jung, Jacqueline Chen, Jack Guo, Davy Brouzet, Mohsen Talei, Bruno Savard, Alexei Poludnenko, and Matthias Ihme. Turbulence in Focus: Benchmarking Scaling Behavior of 3D Volumetric Super-Resolution with

BLASTNet 2.0 Data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77430–77484, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f458af2455b1e12608c2a16c308d663d-Abstract-Datasets_and_Benchmarks.html.

Emily Clark, Travis Askham, Steven L. Brunton, and J. Nathan Kutz. Greedy Sensor Placement With Cost Constraints. *IEEE Sensors Journal*, 19(7):2642–2656, 2019. doi: 10.1109/JSEN.2018.2887044.

Andrew Collette. *Python and HDF5: Unlocking Scientific Data*. O'Reilly Media, 2013. ISBN 9781491944981.

J. Crank and P. Nicolson. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Mathematical Proceedings of the Cambridge Philosophical Society*, 43(1):50–67, 1947. doi: 10.1017/S0305004100023197.

D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. doi: 10.1109/TIT.2006.871582.

Zlatko Drmač and Serkan Gugercin. A New Selection Operator for the Discrete Empirical Interpolation Method—Improved A Priori Error Bound and Extensions. *SIAM Journal on Scientific Computing*, 38(2):A631–A648, 2016. doi: 10.1137/15M1019271.

Pierre Dubois, Thomas Gomez, Laurent Planckaert, and Laurent Perret. Machine learning for fluid flow reconstruction from limited measurements. *Journal of Computational Physics*, 448:110733, 2022. doi: 10.1016/j.jcp.2021.110733.

N. Benjamin Erichson, Lionel Mathelin, Zhewei Yao, Steven L. Brunton, Michael W. Mahoney, and J. Nathan Kutz. Shallow neural networks for fluid flow reconstruction with limited sensors. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2238): 20200097, 2020. doi: 10.1098/rspa.2020.0097.

Geir Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5): 10143–10162, 1994. doi: 10.1029/94JC00572.

Richard Everson and Lawrence Sirovich. Karhunen–Loève procedure for gappy data. *Journal of the Optical Society of America A*, 12(8):1657–1664, 1995. doi: 10.1364/JOSAA.12.001657. URL https://doi.org/10.1364/JOSAA.12.001657.

Kai Fukami, Koji Fukagata, and Kunihiko Taira. Super-resolution reconstruction of turbulent flows with machine learning. *Journal of Fluid Mechanics*, 870:106–120, 2019. doi: 10.1017/jfm.2019.238.

Kai Fukami, Romit Maulik, Nesar Ramachandra, Koji Fukagata, and Kunihiko Taira. Global field reconstruction from sparse sensors with Voronoi tessellation-assisted deep learning. *Nature Machine Intelligence*, 3(11):945–951, 2021. doi: 10.1038/s42256-021-00402-2.

Kai Fukami, Koji Fukagata, and Kunihiko Taira. Super-resolution analysis via machine learning: a survey for fluid flows. *Theoretical and Computational Fluid Dynamics*, 37(4):421–444, 2023. doi: 10.1007/s00162-023-00663-0.

Divyam Goel, Nithin Chalapathi, Sanjeev Raja, and Aditi S. Krishnapriyan. PDEInvBench: A comprehensive dataset and design space exploration of neural networks for PDE inverse problems. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=MSjhqRnNyZ.

Sigal Gottlieb and Chi-Wang Shu. Total variation diminishing Runge-Kutta schemes. *Mathematics of Computation*, 67(221):73–85, 1998. doi: 10.1090/S0025-5718-98-00913-2.

Ishaan Gulrajani and David Lopez-Paz. In Search of Lost Domain Generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Hasan Gunes, Sirod Sirisup, and George Em Karniadakis. Gappy data: To Krig or not to Krig? *Journal of Computational Physics*, 212(1):358–382, 2006. doi: 10.1016/j.jcp.2005.06.023.

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

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. GNOT: A general neural operator transformer for operator learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12556–12569. PMLR, 2023. URL https://proceedings.mlr.press/v202/hao23c.html.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. doi: 10.1038/s41586-020-2649-2.

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient Foundation Models for PDEs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 72525–72624, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/84e1b1ec17bb11c57234e96433022a9a-Abstract-Conference.html.

M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952. doi: 10.6028/jres.049.044.

Jingren Hou, Hong Wang, Pengyu Xu, Chang Gao, Huafeng Liu, and Liping Jing. Learning neural operators from partial observations via latent autoregressive modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(1):390–398, 2026. doi: 10.1609/aaai.v40i1.37001. URL https://ojs.aaai.org/index.php/AAAI/article/view/37001.

Peiyan Hu, Haodong Feng, Hongyuan Liu, Tongtong Yan, Wenhao Deng, Tianrun Gao, Rong Zheng, Haoren Zheng, Chenglei Yu, Chuanrui Wang, Kaiwen Li, Zhi-Ming Ma, Dezhi Zhou, Xingcai Lu, Dixia Fan, and Tailin Wu. RealPDEBench: A benchmark for complex physical systems with real-world data. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=y3oHMcoItR. Oral Presentation.

Jiahe Huang, Guandao Yang, Zichen Wang, and Jeong Joon Park. DiffusionPDE: Generative PDE-solving under partial observation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130291–130323, 2024. doi: 10.52202/079017-4140. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/eb3878c1dcbfff9ee95d5d033e5f5942-Abstract-Conference.html.

Xiaowei Jin, Shengze Cai, Hui Li, and George Em Karniadakis. NSFnets (Navier-Stokes flow nets): Physics-informed neural networks for the incompressible Navier-Stokes equations. *Journal of Computational Physics*, 426:109951, 2021. doi: 10.1016/j.jcp.2020.109951.

J. Emmanuel Johnson, Quentin Febvre, Anastasiia Gorbunova, Sam Metref, Maxime Ballarotta, Julien Le Sommer, and Ronan Fablet. OceanBench: The Sea Surface Height Edition. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78275–78295, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6ccbf94fa57c2ae372ece91b537574d-Abstract-Datasets_and_Benchmarks.html.

S. Joshi and S. Boyd. Sensor Selection via Convex Optimization. *IEEE Transactions on Signal Processing*, 57(2):451–462, 2009. doi: 10.1109/TSP.2008.2007095.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021. doi: 10.1038/s42254-021-00314-5.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Felix Koehler, Simon Niedermayr, Rüdiger Westermann, and Nils Thuerey. APEBench: A benchmark for autoregressive neural emulators of PDEs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 120252–120310, 2024. doi: 10.52202/079017-3822. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d9875ebcf74bccdc5076acab0dbee62c-Abstract-Datasets_and_Benchmarks_Track.html.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural Operator: Learning Maps Between Function Spaces With Applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023. URL https://jmlr.org/papers/v24/21-1524.html.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies. *Journal of Machine Learning Research*, 9(8):235–284, 2008. URL https://jmlr.org/papers/v9/krause08a.html.

Randall J. LeVeque. *Finite Difference Methods for Ordinary and Partial Differential Equations*. Society for Industrial and Applied Mathematics, 2007. doi: 10.1137/1.9780898717839.

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

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. *ACM / IMS Journal of Data Science*, 1(3):1–27, 2024. doi: 10.1145/3648506. URL https://doi.org/10.1145/3648506.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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

Yining Luo, Yingfa Chen, and Zhen Zhang. CFDBench: A Large-Scale Benchmark for Machine Learning Methods in Fluid Dynamics. *arXiv preprint arXiv:2310.05963*, 2023. URL https://arxiv.org/abs/2310.05963.

Krithika Manohar, Bingni W. Brunton, J. Nathan Kutz, and Steven L. Brunton. Data-Driven Sparse Sensor Placement for Reconstruction: Demonstrating the Benefits of Exploiting Known Patterns. *IEEE Control Systems*, 38(3):63–86, 2018. doi: 10.1109/MCS.2018.2810460.

Daniel Musekamp, Marimuthu Kalimuthu, David Holzmüller, Makoto Takamoto, and Mathias Niepert. Active learning for neural PDE solvers. In *The Thirteenth International Conference on Learning Representations*. OpenReview.net, 2025. URL https://openreview.net/forum?id=x4ZmQaumRg.

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

Karl Otness, Arvi Gjoka, Joan Bruna, Daniele Panozzo, Benjamin Peherstorfer, Teseo Schneider, and Denis Zorin. An Extensible Benchmark Suite for Learning to Simulate Physical Systems. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/3def184ad8f4755ff269862ea77393dd-Abstract-round1.html.

C. C. Paige and M. A. Saunders. Solution of Sparse Indefinite Systems of Linear Equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975. doi: 10.1137/0712047.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1912.01703.

G. S. Patterson and Steven A. Orszag. Spectral Calculations of Isotropic Turbulence: Efficient Removal of Aliasing Interactions. *The Physics of Fluids*, 14(11):2538–2541, 1971. doi: 10.1063/1.1693365.

Benjamin Peherstorfer, Zlatko Drmač, and Serkan Gugercin. Stability of Discrete Empirical Interpolation and Gappy Proper Orthogonal Decomposition with Randomized and Deterministic Sampling Points. *SIAM Journal on Scientific Computing*, 42(5):A2837–A2864, 2020. doi: 10.1137/19M1307391.

Maziar Raissi. Deep Hidden Physics Models: Deep Learning of Nonlinear Partial Differential Equations. *Journal of Machine Learning Research*, 19(25):1–24, 2018. URL https://jmlr.org/papers/v19/18-046.html.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Numerical Gaussian Processes for Time-Dependent and Nonlinear Partial Differential Equations. *SIAM Journal on Scientific Computing*, 40(1):A172–A198, 2018. doi: 10.1137/17M1120762.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045. URL https://doi.org/10.1016/j.jcp.2018.10.045.

Maziar Raissi, Alireza Yazdani, and George Em Karniadakis. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science*, 367(6481):1026–1030, 2020. doi: 10.1126/science.aaw4741.

Bogdan Raonić, Roberto Molinaro, Tim De Ryck, Tobias Rohner, Francesca Bartolucci, Rima Alaifari, Siddhartha Mishra, and Emmanuel de Bézenac. Convolutional neural operators for robust and accurate learning of PDEs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77187–77200, 2023. doi: 10.52202/075280-3376. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f3c1951b34f7f55ffaecada7fde6bd5a-Abstract-Conference.html.

Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020. doi: 10.1029/2020MS002203.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024. doi: 10.1029/2023MS004019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28. URL https://doi.org/10.1007/978-3-319-24574-4_28.

Clarence W. Rowley and Scott T.M. Dawson. Model Reduction for Flow Analysis and Control. *Annual Review of Fluid Mechanics*, 49(1):387–417, 2017. doi: 10.1146/annurev-fluid-010816-060042.

Yousef Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, second edition, 2003. doi: 10.1137/1.9780898718003.

Javier E. Santos, Zachary R. Fox, Arvind Mohan, Daniel O'Malley, Hari Viswanathan, and Nicholas Lubbers. Development of the Senseiver for efficient field reconstruction from sparse observations. *Nature Machine Intelligence*, 5:1317–1325, 2023. doi: 10.1038/s42256-023-00746-x. URL https://www.nature.com/articles/s42256-023-00746-x.

Louis Serrano, Lise Le Boudec, Armand Kassaï Koupaï, Thomas X Wang, Yuan Yin, Jean-Noël Vittaut, and Patrick Gallinari. Operator learning with neural fields: Tackling PDEs on general geometries. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3093. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/df54302388bbc145aacaa1a54a4a5933-Abstract-Conference.html.

Leslie N. Smith and Nicholay Topin. Super-convergence: very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*. SPIE, 2019. doi: 10.1117/12.2520589.

Gilbert Strang. On the Construction and Comparison of Difference Schemes. *SIAM Journal on Numerical Analysis*, 5(3):506–517, 1968. doi: 10.1137/0705041.

A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010. doi: 10.1017/S0962492910000061.

Luning Sun and Jian-Xun Wang. Physics-constrained bayesian neural network for fluid flow reconstruction with sparse and noisy data. *Theoretical and Applied Mechanics Letters*, 10(3): 161–169, 2020. doi: 10.1016/j.taml.2020.01.031.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1596–1611, 2022. doi: 10.52202/068431-0117. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0a9747136d411fb83f0cf81820d44afb-Abstract-Datasets_and_Benchmarks.html.

Lloyd N. Trefethen. *Spectral Methods in MATLAB*. Society for Industrial and Applied Mathematics, 2000. doi: 10.1137/1.9780898719598.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Gege Wen, Zongyi Li, Kamyar Azizzadenesheli, Anima Anandkumar, and Sally M. Benson. U-FNO: An enhanced Fourier neural operator-based deep-learning model for multiphase flow. *Advances in Water Resources*, 163:104180, 2022. doi: 10.1016/j.advwatres.2022.104180. URL https://doi.org/10.1016/j.advwatres.2022.104180.

Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. *ACM Computing Surveys*, 55(4):1–37, 2022. doi: 10.1145/3514228.

K. Willcox. Unsteady flow sensing and estimation via the gappy proper orthogonal decomposition. *Computers & Fluids*, 35(2):208–226, 2006. doi: 10.1016/j.compfluid.2004.11.006.

Jan P. Williams, Olivia Zahn, and J. Nathan Kutz. Sensing with shallow recurrent decoder networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 480(2298): 20240054, 2024. doi: 10.1098/rspa.2024.0054.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for PDEs on general geometries. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53681–53705. PMLR, 2024. URL https://proceedings.mlr.press/v235/wu24r.html.

Jiachen Yao, Abbas Mammadov, Julius Berner, Gavin Kerrigan, Jong Chul Ye, Kamyar Azizzadenesheli, and Anima Anandkumar. Guided diffusion sampling on function spaces with applications to PDEs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 140872–140909, 2025. doi: 10.52202/085713-4233. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b8a28baaa25cebfa9345a8a2edf7f19d-Abstract-Conference.html.

B. Yildirim, C. Chryssostomidis, and G.E. Karniadakis. Efficient sensor placement for ocean measurements using low-dimensional concepts. *Ocean Modelling*, 27(3-4):160–173, 2009. doi: 10.1016/j.ocemod.2009.01.001.

Yuan Yin, Matthieu Kirchmeyer, Jean-Yves Franceschi, Alain Rakotomamonjy, and Patrick Gallinari. Continuous PDE dynamics forecasting with implicit neural representations. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=B73niNjbPs.

PDE-OBS: CONTROLLED EVALUATION ACROSS OBSERVATION PATTERNS

Supplemental Material

A Benchmark Specification and Implementations **19**

A.1 Task Contract and Information Permissions 19

A.2 Physical Systems, Numerical Routes, and Regimes 19

A.3 Data Design, Splits, and Storage 23

A.4 Observation Construction 24

A.5 Metric Definitions and Scoring Semantics 24

A.6 Baseline Adaptations and Training Recipes 25

A.7 Software and Artifact Access 27

A.8 Baseline Adaptation Notes 27

B Using and Extending the Released Code **27**

B.1 Records and Physical Inputs 28

B.2 Observation Namespaces 28

B.3 Models, Budgets, and Splits 29

B.4 Inference Inputs and Checkpoint Loading 29

B.5 Small End-to-End Example 30

B.6 Formal Evaluation and Reproduction Boundary 30

B.7 Extension Contracts 30

B.8 Local Inspection and Tests 31

B.9 Data Format, Benchmark Access, Maintenance, and Extensibility 31

C Metric Validity and Reproducibility **32**

C.1 A new prediction-level result version 32

C.2 Score changes and uncertainty 32

C.3 Data, weights, and anonymous availability 33

C.4 Training provenance and reproducible arithmetic 33

D Configuration-Matched Original-Budget Subset **34**

E Completed Diagnostics and Further Controls **35**

F Complete Prediction-Level Results **36**

F.1 PDE-level overview 36

F.2 Training cohorts and aggregation 37

F.3 All pair-level summaries 38
F.4 All forecasting horizons 40
F.5 All 49 train-view by test-view matrices 41

G Mixed-Pattern Training **58**

A BENCHMARK SPECIFICATION AND IMPLEMENTATIONS

This appendix specifies the tasks, generators, observations, losses, and baseline adaptations used to interpret the reported results. Implementation references are paths within the anonymous repository linked in the main text. We distinguish the configurable software, the selected experimental protocol, and the newly verified outcomes of retained checkpoints. An available configuration is not evidence that its experiment was executed; a complete score table is not evidence of a new prediction audit. Appendix C gives the artifact and scoring limitations.

A.1 TASK CONTRACT AND INFORMATION PERMISSIONS

Let c denote a complete physical condition, u_t a complete state, M a binary observation mask, and G the geometry channel. Here $M = 1$ means observed and $G = 1$ means solid or boundary geometry. Recovery observes the solution itself, not the coefficient or source that generated it. Rollout observes an initial state and predicts future states; for the selected temporal records, the stored condition is that initial state.

Task	Predictor input	Supervised target	Role in this paper
Recovery	$M \odot u_t$, M, G	complete u_t	stationary experiments
Rollout	masked history, masks, G	complete future states	temporal experiments
Forward	$M \odot c$, M, G	complete solution	implemented extension
Inverse	$M \odot u_t$, M, G	complete condition c	implemented extension

The complete record is available to the dataset and loss code, but is not passed wholesale to the predictor. Unobserved input entries are zero-filled; the mask distinguishes missing data from a measured physical zero. Complete targets provide training supervision and are held separately for scoring at inference. PINO additionally uses privileged *training-loss* information: the coefficient or source for stationary residuals, and physical parameters and frame times for temporal residuals. This does not give PINO complete inputs or future targets at inference, but makes its training information different from a purely data-driven baseline. Third-party adapters must declare their own access; the interface does not audit arbitrary plugins.

The reported rollout is

$$\widehat{u}_1 = F(M \odot u_0, M, G), \qquad \widehat{u}_{h+1} = F(\widehat{u}_h, \mathbf{1}, G), \quad h = 1, 2. \tag{4}$$

The autoregressive wrapper in `src/pdeobs/methods/neural.py` clears the initial mask after the first prediction; the one-step models supply an all-one mask when it is absent. Predictions are fed back without detaching the computation graph. The selected protocol fixes teacher forcing to zero: future truth enters the loss, never the recurrent input. Three transitions between saved frames do not establish long-time forecasting stability.

A.2 PHYSICAL SYSTEMS, NUMERICAL ROUTES, AND REGIMES

Equations and variables. On the unit square, the selected stationary systems are

$$-\nabla\cdot(a\nabla u) = f \quad \text{(Darcy)}, \qquad -\Delta u = f \quad \text{(Poisson)}, \qquad (-\Delta - k^2)u = f \quad \text{(Helmholtz)}. \tag{5}$$

The temporal systems use periodic boundaries in the reported subset:

$$\begin{aligned} \partial_t u &= \alpha \Delta u, && \text{Heat,} \\ \partial_t u &= 0.008\,\Delta u + \rho(u - u^3), && \text{reaction–diffusion,} \\ \partial_t u + u(\partial_{x_1} u + \partial_{x_2} u) &= \nu \Delta u, && \text{scalar Burgers,} \\ \partial_t \omega + \boldsymbol{v}\cdot\nabla\omega &= \nu\Delta\omega + f, && \text{Navier–Stokes.} \end{aligned} \tag{6}$$

Reaction–diffusion is Allen–Cahn-type and scalar, not a two-species system. Burgers is a two-dimensional scalar equation, not a two-component velocity equation. Navier–Stokes is stored and predicted as vorticity, with $-\Delta\psi = \omega$ and $\boldsymbol{v} = (\partial_{x_2}\psi, -\partial_{x_1}\psi)$. Its fixed forcing is $0.1[\sin(2\pi(x_1 + x_2)) + \cos(2\pi(x_1 + x_2))]$. Darcy uses $f = \sin(2\pi x_1)\sin(2\pi x_2) + 0.2\sin(4\pi x_1 + 0.3)\sin(2\pi x_2)$, centered to zero discrete mean. The family modules under `src/pdeobs/pdes/` define these equations and parameters.

Table 1: Evaluated task definitions. Each family uses 2,000 scalar fields or trajectories on a 128×128 grid. Temporal targets are frames 1–3 of 15 equally spaced frames, $t_h = hT/14$. All inputs are noise-free partial observations.

Family	Prediction	Boundary	T
Darcy	Solution-field recovery	Dirichlet	–
Poisson	Solution-field recovery	Dirichlet	–
Helmholtz	Solution-field recovery	Dirichlet	–
Heat	Initial field to three future states	Periodic	.25
Reaction–diffusion	Initial field to three future states	Periodic	.8
Burgers	Initial field to three future states	Periodic	.35
Navier–Stokes	Initial vorticity to three future states	Periodic	.4

Table 2: Parameters in low/medium/high order. Regime names are family-specific, not a common difficulty scale.

Family	Parameter	Values
Darcy	coefficient contrast χ	2, 8, 32
Poisson	source amplitude	0.5, 1, 2
Helmholtz	wavenumber k; damping	0.5, 1, 2; zero
Heat	diffusivity α	0.005, 0.02, 0.08
Reaction–diffusion	rate ρ; diffusivity	0.5, 1.5, 4; 0.008
Burgers	viscosity ν	0.04, 0.015, 0.004
Navier–Stokes	viscosity; initial-vorticity scale	0.02, 0.008, 0.0025; 2, 3, 4

Condition generation. The selected construction, `smooth_grf` in `src/pdeobs/settings.py`, filters standard-normal grid noise in Fourier space with

$$A(q_1, q_2) = \exp[-2(\pi\ell)^2 HW(q_1^2 + q_2^2)], \qquad \ell = 0.22, \tag{7}$$

where q is frequency in cycles per grid cell, then centers and standardizes the field. Darcy range-normalizes the latent field to $s \in [-1, 1]$ and sets $a = \exp[\frac{1}{2}\log(\chi)s]$. Poisson multiplies it by the source amplitude; Helmholtz uses it directly as the source. Heat and Burgers use it as the initial state, reaction–diffusion multiplies it by 0.8, and Navier–Stokes applies the regime-specific vorticity scale and removes the mean. These are generation transforms, not normalizations fitted to test data.

Grid, integration, and stored times. The selected records use 128×128 cell-center coordinates $x_j = (j + \frac{1}{2})/128$. For stationary Dirichlet cases, the outer stored rows and columns are zero and the system is solved on the interior. Geometry marks this ring; the reported subset has no interior obstacle. Numerical coordinates must not be confused with the coordinate features constructed internally by a neural adapter.

The kernels in `src/pdeobs/pdes/numerics.py` assemble a sparse second-order flux operator with arithmetic face coefficients. Darcy and Poisson use diagonally preconditioned conjugate gradients; Helmholtz uses MINRES for the negative reaction term. The selected small Dirichlet wavenumbers do not establish a high-frequency, indefinite Helmholtz benchmark merely because MINRES is used. The solvers request relative tolerance 10^{-9}, record the attained residual, and raise on iterative failure. Periodic Heat uses Fourier diffusion; reaction–diffusion uses Strang splitting with an analytic reaction update. The selected smooth periodic Burgers cases use dealiased spectral advection with SSPRK2 stages and diffusion updates. Periodic Navier–Stokes uses dealiased vorticity advection and Crank–Nicolson diffusion, with internal increments bounded by 10^{-4}. Numerical arrays are computed in float64 and the selected storage dtype is float32.

