BROC: BLOCK-RECURSIVE OUTPUT CORRECTION FOR FROZEN NEURAL PDE SURROGATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Simulation-trained neural PDE surrogates often suffer systematic performance degradation when deployed on measured physical streams or subjected to shifted target dynamics. We introduce *Block-Recursive Output Correction* (BROC), a residual observer that calibrates a trained and frozen surrogate using matured sparse measurements. Unlike conventional test-time adaptation (TTA) methods that rely on iterative gradient steps, BROC updates a spatially shared temporal component error map via a recursive block, requiring neither backpropagation nor neural parameter updates. Theoretically, we prove the exact optimality of the recursive update and strict temporal non-anticipation. We also establish a finite-sample certificate for full-field improvement when the shared residual signal dominates delay drift, probe error, and measurement noise. Across 20 system–surrogate–checkpoint cells in RealPDEBench, BROC achieves the best accuracy in every cell, reducing MSE by 32.56% relative to the Frozen surrogate and outperforming recent adaptation baselines by 23.84%–29.84%. Under controlled shifts of physical parameters across 12 PDEBench cells, it removes 38.31%–80.47% of Frozen error and outperforms recent baselines by 25.05%–65.87%. Moreover, its analytic calibration core executes in 5.317 ms on RealPDEBench and 4.348 ms on PDEBench, yielding respective 5.02×–11.97× and 7.74×–17.08× speedups over iterative output adapters. Compared with the baseline requiring surrogate re-execution, BROC achieves speedups of 74.90× and 26.75×. Comprehensive stress tests show that BROC remains effective under sparse, dynamic, delayed, and noisy sensing, maintains substantial gains in autoregressive rollouts, supports protective gating, and improves physical fidelity.

1 INTRODUCTION

Neural PDE surrogates provide fast predictions of evolving physical fields, reducing the cost of repeated numerical simulation (Li et al., 2021; Lu et al., 2021). Their practical usefulness, however, depends on maintaining accuracy under the physical conditions encountered during deployment. A surrogate trained on simulations can exhibit systematic prediction errors on measured trajectories. Measurements collected during operation provide information about these errors and an opportunity to improve subsequent forecasts. We study how to use this information while keeping the pretrained surrogate frozen and preserving the computational advantages of surrogate-based prediction. RealPDEBench exposes this sim-to-real gap in measured systems (Hu et al., 2026), while PDEBench permits controlled tests of physical-parameter shifts (Takamoto et al., 2022).

Turning incoming measurements into useful forecast corrections presents a coupled spatial and temporal challenge. Measurements may cover only a small subset of locations, whereas predictions are required across the full field. Moreover, a correction fitted to these readings must remain useful beyond both the sensors and the observed interval. Reducing error at observed locations and past times therefore does not by itself establish improvement at unobserved locations and future times. The task requires identifying error relationships that generalize across space and time from limited evidence at a modest computational cost.

Our premise is that spatially varying errors may follow a common relationship with local forecast values. If this relationship changes sufficiently slowly, measurements can constrain corrections at other positions and later times. This motivates a compact model linking predictions to their

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.

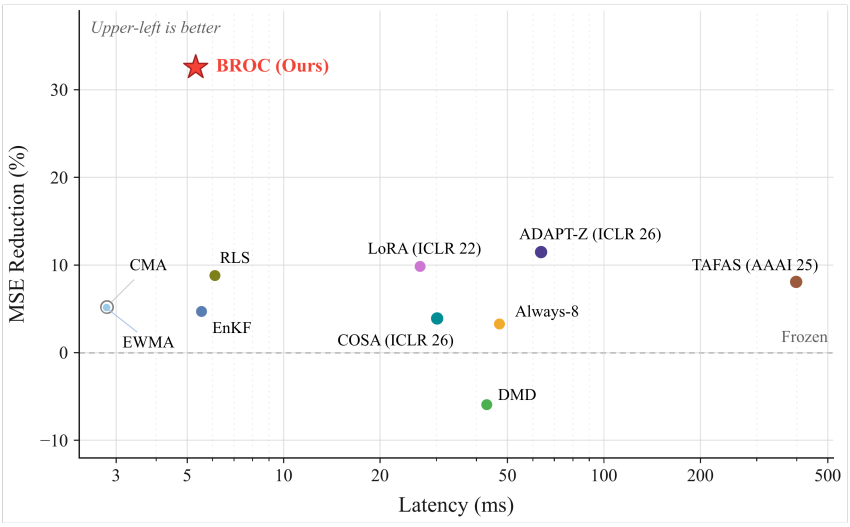


Figure 1: Accuracy and efficiency comparison on RealPDEBench; higher accuracy and lower latency are better.

errors, with interactions across prediction steps and physical variables. The validity of this premise is examined through residual diagnostics and an analysis of the effects of drift and observation uncertainty.

We introduce *Block-Recursive Output Correction* (BROC), a residual observer that estimates a spatially shared affine map coupling future prediction blocks and physical variables. Once an observation block matures, BROC pairs it with the cached surrogate forecast to update two sufficient statistics. A regularized linear solve yields the correction map for subsequent full-field forecasts. Each forecast uses only previously matured observations, without backpropagation, neural-parameter updates, or additional surrogate evaluations. The state dimension depends on the forecast horizon and corrected variables, not the spatial grid size.

Our contributions are threefold:

- We introduce BROC, an observer that converts sparse measurements into spatially shared corrections for frozen neural PDE surrogates without backpropagation.
- We prove exact recursive optimality and temporal non-anticipation, and derive a finite-sample sparse-to-field certificate exposing delay, sampling, and noise penalties.
- Across two benchmarks, BROC provides the strongest accuracy–efficiency operating point among the compared methods. Its gains persist across a broad range of deployment conditions.

2 Related Work

PDE surrogates have been widely studied. They approximate mappings from physical inputs to evolving solution fields, enabling fast inference in place of repeated numerical solves. Convolutional encoder–decoder architectures such as U-Net have been used for PDE forecasting (Ronneberger et al., 2015; Takamoto et al., 2022), while FNO and DeepONet learn solution operators as function-to-function maps (Li et al., 2021; Lu et al., 2021). CNO extends convolutional models to continuous operator learning (Raonic et al., 2023), and Transolver uses physics-aware attention to handle general geometries (Wu et al., 2024). Despite these advances, a surrogate trained on simulations can retain systematic errors on measured trajectories. This deployment gap motivates adaptation methods that use available measurements from the target system to improve subsequent forecasts.