`configs/dataset/numerics_full_t15.yaml` specifies dense temporal outputs for residual checks and selects 15 exact frames by an integer stride, not interpolation. With final stored time T in Table 1, the saved times are $t_h = hT/14$. The benchmark observes saved frame 0 and predicts saved frames 1–3; solver-frame indices and physical times are retained in metadata. It does not evaluate a forecast over the entire stored trajectory.

What quality checks establish. Generation checks array contracts, finiteness, geometry, boundary and initial-condition defects, transition replay, and family-specific discrete PDE residuals. Default tolerances include finite fraction 1, geometry error 10^{-6}, initial-condition defect 10^{-6}, boundary defect 10^{-4}, and first-transition replay defect 5×10^{-6}. The selected numerical configuration inherits a `strict` profile with normalized PDE-residual threshold 0.05 and divergence threshold 10^{-6} where applicable. These are discrete consistency checks, not bounds on error against an independent continuum solution.

The profiles are not interchangeable. `report` records threshold outcomes but still rejects hard array, domain, or residual-contract failures; `strict` also rejects failing quality status. The current `publication` path requires independently verified solver/calibration evidence and keeps publication readiness false because no trusted verification mechanism is configured. Neither the profile name nor a checksum-shaped metadata field certifies publication quality.

Recorded production quality. The original generation summary, whose digest matches the dataset pin of all 441 archived settings, records 560,000 passing records, with no missing or invalid quality entries and 80,000 records per family. Filtering its disjoint calibration groups to the evaluated boundary and `smooth_grf` construction gives exactly 2,000 records per family (regimes 667/667/666). Table 3 reports their recorded normalized discrete PDE residuals; every maximum is below 0.05. The anonymous artifact includes the full-precision extraction, source digest, operator identifiers and aggregation procedure in `results/numerical_quality/`. These are verified generation-time records, not a new solution of the full corpus or independent reference-solution error bounds. No grid-convergence study is inferred from a small residual.

Table 3: Recorded production consistency for the evaluated slice (2,000 records per family). Residuals use the family-specific discrete operator and normalization, not a common continuum-error scale.

PDE	Mean normalized residual	Maximum normalized residual
Darcy	1.14×10^{-5}	1.57×10^{-5}
Poisson	1.69×10^{-5}	2.84×10^{-5}
Helmholtz	1.80×10^{-5}	3.52×10^{-5}
Heat	1.32×10^{-8}	1.51×10^{-8}
Reaction–diffusion	3.39×10^{-5}	1.55×10^{-4}
Burgers	1.26×10^{-2}	3.85×10^{-2}
Navier–Stokes	5.08×10^{-5}	1.37×10^{-4}

A.2.1 Discrete Updates and Production Residuals

The parameters below describe the seven routes used in the evaluated slice, not every optional solver in the package. They are traced to the family generators and `src/pdeobs/pdes/numerics.py`, with the original production calibration contexts retained in `results/numerical_quality/`. The residual summary predates model training. No additional numerical solutions are generated for this analysis.

Elliptic systems. Let $h = 1/128$. For Darcy, a face coefficient is the arithmetic average of its neighboring cell values, for example $a_{i+1/2,j} = (a_{i+1,j} + a_{i,j})/2$. The discrete operator (LeVeque, 2007) is

$$\begin{aligned}(A_a u)_{i,j} = &- \frac{a_{i+1/2,j}(u_{i+1,j} - u_{i,j}) - a_{i-1/2,j}(u_{i,j} - u_{i-1,j})}{h^2} \\ &- \frac{a_{i,j+1/2}(u_{i,j+1} - u_{i,j}) - a_{i,j-1/2}(u_{i,j} - u_{i,j-1})}{h^2}.\end{aligned} \tag{8}$$

The unknowns are the interior samples; prescribed zero outer-layer values enter the assembled boundary-value problem. Darcy solves $A_a u = f$, Poisson solves $A_1 u = f$, and Helmholtz solves $(A_1 - k^2 I)u = f$. The diagonal preconditioner (Saad, 2003) divides by $\max(|A_{ii}|, 10^{-12})$. The requested relative tolerance is 10^{-9}; iteration caps are 6,000, 4,000, and 6,000, respectively. Conjugate gradients (Hestenes & Stiefel, 1952) are used for Darcy/Poisson and MINRES (Paige & Saunders, 1975) for Helmholtz, whose small selected wavenumbers remain coercive on this Dirichlet domain.

The stored solution is float32, whereas assembly and iteration use float64; the reported post-storage residual is therefore distinct from the solver's stopping tolerance.

Diffusion and reaction–diffusion. For periodic diffusion, a substep of length δ multiplies each Fourier coefficient by $\exp(-\alpha\delta|\xi|^2)$, with $\xi = 2\pi\,\mathrm{fftfreq}(128, h)$. Heat uses this exact semidiscrete propagation (Trefethen, 2000; Boyd, 2001). Reaction–diffusion alternates exact half reaction, exact Fourier diffusion, and exact half reaction (Strang, 1968). The reaction flow is

$$\mathcal{R}_s(u) = \frac{u}{\sqrt{u^2 + (1-u^2)e^{-2\rho s}}}, \qquad u^{n+1} = \mathcal{R}_{\delta/2}\big(e^{\kappa\delta\Delta_h}\mathcal{R}_{\delta/2}(u^n)\big)\,. \tag{9}$$

The code floors the square-root argument at float64 machine epsilon. Each dense output interval Δ_q is divided into $m = \max(1, \lceil \rho\Delta_q/0.18 \rceil)$ substeps with $\delta = \Delta_q/m$. The selected regimes and output intervals give $m = 1$; no state clipping is applied.

Burgers. The selected smooth periodic route differentiates spectrally and filters the nonlinear product to integer Fourier modes satisfying $|k_x|, |k_y| \le \lfloor 128/3 \rfloor$ (Patterson & Orszag, 1971). Writing the negative dealiased advection as $N(u)$, a substep forms (Gottlieb & Shu, 1998) $u^* = u^n + \delta N(u^n)$ and $\widetilde{u} = u^n + \frac{\delta}{2}[N(u^n) + N(u^*)]$, followed by $u^{n+1} = e^{\nu\delta\Delta_h}\widetilde{u}$. The step is

$$\delta = \min\left(\text{remaining output interval}, \frac{0.35}{\max(\|u^n\|_\infty, 10^{-12})(h^{-1} + h^{-1})}\right). \tag{10}$$

The controller raises on nonfinite states or more than 10,000 substeps in one output interval. It does not clip the state. The package's Rusanov alternative is used for other constructions or boundaries, not for the smooth periodic slice plotted here.

Periodic vorticity. With $\widehat{\psi} = \widehat{\omega}/|\xi|^2$ and $\widehat{\psi}(0) = 0$, velocity is obtained by Fourier differentiation. Let $\widehat{N}$ be the two-thirds-filtered transform of $\boldsymbol{v} \cdot \nabla\omega$. The update is

$$\widehat{\omega}^{n+1} = \frac{(1 - \nu\delta|\xi|^2/2)\widehat{\omega}^n - \delta\widehat{N}^n + \delta\widehat{f}}{1 + \nu\delta|\xi|^2/2}. \tag{11}$$

An output interval uses $m = \max(1, \lceil \Delta_q/10^{-4} \rceil)$ substeps, $\delta = \Delta_q/m$; mean vorticity is removed at the interval boundary. Nonlinear advection and forcing are explicit, while diffusion is Crank–Nicolson (Crank & Nicolson, 1947). The forcing and regime-dependent initial amplitudes are given above; this is the vorticity route, not a velocity–pressure or obstacle-flow simulation.

Table 4: Time grids in the evaluated periodic slice. Dense quality frames include both endpoints; stored data select every indicated stride. Internal substeps are separate from both grids.

PDE	Final T	Dense frames	Stride	Stored frames	Internal rule
Heat	0.25	155	11	15	Fourier propagation
Reaction–diffusion	0.8	267	19	15	$\rho\delta \le 0.18$
Burgers	0.35	267	19	15	CFL bound above
Navier–Stokes	0.4	71	5	15	$\delta \le 10^{-4}$

What the plotted PDE loss measures. For each record, `src/pdeobs/quality.py` computes

$$q = \frac{\mathrm{RMS}_{\Omega_a}(r)}{\sum_j \mathrm{RMS}_{\Omega_a}(a_j)}, \qquad r = \sum_j s_j a_j, \tag{12}$$

where $s_j \in \{-1, 1\}$ are the equation signs and Ω_a contains valid stencils. If the denominator is at most 10^{-12}, the code returns zero when the numerator is also at most 10^{-12}, and otherwise divides by 10^{-12}. For the selected stationary wall geometry, the outer ring and its one-cell stencil halo are excluded from quality scoring; periodic scoring uses all cells. This domain differs from the full-grid neural prediction metric.

The stationary components are the complete elliptic operator and its source. For the nonlinear temporal families, the quality checker uses a finite difference in time and midpoint states for diffusion,

reaction or dealiased advection, with the original physical coefficients and forcing. The principal reported residual aggregates post-initial *dense quality* intervals; the first dense transition is checked separately by replaying the same numerical update, with threshold 5×10^{-6}. For Heat the principal operator is instead the spectral semigroup defect $u^{n+1} - e^{\alpha \Delta_q \Delta_h} u^n$, normalized by the two state RMS values; the finite-difference strong-form defect is auxiliary. This explains why its scale should not be interpreted as a cross-family solution-accuracy ranking.

The bars in Figure 4 show the arithmetic mean of these per-record values over all 2,000 selected records per family, including both training and test identities. Six family means are below 1%; Burgers has mean 1.26%. The 1% reference is not the configured 5% generation gate, and no value is clipped to it. Extrema remain reported in Table 3. These are neither training losses nor independent discretization-error estimates. Calibration groups are disjoint: counts and first moments are summed with their recorded sample weights, second moments recover the pooled variance, and extrema are taken across groups. The separate initial replay verifies implementation consistency, not an independent reference solution, and is not a replay over the longer stored $t_0 \to t_1$ interval. Analytic/reference error and grid-convergence evidence for all seven production routes remains outside this result cut.

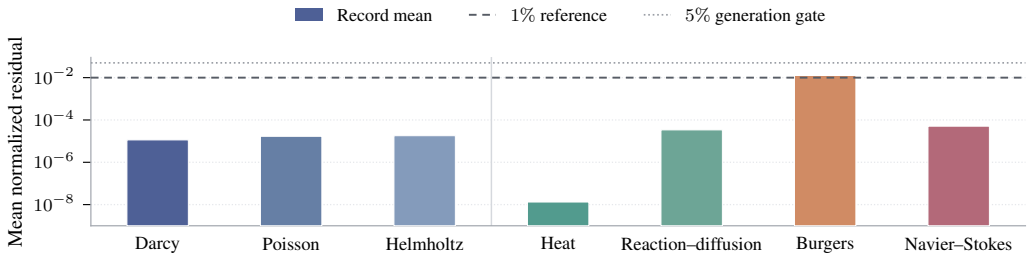

Figure 4: **Mean recorded PDE residuals.** Each bar is the arithmetic mean over 2,000 selected records, shown on a logarithmic axis. Six family means are below 1%; Burgers is 1.26%. The dashed line marks the 1% reference and the dotted line the original 5% generation gate. Residual operators and normalizations are family-specific, with a semigroup defect for Heat; these values measure production discrete consistency, not independent solution error.

A.3 Data Design, Splits, and Storage

Design versus evaluated slice. The full software design has seven families, four boundary protocols, and ten condition constructions, with 2,000 records per combination. The regimes *partition* those records as 667/667/666; they do not multiply the count. The design therefore has $7 \times 4 \times 10 \times 2000 = 560{,}000$ records. The paper uses seven combinations: Dirichlet with `smooth_grf` for stationary families and periodic with `smooth_grf` for temporal families, totaling 14,000 records. Other boundaries and constructions are implemented capabilities, not additional reported experiments. Release documentation lists 3,360 HDF5 shards for the full design and 84 for the selected slice; these are inventory declarations, not a fresh full-corpus verification here.

Exact split. `src/pdeobs/one_setting.py::stable_split` requires 2,000 records and all three regimes. Within each regime it ranks identities by SHA-256 of `seed|pde|boundary|setting|sample_id`, using seed 20260804. Largest-remainder allocation assigns 67/67/66 test identities and leaves 600 training identities per regime: 1,800 training, no validation, and 200 test records. The same identity lists are used across methods and observation views. A loader label such as `iid` does not reconstruct this partition; the identity lists and split receipt must be retained. Test identities must not be used for early stopping, hyperparameter selection, or checkpoint selection.

Canonical storage. Each HDF5 shard stores conditions of shape $[N, H, W, C_c]$, trajectories of shape $[N, T, H, W, C_s]$, and geometry of shape $[N, H, W, 1]$. The array keys are `condition`, `trajectory`, and `geometry`; each identity also has a UTF-8 JSON metadata record. Here $C_c = C_s = 1$, with $T = 1$ for stationary cases and $T = 15$ for temporal cases. Metadata

Table 5: Exact observation counts out of 16,384 locations. Equal-count comparisons use R50/H/V/CL or BL/LI/BD, not all nominal 50% masks together. A perimeter band on a periodic grid is not a physical wall.

Code	Pattern	Count K	Construction
R50	Random	8192	Uniform distinct locations
R65	Random	10650	Uniform distinct locations
R80	Random	13107	Uniform distinct locations
BL	Missing block	8284	Outside a missing 90×90 square
LI	Crossed lines	8284	Union of 38 rows and 38 columns
H	Horizontal lines	8192	64 distinct rows
V	Vertical lines	8192	64 distinct columns
BD	Perimeter band	8284	A 19-cell band
CL	Clustered points	8192	Four Gaussian clusters; distinct locations

includes family, boundary, construction, regime, seeds, identity, frame indices, and physical times. `src/pdeobs/storage.py` writes compressed arrays and atomically finalizes temporary shards, with checksum, manifest, metadata, and quality sidecars. Default compression is gzip level 4. Identity derives from the generation specification, not a row's file position. Regenerated records are not automatically the historical evaluated records: their manifests and identity receipts must agree.

A.4 Observation Construction

Views are frozen in `configs/paper/observations.yaml` and implemented by `src/pdeobs/masks.py`. For grid Ω_h, the realized count is $K_{i,w} = \sum_{x \in \Omega_h} M_{i,w}(x)$ and density is $r_{i,w} = K_{i,w}/|\Omega_h|$. Integer counts, rather than nominal percentages, define the equal-count groups. R50, R65, and R80 sample distinct cells uniformly without replacement with the requested counts. These are separate calls under the deterministic seed policy, not a constructed nested sequence; neither nesting nor statistical independence is assumed. H and V choose 64 distinct rows or columns. LI observes 38 rows and 38 columns, giving $38 \cdot 128 + 38 \cdot 128 - 38^2 = 8{,}284$ cells. BL observes outside a non-wrapping 90×90 missing square whose top-left corner is randomly drawn from valid positions. BD observes the deterministic width-19 band, also giving 8,284 cells. CL samples 8,192 distinct cells using four Gaussian clusters, random centers in array coordinates, and standard deviation $0.08 \cdot 128$ cells. Blocks and cluster distances do not wrap across periodic edges.

Factories operate on array locations and do not intersect observations with geometry. A stationary Dirichlet record has 508 prescribed-zero outer-ring cells; selecting one counts as an observation. BD includes that ring and therefore observes 7,776 interior cells. Other views contain different boundary-cell counts, so equal totals need not imply equal numbers of unknown interior values. The metric includes boundary entries; this fixed convention does not remove that information-content confound. On periodic data, BD is a perimeter sampling layout, not a new physical wall.

`BenchmarkDataset` derives a stable mask seed from the campaign seed, `mask`, sample identity, and canonical factory name. There is no epoch or method index, so a record's mask stays fixed across epochs and methods under a fixed view. The `paper-row` path separates this seed from the model training seed. The runner configuration exposes the same separation: `data.mask_seed` pins the sensors independently of the campaign seed and `training.seed` seeds only initialization and batch order. Specifications carry namespace, parameters, version, and a content-derived identifier. Changing a general API seed is not equivalent to changing only initialization. The selected views are noiseless, stationary point observations; noise, moving sensors, and active sampling are not reported controls.

A.5 Metric Definitions and Scoring Semantics

For finite predictions and targets, the intended per-identity error and block mean are

$$e_i^{v,w} = \frac{\|\widehat{y}_i^{v,w} - y_i\|_2}{\max(\|y_i\|_2, 10^{-12})}, \qquad E_{v,w} = \frac{1}{N_{\text{test}}} \sum_{i \in \mathcal{I}_{\text{test}}} e_i^{v,w}, \qquad N_{\text{test}} = 200. \tag{13}$$

The norm includes every stored spatial location, including observed cells, and jointly includes all three target frames for rollout. Ratios are not percentages unless labelled. For nine views $\mathcal{V}$,

$$D = \frac{1}{9}\sum_{v} E_{v,v}, \qquad C = \frac{1}{72}\sum_{v \neq w} E_{v,w}, \tag{14}$$

and the destination-referenced contrast is

$$\Delta_{v \to w} = E_{v,w} - E_{w,w}. \tag{15}$$

Both terms use the same test view and identities. Over all transfers, mean $\Delta = C - D$; the contrast is useful directionally or within restricted groups, not as an independent third whole-matrix summary. Complete-matrix summaries require every cell.

Separate diagnostics. Horizon-wise errors are

$$e_{i,h}^{v,w} = \frac{\|\widehat{u}_{i,h}^{v,w} - u_{i,h}\|_2}{\max(\|u_{i,h}\|_2, 10^{-12})}, \quad h = 1, 2, 3. \tag{16}$$

Their mean is generally not the joint error. For stationary recovery, set $R_i = \widehat{u}_i - u_i$ and $q_i = \max(\|u_i\|_2, 10^{-12})$. Then

$$e_{\mathrm{obs},i} = \frac{\|M_i \odot R_i\|_2}{q_i}, \quad e_{\mathrm{hid},i} = \frac{\|(1 - M_i) \odot R_i\|_2}{q_i}, \quad e_i^2 = e_{\mathrm{obs},i}^2 + e_{\mathrm{hid},i}^2. \tag{17}$$

This identity holds per record, not after separately averaging the errors. Hidden-only normalization by $\|(1 - M_i) \odot u_i\|_2$ is a different diagnostic. Replacing observed predictions with supplied noiseless values leaves hidden predictions unchanged and cannot increase full-field recovery error. The stationary projection results are reported separately in Appendix E; projection cannot inject future truth into a forecast.

Strict scorer versus archive. The `pdeobs-strict-v1` scorer uses an explicit contract fixing expected identities, shape, task, and target-frame indices. It aligns by identity and rejects missing, extra, duplicate, nonfinite, or mismatched entries before float64 norms and ordinary arithmetic means. Recovery also checks raw observations and masks against targets before projection. These checks do not authenticate provenance, establish numerical fidelity, or prove generalization. The implementation is in `src/pdeobs/strict_score.py`. The original campaign metrics remain a separate legacy archive. The current manuscript instead uses new predictions and strict scores for every retained checkpoint, with the terminal audit described in Appendix C; original arrays are not retrospectively recovered.

A.6 Baseline Adaptations and Training Recipes

Reference architectures. The baselines are PDE-OBS adaptations, not interchangeable copies of upstream pipelines. Their reference registry is `configs/method/all_pde_source_faithful_registry.yaml`; exposed schemas and presets are in `src/pdeobs/api/specs.py`. U-FNO-2D has three Fourier and three U-Fourier blocks, width 36, 12 modes, and padding 8. FNO uses spectral convolution with a parallel local path. The CNO-inspired CNN uses fixed binomial filtering in an encoder–decoder; its scores are not those of an exact upstream CNO implementation. DeepONet uses an ordered spatial CNN branch and a branch–trunk product. GNOT uses linear attention and geometry-gated experts without upstream unit normalization. Transolver uses physics-slice attention and a larger temporal preset. PINO combines an FNO backbone with a trainer-side residual. Registry revisions, licenses, and adaptation notes provide attribution, not a new independent equivalence or license audit.

Original recipe and later attempts. The original reference budget is 500 epochs over 1,800 training records, with no validation partition, plateau stopping, or mixed precision, and with the final checkpoint retained. Its initial or peak learning rate is 10^{-3}. For minibatch $\mathcal{B}$,

$$\mathcal{L}_{\mathrm{data}} = \frac{1}{|\mathcal{B}|}\sum_{i \in \mathcal{B}} \frac{\|\widehat{y}_i - y_i\|_2}{\max(\|y_i\|_2, 10^{-12})}, \tag{18}$$

Table 6: Frozen reference model sizes and nominal per-device batch sizes. These describe the reference registry, not a proof of identical effective batches or capacities in every execution attempt. Complex weights are counted as real scalars. Attempt-specific configurations remain necessary for reproduction.

Adaptation	Reference configuration	Parameters	Batch
U-FNO-2D	Width 36; Fourier and U blocks	5,049,545	4
FNO	Width 32; 12 modes; 4 layers	2,366,145	8
CNO-inspired CNN	Width 32; filtered convolution	103,617	8
DeepONet	Spatial CNN branch; latent 128	221,121	4
GNOT	Width 128; 3 layers; 2 experts	2,099,975	4
Transolver (static)	Width 128; 8 layers; 64 slices	2,811,457	4
Transolver (temporal)	Width 256; 8 layers; 32 slices	11,196,737	2
PINO	FNO width 64; 20 modes; 5 layers	32,798,273	4

Table 7: Original optimization defaults. Step schedules halve the rate every 100 epochs; one-cycle schedules advance per batch with cosine annealing. Parentheses denote temporal Transolver settings.

Method	Optimizer	Weight decay	Batch	Schedule	Grad. clip
U-FNO-2D	Adam	10^{-4}	4	step	1.0
FNO	Adam	10^{-4}	8	step	1.0
CNO-inspired CNN	AdamW	10^{-6}	8	step	1.0
DeepONet	Adam	0	4	step	1.0
GNOT	AdamW	5×10^{-5}	4	one-cycle, 20% warm-up	1000
Transolver	AdamW	10^{-5}	4 (2)	one-cycle, 30% warm-up	0.1 (none)
PINO	Adam	0	4	step	1.0

computed on float32 predictions and targets after flattening non-batch dimensions. Table 7 gives reference defaults, not a universal description of the archive. Actual configurations take precedence for later learning-rate, clipping, continuation, and stopping amendments. Actual epochs, cohorts, stop reasons, and checkpoint identities remain in the results. The public `paper-row` launcher executes `original500` and `demo`; historical fixed-200, patience, and recovery labels are not executable replays of those attempts in that launcher.

PINO residual. The reference objective is $5\mathcal{L}_{\text{data}} + \mathcal{L}_{\text{phys}}$. Complete targets enter the data term; the residual uses predictions and declared physical context. For active cells $\mathcal{A}$,

$$\mathcal{L}_{\text{phys}} = \frac{1}{|\mathcal{A}|} \sum_{p \in \mathcal{A}} (r_p/\sigma)^2, \qquad \sigma = \max\left(\sum_j \text{RMS}_{\mathcal{A}}(\tau_j), 10^{-12} \right), \tag{19}$$

where a signed sum of PDE terms τ_j forms r; contributions are averaged over records and temporal pairs. Stationary residuals use finite-difference/flux stencils, stored coefficients or sources, and a detached normalization scale. The active mask excludes solid cells and their stencil neighbors, including the prescribed wall ring in the selected Dirichlet data. Temporal residuals use spectral spatial derivatives of the midpoint between two predicted states and their finite-difference time derivative. Only pairs $(\widehat{u}_1, \widehat{u}_2)$ and $(\widehat{u}_2, \widehat{u}_3)$ contribute, using recorded physical times; their normalization remains differentiable. `src/pdeobs/pino.py` implements distinct `pino_static_fd_v1` and `pino_rollout_spectral_v1` contracts, the latter periodic-only. Mixed precision is rejected while a physics residual is active.

Unreported comparators. The software also contains interpolation, gappy POD, zero-fill, observed-mean, and persistence methods. The method named `mean` fills using an example's observed mean, not a training-population mean. These methods and forward/inverse interfaces contribute no result to the 441-setting learned grid.

A.7 Software and Artifact Access

The inspected package is version 0.2.1 (0.2.0 plus the observation-spec interface of Appendix B.2) with review overlays and MIT-licensed code. It requires Python 3.10 or newer, NumPy, SciPy ≥ 1.12, h5py, and PyYAML; learned training adds PyTorch through the `train` extra. The package includes generators, adapters, configurations, tests, demonstrations, new per-record and aggregate scores, and the separate archived metrics, not the full tensor data or weight corpus. Data licensing is separate. Release documents point to separate data and model deposits, whose anonymous accessibility and inventory require independent checks; a README badge is not verification. The dated 26 September 2026 release report records unauthenticated access and file-digest agreement for the data and model deposits; inventory scope and reproduction limits are stated in Appendix C. `pdeobs download` requires an explicit checksum manifest and has no default endpoint. Checksums verify bytes against a manifest, not the identity or anonymity of its publisher. Appendices B and C distinguish runnable interfaces, aggregate arithmetic, inference, and training reproducibility.

A.8 Baseline Adaptation Notes

U-FNO. U-FNO (Wen et al., 2022) combines Fourier blocks with U-Net-style local branches (Ronneberger et al., 2015). We adapt its six-block composition from three-dimensional space–time to two spatial dimensions, adding mask, coordinate and geometry channels. It is explicitly a two-dimensional adaptation, not a replication of the original model.

FNO. FNO (Li et al., 2021) combines learned truncated Fourier multipliers with a pointwise path. We use a compact mask-channel variant; PINO below uses the same backbone family with a different width.

CNO-inspired CNN. Inspired by CNO (Raonić et al., 2023), our encoder–bottleneck–decoder applies a fixed binomial low-pass filter before each learned convolution. It does not reproduce the original continuous–discrete construction and is therefore labelled an inspired variant.

DeepONet. DeepONet (Lu et al., 2021) combines an input branch and coordinate trunk by an inner product. A small convolutional branch on the masked field retains sensor ordering, replacing the original fixed sensor vector; a set branch remains available for legacy checkpoints.

GNOT. GNOT (Hao et al., 2023) uses normalized linear cross-attention and geometry-gated mixture-of-expert layers. The structured-grid adaptation retains these mechanisms but omits upstream unit normalization, optimizing the physical relative-L^2 objective.

Transolver. Transolver (Wu et al., 2024) attends over learned physics-aware slices instead of grid points. We use structured-grid slice attention with the published Darcy and Navier–Stokes dimensions; it is the only baseline with a separate temporal structure.

PINO. PINO (Li et al., 2024) augments FNO's data loss with stationary finite-difference or temporal spectral residuals. Coefficients, sources and physical parameters enter the training residual, not inference. Complete targets enter only the data term. Its wider backbone prevents interpreting the FNO–PINO contrast as an isolated effect of physics supervision.

B Using and Extending the Released Code

This appendix separates a small software example from a formal benchmark run. The public entry points are `from pdeobs import api` and the `pdeobs` command. `configs/paper/protocol.yaml` describes the study; it is not itself a pipeline configuration or job launcher. Likewise, `preset="paper"` selects reference defaults, not a certified reproduction of a historical checkpoint.

Table 8: General observation controls. Record realized locations and counts; a requested ratio need not equal realized density. Alternatives separated by *or* are mutually exclusive.

Factory	Controls
Random	`ratio` or `count`; distinct cells
Grid	`ratio` or `spacing`; optional `random_offset`
Block	`missing_fraction` or `block_shape`; observe outside the block
Line / horizontal / vertical	`ratio` or `num_lines`; orientation for general lines
Boundary	`width`; optional deterministic `count` within the band
Clustered	`ratio` or `count`; `clusters` and `spread`
Full	no pattern parameters; diagnostic complete observation
Mixture	components joined by `+`, optional integer weights (`x2`); one component per record by stratum position

B.1 Records and Physical Inputs

Local generation. `api.create_dataset` takes explicit `pde` and `out` arguments. Options include boundary, construction, regime, count, resolution, seed, temporal frame counts, shard size, dtype, and numerical solver options. Static records have one frame; temporal calls must specify at least two. This convenience function uses the `report` quality profile, not the full numerical configuration's `strict` profile. Advanced generation resolves quality and dense/stored frame choices from `configs/dataset/numerics_full_t15.yaml` and its included configuration. A short count or failed solve is an error, not permission to substitute new identities.

The ten constructions comprise three Gaussian-field smoothness levels; low-frequency and multi-frequency Fourier fields; Gaussian blobs; piecewise blocks; threshold level sets; dipole vortex pairs; and front/ring/shock fields. Their exact registry keys are defined in `src/pdeobs/settings.py`. Gaussian-field correlation parameters are 0.22, 0.09, and 0.025. Only `smooth_grf` occurs in the reported grid. The boundary labels `periodic`, `dirichlet`, `neumann`, and `robin_obstacle` have family-dependent numerical meanings; they are not interchangeable solver settings.

Loading and verification. With `verify=True`, `api.load_dataset` verifies local shards against completion sidecars. With `source="manifest"`, it uses an identity manifest and verifies its shard hashes. The separate `pdeobs download --manifest <manifest>` route consumes a release download manifest, a different format from an identity manifest. The inspected `api.load_dataset(..., source="download")` route raises rather than downloading. Loading neither trains a model nor regenerates missing records.

External conditions. `api.create_dataset_from_arrays` accepts a family, boundary, output path, complete physical arrays, and physical parameters. It solves these specified problems with the existing kernels and writes canonical records. It requires a complete problem specification; it does not infer missing coefficients. These outputs are new data, not additional historical evidence.

B.2 Observation Namespaces

`api.make_observation` resolves four namespaces. `paper:R50` and its eight companions replay frozen definitions and reject overrides. The paper namespace requires a 128×128 grid in the training interface. The general namespace provides parameterized factories; custom delegates to a registered factory; stored reuses a supplied inference mask. Names, namespaces, versions, and resolved parameters contribute to the observation identifier. A general mask matching a paper mask numerically does not acquire paper provenance. `api.realized_count(spec, (H,W))` evaluates the discrete count. A mask spec is a protocol name with the observed fraction in percent (`uniform 20`, `block 50`, `boundary 50`) or with explicit arguments (`line_sensors(num_lines=64, orientation=horizontal)`), optionally weighted and joined by `+` into a mixture. The weight-expanded component cycle is rotated by a seed derived from the mask seed, the canonical spec and the record's stratum and indexed by the original generation-time stratum position. Assignments are fixed across epochs; selecting the training subset does not preserve exact balance (Appendix G). The per-record mask

seed uses the resolved protocol name only, hence `uniform 50` equals R50 and a component's mask does not depend on the other components. Specs are accepted in `data.mask.protocol`, `evaluation.observation_protocols`, `--obs` and `api.make_observation`; the paper namespace still rejects them (`docs/observation_specs.md`). The legacy key `random_3pct` can carry ratio 0.5; its string alone does not determine density. For a nested-density control, retain explicitly constructed nested masks and use stored inputs. A mixture assigns one component per identity by stratum position and does not resample identities across epochs; the mixed-pattern rows of Appendix G were trained with the nine frozen views as components, each with its own arguments, so every record's sensors are those of the corresponding single-pattern row. No nested-mask experiment is reported.

B.3 Models, Budgets, and Splits

Model schemas are inspectable with `api.list_models()`, `api.describe_model(name)`, and command-line description tools. `api.resolve_model` validates names, parameters, task compatibility, presets, and structural constraints. `api.check_grid_compatibility` identifies incompatible spectral modes or branch grids. Overrides yield a custom configuration rather than an unchanged paper preset. The training facade enforces batch size at least two for U-FNO's normalization-based branches. Equal input/output contracts do not imply equal parameter counts.

```python
from pdeobs import api
cfg = api.resolve_model("transolver", task="rollout",
                        preset="paper")
api.describe_model("gnot")
api.check_grid_compatibility(cfg, (128, 128))
```

Listing 1: Configuration-only inspection; no trained weights are loaded.

`api.train` requires an explicit epoch or optimizer-step budget. The archive's epoch budgets must not be relabelled as optimizer steps. The generic default is `split="all"`, so held-out evidence requires an explicit split. Routes include the exact `paper` split, stored metadata, deterministic `holdout:<fraction>`, and explicit train/test lists. The facade records those identities and does not evaluate its held-out subset during training. The base trainer can accept a validation loader directly, but `api.train` supplies none; naming a monitor `val_loss` does not create a validation set. New development subsets must come from training identities, never the paper test set.

`TrainingConfig` exposes objective weights, optimizer, learning rate, weight decay, clipping, schedules, precision, checkpointing, and health checks. These are capabilities, not permission to reinterpret old attempts. Retain a historical model's recorded final checkpoint rather than selecting one from test performance. Three seeds are independent: the campaign `seed` fixes the identity split and, by default, the sensors; `data.mask_seed` pins the sensors on its own; `training.seed` seeds initialization and batch order. Moving `training.seed` alone keeps the split and every mask of a row; the convenience API seed still affects masks and splits as well as initialization, so use the explicit keys or the formal `paper-row` separation when changing only the training seed. `configs/paper/training_cohorts.yaml` describes historical stopping semantics but does not implement every continuation controller.

B.4 Inference Inputs and Checkpoint Loading

`api.inference_input_from_dataset` returns an observation package and targets separately. The standard predictor receives masked values, masks, and required geometry, not hidden targets or PINO's training-loss condition. Custom `InferenceInput` arrays are channel-last: $[N, H, W, C]$ for recovery or $[N, T_{\text{hist}}, H, W, C]$ for rollout, with aligned one-channel binary masks. The constructor can generate placeholder identities or accept absent times; these fallbacks cannot establish correspondence with paper-test identities and source-frame indices. An executable grid/horizon or extrapolated time label is not a newly validated forecasting range.

`api.load_predictor` restores a new-format artifact from `model.json` and its weights, checking recorded weight hashes and refusing missing weights or mismatched structures by default. Legacy

`last.pt` files use `api.load_legacy_checkpoint` with explicit architecture and task. That adapter strictly matches model-state keys, cross-checks available task/physics metadata, and records unknown provenance fields. It does not authenticate missing identities or reconstruct omitted optimizer state. The release documentation describes stripped weights-only checkpoints: these support inference, not exact full-state continuation. A successful load is necessary but insufficient to reproduce a historical score.

B.5 Small End-to-End Example

Listing 2 checks the local CPU workflow on eight new records and two optimizer steps. It uses a general mask because frozen paper views are not defined on the demonstration grid. It is not a member of the 441-setting experiment; its score must not enter a paper table. Use fresh output directories, which are not overwritten by default.

```python
import torch
from pdeobs import api
torch.set_num_threads(1)
data = api.create_dataset(
    pde="poisson", boundary="dirichlet", out="demo-records",
    num_samples=8, resolution=16, seed=17)
obs = api.make_observation("general:random", ratio=0.5)
run = api.train(
    data, task="recovery", model={"name": "fno", "preset": "smoke"},
    observation=obs, split="holdout:0.25",
    budget={"max_steps": 2}, seed=17, device="cpu",
    out="demo-training")
predictor = api.load_predictor(run.artifact.path, device="cpu")
inputs, targets = api.inference_input_from_dataset(
    data, obs, task="recovery", sample_ids=run.split.test_ids, seed=17)
bundle = api.predict(predictor, inputs)
score = api.evaluate(bundle, targets, target_ids=inputs.sample_ids,
                     target_time_indices=bundle.time_indices)
assert score["status"] == "valid"
```

Listing 2: A small CPU example. Its two held-out demo records are not the 200 paper-test identities.

B.6 Formal Evaluation and Reproduction Boundary

The convenience `api.evaluate` call builds its default contract from the supplied bundle. It checks that supplied set, but cannot detect paper identities that were never declared. Its default scope is a user evaluation, not a paper performance estimate. For a formal block, freeze an independent contract with all 200 test identities, expected shape, frame coordinates, dataset/identity-manifest binding, observation definition, and checkpoint identity before prediction. For legacy loading, explicitly supply the loaded file's digest; a convenience `checkpoint_id` is not automatically populated from every legacy provenance format.

`pdeobs strict-infer` and `pdeobs strict-score` consume explicit materials and retain predictions for rescoring. The `paper-row` original-500 path writes the split, completion records, contracts, predictions, and nine test blocks. It requires all 2,000 identities across three regimes, not one partial shard. Its template is `configs/paper/row_original500.yaml`; campaign/registry paths must remain valid if copied. Running it trains a new model, rather than verifying an archived score. The other executable cohort, `demo`, is deliberately reduced. Historical fixed-200, min-120 patience, and recovered attempts require separately versioned attempt logic; their labels alone are not executable reproduction instructions.

B.7 Extension Contracts

Registries expose masks, condition constructions, PDE generators, methods, and metrics. Extensions must declare tasks and information permissions, use supplied randomness, return the expected schema, and preserve resolved parameters. Registration does not certify numerical correctness or leakage

freedom. Listing 3 defines a custom observation; other contracts are in `docs/extending.md` and `src/pdeobs/registry.py`. Changed layouts must not reuse frozen paper labels.

```python
import numpy as np
from pdeobs import api
from pdeobs.registry import MASK_REGISTRY
@MASK_REGISTRY.register("alternating_rows")
def alternating_rows(shape, rng, **kwargs):
    mask = np.zeros(shape, dtype=bool)
    mask[::2, :] = True
    return mask
spec = api.make_observation("alternating_rows", namespace="custom")
assert api.realized_count(spec, (128, 128)) == 8192
```

Listing 3: A custom mask, not a new result under a frozen paper-view label.

B.8 Local Inspection and Tests

`AI_START_HERE.md` is the implementation and validation index; `AGENTS.md` states the repository rules. These inspection commands do not train models:

```
python tools/ai_assist.py doctor --json
python tools/ai_assist.py context --task scoring --max-chars 24000
```

Listing 4: Read-only implementation and evidence lookup.

Regression tests cover facade validation, scoring, observation replay, and evidence bundles. Report the environment and passed/failed/skipped scope; an archived receipt is not a newly rerun test. CPU examples do not certify GPU behavior, full-corpus fidelity, or historical model performance.

B.9 Data Format, Benchmark Access, Maintenance, and Extensibility

Data format. To reuse physical records across patterns, HDF5 shards (Collette, 2013) store channel-last condition, trajectory, and geometry arrays separately from observation masks. JSON metadata records identity, family, boundary, condition setting, regime, seeds, and times; atomic writes and hash sidecars track completed shards. Masks are then constructed at access time from the identity and protocol. Two versioned packages keep inference separate from scoring: an observation-only input produces a prediction bundle, which is scored against separately supplied targets. External sensor data can use the prediction interface, but quantitative scoring still requires corresponding targets (Appendix A.3, Appendix B.4).

Benchmark access. Code, configurations, tests and documentation are available at our anonymous repository under the MIT code license; data licensing is separate. The result release contains full-precision scores, per-record errors and hash bindings, with historical scores archived separately. Raw tensors are not embedded in Git. Reproducing this evaluation requires the bound records and weights, not merely newly generated data. Their integrity was checked in the verification environment, but anonymous external access remains a separate release requirement (Appendix C).

Maintenance. Python and command-line interfaces use NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020), h5py, and PyYAML, with optional PyTorch training support (Paszke et al., 2019). Exploration uses a convenience configuration resolver, whereas frozen experimental rows have stricter guards. Versioned schemas, resolved configurations, regression tests, CPU demonstrations, and a `doctor` command support inspection (Appendix A.7). These checks establish software behavior, not the availability of every artifact needed to replay a historical run.

Extensibility. Public registries extend observation views, condition constructions, prediction methods, PDE families, and metrics. A view returns a boolean mask, a method implements the observation-to-prediction interface, and a PDE generator returns a complete record. This separation lets new

methods and views reuse existing records and scoring without redefining the physical task. To protect the reference protocol, the nine frozen labels cannot be redefined; optional-plugin failures are recorded separately from built-in baselines. Registration itself does not certify a plugin's scientific validity. API contracts, entry-point discovery, and read-only inspection tools, including AI-assisted commands, are documented in Appendix B.

C Metric Validity and Reproducibility

C.1 A new prediction-level result version

All main-grid results use the prediction evaluation identified as `20260925-v1`, covering 441 retained checkpoints and 3,969 observation-pattern blocks with 200 held-out records per PDE. The additional five MIX models use the 26 September strict prediction rescoring and main-grid specialist references described in Appendix G. The verification made no optimizer updates, selected no new checkpoints, and changed no training setting. The original scalar archive remains preserved for provenance, but supplies no value to the current tables, figures, or conclusions.

The frozen scorer checks all prediction and target entries for finiteness and recomputes relative-L^2 errors in float64 without NaN-skipping reductions. Complete contracts bind record identities, mask, geometry, shape, checkpoint and configuration digests, data split, and frame mapping. For forecasting, stored positions 1–3 are distinguished from each record's internal solver indices and physical times. All 3,969 blocks contain their full 200-record population; no invalid block or failed model is omitted. Large finite errors remain valid poor outcomes, not an exclusion criterion.

The terminal audit freshly hashes every prediction HDF5 file and authenticates its full-array finiteness receipt against those identical bytes. It independently checks all identities and shapes, regenerates masks from the frozen identity/seed chain, compares geometry with the source, and verifies physical-time mappings. It also rehashes all 84 selected data shards, released checkpoints and resolved configurations against their pinned manifests. Thus the audit does not substitute a scalar-finiteness check for the scorer's array check. It performs neither another model forward pass nor another training run. Prediction-file validity, score reproduction, and scientific quality remain separate outcomes.

C.2 Score changes and uncertainty

Of the 3,969 blocks, 3,646 meet the original score-comparison tolerance ($\mathrm{atol} = 10^{-7}$, $\mathrm{rtol} = 10^{-4}$); an additional 314 differ by at most 1% relative to the historical value, and 9 differ by more than 1%. The 1% reporting threshold was adopted after initial verification results, not preregistered as a scientific validity gate. All new valid scores are used, whether better or worse, with no repeat inference to obtain closer agreement. The full old/new comparison remains an audit artifact. The specific numerical causes of the differences have not been established.

Each cell reports mean $\pm$ sample standard deviation of 200 per-record errors, with $\mathrm{ddof} = 1$. Pair-level D and C instead summarize the means of 9 matched and 72 cross-pattern cells. The main table pools 63/504 such cells per PDE across seven methods and labels that different population explicitly. These standard deviations describe record or view variability, not multiple training seeds, confidence intervals, or independent replications. Dynamic joint error uses one norm over all three frames; the three auxiliary horizon errors are never averaged to replace it. Counts, medians and derived ratios receive no invented standard deviation. Full precision is retained before uniform display rounding.

Paired test-identity bootstrap. We resample the 200 test identities within each PDE's low/medium/high regimes (67/67/66 draws with replacement), using one shared draw across all its models, observation views and forecasting horizons. Cell means and derived statistics are re-computed in each of 5,000 replicates (seed 20260926). The 95% percentile interval for the median C/D over 49 pairs is [10.07, 10.71], with point estimate 10.46; the fixed 13-pair subset gives 10.74 [10.44, 11.12]. Each of the 49 pair-level C/D intervals lies above one. These are marginal, conditional test-identity intervals for fixed trained models, not simultaneous coverage guarantees, training-seed uncertainty, or corrections for retrospective selection. The full 7,542-statistic output and resampling specification are in `results/revision_v2/bootstrap_intervals.csv` and `bootstrap_config.json`; no new model inference is used by this resampling.

C.3 DATA, WEIGHTS, AND ANONYMOUS AVAILABILITY

Released-file manifests map each original checkpoint digest to the stripped release-file digest, including the retained Heat/Transolver/R50 checkpoint at epoch 257. Removing optimizer, scaler, RNG, or identifying metadata changes file bytes; original and released digests must therefore not be treated as interchangeable. Verified mappings, compatible model structures, exact weights, observations, splits and physical records are all required for inference reproducibility. Weights-only files cannot restore full-state training.

The PDE-OBS data deposit and model deposit were republished from de-identified copies on 26 September 2026. Direct manifests identify the paper data subset, full data resource, and model groups A, B, and C. The dated release report records unchanged numerical arrays and model parameters, unauthenticated access, and agreement with declared file digests for 20,167 dataset files and 6,477 model-deposit files (large objects by SHA-256, other files by Git blob identifier). The `scrub-manifest.json` files map original to republished digests where metadata changes alter bytes; the data mapping is linked here. This dated availability check is distinct from permanent access or an end-to-end reproduction from downloaded artifacts; the latter smoke test is not reported as completed. The checkpoint inventory covers the 441 main-grid models, not evidence of download coverage for the five additional MIX checkpoints. New aggregate scores, per-record errors and bindings accompany the result release; raw prediction tensors are not embedded in Git.

C.4 TRAINING PROVENANCE AND REPRODUCIBLE ARITHMETIC

The common framework fixes physical records, observation rules, the supervised objective, rollout interface and final-checkpoint convention. Per-setting budgets, stopping rules, continuations and exceptions remain documented in `docs/training_template_and_records.md`; they are not retrospectively harmonized. The fixed 13-pair, 117-model original500 sensitivity subset additionally matches complete recorded configurations within each pair, excluding only the intended observation mask and run-name/data-root fields (Appendix D). Equal epochs alone, used by a broader 23-pair historical summary, do not establish this stronger match. The analysis remains single-seed and retrospective.