Adaptation methods. Sequential estimation incorporates observations into forecast states, while learned dynamical representations and model-error estimators address evolving dynamics and pre-

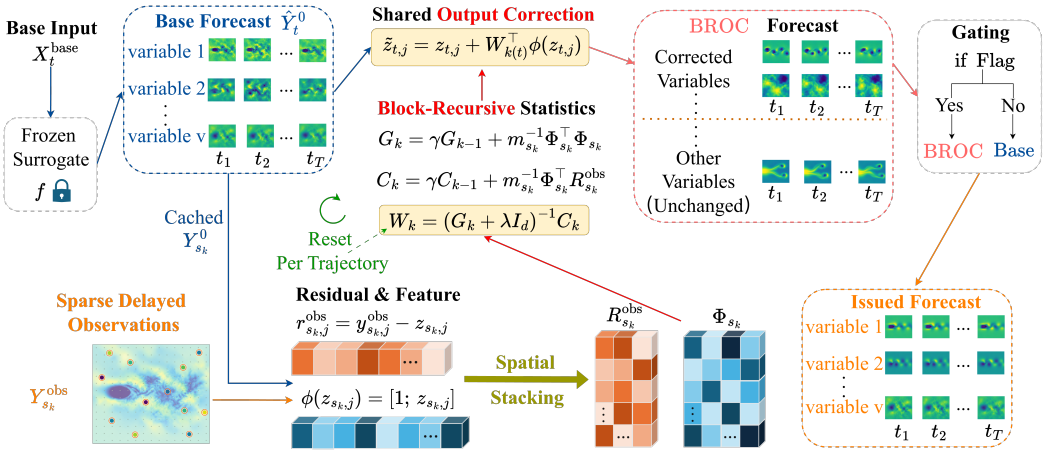


Figure 2: Overview of Block-Recursive Output Correction (BROC).

diction discrepancies (Evensen, 1994; Williams et al., 2015; Farchi et al., 2021). Methods for transfer across physical settings adapt contextual representations or compose pretrained operators at test time (Kirchmeyer et al., 2022; Kassaï Koupaï et al., 2024; Serrano et al., 2026). Test-time adaptation adjusts normalization, selected parameters, or activations under distribution shift (Wang et al., 2021; Niu et al., 2022; 2024). In forecasting, revealed targets support input calibration and latent-feature adjustment (Kim et al., 2025; Huang et al., 2026). Other methods update lightweight adapters online through gradient-based optimization, including low-rank, output-space, and frequency-aware designs (Medeiros et al., 2025; Im & Kwon, 2026; Wang et al., 2026). Running and exponentially weighted residual averages are computationally inexpensive (Roberts, 1959), but their uncoupled summaries cannot capture interactions among forecast leads and physical components, which can limit overall accuracy.

3 Block-Recursive Output Correction

BROC adds an observer to the output of a frozen surrogate. It estimates a spatially shared residual map from delayed, sparse measurements and applies it across the full field. The online state contains two sufficient statistics and an affine coefficient matrix, requiring no backpropagation or neural-parameter updates. Figure 2 illustrates the overall sequence.

3.1 Forecast and Residual Representation

At decision t, a frozen surrogate issues a block forecast

$$\widehat{Y}_t^0 = f(X_t^{\text{base}}) \in \mathbb{R}^{T \times N_t \times C}, \tag{1}$$

where X_t^{base} is the surrogate input and $f(\cdot)$ denotes the frozen surrogate. Here T is the number of time steps within each forecast block, N_t is the number of spatial locations, and C is the number of physical variables. The base forecast is cached. The corresponding target is observed through

$$Y_t^{\text{obs}} = M_t P_{\text{corr}} Y_t + \varepsilon_t. \tag{2}$$

Here $Y_t \in \mathbb{R}^{T \times N_t \times C}$ is the ground-truth block corresponding to $\widehat{Y}_t^0$. M_t selects the observed spatial locations from the N_t locations, and P_{corr} selects the correctable physical variables from the C variables. The term ε_t denotes measurement noise.

At each observed location j, define the clean and observed residuals as

$$r_{t,j}^\star = y_{t,j} - z_{t,j}, \qquad r_{t,j}^{\text{obs}} = y_{t,j}^{\text{obs}} - z_{t,j} = r_{t,j}^\star + \varepsilon_{t,j}. \tag{3}$$

Here $y_{t,j} \in \mathbb{R}^q$ is the noise-free target vector, $y_{t,j}^{\text{obs}} \in \mathbb{R}^q$ is its measured counterpart, and $z_{t,j} \in \mathbb{R}^q$ is the cached forecast vector from $\widehat{Y}_t^0$. These vectors concatenate values at location j across the T time steps in the forecast block and the variables selected by P_{corr}. Let C_{corr} denote the number

of selected variables; then $q = TC_{\mathrm{corr}}$. We define the feature vector with an intercept as $\phi(z_{t,j}) = [1; z_{t,j}] \in \mathbb{R}^d$; then $d = q + 1$. Let m_t be the number of observed locations. Stacking the observed residuals and feature vectors as rows gives $R_t^{\mathrm{obs}} \in \mathbb{R}^{m_t \times q}$ and $\Phi_t \in \mathbb{R}^{m_t \times d}$, respectively.

3.2 BROC UPDATE AND PREDICTION

Let s_k denote the forecast block incorporated by the k-th committed update. Starting from $G_0 = C_0 = 0$, BROC updates

$$G_k = \gamma G_{k-1} + m_{s_k}^{-1} \Phi_{s_k}^\top \Phi_{s_k}, \qquad C_k = \gamma C_{k-1} + m_{s_k}^{-1} \Phi_{s_k}^\top R_{s_k}^{\mathrm{obs}}, \tag{4}$$

where $\gamma \in (0, 1]$ is the forgetting factor. The matrices $G_k \in \mathbb{R}^{d \times d}$ and $C_k \in \mathbb{R}^{d \times q}$ store the accumulated feature and feature–residual statistics, respectively. The correction matrix $W_k \in \mathbb{R}^{d \times q}$ is given by

$$W_k = (G_k + \lambda I_d)^{-1} C_k, \tag{5}$$

where $\lambda > 0$ is the regularization parameter and I_d is the $d \times d$ identity matrix. As G_k and C_k summarize the matured blocks, computing W_k requires only one linear solve. This update solves the following objective:

$$\mathcal{J}_k(W) = \sum_{i=1}^{k} \gamma^{k-i} m_{s_i}^{-1} \| R_{s_i}^{\mathrm{obs}} - \Phi_{s_i} W \|_{\mathrm{F}}^2 + \lambda \| W \|_{\mathrm{F}}^2. \tag{6}$$

Here i indexes committed updates and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. The first term measures residual-fitting error. $m_{s_i}^{-1}$ normalizes each block by its number of observed locations and γ^{k-i} discounts earlier updates. The second term controls the magnitude of W.