The anonymous artifact's `results/prediction_verification_20260924/` contains the new index, per-record errors, full-precision statistics, dataset and checkpoint hash bindings, configuration-subset recheck, and a file manifest. The public `tools/prediction_analysis/` scripts rebuild D, C, directional differences, ratios, horizon summaries and tables from this version without reopening models or rerunning inference. Every finite adverse outcome is retained. The immutable old archive is not a fallback for a missing new result. Independent reference-solution error and grid-convergence checks remain separate from these prediction and production-residual checks.

D CONFIGURATION-MATCHED ORIGINAL-BUDGET SUBSET

To examine sensitivity to heterogeneous training histories, we select pairs whose nine checkpoints all belong to the original500 cohort and actually reached epoch 500. This metadata-based selection predates the new prediction scores and fixes 13 pairs, 117 models and 1,053 blocks. No model is retrained and no pair is selected for a favorable test outcome.

Complete configuration audit. The released `resolved.yaml` of each selected model is freshly hashed and checked against the earlier release mapping and configuration audit. After removing only the row name, observation-mask specification and filesystem data root, the complete remaining configuration agrees across all nine views within each pair. This includes architecture, batch, loss, optimizer, learning rate and schedule, precision, gradient controls, stopping and loader settings. Dataset, split, seed and repository bindings also agree. It does not establish identical hardware, execution time or bitwise-deterministic training. The earlier field-level audit remains in `results/training_sensitivity/`; the new result cut records its hash and the fresh input recheck.

Selection and new-score results. The subset includes CNO on Burgers, Darcy, Heat, Helmholtz, Navier–Stokes and reaction–diffusion; FNO and PINO on Darcy and Helmholtz; U-FNO on Darcy; and DeepONet and Transolver on Helmholtz. Table 9 compares the same descriptive statistics, using only new predictions in both columns. The 13 blue rows in the consolidated Table 13 supply every selected pair's absolute means and standard deviations without duplicating them in a separate table.

Table 9: Sensitivity to fixed original500 training provenance. Counts and medians summarize dependent comparisons, not independent experimental replications. The metadata-defined subset is unchanged.

Statistic	Full grid	Fixed subset
Pairs	49	13
Pairs with $C > D$	49/49	13/13
Median C/D	10.46	10.74
Positive layout directions, group I	545/588	151/156
Positive layout directions, group II	276/294	76/78
R80 improves over R50, R50-trained	13/49	0/13
R80 improves over R50, R65-trained	37/49	8/13
R80 improves over R50, R80-trained	49/49	13/13

The matched-versus-transfer direction holds in 13 of the 13 pairs. Effect magnitudes are not identical to the full grid: the subset's median R80/R50 ratio for R50-trained checkpoints is 9.93, versus 1.58 over all 49 pairs. This is a retrospective, single-seed, completion-conditioned sensitivity analysis, not a randomized architecture comparison or evidence that omitted models behave identically.

E Completed Diagnostics and Further Controls

Prediction-level verification enables the stationary diagnostics below without retraining. We distinguish these completed analyses and the MIX study from the further controls still needed for stronger interpretation.

Stationary observed and hidden errors. For a noise-free stationary record, let $r = \widehat{u} - u$ and $d = \max(\|u\|_2, 10^{-12})$. Define $e_{\text{obs}} = \|M \odot r\|_2/d$ and $e_{\text{hid}} = \|(1 - M) \odot r\|_2/d$. Then $e^2 = e_{\text{obs}}^2 + e_{\text{hid}}^2$ per record. Squaring the separately averaged errors does not preserve this identity. A hidden-only relative error normalized by $\|(1 - M) \odot u\|_2$ is a different metric and must be labelled separately.

The projection

$$\widehat{u}_{\text{dc}} = M \odot u + (1 - M) \odot \widehat{u} \tag{20}$$

leaves hidden predictions unchanged and eliminates observed-point error. Full-field squared error, and thus the per-record full-field relative error, cannot increase. This algebraic diagnostic does not use hidden values to repair predictions. It applies to stationary noiseless recovery only: injecting true future values into a forecast would violate the task.

These diagnostics were computed from the retained predictions for all 1,701 stationary blocks (189 matched, 1,512 cross-pattern), each with 200 records. For each block the observed squared-error share is $\sum_i e_{\text{obs},i}^2 / \sum_i e_i^2$, not a fraction of incorrect grid points. Its median is 0.43 across matched blocks and 0.46 across cross-pattern blocks. The median ratio of projected to raw block-mean error is 0.74 and 0.73, respectively, with an overall range of 0.19–1.00. Matched Transolver blocks show a smaller projection effect (median ratio 0.96; observed squared-error share 0.04). Record-level checks verify the squared-error decomposition, equality of projected error to the hidden contribution, and non-increasing error. Thus some raw error lies at already observed points, but projection leaves hidden predictions unchanged and does not explain their error. Raw predictions remain the headline results; projection is a separately labelled diagnostic, recorded in `results/revision_v2/stationary_diagnostics_summary.csv` and `stationary_diagnostics_checks.json`.

Simple and mixed-pattern baselines. A training-population mean, interpolation, or a training-only POD basis can test whether smooth stationary fields are easily reconstructed without the learned interface. The software's existing `mean` method is instead an observed-mean fill, and must not be labelled a training-population baseline. Interpolation alone does not predict temporal evolution; a forecasting comparator must state its temporal rule. Five completed MIX models, their actual epochs and update counts, and the distinction between all-nine descriptive and Poisson recipe-matched comparisons are reported in Appendix G. The formal-protocol simple baselines remain unexecuted; no result is inferred from their implementation.

Forecast ambiguity and density. A full-input versus partial-input, hidden-parameter versus parameter-conditioned comparison can separate sources of temporal uncertainty. Conditioning on a recorded physical coefficient is an explicitly different task, not an unreported change to ordinary inference. Nested random masks isolate the addition of observations by preserving the smaller set. These controls require new evidence before any mechanistic claim.

Uncertainty and sensitivity. The paired test-identity bootstrap is reported in Appendix C.2; independent training seeds are still needed for training-seed uncertainty. Output variance and sensitivity to changed visible values can diagnose an input-insensitive model. Near-equal aggregate scores alone neither prove collapse nor demonstrate useful robustness.

F Complete Prediction-Level Results

All 49 PDE–method matrices use the new complete prediction-level evaluation, retaining every finite adverse result. The tables expand Figure 2 without selecting a different subset. Each block contains 200 held-out identities and passes the file-validity checks in Appendix C; this does not imply small prediction error. Means are arithmetic, standard deviations use $\mathrm{ddof} = 1$, and the 90th percentile uses linear interpolation. These tables use two decimal places with matching mean/SD precision; counts remain integers. MIX errors are explicitly scaled to percentages in Appendix G, while these grid errors and all ratios retain their stated units. All calculations precede display rounding. A displayed 0.00 can represent a small positive value, not exact zero; the released scores retain full precision.

F.1 PDE-level overview

Table 10 complements the shared absolute-error color scale in Figure 2. It retains variability across all seven methods and observation cells rather than assigning a winner. These pooled descriptive summaries do not equalize training histories or model capacities.

Table 10: Absolute error and transfer sensitivity from new predictions. Matched/cross-pattern entries are mean $\pm$ sample SD over 63/504 cell means per PDE (seven methods); the ratio is the median of seven pairwise C/D values. These are not error percentages, seed uncertainty, or an architecture ranking.

	Relative-L^2 test error		Transfer ratio
PDE	Matched	Cross-pattern	Median C/D
Stationary solution recovery			
Darcy	0.08 ± 0.10	0.80 ± 2.40	13.45
Poisson	0.06 ± 0.07	1.50 ± 5.71	16.58
Helmholtz	0.05 ± 0.14	4.91 ± 30.45	28.61
Temporal state forecasting			
Heat	0.09 ± 0.07	1.16 ± 6.77	3.62
Reaction–diffusion	0.11 ± 0.12	1.19 ± 5.60	3.36
Burgers	0.10 ± 0.11	0.51 ± 1.65	3.63
Navier–Stokes	0.04 ± 0.04	0.34 ± 0.50	7.39

F.2 TRAINING COHORTS AND AGGREGATION

Table 11: Retained training-provenance groups. Codes identify histories, not scientific quality; every retained model contributes all nine test views.

Code	Recorded cohort	Models
O500	original500	257
INT	interrupted_or_recovered	28
M120	budget200_min120	44
P10	budget200_patience10_no_floor	10
R500	original500_recovered	66
F200	fixed200_recovered	36

For each matrix, D averages nine matched cell means and C averages 72 cross-pattern cell means. Their standard deviations describe those 9/72 values, not seed uncertainty. The main-text median C/D weights each of the 49 matrices equally, not each physical record pooled across PDEs. Each colored cell in Figure 2 is one $E_{v,w}$, not a ratio or a median across methods. The logarithmic color scale preserves all 3,969 positive finite cell means without clipping; their rounded values and sample SDs appear below. No error or ratio is winsorized.

Table 12: Layout and density contrasts. The first group counts $E_{v,w} > E_{w,w}$ and summarizes $E_{v,w}/E_{w,w}$; the second counts $E_{v,w} < E_{v,\mathrm{R50}}$ and summarizes $E_{v,w}/E_{v,\mathrm{R50}}$. Density masks are sampled separately and are not constructed as nested sets; statistical independence is not assumed. Counts are integers; ratios are dimensionless.

Condition	Comparison	Count / total	Median ratio
Equal-count layout: positive destination-referenced difference			
$K = 8,192$	Recovery	238/252	7.53
$K = 8,192$	Rollout	307/336	2.97
$K = 8,284$	Recovery	122/126	12.37
$K = 8,284$	Rollout	154/168	3.76
Density: lower error than the same checkpoint at R50			
Train R50	Test R65	15/49	1.13
Train R50	Test R80	13/49	1.58
Train R65	Test R65	48/49	0.61
Train R65	Test R80	37/49	0.93
Train R80	Test R65	49/49	0.56
Train R80	Test R80	49/49	0.29

F.3 ALL PAIR-LEVEL SUMMARIES

Table 13: All 49 PDE–method summaries. Relative-L^2 errors are displayed to two decimals. D/C are mean $\pm$ sample SD across 9/72 cell means; the last three columns describe the 72 cross-pattern means. Blue retains the 13 fixed original500 pairs, not a winner or a uniform-budget architecture ranking. All medians, ratios, and tails are calculated before rounding.

	Mean $\pm$ SD			Cross-pattern cell means		
Method	Matched D	Cross C	C/D	Median	90th pct.	Max.
Darcy						
U-FNO	0.26 ± 0.10	0.51 ± 0.32	2.01	0.36	1.07	1.49
FNO	0.08 ± 0.11	0.81 ± 0.86	10.46	0.54	1.80	4.39
CNO	0.04 ± 0.05	0.48 ± 0.53	10.74	0.37	0.82	3.63
DeepONet	0.06 ± 0.04	1.58 ± 2.95	26.63	0.44	3.67	14.26
GNOT	0.06 ± 0.04	0.74 ± 3.21	13.45	0.10	0.54	26.18
Transolver	0.03 ± 0.05	0.51 ± 0.46	15.30	0.46	1.10	2.33
PINO	0.03 ± 0.04	0.99 ± 4.42	28.83	0.36	1.04	37.80
Poisson						
U-FNO	0.12 ± 0.10	0.93 ± 1.14	8.01	0.47	3.07	4.11
FNO	0.04 ± 0.05	0.70 ± 0.70	16.58	0.49	1.45	3.01
CNO	0.04 ± 0.05	0.48 ± 0.45	11.96	0.35	1.13	1.85
DeepONet	0.06 ± 0.07	1.79 ± 3.10	30.16	0.53	6.49	13.65
GNOT	0.06 ± 0.09	5.20 ± 14.19	91.12	0.18	9.66	85.57
Transolver	0.04 ± 0.05	0.80 ± 0.62	18.85	0.67	1.70	2.60
PINO	0.04 ± 0.04	0.60 ± 0.46	14.90	0.50	1.06	2.25
Helmholtz						
U-FNO	0.17 ± 0.34	0.78 ± 0.67	4.70	0.53	1.63	2.89
FNO	0.02 ± 0.02	0.61 ± 0.56	28.61	0.53	1.38	2.12
CNO	0.03 ± 0.04	0.45 ± 0.36	15.35	0.40	0.90	1.15
DeepONet	0.05 ± 0.05	1.71 ± 3.12	34.16	0.48	4.54	15.28
GNOT	0.05 ± 0.07	29.75 ± 76.34	575.98	0.16	127.29	494.98
Transolver	0.02 ± 0.02	0.58 ± 0.39	32.52	0.52	0.99	1.93
PINO	0.04 ± 0.03	0.53 ± 0.39	13.00	0.35	1.13	1.42
Heat						
U-FNO	0.17 ± 0.12	0.33 ± 0.23	1.94	0.24	0.65	0.93
FNO	0.05 ± 0.01	0.20 ± 0.20	3.60	0.09	0.53	0.80
CNO	0.07 ± 0.02	0.27 ± 0.23	3.83	0.18	0.63	0.79
DeepONet	0.10 ± 0.04	6.40 ± 17.06	65.14	0.73	17.37	93.63
GNOT	0.10 ± 0.02	0.39 ± 0.78	4.01	0.10	0.63	4.59
Transolver	0.10 ± 0.09	0.36 ± 0.24	3.62	0.34	0.74	0.87
PINO	0.06 ± 0.01	0.22 ± 0.20	3.53	0.11	0.55	0.66
Reaction–diffusion						
U-FNO	0.14 ± 0.07	0.31 ± 0.31	2.25	0.22	0.58	1.72
FNO	0.06 ± 0.01	0.20 ± 0.21	3.36	0.10	0.43	0.97
CNO	0.07 ± 0.02	0.25 ± 0.24	3.58	0.12	0.62	0.84
DeepONet	0.11 ± 0.03	4.72 ± 12.99	42.79	0.53	12.91	77.24
GNOT	0.11 ± 0.04	2.20 ± 5.96	20.05	0.13	6.23	33.23
Transolver	0.20 ± 0.29	0.44 ± 0.32	2.19	0.36	0.96	1.06
PINO	0.07 ± 0.01	0.22 ± 0.20	3.29	0.10	0.57	0.67

Continued on the next page

Table 13 (continued)

	Mean ± SD			Cross-pattern cell means		
Method	Matched D	Cross C	C/D	Median	90th pct.	Max.
Burgers						
U-FNO	0.21 ± 0.26	0.32 ± 0.31	1.50	0.16	0.80	1.23
FNO	0.05 ± 0.01	0.19 ± 0.18	3.70	0.08	0.49	0.61
CNO	0.07 ± 0.03	0.26 ± 0.24	3.89	0.14	0.66	0.71
DeepONet	0.12 ± 0.03	1.99 ± 4.03	17.11	0.67	5.32	20.33
GNOT	0.09 ± 0.03	0.28 ± 0.35	3.06	0.10	0.73	2.15
Transolver	0.10 ± 0.06	0.33 ± 0.23	3.18	0.30	0.69	0.86
PINO	0.05 ± 0.01	0.20 ± 0.19	3.63	0.10	0.54	0.58
Navier–Stokes						
U-FNO	0.03 ± 0.03	0.18 ± 0.18	7.09	0.07	0.47	0.55
FNO	0.01 ± 0.01	0.18 ± 0.19	17.47	0.08	0.51	0.56
CNO	0.04 ± 0.04	0.29 ± 0.28	7.18	0.14	0.69	1.01
DeepONet	0.07 ± 0.05	0.76 ± 0.96	11.31	0.52	1.99	4.18
GNOT	0.06 ± 0.06	0.17 ± 0.17	2.65	0.08	0.45	0.63
Transolver	0.05 ± 0.04	0.40 ± 0.30	7.39	0.37	0.79	0.94
PINO	0.01 ± 0.02	0.43 ± 0.56	38.18	0.20	1.27	2.12

F.4 ALL FORECASTING HORIZONS

The 28 temporal matrices each include three horizon diagnostics. The mean ± sample SD of 9 matched or 72 cross-pattern cell means is reported at every horizon. These auxiliary errors are not averaged to obtain the joint three-frame score. Three predicted transitions do not establish long-term stability.

Table 14: All forecasting horizons in one grouped table. Each entry reports mean ± sample SD of 9 matched or 72 cross-pattern cell means on one line, with two decimals in both terms. Gray retains the method grouping. These relative-L^2 errors are auxiliary per-frame scores, not components to average into the joint rollout score.

	Horizon 1		Horizon 2		Horizon 3	
Method	D_1	C_1	D_2	C_2	D_3	C_3
Heat						
U-FNO	0.16±0.12	0.33±0.24	0.16±0.12	0.32±0.23	0.18±0.11	0.32±0.22
FNO	0.03±0.01	0.19±0.21	0.05±0.01	0.19±0.21	0.08±0.01	0.20±0.20
CNO	0.05±0.03	0.26±0.25	0.06±0.02	0.26±0.23	0.09±0.02	0.27±0.22
DeepONet	0.08±0.04	1.40±1.97	0.09±0.04	2.89±6.12	0.12±0.04	10.66±29.76
GNOT	0.08±0.02	0.25±0.29	0.09±0.01	0.30±0.46	0.11±0.01	0.50±1.24
Transolver	0.08±0.12	0.36±0.24	0.09±0.09	0.35±0.24	0.11±0.08	0.35±0.23
PINO	0.03±0.01	0.20±0.21	0.06±0.01	0.21±0.20	0.08±0.01	0.23±0.19
Reaction–diffusion						
U-FNO	0.14±0.06	0.30±0.27	0.14±0.07	0.31±0.33	0.14±0.07	0.32±0.33
FNO	0.04±0.01	0.21±0.22	0.06±0.01	0.19±0.20	0.07±0.01	0.19±0.20
CNO	0.06±0.03	0.25±0.25	0.07±0.02	0.25±0.24	0.08±0.02	0.25±0.23
DeepONet	0.09±0.04	1.45±2.64	0.11±0.03	2.72±6.56	0.13±0.03	7.48±21.69
GNOT	0.10±0.05	0.63±1.25	0.11±0.04	2.93±8.66	0.12±0.04	1.94±5.05
Transolver	0.19±0.28	0.44±0.30	0.20±0.30	0.43±0.32	0.21±0.30	0.44±0.33
PINO	0.05±0.01	0.20±0.21	0.07±0.01	0.21±0.20	0.08±0.01	0.23±0.21
Burgers						
U-FNO	0.16±0.15	0.28±0.26	0.19±0.23	0.30±0.29	0.27±0.38	0.36±0.38
FNO	0.02±0.01	0.18±0.19	0.04±0.01	0.17±0.18	0.07±0.01	0.19±0.17
CNO	0.04±0.04	0.26±0.25	0.06±0.03	0.25±0.24	0.09±0.02	0.27±0.22
DeepONet	0.07±0.04	1.28±1.79	0.10±0.04	1.68±3.12	0.16±0.03	2.67±6.15
GNOT	0.07±0.03	0.25±0.30	0.08±0.03	0.26±0.35	0.12±0.02	0.30±0.37
Transolver	0.07±0.05	0.33±0.23	0.10±0.06	0.32±0.23	0.13±0.07	0.34±0.23
PINO	0.02±0.01	0.18±0.20	0.05±0.01	0.19±0.19	0.08±0.01	0.22±0.18
Navier–Stokes						
U-FNO	0.03±0.03	0.19±0.18	0.02±0.03	0.18±0.18	0.03±0.03	0.18±0.18
FNO	0.01±0.01	0.19±0.19	0.01±0.01	0.17±0.19	0.01±0.01	0.17±0.19
CNO	0.04±0.05	0.30±0.30	0.04±0.04	0.28±0.28	0.04±0.04	0.28±0.27
DeepONet	0.06±0.05	0.79±1.02	0.07±0.05	0.71±0.85	0.07±0.05	0.75±0.97
GNOT	0.06±0.06	0.18±0.18	0.06±0.05	0.16±0.16	0.07±0.06	0.16±0.15
Transolver	0.05±0.04	0.39±0.29	0.05±0.04	0.40±0.30	0.06±0.04	0.41±0.31
PINO	0.01±0.02	0.34±0.46	0.01±0.02	0.43±0.58	0.01±0.02	0.48±0.63

F.5 ALL 49 TRAIN-VIEW BY TEST-VIEW MATRICES

Rows are training patterns and columns are test patterns, ordered R50, H, V, CL, BL, LI, BD, R65, R80. Each complete 9×9 matrix is shown in one table without splitting its columns. Each cell reports mean $\pm$ sample SD of 200 per-record relative-L^2 errors on one line, with two decimal places in both terms. Blue identifies matched observations, not a best-method ranking. Captions record the actual epoch and cohort of every training-pattern checkpoint using Table 11; row order also determines the order of the epoch/cohort list. No budgets are silently mixed. CNO denotes the CNO-inspired adaptation. Full-precision per-record scores accompany the artifact; no cell is an estimate or placeholder.

Table 15: Darcy / U-FNO. All 81 cells: mean $\pm$ SD on one line ($n = 200$); blue marks matched views. All training views: 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.18±0.03	0.25±0.06	0.24±0.06	0.76±0.15	0.76±0.07	0.24±0.05	0.73±0.06	0.31±0.02	0.57±0.04
H	0.32±0.05	0.32±0.05	0.38±0.07	1.08±0.30	1.49±0.30	0.35±0.07	1.47±0.26	0.31±0.06	0.32±0.05
V	0.36±0.18	0.36±0.18	0.36±0.18	0.36±0.18	0.36±0.18	0.36±0.18	0.36±0.18	0.36±0.18	0.36±0.18
CL	0.27±0.14	0.27±0.14	0.27±0.14	0.30±0.14	0.30±0.15	0.27±0.14	0.31±0.16	0.29±0.12	0.34±0.10
BL	0.36±0.18	0.36±0.18	0.36±0.18	0.36±0.18	0.36±0.18	0.36±0.18	0.36±0.18	0.36±0.18	0.36±0.18
LI	0.08±0.02	0.09±0.02	0.09±0.02	0.57±0.14	0.62±0.09	0.09±0.02	0.66±0.08	0.28±0.02	0.53±0.03
BD	1.07±0.04	1.07±0.04	1.07±0.04	1.37±0.25	0.70±0.19	1.07±0.04	0.36±0.19	1.08±0.04	1.10±0.04
R65	0.27±0.02	0.31±0.06	0.30±0.05	0.63±0.12	0.71±0.09	0.30±0.06	0.75±0.06	0.13±0.02	0.26±0.02
R80	0.44±0.02	0.46±0.05	0.45±0.05	0.64±0.14	0.74±0.10	0.45±0.05	0.81±0.06	0.29±0.03	0.21±0.04

Table 16: Darcy / FNO. All 81 cells: mean $\pm$ SD on one line ($n = 200$); blue marks matched views. All training views: 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.02±0.01	0.16±0.07	0.20±0.09	1.87±0.50	1.81±0.33	0.14±0.05	1.73±0.29	0.39±0.04	1.23±0.16
H	0.15±0.07	0.03±0.02	1.44±0.81	2.98±0.98	4.28±0.80	0.99±0.61	4.39±0.81	0.34±0.06	0.48±0.09
V	0.07±0.02	0.48±0.23	0.02±0.01	2.52±0.61	2.17±0.41	0.36±0.19	2.21±0.40	0.88±0.12	0.95±0.16
CL	0.20±0.05	0.20±0.06	0.19±0.06	0.14±0.09	0.26±0.10	0.20±0.06	0.35±0.07	0.26±0.03	0.35±0.05
BL	0.48±0.12	1.05±0.45	0.80±0.30	0.60±0.16	0.08±0.04	0.76±0.31	0.07±0.04	0.42±0.12	0.37±0.09
LI	0.02±0.01	0.03±0.02	0.03±0.01	0.59±0.18	0.56±0.12	0.03±0.01	0.61±0.09	0.16±0.02	0.22±0.02
BD	0.88±0.03	0.88±0.02	0.88±0.02	1.47±0.37	0.69±0.16	0.88±0.02	0.35±0.18	0.85±0.03	0.83±0.03
R65	0.20±0.02	0.28±0.07	0.28±0.07	1.07±0.23	1.11±0.18	0.27±0.05	1.17±0.16	0.02±0.01	0.20±0.03
R80	0.51±0.07	0.54±0.08	0.55±0.08	1.01±0.26	1.09±0.17	0.53±0.08	1.13±0.13	0.25±0.03	0.01±0.01

Table 17: Darcy / CNO. All 81 cells: mean $\pm$ SD on one line ($n = 200$); blue marks matched views. All training views: 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.01±0.00	0.10±0.05	0.09±0.04	0.56±0.15	0.73±0.11	0.08±0.04	0.80±0.06	1.77±0.76	3.63±0.45
H	0.06±0.01	0.01±0.01	0.19±0.04	0.38±0.19	0.55±0.13	0.12±0.03	0.59±0.07	0.93±0.09	0.96±0.13
V	0.02±0.00	0.08±0.03	0.02±0.01	0.36±0.15	0.48±0.10	0.05±0.01	0.51±0.07	0.05±0.01	0.08±0.01
CL	0.07±0.01	0.07±0.02	0.08±0.02	0.11±0.09	0.29±0.16	0.07±0.02	0.36±0.15	0.08±0.01	0.10±0.02
BL	0.48±0.11	0.38±0.08	0.37±0.08	0.26±0.10	0.15±0.09	0.39±0.08	0.14±0.09	0.57±0.14	0.63±0.16
LI	0.01±0.00	0.01±0.01	0.01±0.00	0.38±0.15	0.57±0.09	0.01±0.00	0.61±0.06	0.82±0.17	2.03±0.27
BD	0.69±0.14	0.76±0.21	0.71±0.16	0.72±0.34	0.35±0.14	0.70±0.17	0.08±0.05	0.69±0.21	0.80±0.34
R65	0.13±0.02	0.24±0.05	0.24±0.04	0.65±0.15	0.82±0.11	0.23±0.04	0.89±0.06	0.01±0.00	0.15±0.03
R80	0.23±0.03	0.34±0.05	0.33±0.04	0.69±0.14	0.78±0.09	0.32±0.04	0.83±0.05	0.10±0.01	0.01±0.00

Table 18: Darcy / DeepONet. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 500/O500, 290/INT, 500/O500, 500/O500, 500/O500, 500/O500, 500/O500, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.04±0.02	0.14±0.04	0.16±0.04	0.58±0.11	0.71±0.08	0.11±0.03	0.74±0.04	0.14±0.01	0.20±0.02
H	0.26±0.06	0.04±0.02	0.69±0.24	0.80±0.26	0.80±0.17	0.36±0.13	0.69±0.12	0.34±0.07	0.40±0.07
V	0.33±0.02	0.80±0.32	0.05±0.02	0.48±0.15	0.44±0.14	0.45±0.20	0.32±0.10	0.18±0.02	0.15±0.03
CL	0.24±0.06	2.88±1.08	1.59±0.44	0.14±0.07	0.29±0.10	0.58±0.19	0.31±0.07	0.22±0.06	0.23±0.06
BL	8.56±1.10	12.22±2.47	14.26±2.39	1.87±0.77	0.10±0.06	13.56±2.00	0.08±0.04	7.52±0.96	4.92±0.68
LI	0.53±0.03	0.16±0.03	0.17±0.03	0.42±0.11	0.43±0.10	0.04±0.02	0.43±0.07	0.41±0.02	0.25±0.02
BD	2.53±0.39	3.22±0.98	2.93±0.87	3.70±0.96	2.12±1.52	2.60±0.90	0.05±0.03	3.40±0.46	4.44±0.53
R65	0.13±0.01	0.24±0.05	0.24±0.05	0.55±0.12	0.64±0.08	0.22±0.04	0.67±0.05	0.03±0.02	0.08±0.01
R80	0.32±0.02	0.38±0.05	0.38±0.05	0.58±0.13	0.67±0.10	0.37±0.04	0.71±0.05	0.15±0.01	0.04±0.02

Table 19: Darcy / GNOT. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 185/INT, 182/INT, 184/INT, 500/O500, 500/O500, 185/INT, 500/O500, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.01±0.01	0.03±0.01	0.03±0.01	0.54±1.57	2.47±7.11	0.03±0.01	0.33±0.24	0.01±0.01	0.01±0.01
H	0.08±0.01	0.08±0.01	0.10±0.02	0.31±0.20	0.35±0.12	0.09±0.02	0.35±0.09	0.08±0.01	0.08±0.01
V	0.04±0.01	0.06±0.02	0.05±0.01	0.26±0.18	0.33±0.12	0.05±0.02	0.34±0.07	0.04±0.01	0.04±0.01
CL	0.06±0.03	0.06±0.03	0.06±0.03	0.10±0.07	0.23±0.10	0.06±0.03	0.24±0.07	0.06±0.03	0.06±0.03
BL	0.27±0.07	0.28±0.07	0.27±0.07	0.41±0.17	0.11±0.08	0.28±0.07	0.08±0.07	0.27±0.07	0.27±0.07
LI	0.01±0.01	0.02±0.01	0.02±0.01	0.84±4.49	0.74±0.83	0.02±0.01	0.40±0.11	0.01±0.01	0.01±0.01
BD	0.48±0.10	0.49±0.11	0.49±0.11	0.69±0.27	0.51±0.17	0.48±0.11	0.09±0.04	0.49±0.10	0.48±0.10
R65	0.02±0.01	0.04±0.02	0.04±0.02	0.37±0.52	0.32±0.10	0.04±0.01	0.28±0.05	0.01±0.01	0.01±0.01
R80	0.03±0.02	0.10±0.04	0.10±0.04	8.26±27.66	26.18±94.73	0.10±0.04	2.32±6.55	0.03±0.02	0.02±0.02

Table 20: Darcy / Transolver. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 500/O500, 500/O500, 500/O500, 500/O500, 500/O500, 200/M120, 500/O500, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.00±0.00	0.52±0.16	0.60±0.20	1.31±0.36	1.53±0.31	0.53±0.11	1.55±0.26	0.05±0.01	0.13±0.02
H	0.19±0.03	0.00±0.01	0.29±0.04	0.33±0.13	0.43±0.10	0.10±0.02	0.46±0.08	0.17±0.03	0.15±0.03
V	0.40±0.04	0.46±0.05	0.01±0.00	0.64±0.19	0.73±0.15	0.31±0.04	0.80±0.10	0.34±0.04	0.27±0.03
CL	0.02±0.00	0.04±0.02	0.04±0.02	0.06±0.06	0.19±0.10	0.04±0.01	0.18±0.09	0.04±0.01	0.06±0.01
BL	0.69±0.17	0.82±0.21	0.66±0.18	0.54±0.24	0.08±0.05	0.57±0.16	0.06±0.04	0.61±0.15	0.55±0.13
LI	0.09±0.02	0.01±0.01	0.01±0.01	0.37±0.16	0.70±0.10	0.00±0.00	0.75±0.09	0.07±0.01	0.15±0.03
BD	0.53±0.10	0.49±0.07	0.59±0.14	0.73±0.23	0.48±0.24	0.54±0.13	0.15±0.09	0.49±0.07	0.50±0.07
R65	0.03±0.01	0.26±0.09	0.30±0.10	0.77±0.18	0.83±0.10	0.28±0.06	0.87±0.06	0.00±0.00	0.05±0.00
R80	0.41±0.11	0.86±0.12	2.33±0.78	1.12±0.24	1.59±0.26	1.36±0.55	1.62±0.20	0.07±0.01	0.00±0.00

Table 21: Darcy / PINO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. All training views: 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.02±0.01	0.08±0.06	0.08±0.05	0.92±0.36	1.34±0.24	0.07±0.05	1.50±0.22	0.29±0.01	0.57±0.03
H	0.10±0.02	0.02±0.01	0.24±0.07	1.05±0.28	1.09±0.15	0.17±0.06	1.10±0.08	0.29±0.02	0.53±0.04
V	0.50±0.12	0.25±0.08	0.02±0.01	37.80±27.49	2.20±1.19	0.27±0.11	2.73±0.67	0.53±0.11	0.63±0.09
CL	0.18±0.08	0.15±0.07	0.15±0.07	0.14±0.09	0.29±0.10	0.15±0.07	0.34±0.08	0.13±0.05	0.16±0.05
BL	0.38±0.07	0.30±0.09	0.31±0.10	0.40±0.12	0.04±0.04	0.29±0.08	0.03±0.03	0.31±0.07	0.24±0.06
LI	0.04±0.01	0.02±0.02	0.02±0.01	0.45±0.13	0.44±0.11	0.02±0.01	0.49±0.08	0.26±0.02	0.53±0.05
BD	0.51±0.11	0.47±0.12	0.46±0.11	0.86±0.20	0.36±0.11	0.45±0.11	0.02±0.01	0.47±0.11	0.44±0.12
R65	0.23±0.01	0.25±0.05	0.25±0.04	0.57±0.15	0.71±0.13	0.24±0.04	0.73±0.08	0.02±0.01	0.24±0.01
R80	0.36±0.01	0.37±0.04	0.37±0.03	0.62±0.16	0.67±0.12	0.36±0.03	0.73±0.06	0.18±0.01	0.02±0.01

Table 22: Poisson / U-FNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 130/M120, 130/M120, 137/M120, 200/M120, 200/M120, 134/M120, 200/M120, 151/M120, 163/M120.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.18±0.11	0.20±0.13	0.21±0.13	2.69±1.76	3.28±2.08	0.19±0.11	3.11±1.98	0.12±0.06	1.00±0.64
H	0.50±0.14	0.08±0.05	0.57±0.11	0.49±0.15	0.73±0.21	0.24±0.06	0.77±0.20	0.46±0.14	0.42±0.12
V	0.62±0.24	0.61±0.24	0.06±0.03	0.74±0.28	0.90±0.36	0.35±0.13	0.93±0.35	0.60±0.24	0.62±0.24
CL	0.06±0.03	0.07±0.02	0.07±0.03	0.05±0.04	0.38±0.15	0.05±0.02	0.39±0.14	0.05±0.02	0.04±0.01
BL	0.31±0.05	0.33±0.10	0.34±0.09	0.36±0.18	0.07±0.04	0.32±0.07	0.05±0.02	0.25±0.05	0.34±0.18
LI	0.20±0.07	0.18±0.08	0.17±0.08	0.77±0.30	1.01±0.44	0.16±0.07	1.05±0.45	0.26±0.12	0.48±0.24
BD	2.57±1.77	4.11±2.16	4.06±2.08	1.19±0.67	1.01±0.35	3.83±2.06	0.04±0.03	2.43±1.91	1.97±1.84
R65	0.33±0.22	0.48±0.32	0.56±0.37	3.30±2.00	3.97±2.38	0.47±0.30	3.92±2.36	0.34±0.21	0.14±0.06
R80	0.19±0.06	0.27±0.09	0.27±0.08	0.94±0.51	1.43±0.87	0.26±0.08	1.51±0.89	0.12±0.05	0.06±0.02

Table 23: Poisson / FNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 433/INT, 500/O500, 500/O500, 262/INT, 130/M120, 480/INT, 130/M120, 500/O500, 461/INT.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.01±0.00	0.05±0.03	0.05±0.03	0.47±0.19	0.81±0.24	0.04±0.02	0.88±0.24	0.15±0.04	0.29±0.08
H	0.06±0.03	0.02±0.01	0.36±0.23	2.14±1.37	2.99±1.92	0.25±0.16	3.01±1.94	0.19±0.10	0.51±0.23
V	0.03±0.01	0.18±0.09	0.02±0.01	0.65±0.31	1.21±0.56	0.12±0.06	1.26±0.58	0.11±0.05	0.25±0.13
CL	0.08±0.03	0.08±0.03	0.09±0.03	0.08±0.04	0.47±0.21	0.08±0.03	0.54±0.18	0.07±0.04	0.11±0.04
BL	0.67±0.17	0.86±0.41	0.98±0.58	0.71±0.33	0.14±0.09	0.86±0.40	0.10±0.05	0.34±0.09	0.29±0.08
LI	0.02±0.00	0.02±0.01	0.02±0.01	0.38±0.14	0.73±0.15	0.01±0.00	0.77±0.14	0.21±0.14	0.53±0.33
BD	1.07±0.50	0.99±0.44	1.97±1.24	1.45±0.83	1.11±0.69	1.42±0.85	0.07±0.04	1.08±0.51	1.11±0.50
R65	0.17±0.08	0.21±0.09	0.20±0.09	0.67±0.28	0.96±0.26	0.19±0.08	0.97±0.24	0.02±0.01	0.20±0.13
R80	1.05±0.49	1.13±0.54	1.12±0.54	2.02±1.13	2.30±1.41	1.09±0.52	2.36±1.44	0.44±0.19	0.01±0.00

Table 24: Poisson / CNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 275/INT, 500/O500, 500/O500, 500/O500, 500/O500, 500/O500, 500/O500, 500/O500, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.02±0.01	0.13±0.09	0.13±0.09	1.14±0.53	1.53±0.64	0.12±0.06	1.53±0.63	0.30±0.07	1.35±0.41
H	0.02±0.01	0.01±0.00	0.06±0.02	0.43±0.21	0.87±0.31	0.04±0.01	0.92±0.31	0.15±0.03	0.34±0.08
V	0.01±0.01	0.05±0.02	0.01±0.00	0.39±0.17	0.80±0.14	0.03±0.01	0.81±0.10	0.12±0.01	0.25±0.03
CL	0.09±0.04	0.08±0.02	0.07±0.02	0.12±0.07	0.74±0.26	0.08±0.02	0.83±0.27	0.14±0.05	0.20±0.06
BL	0.64±0.35	0.68±0.31	0.65±0.29	0.81±0.26	0.12±0.07	0.62±0.31	0.09±0.04	0.65±0.40	0.66±0.41
LI	0.01±0.01	0.01±0.01	0.01±0.01	0.50±0.25	1.06±0.40	0.01±0.01	1.08±0.39	0.06±0.03	0.10±0.05
BD	0.51±0.11	0.58±0.17	0.58±0.19	1.85±0.87	1.23±0.96	0.55±0.15	0.06±0.04	0.40±0.13	0.36±0.19
R65	0.07±0.04	0.15±0.11	0.15±0.11	0.66±0.20	0.95±0.18	0.17±0.10	0.96±0.14	0.01±0.00	0.05±0.03
R80	0.11±0.03	0.28±0.13	0.29±0.14	0.93±0.37	1.13±0.42	0.27±0.12	1.15±0.40	0.04±0.01	0.01±0.00

Table 25: Poisson / DeepONet. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 500/O500, 500/O500, 66/INT, 500/O500, 500/O500, 500/O500, 500/O500, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.02±0.01	0.16±0.04	0.15±0.05	0.59±0.11	0.89±0.09	0.13±0.03	0.91±0.09	0.06±0.02	0.19±0.04
H	0.19±0.08	0.03±0.01	0.29±0.14	0.65±0.27	0.83±0.17	0.16±0.11	0.81±0.15	0.10±0.05	0.09±0.03
V	0.17±0.06	0.24±0.09	0.03±0.01	0.54±0.19	0.78±0.16	0.11±0.07	0.76±0.16	0.12±0.06	0.10±0.06
CL	0.60±0.24	1.44±0.55	0.29±0.12	0.23±0.09	0.61±0.24	0.94±0.36	0.65±0.24	0.52±0.20	0.38±0.16
BL	7.26±2.82	13.65±7.26	11.65±5.08	1.62±0.95	0.13±0.08	11.44±5.05	0.07±0.03	6.52±2.16	4.46±1.19
LI	0.34±0.06	0.11±0.03	0.13±0.03	0.47±0.15	0.77±0.14	0.03±0.01	0.77±0.13	0.26±0.05	0.16±0.04
BD	5.87±2.70	6.23±3.20	5.95±3.09	7.49±4.24	1.16±0.73	5.96±3.03	0.04±0.02	7.16±3.36	8.56±4.00
R65	0.06±0.02	0.17±0.04	0.16±0.05	0.49±0.15	0.86±0.10	0.14±0.04	0.87±0.09	0.02±0.01	0.08±0.03
R80	0.15±0.01	0.26±0.05	0.26±0.05	0.61±0.23	0.90±0.25	0.24±0.05	0.92±0.24	0.05±0.01	0.02±0.01

Table 26: Poisson / GNOT. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 500/O500, 200/P10, 500/O500, 200/M120, 500/O500, 200/M120, 200/M120, 135/INT.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.01±0.00	0.02±0.01	0.02±0.01	15.43±33.93	71.17±92.13	0.02±0.01	85.57±55.23	0.01±0.00	0.01±0.00
H	0.01±0.00	0.01±0.00	0.05±0.04	6.17±17.29	26.94±23.73	0.02±0.01	33.87±19.62	0.01±0.00	0.01±0.00
V	0.02±0.01	0.08±0.04	0.01±0.01	6.43±9.06	4.12±7.29	0.06±0.03	0.92±0.38	0.01±0.00	0.01±0.00
CL	0.02±0.01	0.02±0.01	0.02±0.01	0.04±0.03	0.38±0.16	0.02±0.01	0.34±0.12	0.02±0.01	0.02±0.01
BL	0.27±0.14	0.30±0.15	0.29±0.15	0.60±0.27	0.29±0.16	0.29±0.16	0.26±0.16	0.27±0.14	0.27±0.15
LI	0.01±0.00	0.01±0.00	0.01±0.00	6.39±13.21	20.64±25.29	0.01±0.00	8.91±5.64	0.01±0.00	0.01±0.00
BD	6.05±4.26	6.58±4.66	6.92±5.17	27.02±35.01	9.75±13.34	6.73±5.06	0.07±0.05	6.07±4.26	6.07±4.27
R65	0.01±0.01	0.05±0.02	0.05±0.02	0.40±0.27	2.51±1.82	0.05±0.02	3.02±1.92	0.01±0.00	0.01±0.00
R80	0.07±0.02	0.10±0.03	0.10±0.03	0.43±0.17	0.84±0.39	0.10±0.03	0.94±0.41	0.07±0.02	0.07±0.02

Table 27: Poisson / Transolver. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 500/O500, 189/INT, 500/O500, 500/O500, 191/INT, 500/O500, 500/O500, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.00±0.00	0.29±0.11	0.36±0.17	0.66±0.22	1.98±1.31	0.29±0.11	2.17±1.37	0.08±0.02	0.31±0.07
H	1.74±0.86	0.00±0.01	0.73±0.14	0.62±0.20	1.00±0.43	0.32±0.08	1.02±0.33	1.88±0.96	1.44±0.87
V	0.50±0.08	0.65±0.12	0.06±0.03	0.83±0.36	1.05±0.39	0.34±0.07	1.09±0.38	0.40±0.08	0.35±0.09
CL	0.07±0.04	0.13±0.06	0.12±0.05	0.07±0.04	0.69±0.19	0.08±0.02	0.68±0.15	0.08±0.02	0.11±0.04
BL	1.25±0.59	1.04±0.43	0.99±0.39	0.93±0.47	0.15±0.09	0.99±0.37	0.10±0.06	1.32±0.62	1.35±0.64
LI	0.15±0.04	0.05±0.02	0.05±0.03	0.55±0.21	0.80±0.14	0.05±0.02	0.83±0.10	0.17±0.07	0.21±0.08
BD	1.48±0.63	1.12±0.51	1.14±0.49	1.65±0.98	0.98±0.65	1.09±0.52	0.04±0.03	1.70±0.68	1.72±0.64
R65	0.06±0.03	0.58±0.28	0.61±0.26	1.49±0.85	2.55±1.52	0.57±0.25	2.60±1.51	0.00±0.00	0.07±0.01
R80	0.57±0.15	0.97±0.48	0.46±0.14	0.58±0.14	1.02±0.27	0.57±0.19	1.02±0.22	0.08±0.01	0.00±0.00

Table 28: Poisson / PINO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 500/O500, 500/O500, 130/M120, 500/O500, 500/O500, 500/O500, 500/O500, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.04±0.01	0.16±0.07	0.15±0.06	0.66±0.16	0.95±0.20	0.14±0.05	1.05±0.18	0.28±0.05	0.55±0.14
H	0.14±0.08	0.01±0.01	0.23±0.08	1.05±0.57	1.73±1.19	0.16±0.07	2.25±1.45	0.30±0.07	0.52±0.10
V	0.14±0.08	0.25±0.12	0.02±0.01	1.31±0.78	1.68±0.83	0.16±0.08	2.02±1.00	0.29±0.08	0.52±0.12
CL	0.38±0.03	0.37±0.05	0.38±0.06	0.13±0.08	0.45±0.16	0.37±0.04	0.45±0.18	0.24±0.03	0.11±0.03
BL	0.76±0.08	0.63±0.20	0.83±0.23	0.83±0.21	0.07±0.05	0.61±0.19	0.05±0.03	0.69±0.09	0.64±0.10
LI	0.05±0.02	0.04±0.02	0.04±0.02	0.52±0.16	0.77±0.19	0.03±0.02	0.84±0.19	0.24±0.05	0.49±0.12
BD	0.86±0.15	1.18±0.53	0.91±0.24	0.87±0.15	1.15±0.51	1.06±0.40	0.02±0.01	0.82±0.15	0.76±0.12
R65	0.24±0.01	0.24±0.07	0.24±0.06	0.57±0.16	0.90±0.15	0.23±0.06	1.02±0.11	0.03±0.01	0.24±0.04
R80	0.36±0.01	0.38±0.05	0.38±0.05	0.71±0.25	0.97±0.28	0.37±0.04	1.01±0.23	0.18±0.01	0.01±0.01