For subsequent forecasts, W_k corrects a local vector z:

$$\widetilde{z} = z + W_k^\top \phi(z). \tag{7}$$

This correction jointly models residuals across the T time steps and C_{corr} selected variables within each forecast block. Applying the same W_k at every spatial location yields the corrected full-field forecast. This assumes that observed and unobserved locations share the same local feature–residual relation. Variables not selected by P_{corr} retain their frozen predictions.

An optional gate selects between the corrected and Frozen forecasts using the most recent matured loss comparison. Let $s^\star(t)$ denote the most recent block available before forecast t. The gate is

$$a_t = \mathbf{1}\left\{ \ell_{s^\star(t)}^{\mathrm{BROC}} < \ell_{s^\star(t)}^{0} \right\}, \qquad \widehat{Y}_t^{\mathrm{gate}} = a_t \widetilde{Y}_t + (1 - a_t) \widehat{Y}_t^{0}. \tag{8}$$

The losses compare BROC with Frozen on $s^\star(t)$.

Let $k(t)$ be the number of updates committed before forecast t is issued. That forecast uses $W_{k(t)}$. Newly matured observations affect only later forecasts, and the state resets between physical trajectories. Appendix A provides the supporting derivations and deployment details.

4 GUARANTEES

Equation 5 uniquely solves the regularized objective in Equation 6 for arbitrary observation delays and $\gamma \in (0, 1]$ (Theorem 1). The prediction-before-update protocol ensures temporal non-anticipation: each corrected forecast depends only on the frozen forecast and previously matured observations (Theorem 2).

Theorem 7 gives a deterministic finite-sample full-field bound. At decision t, let W_t^F denote the correction matrix obtained from the complete clean fields of the matured blocks, and let $\widehat{W}_t = W_{k(t)}$ denote the matrix estimated from sparse, noisy observations. Then

$$\begin{aligned} \mathcal{R}_{t,h}^{+}(0) - \mathcal{R}_{t,h}^{+}(\widehat{W}_t) \geq \mathsf{S}_t - 2\|W_t^F\|_{\mathrm{F}} d_{K,t,h} - \|W_t^F\|_{\mathrm{F}}^2 d_{\Sigma,t,h} \\ - 2\eta_t \zeta_{t,h} - \|\Sigma_{t,h}^{+}\|_2 \eta_t^2. \end{aligned} \tag{9}$$

Here S_t is the clean full-field residual signal captured by W_t^F. The terms $d_{K,t,h}$ and $d_{\Sigma,t,h}$ measure temporal changes in the residual statistics. The term η_t bounds estimation error from sparse and noisy observations. The term $\zeta_{t,h}$ measures future residual misalignment. A positive right-hand side guarantees lower future clean full-field risk than Frozen for the corrected variables. Under uniform sampling without replacement and conditionally sub-Gaussian measurement noise, Corollary 3 provides an explicit bound for η_t that holds with probability at least $1 - \delta$ jointly over all matured blocks. Appendix A states the assumptions, concentration terms, and proofs.

5 Experiments

5.1 Benchmarks, Compared Models, and Protocol

Benchmarks. We evaluate BROC on two complementary benchmarks. From **RealPDEBench** (Hu et al., 2026), we use four measured physical systems: Cylinder, Controlled Cylinder, Foil, and Fluid–Structure Interaction (FSI). Each system contributes four or six cells defined by its supported frozen surrogate checkpoints. The 20 cells contain 138 unique physical trajectories and 686 model records. **PDEBench** (Takamoto et al., 2022) provides a controlled benchmark for evaluating shifts in governing physics. We use 2D Diffusion–Reaction, 1D Diffusion–Sorption, and 1D Compressible Navier–Stokes, together with frozen FNO and U-Net surrogates. For each PDEBench system, we construct mild and strong shifts in its governing physical parameters to emulate changes in physical conditions during deployment. The 12 target-shift cells contain 288 physical trajectories and 576 model records.

For each benchmark, we construct the Development and Evaluation sets with disjoint physical trajectories. Evaluation trajectories are processed sequentially. At each decision, four future blocks are predicted, and the corresponding observations become available after a declared delay.

Compared models. We compare BROC with an unadapted surrogate (**Frozen**) as the fundamental reference and ten adaptive baselines. We first include **Always-8**, a persistent control that performs eight gradient-adaptation steps at every update. Other baselines include classical methods: **CMA** (Welford, 1962), **EWMA** (Roberts, 1959), **RLS** (Plackett, 1950), **EnKF** (Evensen, 1994), and **DMD** (Schmid, 2010); and recent methods: **LoRA** (Hu et al., 2022), **COSA** (Im & Kwon, 2026), **ADAPT-Z** (Huang et al., 2026), and **TAFAS** (Kim et al., 2025).

All methods follow the deployment contract and are evaluated on the same trajectories and scored decisions. Each adaptive method issues its forecast before updating from matured observations, and its state resets between physical trajectories. We tune the baselines on the Development set and lock the best global configuration for each method before Evaluation. BROC is not tuned and uses its default hyperparameters $(\lambda, \gamma) = (10^{-2}, 1)$ throughout Evaluation. Appendix B.1 details the benchmark partitions and tuning information.

Other Protocols. All experiments are run on NVIDIA L20 GPUs, and latency is measured on a single idle device. For stochastic methods, results from three seeds are averaged within each physical trajectory. Random sensing interventions are likewise averaged over three seeds within each trajectory. Throughout the paper, mean squared error (MSE) follows a macro-averaging protocol. We define $M_{m,c}$ as the MSE of method m in cell c. Scored decisions are first averaged within each trajectory, and all trajectories are then weighted equally within the cell. We define $M_{m,s}$ as the mean of $M_{m,c}$ across all cells belonging to system s, with all cells receiving equal weight. MSE scales differ across physical systems. We therefore use normalized ratios for method comparisons across each benchmark. To calculate the MSE reduction of method m relative to comparator b, we divide $M_{m,s}$ by $M_{b,s}$ within each system and average the resulting ratios equally across the benchmark. One minus this average gives the MSE reduction.

Online latency measures one matured-observation update followed by four future corrections on cached forecasts. TAFAS is measured separately on its input-adaptation path, including the required surrogate evaluations. For each cell, we select two physical trajectories and time their first, middle, and last selected states. After five warm-up repetitions, we record 20 timed repetitions per state. We take medians successively over repetitions, states, and trajectories, then take the median across cells for the benchmark result. The timed region includes all computations affecting the method output, with device synchronization immediately before and after each repetition. All reported 95% confi-

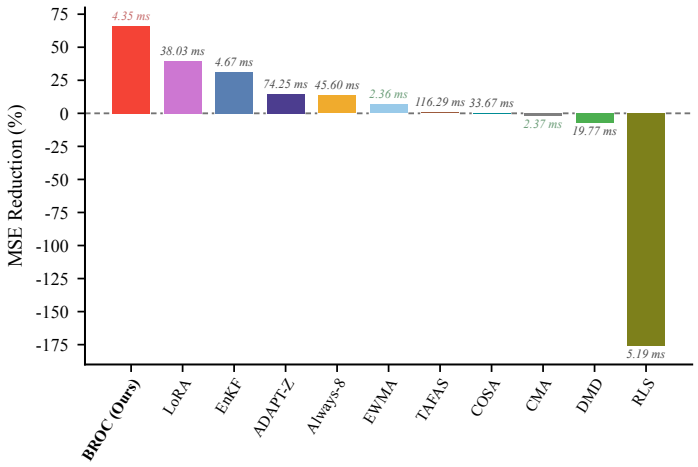


Figure 3: Accuracy and latency comparison on PDEBench; higher reduction and lower latency are better.

dence intervals use 10,000 paired bootstrap replicates with physical trajectories as the independent resampling units.

5.2 Accuracy and Efficiency

RealPDEBench. Figure 1 compares accuracy and online cost on RealPDEBench. The vertical coordinate reports the reduction in MSE relative to Frozen, and the horizontal coordinate reports online latency. BROC is the most accurate method across all baselines. It reduces MSE by 32.56% and 30.31% relative to Frozen and Always-8, respectively. Among the classical baselines, RLS achieves the lowest MSE, which BROC reduces by 25.90%. Among the recent baselines, ADAPT-Z achieves the lowest MSE, which BROC reduces by 23.84%.

BROC requires 5.317 ms on RealPDEBench. Two classical residual-estimation baselines are faster: EWMA and CMA require 2.801 and 2.810 ms, respectively, but their MSEs are 40.69% and 40.61% higher than BROC's. Always-8 requires 47.186 ms. Among the recent baselines, LoRA is the fastest at 26.675 ms, but it remains 5.02× slower than BROC and incurs a 33.68% higher MSE. ADAPT-Z is the most accurate recent baseline, yet it requires 11.97× the latency of BROC and incurs a 31.32% higher MSE.

PDEBench. Figure 3 reports the accuracy–latency analysis on PDEBench. The trend is consistent with RealPDEBench. BROC reduces Frozen MSE by 65.91% and requires 4.348 ms on this benchmark. In comparison, the most accurate baseline, LoRA, reduces Frozen MSE by 39.60%, while BROC is 8.75× faster. EWMA is the fastest baseline at 2.359 ms, while BROC reduces its MSE by 38.39%.

Overall, BROC delivers the strongest overall combination of predictive accuracy and online efficiency among the compared methods. The complete MSE matrices and latency measurements for both benchmarks are provided in Appendix B.2.

5.3 Sparse Sensing, Dynamic Sensing, and Noise

Sparse sensing is evaluated from the second decision onward because the first decision reflects one-block initialization rather than recursive sparse-to-field transfer. Dynamic sensing and noise experiments retain all decisions because their interventions begin with the initial update and are evaluated over the complete deployment. For RealPDEBench, this section uses a balanced Development panel of eight numerical-checkpoint cells, with two cells per system.

Sparse Sensing. We evaluate BROC using dense observations and uniform or random probes at 5% and 1% spatial coverage. Table 1 reports the full-field MSE ratios relative to Frozen. Values below one favor BROC.

Table 1: Sparse sensing on RealPDEBench and PDEBench; lower MSE ratios are better.

Benchmark	Dense	Uniform 5%	Random 5%	Uniform 1%	Random 1%
RealPDEBench	.628	.631	.635	.652	.670
PDEBench	.218	.220	.237	.272	.409

On RealPDEBench, the ratio increases only from .628 under dense sensing to .670 under random 1% sensing. PDEBench is more sensitive to probe placement, with the ratio reaching .409 under random 1% coverage. Across both benchmarks, every full-field ratio remains below one, supporting transfer from sparse observations to the full field. Complete results are provided in Appendix B.3.

Dynamic Sensing. In practical sparse deployments, sensor configurations may not remain static: physical probes may move between measurement intervals, and individual readings may be lost. To assess BROC's robustness to such changes, we consider two scenarios based on a random 1% sensor configuration. To simulate sensor mobility, we resample the probe locations at every measurement interval. To simulate sensor failure, we randomly drop 25% or 50% of the stationary probes. Figure 4 reports the MSE of each condition relative to the corresponding static baseline. Ratios below one indicate improvement.

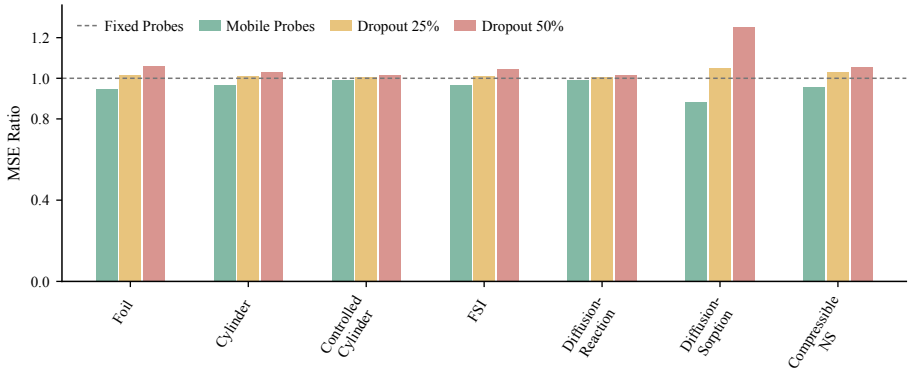


Figure 4: Dynamic sensing on RealPDEBench and PDEBench; lower MSE ratios are better.

Mobile sensors improve upon the static baseline in every system across both benchmarks, with MSE ratios ranging from 0.884 to 0.993. This result indicates that BROC does not require the same spatial locations to remain continuously observable. Dropping 25% of the probes causes only a marginal error increase, with ratios ranging from 1.005 to 1.051. Under 50% dropout, the MSE of Diffusion–Sorption increases by 25.1%, while those of the other systems remain within 6.2% of their corresponding static-baseline values.