Table 29: Helmholtz / U-FNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 132/M120, 133/M120, 135/M120, 138/M120, 200/M120, 134/M120, 200/M120, 155/M120, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	1.07±0.29	1.01±0.27	1.02±0.28	1.14±0.37	1.37±0.51	1.03±0.28	1.34±0.51	0.95±0.26	0.45±0.11
H	0.47±0.01	0.02±0.01	0.68±0.06	0.52±0.16	0.76±0.23	0.26±0.03	0.80±0.20	0.44±0.01	0.40±0.02
V	0.47±0.02	0.67±0.06	0.03±0.01	0.60±0.18	0.78±0.27	0.28±0.05	0.81±0.25	0.46±0.02	0.47±0.03
CL	0.04±0.02	0.09±0.03	0.12±0.07	0.07±0.03	0.38±0.17	0.08±0.03	0.33±0.13	0.06±0.02	0.08±0.03
BL	0.63±0.43	1.36±0.47	1.42±0.48	0.29±0.10	0.06±0.03	1.24±0.52	0.05±0.02	0.52±0.37	0.27±0.21
LI	0.17±0.05	0.04±0.01	0.04±0.01	0.48±0.15	0.74±0.13	0.03±0.01	0.77±0.12	0.14±0.04	0.17±0.06
BD	1.14±0.43	1.67±0.59	1.65±0.55	1.13±0.37	0.99±0.31	1.53±0.57	0.04±0.02	1.26±0.55	1.47±0.66
R65	0.25±0.07	0.35±0.11	0.35±0.10	1.93±0.57	2.32±0.80	0.30±0.09	2.31±0.81	0.14±0.04	0.24±0.07
R80	0.53±0.24	0.62±0.27	0.45±0.19	2.12±0.68	2.89±0.97	0.44±0.16	2.88±0.98	0.49±0.19	0.03±0.01

Table 30: Helmholtz / FNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. All training views: 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.01±0.00	0.07±0.04	0.08±0.05	1.16±0.41	1.60±0.65	0.05±0.04	1.59±0.65	0.13±0.03	0.44±0.10
H	0.02±0.00	0.01±0.00	0.12±0.05	0.70±0.29	1.25±0.40	0.08±0.03	1.29±0.39	0.27±0.09	0.59±0.20
V	0.03±0.01	0.19±0.08	0.01±0.00	0.68±0.22	1.11±0.29	0.12±0.05	1.12±0.28	0.10±0.02	0.26±0.07
CL	0.04±0.00	0.04±0.01	0.04±0.01	0.04±0.02	0.33±0.17	0.04±0.01	0.42±0.21	0.07±0.02	0.13±0.04
BL	0.53±0.14	0.57±0.20	0.70±0.32	0.79±0.28	0.06±0.03	0.55±0.18	0.04±0.02	0.56±0.15	0.69±0.19
LI	0.02±0.01	0.02±0.01	0.02±0.01	0.53±0.21	0.81±0.30	0.02±0.01	0.82±0.29	0.12±0.03	0.36±0.08
BD	1.26±0.44	1.29±0.45	1.52±0.83	1.46±0.52	1.08±0.72	1.17±0.60	0.03±0.01	1.34±0.44	1.39±0.41
R65	0.12±0.04	0.18±0.06	0.18±0.06	0.74±0.24	1.10±0.45	0.17±0.05	1.16±0.46	0.01±0.00	0.22±0.09
R80	0.35±0.05	0.58±0.16	0.58±0.17	1.71±0.51	2.04±0.73	0.54±0.15	2.12±0.76	0.17±0.02	0.01±0.00

Table 31: Helmholtz / CNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. All training views: 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.00±0.00	0.07±0.05	0.08±0.05	0.49±0.15	0.85±0.09	0.07±0.03	0.90±0.06	0.11±0.01	0.22±0.04
H	0.02±0.01	0.01±0.00	0.09±0.03	0.42±0.21	0.82±0.23	0.05±0.01	0.86±0.22	0.20±0.03	0.43±0.07
V	0.02±0.00	0.07±0.02	0.01±0.00	0.44±0.20	0.85±0.24	0.04±0.01	0.90±0.23	0.12±0.02	0.20±0.04
CL	0.15±0.05	0.15±0.05	0.14±0.04	0.09±0.05	0.57±0.22	0.14±0.05	0.63±0.16	0.17±0.04	0.22±0.05
BL	0.95±0.34	1.00±0.34	1.08±0.40	0.74±0.30	0.09±0.05	1.00±0.34	0.07±0.03	0.90±0.32	0.90±0.32
LI	0.01±0.00	0.01±0.00	0.01±0.00	0.37±0.17	0.78±0.10	0.01±0.00	0.83±0.07	0.13±0.01	0.25±0.02
BD	0.73±0.10	0.67±0.11	0.73±0.12	0.89±0.29	0.58±0.33	0.68±0.10	0.05±0.03	0.67±0.10	0.66±0.13
R65	0.14±0.02	0.24±0.05	0.24±0.05	0.54±0.15	0.86±0.09	0.22±0.04	0.90±0.06	0.00±0.00	1.15±0.27
R80	0.04±0.01	0.15±0.05	0.15±0.06	0.58±0.17	0.91±0.13	0.14±0.04	0.95±0.11	0.02±0.00	0.01±0.00

Table 32: Helmholtz / DeepONet. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. All training views: 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.02±0.01	0.14±0.04	0.13±0.05	0.63±0.13	0.92±0.10	0.11±0.03	0.93±0.09	0.08±0.03	0.17±0.06
H	0.22±0.03	0.02±0.01	0.34±0.08	0.53±0.16	0.84±0.14	0.18±0.08	0.81±0.12	0.11±0.02	0.04±0.01
V	0.31±0.09	0.38±0.14	0.02±0.01	0.51±0.16	0.82±0.22	0.19±0.09	0.82±0.23	0.30±0.09	0.30±0.09
CL	0.28±0.09	3.01±0.82	3.12±1.11	0.16±0.06	0.78±0.28	1.43±0.50	0.89±0.19	0.23±0.06	0.23±0.08
BL	7.60±3.38	15.28±7.37	13.81±6.76	1.60±0.80	0.10±0.07	12.74±5.56	0.06±0.02	6.77±3.09	4.59±2.21
LI	0.37±0.04	0.13±0.03	0.14±0.03	0.48±0.14	0.77±0.13	0.03±0.01	0.78±0.13	0.25±0.03	0.11±0.02
BD	3.04±1.32	4.12±1.78	3.34±1.16	7.10±4.58	1.00±0.87	3.44±1.25	0.07±0.04	4.14±2.42	6.12±3.95
R65	0.04±0.01	0.17±0.04	0.14±0.05	0.48±0.15	0.85±0.09	0.13±0.04	0.86±0.08	0.02±0.01	0.07±0.02
R80	0.11±0.01	0.23±0.04	0.24±0.05	0.48±0.17	0.83±0.11	0.21±0.04	0.84±0.09	0.04±0.01	0.02±0.01

Table 33: Helmholtz / GNOT. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 500/O500, 500/O500, 500/O500, 200/M120, 45/P10, 500/O500, 200/P10, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.01±0.00	0.04±0.12	0.05±0.33	132.35±117.02	81.74±74.92	0.02±0.01	13.44±11.45	0.01±0.00	0.00±0.00
H	0.01±0.00	0.01±0.00	0.07±0.14	65.50±98.78	62.51±54.10	0.03±0.01	51.71±16.29	0.01±0.00	0.01±0.00
V	0.01±0.00	0.16±0.44	0.01±0.00	20.08±37.63	13.43±8.04	0.03±0.01	11.85±4.28	0.01±0.00	0.01±0.00
CL	0.02±0.01	0.02±0.01	0.02±0.01	0.03±0.02	0.49±0.13	0.02±0.01	0.49±0.13	0.02±0.01	0.02±0.01
BL	0.28±0.11	0.29±0.11	0.29±0.11	0.51±0.39	0.20±0.15	0.29±0.11	0.16±0.12	0.29±0.11	0.28±0.11
LI	0.14±0.04	0.15±0.05	0.14±0.05	0.37±0.15	1.40±0.39	0.14±0.05	1.36±0.36	0.14±0.04	0.14±0.04
BD	174.97±206.76	184.82±225.87	191.47±230.51	494.98±562.86	9.23±22.04	176.12±207.69	0.06±0.05	174.18±205.63	174.11±206.45
R65	0.01±0.00	0.05±0.02	0.04±0.02	0.46±0.40	1.14±0.57	0.04±0.02	0.92±0.18	0.01±0.00	0.01±0.00
R80	0.01±0.00	0.13±0.54	0.04±0.02	35.80±41.00	32.71±20.42	0.03±0.01	30.49±8.73	0.01±0.00	0.01±0.00

Table 34: Helmholtz / Transolver. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. All training views: 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.00±0.00	0.31±0.10	0.31±0.13	0.77±0.21	1.37±0.39	0.30±0.09	1.38±0.37	0.07±0.01	0.12±0.03
H	0.58±0.05	0.00±0.00	0.81±0.22	0.73±0.33	1.93±0.89	0.42±0.05	1.91±0.77	0.54±0.11	0.48±0.25
V	0.49±0.07	0.71±0.08	0.00±0.00	0.55±0.16	0.83±0.21	0.34±0.06	0.84±0.19	0.46±0.08	0.46±0.10
CL	0.34±0.12	0.56±0.19	0.46±0.17	0.05±0.04	0.60±0.20	0.55±0.20	0.60±0.21	0.30±0.11	0.37±0.11
BL	0.56±0.15	0.52±0.15	0.53±0.16	0.58±0.20	0.07±0.05	0.49±0.14	0.05±0.03	0.64±0.17	0.71±0.20
LI	0.29±0.04	0.02±0.01	0.07±0.03	0.73±0.30	1.43±0.60	0.00±0.00	1.57±0.60	0.27±0.04	0.26±0.03
BD	0.53±0.11	0.44±0.10	0.43±0.09	0.64±0.24	0.50±0.32	0.47±0.11	0.03±0.02	0.61±0.13	0.62±0.13
R65	0.09±0.03	0.48±0.18	0.43±0.17	0.67±0.23	0.90±0.19	0.56±0.23	0.94±0.18	0.00±0.00	0.08±0.01
R80	0.09±0.01	0.44±0.12	0.38±0.10	0.66±0.24	1.00±0.28	0.38±0.09	1.00±0.24	0.01±0.00	0.00±0.00

Table 35: Helmholtz / PINO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. All training views: 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.04±0.01	0.12±0.04	0.11±0.05	0.56±0.18	0.98±0.17	0.10±0.04	1.12±0.15	0.20±0.03	0.36±0.05
H	0.09±0.04	0.02±0.01	0.27±0.13	0.79±0.15	1.03±0.20	0.18±0.08	1.14±0.23	0.18±0.03	0.32±0.04
V	0.11±0.04	0.27±0.11	0.02±0.01	0.77±0.16	1.22±0.30	0.19±0.08	1.28±0.32	0.19±0.03	0.34±0.04
CL	0.28±0.03	0.28±0.05	0.27±0.06	0.11±0.05	0.44±0.16	0.27±0.04	0.44±0.17	0.18±0.02	0.11±0.02
BL	0.69±0.11	0.67±0.12	0.74±0.18	0.81±0.15	0.08±0.06	0.65±0.16	0.06±0.03	0.64±0.12	0.61±0.13
LI	0.05±0.02	0.04±0.02	0.04±0.02	0.52±0.18	0.85±0.17	0.04±0.02	0.96±0.15	0.15±0.03	0.30±0.04
BD	0.88±0.20	1.16±0.58	1.42±1.31	0.92±0.21	1.18±0.50	1.13±0.47	0.02±0.01	0.83±0.21	0.78±0.20
R65	0.20±0.02	0.20±0.05	0.20±0.06	0.59±0.18	1.02±0.27	0.20±0.05	1.16±0.32	0.03±0.01	0.18±0.03
R80	0.31±0.02	0.31±0.04	0.32±0.05	0.64±0.19	0.94±0.15	0.31±0.05	1.06±0.10	0.15±0.02	0.02±0.01

Table 36: Heat / U-FNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/R500, 500/R500, 500/R500, 500/R500, 500/O500, 500/R500, 500/R500, 500/R500, 500/R500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.36±0.07	0.31±0.06	0.33±0.06	0.55±0.10	0.61±0.10	0.32±0.06	0.60±0.11	0.19±0.05	0.10±0.02
H	0.33±0.02	0.35±0.03	0.42±0.03	0.61±0.09	0.67±0.10	0.37±0.03	0.65±0.10	0.15±0.02	0.06±0.04
V	0.15±0.04	0.26±0.05	0.07±0.03	0.89±0.25	0.93±0.19	0.16±0.04	0.92±0.19	0.14±0.05	0.10±0.06
CL	0.13±0.07	0.13±0.06	0.14±0.07	0.15±0.06	0.14±0.05	0.13±0.06	0.14±0.05	0.08±0.06	0.09±0.04
BL	0.31±0.04	0.27±0.05	0.28±0.05	0.24±0.07	0.11±0.03	0.27±0.05	0.11±0.03	0.17±0.05	0.11±0.05
LI	0.20±0.06	0.24±0.06	0.22±0.06	0.37±0.09	0.36±0.07	0.22±0.06	0.35±0.06	0.09±0.06	0.06±0.05
BD	0.68±0.17	0.63±0.14	0.64±0.17	0.45±0.12	0.35±0.14	0.60±0.15	0.09±0.05	0.21±0.04	0.14±0.04
R65	0.13±0.04	0.20±0.04	0.20±0.04	0.50±0.09	0.55±0.09	0.18±0.03	0.54±0.10	0.09±0.05	0.06±0.04
R80	0.17±0.04	0.21±0.03	0.21±0.04	0.63±0.14	0.70±0.15	0.19±0.03	0.68±0.15	0.10±0.05	0.06±0.05

Table 37: Heat / FNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 500/O500, 500/O500, 500/O500, 500/O500, 500/INT, 500/O500, 500/O500, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.05±0.05	0.06±0.04	0.06±0.04	0.47±0.17	0.56±0.16	0.05±0.05	0.56±0.17	0.11±0.04	0.08±0.04
H	0.05±0.05	0.05±0.05	0.07±0.04	0.42±0.13	0.49±0.10	0.06±0.04	0.49±0.11	0.05±0.05	0.06±0.04
V	0.05±0.04	0.07±0.04	0.05±0.04	0.38±0.11	0.46±0.11	0.06±0.04	0.45±0.11	0.05±0.04	0.06±0.04
CL	0.07±0.06	0.07±0.05	0.07±0.04	0.07±0.04	0.08±0.03	0.06±0.05	0.08±0.03	0.11±0.05	0.08±0.06
BL	0.08±0.03	0.11±0.04	0.11±0.04	0.17±0.05	0.06±0.05	0.10±0.03	0.06±0.05	0.07±0.03	0.06±0.04
LI	0.05±0.05	0.05±0.05	0.05±0.05	0.31±0.10	0.39±0.10	0.05±0.05	0.38±0.10	0.07±0.05	0.08±0.05
BD	0.09±0.03	0.10±0.03	0.11±0.03	0.24±0.08	0.07±0.04	0.10±0.03	0.06±0.04	0.09±0.03	0.09±0.02
R65	0.08±0.03	0.11±0.03	0.11±0.04	0.71±0.25	0.80±0.26	0.10±0.03	0.80±0.27	0.05±0.05	0.09±0.03
R80	0.13±0.03	0.16±0.03	0.16±0.03	0.53±0.11	0.62±0.09	0.15±0.03	0.61±0.09	0.08±0.04	0.05±0.05

Table 38: Heat / CNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. All training views: 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.06±0.04	0.18±0.09	0.21±0.10	0.62±0.09	0.66±0.08	0.19±0.07	0.65±0.08	0.06±0.04	0.06±0.04
H	0.05±0.05	0.05±0.05	0.14±0.07	0.46±0.12	0.59±0.10	0.08±0.03	0.59±0.11	0.05±0.04	0.05±0.04
V	0.06±0.04	0.18±0.08	0.06±0.05	0.48±0.11	0.55±0.10	0.13±0.06	0.54±0.10	0.06±0.04	0.06±0.05
CL	0.09±0.04	0.09±0.05	0.08±0.04	0.12±0.05	0.30±0.12	0.08±0.04	0.33±0.13	0.08±0.04	0.07±0.05
BL	0.10±0.04	0.14±0.04	0.18±0.06	0.54±0.16	0.10±0.04	0.15±0.04	0.09±0.03	0.08±0.04	0.07±0.04
LI	0.05±0.04	0.06±0.04	0.06±0.04	0.43±0.11	0.52±0.09	0.05±0.04	0.51±0.09	0.05±0.04	0.05±0.04
BD	0.10±0.02	0.23±0.06	0.18±0.05	0.63±0.15	0.32±0.15	0.18±0.05	0.08±0.03	0.09±0.03	0.07±0.03
R65	0.05±0.04	0.36±0.13	0.34±0.11	0.72±0.12	0.79±0.12	0.39±0.08	0.79±0.12	0.05±0.04	0.05±0.04
R80	0.06±0.03	0.28±0.07	0.28±0.08	0.60±0.07	0.66±0.07	0.29±0.05	0.66±0.07	0.06±0.04	0.05±0.04

Table 39: Heat / DeepONet. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 500/O500, 500/O500, 500/R500, 500/R500, 500/O500, 500/R500, 500/R500, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.08±0.03	0.52±0.13	0.52±0.13	1.04±0.19	1.12±0.21	0.46±0.10	1.02±0.17	0.21±0.09	0.12±0.07
H	0.26±0.07	0.09±0.03	0.85±0.40	0.63±0.16	0.71±0.12	0.50±0.30	0.64±0.12	0.29±0.13	0.28±0.12
V	0.49±0.13	1.06±0.42	0.09±0.04	0.65±0.14	0.65±0.12	0.61±0.25	0.61±0.11	0.67±0.20	0.53±0.18
CL	0.25±0.11	1.06±1.38	8.75±10.37	0.20±0.07	0.49±0.15	1.18±1.76	0.46±0.14	0.24±0.06	0.21±0.06
BL	6.02±4.26	17.45±9.95	17.62±10.32	0.70±0.29	0.11±0.05	16.67±10.00	0.10±0.04	4.37±3.08	1.84±1.00
LI	0.27±0.06	0.19±0.09	0.18±0.09	0.69±0.16	0.80±0.20	0.08±0.03	0.75±0.15	0.20±0.06	0.21±0.06
BD	36.90±19.85	93.63±36.12	71.98±38.85	2.12±2.88	0.61±0.35	75.21±36.01	0.07±0.03	32.88±18.92	17.67±13.50
R65	0.31±0.14	0.79±0.25	0.89±0.33	7.97±7.00	10.30±6.33	0.70±0.15	5.69±0.70	0.08±0.03	0.28±0.07
R80	1.07±0.24	0.83±0.32	1.01±0.31	1.03±0.24	1.25±0.31	0.76±0.30	1.00±0.21	0.85±0.19	0.08±0.04

Table 40: Heat / GNOT. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 200/P10, 200/F200, 200/F200, 200/F200, 200/F200, 200/F200, 200/F200, 200/P10, 200/F200.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.07±0.03	0.10±0.03	0.10±0.03	0.51±0.17	0.48±0.16	0.10±0.03	0.50±0.16	0.07±0.03	0.07±0.03
H	0.09±0.04	0.10±0.03	0.15±0.05	3.38±3.97	4.59±4.14	0.12±0.04	3.56±3.87	0.09±0.04	0.09±0.04
V	0.09±0.03	0.12±0.04	0.09±0.03	0.74±0.71	0.63±0.24	0.10±0.03	0.63±0.20	0.09±0.03	0.09±0.03
CL	0.07±0.05	0.08±0.05	0.08±0.05	0.13±0.05	0.23±0.08	0.08±0.04	0.29±0.08	0.07±0.05	0.07±0.05
BL	0.08±0.05	0.10±0.04	0.10±0.04	0.34±0.11	0.10±0.04	0.10±0.04	0.10±0.04	0.08±0.05	0.08±0.05
LI	0.10±0.03	0.11±0.04	0.11±0.04	0.82±1.02	1.34±1.35	0.11±0.03	1.33±1.39	0.10±0.03	0.10±0.03
BD	0.07±0.03	0.10±0.03	0.10±0.03	0.47±0.17	0.40±0.17	0.09±0.03	0.08±0.03	0.07±0.03	0.07±0.03
R65	0.10±0.04	0.13±0.04	0.13±0.04	0.51±0.24	0.50±0.14	0.12±0.04	0.47±0.14	0.10±0.04	0.10±0.04
R80	0.09±0.03	0.13±0.04	0.11±0.04	0.54±0.19	0.47±0.12	0.12±0.04	0.42±0.10	0.09±0.03	0.09±0.03

Table 41: Heat / Transolver. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 257/INT, 500/O500, 124/INT, 130/M120, 200/M120, 500/O500, 500/O500, 157/M120, 133/M120.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.05±0.04	0.19±0.08	0.22±0.09	0.74±0.14	0.82±0.10	0.17±0.05	0.82±0.10	0.06±0.04	0.07±0.03
H	0.26±0.04	0.07±0.05	0.48±0.06	0.50±0.14	0.53±0.14	0.16±0.03	0.52±0.13	0.19±0.04	0.14±0.04
V	0.22±0.02	0.43±0.04	0.07±0.04	0.74±0.15	0.82±0.14	0.21±0.04	0.82±0.14	0.15±0.03	0.10±0.03
CL	0.14±0.02	0.17±0.03	0.17±0.03	0.35±0.11	0.51±0.10	0.16±0.02	0.50±0.11	0.13±0.02	0.11±0.02
BL	0.49±0.04	0.35±0.08	0.34±0.07	0.45±0.11	0.08±0.04	0.35±0.07	0.08±0.04	0.49±0.03	0.40±0.02
LI	0.06±0.04	0.09±0.04	0.08±0.04	0.42±0.11	0.48±0.15	0.06±0.05	0.47±0.16	0.06±0.05	0.06±0.05
BD	0.33±0.07	0.34±0.06	0.36±0.06	0.49±0.11	0.25±0.15	0.34±0.06	0.07±0.04	0.41±0.08	0.40±0.04
R65	0.06±0.03	0.28±0.08	0.23±0.06	0.60±0.13	0.68±0.10	0.22±0.04	0.68±0.10	0.06±0.03	0.06±0.03
R80	0.14±0.01	0.44±0.07	0.43±0.05	0.74±0.16	0.87±0.15	0.39±0.04	0.87±0.16	0.08±0.02	0.07±0.02