Noise. We evaluate BROC's robustness to measurement and surrogate-input noise. For **measurement noise**, we perturb matured target observations at a fixed random 1% probe mask using Gaussian noise with $\sigma \in \{.05, .10\}$, variance-normalized Student-t_3 noise with $\sigma = .05$, and Gaussian$^* = .99\mathcal{N}(0, .05^2) + .01\mathcal{N}(0, .5^2)$. Across these conditions, the largest MSE increase is only .336% on RealPDEBench and 7.826% on PDEBench.

For **surrogate-input noise**, we add Gaussian noise with $\sigma \in \{.05, .10\}$ to the dynamic input variables before each frozen-surrogate evaluation. The largest MSE increases are 4.009% on RealPDEBench and 7.992% on PDEBench.

At the highest tested Gaussian noise level, $\sigma = .10$, BROC remains better than the corresponding Frozen in all systems, reducing MSE by 25.8–77.8% under measurement noise and 29.6–78.9% under surrogate-input noise. Appendix B.4 reports the complete results.

5.4 Delayed observations, gating, and autoregressive rollout

Delayed observations. We evaluate BROC on RealPDEBench with observation delays $d \in \{1, 2, 4, 8\}$ under dense and uniform 1% sensing. Its 32 sequential decision blocks support the complete delay range. PDEBench is excluded because Diffusion–Reaction and Diffusion–Sorption

contain only eight decision blocks, while Compressible NS contains seven, preventing a common $d = 8$ evaluation. For a matched comparison, all delay settings are scored on the same forecast decisions after the first eight decisions. Table 2 reports the BROC/Frozen MSE ratios. Ratios below one favor BROC.

Table 2: Delayed observation on RealPDEBench; lower MSE ratios are better.

Observation	$d = 1$	$d = 2$	$d = 4$	$d = 8$
Dense	.668 [.655, .677]	.674 [.662, .683]	.681 [.669, .689]	.697 [.686, .705]
Uniform 1%	.701 [.677, .722]	.708 [.685, .730]	.718 [.695, .740]	.752 [.724, .781]

The error increases gradually with observation delay. The degradation remains moderate, and BROC remains below Frozen under both sensing conditions at every evaluated delay.

Gating. We validate the protective mechanism of the gate on the Diffusion–Sorption system, as its scalar diffusivity permits a continuous trajectory with a reversible schedule: $D = 5 \times 10^{-4} \rightarrow 7.5 \times 10^{-4} \rightarrow 5 \times 10^{-4}$. Direct BROC is more accurate than gated BROC during the initial and shifted regimes. However, after the system returns to the original regime, the gate suppresses stale corrections and reduces MSE relative to direct BROC by 8.3% under dense sensing and 7.4% under uniform 1% sensing.

Autoregressive evaluation. We evaluate autoregressive deployment after all observations become unavailable. Table 3 reports the BROC/Frozen MSE ratios. At the longest available horizon, BROC reduces the Frozen autoregressive MSE by 55.00% on RealPDEBench and 51.89% on PDEBench. We detail the complete results in Appendix B.5.

Table 3: Autoregressive evaluation on RealPDEBench and PDEBench; lower MSE ratios are better.

Benchmark	$h = 2$	$h = 4$	Longest available
RealPDEBench	.738 [.708, .767]	.726 [.675, .929]	.450 [.429, .533]
PDEBench	.410 [.399, .423]	.408 [.393, .425]	.481 [.440, .537]

5.5 Physical Fidelity

We evaluate the physical fidelity of BROC. The available physical information differs between the two benchmarks. We therefore evaluate the diagnostics supported by each. RealPDEBench provides measured velocity fields but not the complete governing-equation states. We compare the predicted and measured velocity-divergence fields. Compared with Frozen, BROC reduces this discrepancy across all four systems by 23.10–47.60%.

PDEBench provides complete state variables and differential operators. We evaluate a broader set of diagnostics, including PDE residuals, global-property errors, boundary-condition and physical-validity violations, and system-specific discrepancies. BROC improves all diagnostics, with relative reductions of 20.05–50.79%, 41.11–87.48%, 15.22–87.33%, and 27.00–79.10%, respectively. Appendix B.6 details the diagnostics and reports the corresponding results.

5.6 Mechanism Evidence

We examine two empirical properties that support the central mechanism of BROC: compact residual structure and temporal persistence. Let $R^{\star}_{s,\mathrm{full}} \in \mathbb{R}^{N_s \times q}$ denote the matrix formed by stacking the clean residual vectors of forecast block s across all spatial locations. Let $\sigma_{s,1} \geq \sigma_{s,2} \geq \cdots$ denote its singular values. The cumulative energy captured by the leading r modes is

$$\mathrm{CE}_s(r) = \frac{\sum_{i=1}^{r} \sigma_{s,i}^2}{\sum_i \sigma_{s,i}^2}. \tag{10}$$

We report the median of $\mathrm{CE}_s(r)$ across clean decision states. This diagnostic measures whether residual variation is concentrated in a small number of shared directions across future blocks and physical variables.

Temporal persistence compares the fit gain on a matured observation block with the full-field gain on subsequent forecasts:

$$\mathcal{G}_s^{\mathrm{fit}} = 1 - \frac{\ell_{s,\mathrm{obs}}^{\mathrm{after}}}{\ell_{s,\mathrm{obs}}^{\mathrm{before}}}, \qquad \mathcal{G}_s^{\mathrm{future}} = 1 - \frac{\ell_{s,\mathrm{future}}^{\mathrm{BROC}}}{\ell_{s,\mathrm{future}}^{\mathrm{Frozen}}}. \tag{11}$$

We first average the gains across decisions within each model record. We then center the resulting gains within each cell and compute their pooled Spearman correlation. The positive-gain rate denotes the percentage of records with $\mathcal{G}_s^{\mathrm{future}} > 0$.

On RealPDEBench, the leading 1, 2, and 4 modes capture 54.88%, 85.96%, and 97.18% of the residual energy. The corresponding values on PDEBench are 91.22%, 99.09%, and 99.93% for the leading 1, 2, and 3 modes ($r = 3$ is the largest rank available for all PDEBench systems). The correlations between matured-block fit gain and future gain are .980 on RealPDEBench and .911 on PDEBench. Positive future gains occur in 98.98% and 96.35% of the corresponding model records. These results show that residual variation is concentrated in a small number of shared modes. Information obtained from matured observations remains useful for subsequent forecasts. Appendix B.7 reports the complete results.

6 Discussion

BROC is designed for the scenario in which the predicted and observed variables of a frozen surrogate are physically aligned, and the residual relationship is sufficiently consistent across observed and unobserved locations. The shared residual relationship enables sparse measurements to correct the full field. Its effectiveness under highly localized residual patterns remains to be evaluated. Because BROC corrects surrogate outputs rather than enforcing the governing equations, improvements in physical diagnostics do not constitute an exact conservation guarantee. The present evaluation covers seven physical systems and four surrogate families. Evaluation under other observation operators remains future work.

7 Conclusion

This work develops BROC, a spatially shared block-recursive residual observer for online correction of frozen neural PDE surrogates without neural-parameter modification. Formal results establish exact recursive optimality and temporal non-anticipation. A finite-sample certificate characterizes sufficient conditions for full-field improvement. Across seven physical systems comprising 32 evaluation cells, BROC attains the highest aggregate accuracy at substantially lower online cost than iterative adaptation methods. Mechanism analyses identify compact residual modes with strong temporal persistence. Performance remains robust under sparse, dynamic, delayed, and noisy sensing, retaining gains during autoregressive rollout. Protective gating suppresses stale corrections after regime recovery. Physical audits show that BROC also improves physical fidelity.

AI use statement

In this work, we used generative AI tools to assist with manuscript editing and language refinement.

References

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.

Alban Farchi, Marc Bocquet, Patrick Laloyaux, Massimo Bonavita, and Quentin Malartic. A comparison of combined data assimilation and machine learning methods for offline and online model error correction. *Journal of Computational Science*, 55:101468, 2021. doi: 10.1016/j.jocs.2021.101468.

David Gross and Vincent Nesme. Note on sampling without replacing from a finite collection of matrices. *arXiv preprint arXiv:1001.2738*, 2010. URL https://arxiv.org/abs/1001.2738.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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.

Xiannan Huang, Shuhan Qiu, Jiayuan Du, and Chao Yang. Online time series prediction using feature adjustment. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=s4U2FWEMTU.

Jeonghwan Im and Hyuk-Yoon Kwon. COSA: Context-aware output-space adapter for test-time adaptation in time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=L7Z5wBMPrW.

Armand Kassaï Koupaï, Jorge Mifsut Benet, Yuan Yin, Jean-Noël Vittaut, and Patrick Gallinari. GEPS: Boosting generalization in parametric PDE neural solvers through adaptive conditioning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 70659–70692, 2024. doi: 10.52202/079017-2258. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/82844e428d9163a9f94830dc03af4f9c-Paper-Conference.pdf.

HyunGi Kim, Siwon Kim, Jisoo Mok, and Sungroh Yoon. Battling the non-stationarity in time series forecasting via test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17868–17876, 2025. doi: 10.1609/aaai.v39i17.33965.

Matthieu Kirchmeyer, Yuan Yin, Jérémie Donà, Nicolas Baskiotis, Alain Rakotomamonjy, and Patrick Gallinari. Generalizing to new physical systems via context-informed dynamics model. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11283–11301, 2022. URL https://proceedings.mlr.press/v162/kirchmeyer22a.html.

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

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.

Heitor R. Medeiros, Hossein Sharifi-Noghabi, Gabriel L. Oliveira, and Saghar Irandoust. Accurate parameter-efficient test-time adaptation for time series forecasting. *arXiv preprint arXiv:2506.23424*, 2025. URL https://arxiv.org/abs/2506.23424.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905, 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-time model adaptation with only forward passes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 38298–38315, 2024. URL https://proceedings.mlr.press/v235/niu24a.html.

Robin L. Plackett. Some theorems in least squares. *Biometrika*, 37(1/2):149–157, 1950. doi: 10.2307/2332158.

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.

Bogdan Raonic, 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, 2023. doi: 10.52202/075280-3376.

S. W. Roberts. Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250, 1959. doi: 10.1080/00401706.1959.10489860.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Peter J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010. doi: 10.1017/S0022112010001217.

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

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Dan MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2022. doi: 10.52202/068431-0117.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012. doi: 10.1007/s10208-011-9099-z.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Haochun Wang, Ruichen Xu, Georgios Kementzidis, Karen Cho, Sebastian Ramirez Villarreal, and Yuefan Deng. Towards principled test-time adaptation for time series forecasting. *arXiv preprint arXiv:2605.17250*, 2026. URL https://arxiv.org/abs/2605.17250.

B. P. Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420, 1962. doi: 10.1080/00401706.1962.10490022.

Matthew O. Williams, Ioannis G. Kevrekidis, and Clarence W. Rowley. A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015. doi: 10.1007/s00332-015-9258-5.

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, 2024. URL https://proceedings.mlr.press/v235/wu24r.html.

A Theoretical Details

A.1 Notation and deployment protocol

We use three indices. The symbol s identifies an issued forecast block, t identifies the current decision, and k counts committed updates. Let $\mathrm{arr}(s)$ be the decision after whose forecast the complete observation for block s becomes available. The blocks available before forecast t form

$$\mathcal{M}_t = \{s : \mathrm{arr}(s) < t\}, \qquad k(t) = |\mathcal{M}_t|. \tag{12}$$

Let s_i be the block incorporated by the i-th committed update, and let $i(s)$ be its update index. This notation follows the actual update order when delays vary or several observations arrive together.

The frozen-surrogate input X_s^{base} contains the information available when block s is issued. The tensor Y_s is the corresponding ground-truth block. Let S_s denote the observed locations and $m_s = |S_s|$. At location j, $z_{s,j}$, $y_{s,j}$, and $y_{s,j}^{\mathrm{obs}}$ denote the vectorized frozen forecast, clean target, and observation. The matrix Φ_s stacks $\phi(z_{s,j})$. The matrices $R_s^\star$ and R_s^{obs} stack the clean and observed residuals. Dense sensing is the special case $m_s = N_s$.

Partitioning the correction matrix as

$$W_k = \begin{bmatrix} b_k^\top \\ A_k \end{bmatrix}, \qquad b_k \in \mathbb{R}^q, \quad A_k \in \mathbb{R}^{q\times q},$$

gives $\widetilde{z} = b_k + (I_q + A_k^\top)z$. The vector b_k supplies the additive correction, and A_k couples the T forecast time steps and C_{corr} selected variables. Let P_{rest} select the variables outside P_{corr}, and let $\Delta(\cdot; W_k)$ reshape the local corrections into a full-field tensor. At decision t, the intervention is

$$\widetilde{Y}_t = \widehat{Y}_t^0 + P_{\mathrm{corr}}^\top \Delta(P_{\mathrm{corr}}\widehat{Y}_t^0; W_{k(t)}), \qquad P_{\mathrm{rest}}\widetilde{Y}_t = P_{\mathrm{rest}}\widehat{Y}_t^0. \tag{13}$$

Gate selection does not interrupt updates to the BROC state. Before any block matures, the gate selects Frozen; simultaneous arrivals are resolved by the largest block index.