Table 42: Heat / PINO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 500/R500, 500/O500, 500/O500, 500/O500, 500/R500, 500/R500, 500/O500, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.06±0.05	0.07±0.06	0.07±0.06	0.55±0.11	0.66±0.10	0.07±0.04	0.66±0.10	0.06±0.04	0.10±0.03
H	0.06±0.03	0.05±0.04	0.12±0.03	0.44±0.11	0.55±0.09	0.09±0.02	0.55±0.09	0.06±0.03	0.07±0.03
V	0.07±0.04	0.15±0.04	0.06±0.04	0.54±0.11	0.63±0.08	0.11±0.03	0.63±0.08	0.06±0.05	0.07±0.05
CL	0.08±0.03	0.09±0.02	0.17±0.07	0.08±0.03	0.09±0.03	0.12±0.05	0.09±0.03	0.08±0.03	0.09±0.04
BL	0.08±0.04	0.18±0.05	0.16±0.04	0.29±0.09	0.07±0.04	0.12±0.04	0.07±0.05	0.09±0.05	0.10±0.05
LI	0.06±0.05	0.07±0.04	0.07±0.04	0.34±0.11	0.42±0.10	0.06±0.04	0.42±0.10	0.06±0.04	0.06±0.05
BD	0.11±0.02	0.18±0.05	0.19±0.04	0.34±0.09	0.06±0.04	0.16±0.03	0.07±0.04	0.11±0.02	0.14±0.03
R65	0.08±0.04	0.10±0.04	0.12±0.04	0.44±0.10	0.51±0.10	0.09±0.03	0.51±0.09	0.05±0.05	0.06±0.04
R80	0.07±0.04	0.31±0.18	0.14±0.07	0.57±0.10	0.63±0.11	0.19±0.11	0.63±0.11	0.07±0.04	0.05±0.04

Table 43: Reaction–diffusion / U-FNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 500/R500, 500/R500, 500/R500, 500/R500, 500/R500, 500/R500, 500/O500, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.16±0.02	0.18±0.03	0.18±0.03	0.47±0.10	0.50±0.10	0.17±0.02	0.49±0.10	0.10±0.02	0.08±0.03
H	0.08±0.03	0.09±0.02	0.12±0.03	0.41±0.12	0.43±0.11	0.08±0.03	0.42±0.11	0.07±0.03	0.07±0.03
V	0.12±0.02	0.19±0.02	0.11±0.01	0.39±0.09	0.40±0.08	0.12±0.02	0.39±0.08	0.09±0.02	0.08±0.03
CL	0.08±0.03	0.09±0.03	0.09±0.04	0.18±0.05	0.26±0.05	0.08±0.03	0.24±0.05	0.08±0.05	0.08±0.06
BL	0.26±0.03	0.22±0.05	0.23±0.05	0.34±0.07	0.25±0.04	0.21±0.05	0.24±0.04	0.12±0.04	0.09±0.04
LI	0.06±0.03	0.07±0.03	0.07±0.03	0.34±0.09	0.33±0.09	0.07±0.03	0.31±0.08	0.06±0.03	0.06±0.04
BD	0.83±0.20	1.72±0.56	1.24±0.41	0.35±0.11	0.46±0.21	1.50±0.38	0.22±0.08	0.36±0.16	0.26±0.07
R65	0.14±0.01	0.19±0.02	0.18±0.02	0.59±0.10	0.66±0.10	0.17±0.02	0.66±0.10	0.10±0.02	0.08±0.03
R80	0.40±0.03	0.43±0.04	0.42±0.04	0.58±0.09	0.58±0.11	0.41±0.04	0.58±0.11	0.17±0.02	0.08±0.03

Table 44: Reaction–diffusion / FNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 500/O500, 500/O500, 469/INT, 500/O500, 500/O500, 500/O500, 500/O500, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.05±0.04	0.06±0.04	0.06±0.04	0.38±0.10	0.44±0.09	0.06±0.04	0.44±0.09	0.07±0.03	0.09±0.03
H	0.06±0.04	0.05±0.04	0.06±0.03	0.32±0.09	0.43±0.11	0.06±0.04	0.43±0.11	0.06±0.04	0.08±0.03
V	0.06±0.04	0.07±0.03	0.05±0.04	0.34±0.09	0.40±0.10	0.06±0.04	0.40±0.10	0.06±0.03	0.07±0.03
CL	0.09±0.06	0.08±0.05	0.08±0.06	0.08±0.04	0.09±0.04	0.08±0.06	0.09±0.03	0.10±0.06	0.10±0.06
BL	0.12±0.04	0.12±0.04	0.15±0.05	0.21±0.07	0.07±0.04	0.12±0.04	0.08±0.04	0.10±0.04	0.09±0.03
LI	0.05±0.04	0.05±0.04	0.05±0.04	0.30±0.09	0.37±0.10	0.05±0.04	0.37±0.10	0.07±0.03	0.09±0.02
BD	0.08±0.03	0.11±0.03	0.12±0.04	0.21±0.06	0.07±0.03	0.10±0.03	0.06±0.04	0.07±0.03	0.06±0.03
R65	0.10±0.02	0.11±0.03	0.11±0.03	0.43±0.08	0.48±0.08	0.11±0.03	0.48±0.08	0.05±0.04	0.07±0.03
R80	0.24±0.01	0.28±0.04	0.29±0.04	0.90±0.21	0.97±0.24	0.27±0.03	0.96±0.24	0.14±0.02	0.05±0.04

Table 45: Reaction–diffusion / CNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. All training views: 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.05±0.04	0.09±0.04	0.10±0.05	0.56±0.09	0.60±0.09	0.09±0.03	0.60±0.09	0.05±0.04	0.06±0.04
H	0.06±0.04	0.06±0.04	0.11±0.05	0.49±0.11	0.56±0.10	0.07±0.03	0.56±0.10	0.07±0.05	0.07±0.06
V	0.06±0.04	0.13±0.06	0.06±0.04	0.56±0.13	0.62±0.12	0.08±0.03	0.63±0.13	0.06±0.04	0.06±0.04
CL	0.08±0.03	0.08±0.03	0.08±0.03	0.11±0.04	0.28±0.12	0.08±0.03	0.32±0.12	0.07±0.03	0.07±0.03
BL	0.09±0.02	0.16±0.04	0.15±0.03	0.55±0.16	0.10±0.04	0.14±0.02	0.09±0.03	0.08±0.03	0.07±0.03
LI	0.06±0.04	0.06±0.04	0.06±0.04	0.45±0.11	0.54±0.08	0.06±0.04	0.53±0.08	0.06±0.04	0.06±0.04
BD	0.11±0.02	0.22±0.05	0.20±0.04	0.63±0.14	0.34±0.14	0.19±0.04	0.08±0.03	0.08±0.02	0.07±0.03
R65	0.06±0.04	0.18±0.08	0.19±0.09	0.76±0.08	0.84±0.07	0.17±0.06	0.84±0.07	0.05±0.04	0.05±0.04
R80	0.06±0.04	0.19±0.08	0.16±0.06	0.60±0.08	0.65±0.07	0.17±0.05	0.64±0.07	0.06±0.04	0.05±0.04

Table 46: Reaction–diffusion / DeepONet. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 500/O500, 500/O500, 500/O500, 500/O500, 500/O500, 500/O500, 251/INT, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.10±0.03	0.62±0.17	0.70±0.14	2.42±0.63	2.53±0.55	0.44±0.10	2.24±0.19	0.11±0.03	0.10±0.03
H	0.33±0.03	0.10±0.03	0.55±0.08	0.65±0.12	0.76±0.11	0.23±0.05	0.65±0.09	0.21±0.03	0.14±0.03
V	0.29±0.03	0.48±0.09	0.11±0.03	0.58±0.11	0.66±0.11	0.22±0.04	0.59±0.08	0.19±0.03	0.12±0.03
CL	0.27±0.06	0.75±0.82	0.45±0.25	0.20±0.07	0.43±0.11	0.34±0.17	0.42±0.10	0.27±0.06	0.23±0.05
BL	4.56±2.63	10.29±4.56	16.34±7.40	0.69±0.42	0.12±0.04	13.20±5.71	0.10±0.04	3.78±2.20	1.72±1.22
LI	0.29±0.03	0.16±0.04	0.16±0.04	0.73±0.17	0.85±0.13	0.10±0.03	0.75±0.10	0.22±0.03	0.16±0.03
BD	31.48±20.72	42.69±26.81	77.24±41.49	5.61±4.73	0.77±0.37	54.39±33.77	0.08±0.04	28.66±19.13	19.48±13.67
R65	0.11±0.03	0.25±0.05	0.24±0.04	0.51±0.09	0.60±0.10	0.19±0.04	0.52±0.09	0.09±0.03	0.10±0.03
R80	0.11±0.03	0.54±0.07	0.53±0.09	0.66±0.13	0.73±0.11	0.37±0.06	0.64±0.11	0.10±0.03	0.09±0.03

Table 47: Reaction–diffusion / GNOT. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 200/P10, 200/F200, 200/F200, 200/F200, 200/F200, 200/F200, 200/F200, 200/F200, 200/F200.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.09±0.03	0.12±0.04	0.32±1.50	11.15±7.82	18.08±2.25	0.12±0.04	18.48±0.52	0.09±0.03	0.09±0.03
H	0.10±0.04	0.11±0.04	0.14±0.05	8.92±7.86	26.61±15.88	0.12±0.04	33.23±15.23	0.10±0.04	0.10±0.04
V	0.10±0.03	0.15±0.06	0.10±0.03	0.66±0.20	0.72±0.19	0.12±0.04	0.68±0.18	0.09±0.03	0.09±0.03
CL	0.14±0.04	0.14±0.04	0.14±0.04	0.22±0.07	0.34±0.11	0.14±0.04	0.38±0.10	0.14±0.04	0.14±0.04
BL	0.10±0.05	0.11±0.05	0.11±0.05	0.33±0.10	0.11±0.04	0.11±0.05	0.10±0.04	0.10±0.05	0.10±0.05
LI	0.08±0.04	0.08±0.04	0.08±0.04	3.61±3.69	6.42±3.83	0.08±0.03	9.22±1.89	0.08±0.04	0.08±0.04
BD	0.10±0.03	0.14±0.04	0.12±0.04	0.59±0.16	0.45±0.20	0.12±0.03	0.11±0.04	0.10±0.03	0.10±0.03
R65	0.08±0.03	0.13±0.05	0.13±0.04	0.71±0.51	1.08±0.43	0.12±0.04	1.21±0.27	0.08±0.03	0.08±0.03
R80	0.09±0.03	0.13±0.04	0.15±0.17	1.88±1.94	3.53±1.85	0.13±0.04	4.53±0.99	0.09±0.03	0.09±0.03

Table 48: Reaction–diffusion / Transolver. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 200/P10, 200/M120, 200/F200, 200/M120, 35/INT, 47/INT, 200/F200, 200/F200, 200/F200.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.07±0.05	0.15±0.05	0.14±0.05	0.54±0.10	0.65±0.08	0.13±0.03	0.65±0.08	0.08±0.05	0.08±0.04
H	0.20±0.02	0.07±0.05	0.56±0.04	0.58±0.10	0.64±0.11	0.21±0.03	0.64±0.11	0.14±0.03	0.11±0.03
V	0.17±0.02	0.30±0.04	0.11±0.03	0.61±0.11	0.68±0.11	0.19±0.03	0.68±0.11	0.14±0.02	0.12±0.02
CL	0.12±0.03	0.10±0.03	0.11±0.03	0.18±0.09	0.40±0.14	0.10±0.03	0.46±0.16	0.12±0.03	0.10±0.03
BL	0.23±0.03	0.22±0.04	0.21±0.04	0.41±0.13	0.17±0.07	0.22±0.03	0.15±0.04	0.19±0.02	0.19±0.02
LI	0.88±0.05	0.97±0.02	0.97±0.03	1.01±0.06	1.06±0.08	0.97±0.02	1.06±0.07	0.84±0.04	0.83±0.04
BD	0.41±0.04	0.35±0.05	0.36±0.05	0.61±0.18	0.45±0.20	0.37±0.04	0.09±0.04	0.43±0.07	0.40±0.03
R65	0.07±0.03	0.65±0.25	0.71±0.24	1.04±0.09	1.06±0.08	0.82±0.21	1.06±0.08	0.06±0.03	0.06±0.03
R80	0.12±0.06	0.33±0.09	0.36±0.10	0.70±0.11	0.76±0.11	0.32±0.06	0.77±0.11	0.08±0.03	0.07±0.03

Table 49: Reaction–diffusion / PINO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 500/O500, 500/R500, 500/R500, 500/R500, 500/R500, 500/O500, 392/INT, 391/INT.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.06±0.04	0.09±0.04	0.08±0.04	0.49±0.10	0.57±0.10	0.08±0.04	0.57±0.10	0.06±0.04	0.63±0.04
H	0.10±0.06	0.06±0.04	0.12±0.04	0.47±0.09	0.58±0.08	0.10±0.04	0.58±0.09	0.10±0.06	0.11±0.05
V	0.08±0.02	0.13±0.04	0.06±0.04	0.46±0.09	0.55±0.08	0.10±0.04	0.56±0.08	0.08±0.02	0.09±0.04
CL	0.08±0.02	0.08±0.03	0.08±0.03	0.08±0.03	0.08±0.03	0.07±0.03	0.08±0.03	0.08±0.02	0.08±0.02
BL	0.08±0.03	0.16±0.05	0.16±0.05	0.27±0.08	0.07±0.04	0.11±0.03	0.07±0.04	0.14±0.03	0.26±0.03
LI	0.06±0.04	0.07±0.04	0.06±0.04	0.32±0.09	0.39±0.10	0.06±0.04	0.39±0.10	0.06±0.04	0.07±0.03
BD	0.08±0.03	0.16±0.05	0.15±0.04	0.26±0.08	0.07±0.04	0.11±0.04	0.07±0.04	0.07±0.03	0.07±0.03
R65	0.06±0.04	0.10±0.04	0.10±0.04	0.49±0.09	0.56±0.10	0.09±0.04	0.56±0.10	0.06±0.04	0.06±0.04
R80	0.07±0.04	0.15±0.05	0.14±0.04	0.62±0.13	0.67±0.15	0.13±0.03	0.67±0.14	0.07±0.04	0.06±0.04

Table 50: Burgers / U-FNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. All training views: 500/R500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.25±0.01	0.30±0.03	0.30±0.03	0.76±0.16	0.80±0.17	0.28±0.02	0.81±0.16	0.11±0.02	0.08±0.02
H	0.91±0.15	0.89±0.12	1.05±0.19	1.20±0.25	1.23±0.33	0.97±0.14	1.21±0.34	0.11±0.04	0.16±0.05
V	0.16±0.03	0.24±0.04	0.19±0.04	0.44±0.10	0.41±0.10	0.20±0.04	0.42±0.09	0.09±0.01	0.10±0.01
CL	0.07±0.03	0.07±0.03	0.07±0.03	0.09±0.03	0.13±0.04	0.07±0.03	0.12±0.05	0.07±0.03	0.07±0.03
BL	0.16±0.05	0.17±0.06	0.17±0.06	0.23±0.06	0.13±0.04	0.16±0.05	0.12±0.04	0.14±0.05	0.14±0.05
LI	0.06±0.02	0.06±0.02	0.06±0.02	0.33±0.10	0.32±0.09	0.06±0.02	0.32±0.08	0.06±0.02	0.06±0.02
BD	0.43±0.08	0.51±0.12	0.51±0.13	0.41±0.12	0.35±0.13	0.48±0.11	0.21±0.05	0.13±0.04	0.07±0.02
R65	0.10±0.04	0.12±0.03	0.12±0.03	0.50±0.11	0.58±0.12	0.12±0.03	0.58±0.11	0.06±0.03	0.05±0.02
R80	0.09±0.03	0.16±0.04	0.16±0.04	0.57±0.13	0.61±0.15	0.14±0.03	0.62±0.14	0.06±0.02	0.05±0.03

Table 51: Burgers / FNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/R500, 500/O500, 500/R500, 500/R500, 500/R500, 500/R500, 500/R500, 500/R500, 500/R500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.04±0.03	0.05±0.03	0.06±0.03	0.42±0.12	0.48±0.11	0.05±0.03	0.49±0.11	0.06±0.03	0.08±0.02
H	0.05±0.03	0.04±0.03	0.08±0.03	0.48±0.12	0.54±0.11	0.06±0.03	0.55±0.11	0.06±0.03	0.07±0.02
V	0.05±0.03	0.06±0.03	0.05±0.03	0.32±0.10	0.37±0.09	0.06±0.03	0.38±0.09	0.06±0.03	0.06±0.03
CL	0.06±0.04	0.06±0.04	0.05±0.04	0.07±0.03	0.08±0.03	0.06±0.04	0.08±0.04	0.06±0.04	0.06±0.04
BL	0.07±0.03	0.10±0.04	0.12±0.05	0.25±0.09	0.05±0.03	0.11±0.04	0.05±0.03	0.07±0.03	0.07±0.02
LI	0.04±0.03	0.04±0.03	0.04±0.03	0.30±0.11	0.35±0.10	0.04±0.03	0.36±0.11	0.06±0.03	0.07±0.02
BD	0.07±0.03	0.17±0.08	0.20±0.09	0.29±0.07	0.07±0.03	0.17±0.06	0.05±0.03	0.07±0.03	0.06±0.03
R65	0.07±0.02	0.09±0.02	0.09±0.02	0.48±0.12	0.53±0.12	0.09±0.02	0.55±0.12	0.05±0.03	0.06±0.02
R80	0.18±0.02	0.23±0.04	0.22±0.03	0.58±0.11	0.60±0.11	0.22±0.03	0.61±0.11	0.10±0.02	0.06±0.03

Table 52: Burgers / CNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. All training views: 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.05±0.03	0.11±0.07	0.12±0.07	0.60±0.13	0.70±0.14	0.11±0.03	0.71±0.13	0.05±0.03	0.05±0.03
H	0.08±0.03	0.05±0.03	0.21±0.09	0.56±0.13	0.65±0.11	0.12±0.03	0.66±0.12	0.07±0.03	0.06±0.03
V	0.08±0.03	0.20±0.08	0.05±0.03	0.51±0.12	0.58±0.10	0.13±0.04	0.58±0.09	0.07±0.03	0.06±0.03
CL	0.07±0.02	0.08±0.03	0.08±0.03	0.13±0.06	0.32±0.13	0.08±0.03	0.36±0.13	0.07±0.02	0.07±0.03
BL	0.13±0.02	0.18±0.03	0.17±0.04	0.61±0.17	0.10±0.04	0.15±0.03	0.08±0.03	0.09±0.02	0.07±0.02
LI	0.05±0.03	0.05±0.03	0.05±0.03	0.48±0.14	0.54±0.12	0.05±0.03	0.55±0.11	0.05±0.03	0.05±0.03
BD	0.13±0.02	0.20±0.04	0.22±0.05	0.65±0.18	0.33±0.16	0.19±0.04	0.08±0.03	0.09±0.02	0.07±0.02
R65	0.06±0.03	0.14±0.06	0.16±0.07	0.60±0.12	0.66±0.11	0.15±0.04	0.66±0.10	0.05±0.03	0.05±0.03
R80	0.06±0.02	0.28±0.10	0.22±0.07	0.68±0.13	0.69±0.12	0.24±0.06	0.69±0.11	0.05±0.03	0.05±0.03

Table 53: Burgers / DeepONet. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 500/R500, 339/INT, 500/R500, 500/O500, 443/INT, 500/O500, 491/INT, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.10±0.04	0.64±0.11	0.51±0.13	0.97±0.10	1.02±0.08	0.56±0.08	1.02±0.06	0.10±0.04	0.11±0.04
H	0.94±0.21	0.11±0.04	2.60±1.43	0.70±0.14	0.69±0.12	1.34±0.60	0.62±0.10	0.81±0.18	0.61±0.14
V	0.89±0.21	1.67±0.45	0.11±0.04	0.65±0.12	0.66±0.18	0.98±0.31	0.59±0.18	0.73±0.19	0.52±0.14
CL	0.32±0.07	1.35±0.90	0.70±0.30	0.20±0.06	0.43±0.12	0.48±0.19	0.40±0.11	0.33±0.07	0.28±0.06
BL	2.86±1.33	7.00±3.12	5.34±2.62	0.68±0.28	0.14±0.05	5.74±2.60	0.12±0.04	2.21±1.08	1.11±0.62
LI	0.32±0.05	0.21±0.08	0.21±0.06	0.54±0.11	0.55±0.11	0.10±0.04	0.50±0.10	0.31±0.06	0.30±0.06
BD	9.42±4.75	20.33±10.52	19.35±8.64	1.69±0.75	0.60±0.34	17.44±8.33	0.10±0.03	8.44±4.32	5.17±2.82
R65	0.19±0.06	0.32±0.07	0.40±0.15	0.74±0.13	0.75±0.11	0.23±0.06	0.69±0.09	0.10±0.04	0.13±0.05
R80	0.53±0.09	0.49±0.11	0.71±0.14	0.96±0.15	0.95±0.19	0.51±0.11	0.74±0.13	0.33±0.07	0.10±0.04

Table 54: Burgers / GNOT. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 200/F200, 500/O500, 200/F200, 200/F200, 500/O500, 200/P10, 200/F200, 200/F200, 200/F200.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.07±0.03	0.10±0.05	0.10±0.03	0.88±0.60	0.79±0.40	0.09±0.03	0.74±0.50	0.07±0.03	0.07±0.03
H	0.08±0.03	0.07±0.03	0.11±0.04	2.15±2.62	1.07±1.11	0.09±0.03	0.72±0.29	0.08±0.03	0.08±0.03
V	0.07±0.03	0.09±0.03	0.07±0.03	0.37±0.13	0.39±0.13	0.08±0.03	0.40±0.12	0.07±0.03	0.07±0.03
CL	0.09±0.04	0.10±0.04	0.10±0.04	0.14±0.05	0.25±0.08	0.10±0.04	0.28±0.08	0.09±0.04	0.09±0.04
BL	0.10±0.04	0.13±0.05	0.12±0.04	0.41±0.14	0.12±0.05	0.12±0.04	0.12±0.05	0.10±0.04	0.10±0.04
LI	0.10±0.04	0.11±0.04	0.11±0.04	0.93±0.58	0.87±0.34	0.11±0.04	0.74±0.27	0.10±0.04	0.10±0.04
BD	0.09±0.03	0.10±0.05	0.10±0.03	0.53±0.38	0.30±0.13	0.10±0.03	0.08±0.03	0.09±0.03	0.09±0.03
R65	0.07±0.03	0.09±0.03	0.09±0.03	0.59±0.37	0.59±0.34	0.08±0.03	0.60±0.34	0.07±0.03	0.07±0.03
R80	0.08±0.03	0.11±0.03	0.11±0.04	0.63±0.37	0.56±0.19	0.10±0.03	0.56±0.18	0.07±0.03	0.07±0.03

Table 55: Burgers / Transolver. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 200/F200, 200/M120, 200/M120, 200/M120, 200/M120, 200/M120, 73/P10, 200/F200, 200/F200.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.20±0.03	0.42±0.06	0.40±0.04	0.71±0.12	0.77±0.12	0.40±0.04	0.77±0.11	0.19±0.03	0.19±0.03
H	0.15±0.02	0.06±0.04	0.31±0.04	0.49±0.13	0.55±0.12	0.16±0.02	0.55±0.12	0.12±0.02	0.10±0.02
V	0.14±0.02	0.29±0.03	0.06±0.03	0.50±0.14	0.60±0.15	0.14±0.03	0.60±0.15	0.11±0.02	0.09±0.02
CL	0.07±0.03	0.08±0.03	0.09±0.03	0.10±0.05	0.26±0.11	0.07±0.03	0.28±0.10	0.06±0.03	0.07±0.03
BL	0.26±0.04	0.23±0.04	0.24±0.04	0.44±0.15	0.08±0.04	0.24±0.05	0.07±0.04	0.17±0.03	0.12±0.02
LI	0.06±0.03	0.06±0.03	0.06±0.04	0.46±0.15	0.56±0.13	0.06±0.04	0.56±0.12	0.06±0.03	0.07±0.02
BD	0.37±0.05	0.38±0.06	0.38±0.06	0.61±0.16	0.32±0.12	0.38±0.06	0.17±0.04	0.32±0.04	0.26±0.04
R65	0.21±0.03	0.44±0.09	0.41±0.10	0.82±0.22	0.85±0.22	0.39±0.06	0.86±0.22	0.15±0.03	0.16±0.02
R80	0.17±0.03	0.35±0.06	0.35±0.07	0.62±0.14	0.70±0.12	0.36±0.07	0.70±0.12	0.06±0.03	0.06±0.04

Table 56: Burgers / PINO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/R500, 500/O500, 500/R500, 500/R500, 500/O500, 500/R500, 500/R500, 500/O500, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.05±0.04	0.06±0.03	0.06±0.03	0.44±0.12	0.50±0.12	0.06±0.03	0.51±0.11	0.05±0.03	0.06±0.03
H	0.07±0.03	0.06±0.04	0.15±0.04	0.49±0.10	0.57±0.10	0.11±0.04	0.58±0.10	0.06±0.03	0.07±0.03
V	0.13±0.07	0.14±0.04	0.05±0.04	0.50±0.10	0.55±0.10	0.11±0.03	0.55±0.09	0.16±0.09	0.17±0.09
CL	0.07±0.03	0.08±0.03	0.08±0.03	0.08±0.03	0.09±0.03	0.07±0.03	0.09±0.03	0.06±0.03	0.06±0.03
BL	0.09±0.04	0.17±0.06	0.16±0.05	0.29±0.09	0.06±0.03	0.12±0.04	0.06±0.03	0.10±0.04	0.10±0.04
LI	0.05±0.03	0.05±0.03	0.05±0.03	0.36±0.11	0.41±0.11	0.05±0.03	0.42±0.11	0.05±0.03	0.06±0.03
BD	0.08±0.03	0.15±0.04	0.15±0.04	0.31±0.09	0.06±0.03	0.11±0.03	0.06±0.03	0.08±0.03	0.07±0.03
R65	0.05±0.03	0.09±0.04	0.08±0.03	0.52±0.12	0.57±0.13	0.08±0.03	0.58±0.12	0.05±0.04	0.06±0.04
R80	0.07±0.03	0.12±0.04	0.12±0.03	0.51±0.11	0.55±0.12	0.11±0.03	0.56±0.11	0.05±0.03	0.05±0.03

Table 57: Navier–Stokes / U-FNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/R500, 500/R500, 500/R500, 500/R500, 500/R500, 500/R500, 500/O500, 500/R500, 500/R500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.01±0.00	0.04±0.02	0.04±0.02	0.45±0.10	0.50±0.10	0.04±0.01	0.50±0.10	0.01±0.00	0.01±0.00
H	0.04±0.00	0.01±0.00	0.07±0.01	0.32±0.10	0.34±0.09	0.03±0.00	0.34±0.09	0.03±0.01	0.03±0.01
V	0.04±0.00	0.11±0.02	0.01±0.00	0.34±0.10	0.37±0.09	0.06±0.01	0.37±0.09	0.03±0.00	0.02±0.00
CL	0.05±0.01	0.05±0.01	0.05±0.01	0.09±0.04	0.14±0.05	0.05±0.01	0.14±0.05	0.05±0.01	0.03±0.01
BL	0.22±0.03	0.24±0.05	0.24±0.05	0.23±0.07	0.05±0.02	0.23±0.04	0.05±0.02	0.07±0.01	0.05±0.01
LI	0.01±0.00	0.01±0.00	0.01±0.00	0.31±0.10	0.36±0.09	0.01±0.00	0.35±0.09	0.01±0.00	0.01±0.00
BD	0.30±0.12	0.36±0.13	0.42±0.13	0.44±0.13	0.37±0.12	0.36±0.12	0.03±0.01	0.09±0.03	0.04±0.01
R65	0.01±0.00	0.06±0.03	0.07±0.03	0.49±0.10	0.55±0.10	0.06±0.02	0.55±0.11	0.01±0.00	0.01±0.00
R80	0.03±0.00	0.08±0.03	0.08±0.03	0.47±0.10	0.52±0.09	0.07±0.02	0.51±0.10	0.01±0.00	0.01±0.00

Table 58: Navier–Stokes / FNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 500/O500, 36/INT, 500/O500, 500/O500, 500/O500, 500/O500, 500/O500, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.00±0.00	0.01±0.01	0.01±0.01	0.36±0.10	0.44±0.10	0.01±0.01	0.45±0.10	0.03±0.01	0.07±0.02
H	0.01±0.00	0.00±0.00	0.05±0.02	0.42±0.12	0.55±0.13	0.04±0.02	0.55±0.13	0.02±0.00	0.03±0.01
V	0.03±0.01	0.09±0.03	0.03±0.01	0.44±0.10	0.51±0.08	0.06±0.02	0.51±0.08	0.05±0.01	0.05±0.01
CL	0.01±0.00	0.02±0.00	0.02±0.00	0.02±0.01	0.05±0.02	0.01±0.00	0.05±0.02	0.02±0.01	0.04±0.02
BL	0.06±0.01	0.09±0.02	0.14±0.08	0.30±0.09	0.01±0.00	0.10±0.03	0.01±0.00	0.05±0.02	0.04±0.02
LI	0.00±0.00	0.00±0.00	0.00±0.00	0.27±0.10	0.34±0.10	0.00±0.00	0.35±0.11	0.01±0.00	0.01±0.00
BD	0.26±0.11	0.37±0.19	0.33±0.16	0.37±0.13	0.03±0.01	0.33±0.15	0.01±0.00	0.16±0.10	0.21±0.13
R65	0.06±0.01	0.09±0.03	0.09±0.03	0.49±0.11	0.56±0.10	0.08±0.02	0.56±0.10	0.00±0.00	0.04±0.00
R80	0.11±0.06	0.14±0.05	0.14±0.05	0.48±0.10	0.53±0.09	0.14±0.05	0.54±0.09	0.04±0.01	0.00±0.00

Table 59: Navier–Stokes / CNO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. All training views: 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.01±0.01	0.11±0.10	0.11±0.09	0.96±0.22	1.00±0.24	0.10±0.05	1.01±0.24	0.02±0.01	0.02±0.01
H	0.09±0.01	0.02±0.01	0.21±0.08	0.50±0.11	0.63±0.10	0.13±0.03	0.62±0.10	0.08±0.01	0.08±0.01
V	0.07±0.02	0.15±0.05	0.02±0.01	0.56±0.18	0.71±0.20	0.11±0.02	0.70±0.19	0.06±0.01	0.05±0.01
CL	0.06±0.01	0.08±0.02	0.08±0.02	0.12±0.05	0.34±0.12	0.07±0.01	0.35±0.12	0.05±0.01	0.05±0.01
BL	0.15±0.06	0.38±0.16	0.42±0.16	0.56±0.15	0.09±0.04	0.35±0.12	0.07±0.03	0.08±0.02	0.05±0.01
LI	0.02±0.01	0.02±0.01	0.02±0.01	0.54±0.15	0.69±0.19	0.02±0.01	0.69±0.19	0.02±0.01	0.02±0.01
BD	0.16±0.09	0.35±0.09	0.51±0.23	0.64±0.14	0.33±0.14	0.38±0.14	0.06±0.02	0.10±0.04	0.06±0.02
R65	0.03±0.01	0.14±0.08	0.12±0.06	0.59±0.09	0.68±0.08	0.12±0.04	0.68±0.08	0.01±0.01	0.02±0.01
R80	0.04±0.01	0.19±0.08	0.20±0.08	0.68±0.14	0.71±0.13	0.19±0.05	0.71±0.12	0.02±0.01	0.01±0.01

Table 60: Navier–Stokes / DeepONet. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 500/O500, 500/O500, 500/O500, 500/O500, 500/O500, 128/INT, 500/O500, 500/O500, 500/O500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.03±0.01	0.13±0.04	0.15±0.04	0.53±0.12	0.64±0.14	0.11±0.03	0.61±0.13	0.04±0.01	0.06±0.02
H	0.60±0.05	0.05±0.02	0.78±0.16	0.54±0.11	0.55±0.12	0.46±0.10	0.51±0.10	0.47±0.05	0.31±0.05
V	0.78±0.10	1.25±0.27	0.05±0.02	0.57±0.11	0.59±0.12	0.78±0.19	0.56±0.10	0.66±0.11	0.47±0.10
CL	0.19±0.04	0.71±0.30	0.50±0.14	0.18±0.05	0.40±0.10	0.33±0.11	0.40±0.09	0.18±0.04	0.16±0.04
BL	2.11±0.69	4.18±1.49	3.89±1.46	0.52±0.14	0.09±0.03	3.80±1.40	0.08±0.03	1.47±0.44	0.58±0.16
LI	0.29±0.04	0.12±0.03	0.10±0.03	0.47±0.10	0.54±0.10	0.09±0.03	0.52±0.11	0.22±0.02	0.15±0.02
BD	2.01±0.49	3.23±0.81	3.24±0.87	0.96±0.28	0.46±0.22	2.95±0.72	0.04±0.01	1.84±0.46	1.42±0.36
R65	0.04±0.01	0.16±0.04	0.17±0.04	0.52±0.11	0.59±0.11	0.14±0.03	0.55±0.10	0.03±0.01	0.04±0.01
R80	0.09±0.02	0.19±0.04	0.22±0.04	0.53±0.11	0.58±0.11	0.19±0.03	0.53±0.10	0.04±0.01	0.03±0.01

Table 61: Navier–Stokes / GNOT. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 200/F200, 200/F200, 200/F200, 200/F200, 200/F200, 200/F200, 200/F200, 500/O500, 200/F200.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.02±0.01	0.04±0.01	0.04±0.01	0.27±0.08	0.36±0.12	0.03±0.01	0.35±0.12	0.02±0.01	0.02±0.01
H	0.06±0.03	0.07±0.03	0.08±0.03	0.36±0.14	0.47±0.18	0.07±0.03	0.46±0.17	0.06±0.03	0.06±0.03
V	0.03±0.01	0.08±0.04	0.03±0.01	0.34±0.13	0.42±0.13	0.06±0.02	0.39±0.12	0.03±0.01	0.03±0.01
CL	0.11±0.04	0.11±0.04	0.12±0.04	0.19±0.06	0.28±0.10	0.11±0.04	0.35±0.09	0.11±0.04	0.11±0.04
BL	0.09±0.03	0.11±0.04	0.11±0.04	0.46±0.12	0.12±0.05	0.11±0.03	0.11±0.05	0.09±0.03	0.09±0.03
LI	0.06±0.03	0.07±0.03	0.06±0.03	0.33±0.13	0.48±0.17	0.07±0.03	0.51±0.19	0.06±0.03	0.06±0.03
BD	0.03±0.01	0.05±0.02	0.05±0.02	0.32±0.11	0.29±0.11	0.05±0.01	0.04±0.02	0.03±0.01	0.03±0.01
R65	0.01±0.01	0.04±0.02	0.04±0.02	0.35±0.12	0.39±0.13	0.04±0.01	0.38±0.15	0.01±0.01	0.01±0.01
R80	0.03±0.01	0.07±0.03	0.07±0.03	0.63±0.50	0.48±0.14	0.07±0.02	0.45±0.15	0.03±0.01	0.03±0.01

Table 62: Navier–Stokes / Transolver. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. Epoch/cohort in row order: 200/M120, 31/P10, 200/M120, 200/M120, 200/M120, 200/M120, 200/M120, 200/M120, 200/M120.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.03±0.01	0.12±0.06	0.12±0.07	0.62±0.12	0.69±0.10	0.11±0.03	0.69±0.10	0.03±0.01	0.04±0.01
H	0.12±0.02	0.12±0.04	0.28±0.10	0.78±0.23	0.93±0.27	0.17±0.03	0.94±0.27	0.11±0.03	0.10±0.03
V	0.82±0.05	0.90±0.01	0.02±0.01	0.78±0.11	0.83±0.08	0.88±0.02	0.83±0.09	0.63±0.16	0.21±0.21
CL	0.08±0.02	0.13±0.05	0.08±0.01	0.13±0.07	0.27±0.10	0.07±0.01	0.27±0.10	0.07±0.01	0.07±0.01
BL	0.41±0.07	0.34±0.06	0.31±0.05	0.45±0.10	0.06±0.04	0.31±0.05	0.05±0.02	0.47±0.11	0.44±0.08
LI	0.03±0.01	0.02±0.01	0.03±0.01	0.45±0.11	0.61±0.11	0.02±0.01	0.60±0.11	0.03±0.01	0.04±0.01
BD	0.79±0.16	0.71±0.10	0.68±0.10	0.70±0.17	0.25±0.16	0.74±0.14	0.04±0.02	0.68±0.14	0.49±0.07
R65	0.03±0.01	0.22±0.08	0.23±0.08	0.66±0.16	0.73±0.15	0.23±0.06	0.73±0.15	0.03±0.01	0.03±0.01
R80	0.14±0.03	0.45±0.10	0.43±0.08	0.67±0.10	0.74±0.09	0.45±0.06	0.75±0.09	0.05±0.01	0.03±0.01

Table 63: Navier–Stokes / PINO. All 81 cells: mean ± SD on one line ($n = 200$); blue marks matched views. All training views: 500/R500.

	Test observation pattern								
Train	R50	H	V	CL	BL	LI	BD	R65	R80
R50	0.00±0.00	0.05±0.02	0.05±0.02	0.55±0.10	0.64±0.10	0.04±0.01	0.64±0.10	0.03±0.02	0.03±0.02
H	0.33±0.19	0.00±0.00	0.81±0.11	0.54±0.11	0.69±0.11	0.73±0.14	0.68±0.11	0.34±0.19	0.35±0.16
V	1.32±0.62	2.12±0.51	0.00±0.00	2.05±0.52	2.12±0.52	2.02±0.51	2.12±0.49	1.34±0.59	1.33±0.57
CL	0.04±0.01	0.07±0.06	0.06±0.02	0.05±0.02	0.06±0.03	0.06±0.02	0.06±0.03	0.04±0.02	0.04±0.03
BL	0.22±0.12	0.21±0.09	0.18±0.07	0.29±0.09	0.02±0.01	0.12±0.04	0.02±0.01	0.18±0.11	0.13±0.10
LI	0.01±0.00	0.01±0.00	0.01±0.00	0.49±0.09	0.57±0.12	0.00±0.00	0.58±0.12	0.02±0.01	0.15±0.13
BD	0.30±0.12	0.17±0.06	0.22±0.07	0.35±0.11	0.03±0.01	0.15±0.06	0.02±0.01	0.29±0.19	0.34±0.31
R65	0.04±0.01	0.05±0.02	0.05±0.02	0.54±0.09	0.60±0.09	0.04±0.02	0.61±0.09	0.00±0.00	0.01±0.00
R80	0.06±0.02	0.07±0.03	0.08±0.03	0.81±0.31	0.73±0.22	0.07±0.02	0.73±0.21	0.02±0.01	0.00±0.00

G MIXED-PATTERN TRAINING

This appendix records five completed mixed-pattern models reported in Figure 3: Helmholtz/FNO, Helmholtz/DeepONet, Heat/DeepONet, Navier–Stokes/U-FNO, and Poisson/U-FNO. One additional model per pair is trained on a deterministic mixture of the nine frozen observation views and scored on those views with the same 200 held-out identities as the single-pattern models. These study rows supplement the frozen 441-model grid; no grid value is replaced.

Training mixture. The training mask is requested as one specification, the nine frozen views joined by + with their frozen arguments:

```
random_3pct(ratio=0.5) + random_3pct(ratio=0.65) +
random_3pct(ratio=0.8) + block_missing(missing_fraction=0.49) +
line_sensors(num_lines=76, orientation=both) +
line_sensors(num_lines=64, orientation=horizontal) +
line_sensors(num_lines=64, orientation=vertical) +
boundary_sensors(width=19) + clustered_sensors(ratio=0.5).
```

Each physical identity is assigned exactly one component: the nine-view cycle is rotated by a seed derived from the campaign mask seed, the canonical specification and the identity's regime, and indexed by its generation-time `regime_sample_index`. Selecting the training identities after this assignment yields 190–208 records per view among the 1,800 training records. The training mixture is therefore approximately balanced, not rebalanced within each training regime. Assignments remain fixed across epochs and methods. Each component derives the per-identity mask seed from its resolved protocol name in the same way as a frozen view, so a given record exposes the same sensors to the mixed model and the corresponding single-pattern model. Components were checked against the frozen views bit for bit before training.

Recorded training budgets and comparison scopes. The MIX models retain the corresponding recorded batch, optimizer, learning-rate schedule, loss and stopping settings, changing the observation assignment. Actual completion receipts record 500 epochs and 112,500 updates for Helmholtz/FNO; 500 epochs and 225,000 updates each for Helmholtz/DeepONet, Heat/DeepONet and Navier–Stokes/U-FNO; and 200 epochs and 90,000 updates for Poisson/U-FNO. Equal stopping rules alone would not establish equal completed budgets. The four 500-epoch pairs have 500-epoch single-pattern references. Poisson uses the settings-v3 one-cycle recipe (learning rate 5×10^{-4}, maximum 200 epochs, train-loss patience after epoch 120); only its BL, BD and CL references use that recipe and complete 200 epochs with 90,000 updates. Its other six references use a step schedule and stop at 130–163 epochs. All-nine Poisson comparisons are consequently descriptive: schedule, completed budget and exposure differ, and these results do not identify which factor dominates.

Scoring. Every MIX result uses strict float64 rescoring of retained prediction arrays with the main-grid scorer, not a legacy scalar. The 45 reference-view and 15 extra-view blocks each contain 200 held-out identities; contracts bind identities, masks, target frames, checkpoints and configurations. All destination-trained denominators and transferred-model comparisons use the main grid's `20260925-v1` prediction scores. The three extra protocols are uniformly random points at 20 % and 35 % density, and a deterministic mixture assigning either line sensors at 30 % or random points at 20 % to a record, not their union. No specialist was trained on these extra protocols, so no specialist ratio is reported for them. Full-precision comparisons and score bindings are recorded in `results/revision_v2/mixed_comparisons.csv`, `mixed_extra_views.csv` and `mixed_summary.json`.

Reading. MIX is below the transferred-model median in 45/45 comparisons and below the best transfer in 34/45; transfer ratios reach 1,040.09. The four 500-epoch pairs give specialist-relative ratios of 0.63–4.07, including Navier–Stokes gains on R50, LI and V (0.63, 0.96 and 0.98). Poisson's recipe-matched range is 2.51–3.49. Extra densities (Figure 5, Table 65) are separately sampled discrete tests, not a continuous or nested-mask density law. These completed reconstruction and forecasting cases are not an architecture ranking or a complete attempt inventory: an additional attempt collapsed at epoch 90 and was not scored. Longer-training and capacity controls remain untested.

Table 64: **Mixed-pattern comparisons with explicit populations.** All columns in a row use the same destination set W. MIX and specialist errors are mean $\pm$ sample SD across the $|W|$ cell means, each based on 200 records, multiplied by 100 (error %, not observation density). The geometric mean of $R_{\mathrm{MIX}\to w}$ is dimensionless. Poisson's all-nine and recipe-matched populations are separate; the full-grid C remains in Appendix F and is not reused for the restricted set.

		Error (%): mean $\pm$ view SD		
Pair	Destinations W	MIX	Specialist	geo. R
Helmholtz / FNO	All nine	5.31 ± 5.81	2.12 ± 1.74	2.07
Helmholtz / DeepONet	All nine	12.14 ± 10.68	5.01 ± 4.82	2.56
Heat / DeepONet	All nine	13.29 ± 6.28	9.83 ± 3.78	1.32
Navier–Stokes / U-FNO	All nine	3.49 ± 3.98	2.55 ± 2.87	1.19
Poisson / U-FNO	All nine (descriptive)	7.12 ± 8.01	11.65 ± 9.77	0.46
Poisson / U-FNO	BL, BD, CL (matched)	17.10 ± 5.69	5.50 ± 1.19	3.05

Table 65: **MIX on protocols absent from training.** Error (%) is $100\times$ per-record relative-L^2, shown as mean $\pm$ sample SD over 200 records (ddof $= 1$). Column percentages instead specify observation density. "Line 30 / random 20" assigns one of the two patterns per record. There is no matched specialist for these protocols.

	Test observation protocol		
Pair	Random 20%	Random 35%	Line 30 / random 20
Helmholtz / FNO	14.22 ± 3.98	4.34 ± 1.44	11.50 ± 4.56
Helmholtz / DeepONet	18.31 ± 4.18	6.73 ± 1.97	15.03 ± 5.82
Heat / DeepONet	23.21 ± 9.93	13.22 ± 5.92	18.92 ± 8.73
Navier–Stokes / U-FNO	8.39 ± 2.23	1.74 ± 0.56	6.27 ± 2.80
Poisson / U-FNO	15.30 ± 1.80	4.47 ± 0.94	12.49 ± 7.14

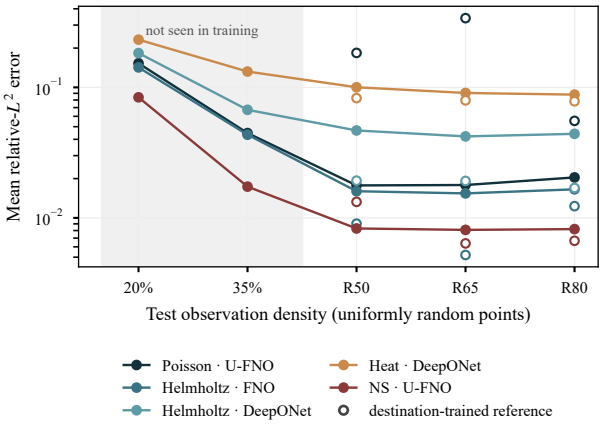

Figure 5: **Density response of five completed MIX models.** Mean relative-L^2 error at random observation densities of 20%, 35%, 50% (R50), 65% (R65) and 80% (R80); shaded densities were absent from training. Open markers show destination-trained references where available. Masks are separately sampled, not nested; connecting lines do not establish a continuous density law.