Assumption 1. The observation for block s is declared mature only after its complete target block is available. Forecast t uses only the state committed from blocks in $\mathcal{M}_t$. Autoregressive evaluation reads no targets or observations after the rollout origin.

Assumption 2. BROC models residuals only for variables selected by P_{corr}. All remaining variables retain their frozen predictions.

Assumption 3. Observed and unobserved locations share the local clean feature–residual relation represented by W_k. All risks below have finite second moments.

Assumption 4. Measurement noise perturbs matured targets while holding the cached frozen forecasts and their features fixed. Thus $R_s^{\mathrm{obs}} = R_s^\star + E_s^\varepsilon$.

A.2 Supporting results

Theorem 1. *For every $k \geq 1$, Equation 5 gives the unique correction matrix solving the objective in Equation 6.*

Theorem 2. *Define the information before and after decision t as*

$$\mathscr{H}_t^- = \sigma\Big(\{X_\tau^{\mathrm{base}}, \widehat{Y}_\tau^0 : \tau \leq t\}, \{(\mathrm{arr}(s), i(s), m_s, S_s, Y_s^{\mathrm{obs}}) : \mathrm{arr}(s) < t\}\Big), \tag{14}$$

$$\mathscr{H}_t^+ = \sigma\big(\mathscr{H}_t^-, \{(\mathrm{arr}(s), i(s), m_s, S_s, Y_s^{\mathrm{obs}}) : \mathrm{arr}(s) = t\}\big). \tag{15}$$

Then $W_{k(t)}$ and the forecast issued at decision t are $\mathscr{H}_t^-$-measurable. The state obtained from observations arriving at t is $\mathscr{H}_t^+$-measurable and enters prediction from the next decision onward.

Theorem 3. *Suppose $R_{s_i}^{\mathrm{obs}} = \Phi_{s_i} W_\star + E_{s_i}$. Define*

$$D_k = \sum_{i=1}^{k} \gamma^{k-i} m_{s_i}^{-1} \Phi_{s_i}^\top E_{s_i}, \qquad \mu_k = \lambda_{\min}(G_k).$$

Then

$$(G_k + \lambda I_d)(W_k - W_\star) = D_k - \lambda W_\star, \qquad \|W_k - W_\star\|_{\mathrm{F}} \leq \frac{\|D_k\|_{\mathrm{F}} + \lambda\|W_\star\|_{\mathrm{F}}}{\mu_k + \lambda}. \tag{16}$$

Theorem 4. *Let $\varphi = \phi(z)$, $\Sigma = \mathbb{E}[\varphi\varphi^\top]$, and $K = \mathbb{E}[\varphi(r^\star)^\top]$. Define $W_\lambda = (\Sigma + \lambda I_d)^{-1}K$ and $\mathcal{R}(W) = \mathbb{E}\|r^\star - W^\top\varphi\|_2^2$. Then*

$$\mathcal{R}(0) - \mathcal{R}(W_\lambda) = \mathrm{tr}(W_\lambda^\top \Sigma W_\lambda) + 2\lambda\|W_\lambda\|_{\mathrm{F}}^2 \geq 0. \tag{17}$$

For a future distribution with $\Sigma' = \Sigma + \Delta_\Sigma$ and $K' = K + \Delta_K$, let $\mathcal{R}'(W)$ denote the corresponding risk. Then

$$\mathcal{R}'(0) - \mathcal{R}'(W_\lambda) \geq \mathcal{R}(0) - \mathcal{R}(W_\lambda) - \|\Delta_\Sigma\|_2\|W_\lambda\|_{\mathrm{F}}^2 - 2\|\Delta_K\|_{\mathrm{F}}\|W_\lambda\|_{\mathrm{F}}. \tag{18}$$

Theorem 5. *At a fixed update, let $W^F = (G^F + \lambda I_d)^{-1}C^F$ use the full clean spatial field and $\widehat{W} = (\widehat{G} + \lambda I_d)^{-1}\widehat{C}$ use the sampled clean locations. Then*

$$\|\widehat{W} - W^F\|_{\mathrm{F}} \leq \frac{\|\widehat{C} - C^F\|_{\mathrm{F}}}{\lambda} + \frac{\|\widehat{G} - G^F\|_2\|C^F\|_{\mathrm{F}}}{\lambda^2}. \tag{19}$$

For any full-field feature vector φ, $\|(\widehat{W} - W^F)^\top\varphi\|_2 \leq \|\varphi\|_2\|\widehat{W} - W^F\|_{\mathrm{F}}$.

Theorem 6. *Suppose $R_{s_i}^{\mathrm{obs}} = R_{s_i}^\star + E_{s_i}^\varepsilon$ while Φ_{s_i} remains fixed. Define*

$$D_k^\varepsilon = \sum_{i=1}^{k} \gamma^{k-i} m_{s_i}^{-1} \Phi_{s_i}^\top E_{s_i}^\varepsilon.$$

Then

$$\|W_k^{\mathrm{obs}} - W_k^\star\|_{\mathrm{F}} \leq \frac{\|D_k^\varepsilon\|_{\mathrm{F}}}{\lambda + \lambda_{\min}(G_k)} \leq \frac{\|D_k^\varepsilon\|_{\mathrm{F}}}{\lambda}. \tag{20}$$

For the full-field result, fix a decision t with $k(t) \geq 1$. Each matured block $s \in \mathcal{M}_t$ has weight

$$\alpha_{t,s} = \gamma^{k(t)-i(s)}, \qquad A_t = \sum_{s\in\mathcal{M}_t} \alpha_{t,s}, \qquad \rho_t = \lambda/A_t. \tag{21}$$

The exponent $k(t) - i(s)$ records the number of committed updates since block s entered the state. For $\varphi_{s,j} = \phi(z_{s,j})$ and the complete spatial set Ω_s, define the clean full-field moments

$$\begin{aligned} \Sigma_t &= A_t^{-1} \sum_{s\in\mathcal{M}_t} \alpha_{t,s}\mathbb{E}_{j\sim\mathrm{Unif}(\Omega_s)}[\varphi_{s,j}\varphi_{s,j}^\top], \\ K_t &= A_t^{-1} \sum_{s\in\mathcal{M}_t} \alpha_{t,s}\mathbb{E}_{j\sim\mathrm{Unif}(\Omega_s)}[\varphi_{s,j}(r_{s,j}^\star)^\top]. \end{aligned} \tag{22}$$

The full-field correction is $W_t^F = (\Sigma_t + \rho_t I_d)^{-1}K_t$. The actual sparse moments are

$$\widehat{\Sigma}_t = A_t^{-1} \sum_{s\in\mathcal{M}_t} \frac{\alpha_{t,s}}{m_s} \sum_{j\in S_s} \varphi_{s,j}\varphi_{s,j}^\top, \tag{23}$$

$$\widehat{K}_t^{\mathrm{obs}} = A_t^{-1} \sum_{s\in\mathcal{M}_t} \frac{\alpha_{t,s}}{m_s} \sum_{j\in S_s} \varphi_{s,j}(r_{s,j}^{\mathrm{obs}})^\top. \tag{24}$$

Write

$$\widehat{\Sigma}_t = \Sigma_t + \Delta_t^\Sigma, \qquad \widehat{K}_t^{\mathrm{obs}} = K_t + \Delta_t^K + \Xi_t, \tag{25}$$

where Δ_t^Σ and Δ_t^K are the clean spatial-sampling deviations. The measurement-noise term is

$$\Xi_t = A_t^{-1} \sum_{s\in\mathcal{M}_t} \frac{\alpha_{t,s}}{m_s} \sum_{j\in S_s} \varphi_{s,j}\varepsilon_{s,j}^\top. \tag{26}$$

Dividing the accumulated statistics in Equation 4 by A_t gives the exact algorithmic state

$$\widehat{W}_t = W_{k(t)} = (\widehat{\Sigma}_t + \rho_t I_d)^{-1}\widehat{K}_t^{\mathrm{obs}}. \tag{27}$$

At future horizon h, let the complete spatial field induce

$$\Sigma_{t,h}^+ = \mathbb{E}[\varphi\varphi^\top], \qquad K_{t,h}^+ = \mathbb{E}[\varphi(r^\star)^\top], \qquad \mathcal{R}_{t,h}^+(W) = \mathbb{E}\|r^\star - W^\top\varphi\|_2^2.$$

Define

$$\begin{aligned}
\mathsf{S}_t &= \mathrm{tr}((W_t^F)^\top \Sigma_t W_t^F) + 2\rho_t \|W_t^F\|_{\mathrm{F}}^2, \\
\mu_t^F &= \rho_t + \lambda_{\min}(\Sigma_t), & \widehat{\mu}_t &= \rho_t + \lambda_{\min}(\widehat{\Sigma}_t), \\
\eta_t &= \frac{\|\Delta_t^K\|_{\mathrm{F}} + \|\Xi_t\|_{\mathrm{F}}}{\widehat{\mu}_t} + \frac{\|\Delta_t^\Sigma\|_2 \|K_t\|_{\mathrm{F}}}{\widehat{\mu}_t \mu_t^F}, \\
d_{K,t,h} &= \|K_{t,h}^+ - K_t\|_{\mathrm{F}}, & d_{\Sigma,t,h} &= \|\Sigma_{t,h}^+ - \Sigma_t\|_2, \\
\zeta_{t,h} &= \|K_{t,h}^+ - \Sigma_{t,h}^+ W_t^F\|_{\mathrm{F}}.
\end{aligned} \tag{28}$$

Theorem 7. *For any decision t with $k(t) \geq 1$ and any future horizon h,*

$$\begin{aligned}
\mathcal{R}_{t,h}^+(0) - \mathcal{R}_{t,h}^+(\widehat{W}_t) \geq \mathsf{S}_t - 2\|W_t^F\|_{\mathrm{F}} d_{K,t,h} - \|W_t^F\|_{\mathrm{F}}^2 d_{\Sigma,t,h} \\
- 2\eta_t \zeta_{t,h} - \|\Sigma_{t,h}^+\|_2 \eta_t^2.
\end{aligned} \tag{29}$$

A positive right-hand side guarantees lower future clean full-field risk than Frozen for the variables selected by P_{corr}.

Corollary 1. *Let $\Sigma^{(s)}$ and $K^{(s)}$ denote the clean full-field moments of block s. Suppose*

$$\|\Sigma_{t,h}^+ - \Sigma^{(s)}\|_2 \leq L_\Sigma(t+h-s), \qquad \|K_{t,h}^+ - K^{(s)}\|_{\mathrm{F}} \leq L_K(t+h-s)$$

for every $s \in \mathcal{M}_t$. Then

$$d_{\Sigma,t,h} \leq L_\Sigma \bar{a}_{t,h}, \qquad d_{K,t,h} \leq L_K \bar{a}_{t,h}, \qquad \bar{a}_{t,h} = A_t^{-1} \sum_{s \in \mathcal{M}_t} \alpha_{t,s}(t+h-s). \tag{30}$$

A.3 Recursive update and temporal validity

Proof of Theorem 1. Expanding Equation 4 from $G_0 = C_0 = 0$ gives

$$G_k = \sum_{i=1}^k \gamma^{k-i} m_{s_i}^{-1} \Phi_{s_i}^\top \Phi_{s_i}, \qquad C_k = \sum_{i=1}^k \gamma^{k-i} m_{s_i}^{-1} \Phi_{s_i}^\top R_{s_i}^{\mathrm{obs}}. \tag{31}$$

Terms in Equation 6 that depend on W are

$$\mathcal{J}_k(W) = \mathrm{tr}(W^\top G_k W) - 2\,\mathrm{tr}(W^\top C_k) + \lambda\,\mathrm{tr}(W^\top W) + \mathrm{const}. \tag{32}$$

Its gradient is

$$\nabla_W \mathcal{J}_k(W) = 2(G_k + \lambda I_d)W - 2C_k.$$

Since $G_k \succeq 0$ and $\lambda > 0$, $G_k + \lambda I_d \succ 0$; hence the objective is strictly convex with the unique solution

$$W_k = (G_k + \lambda I_d)^{-1} C_k.$$

□

Proof of Theorem 2. The initial state is deterministic. For every $s \in \mathcal{M}_t$, $\mathrm{arr}(s) < t$, so Φ_s and R_s^{obs} are measurable with respect to $\mathscr{H}_t^-$. Repeated application of Equations 4 and 5 makes $W_{k(t)}$ $\mathscr{H}_t^-$-measurable, and Equation 7 makes the forecast issued at t $\mathscr{H}_t^-$-measurable. Observations with $\mathrm{arr}(s) = t$ belong to $\mathscr{H}_t^+$ and enter the state after forecast t. □

Proposition 1. *For a fixed issued frozen forecast $\widehat{Y}_t^0$, the full-field correction in Equation 13 satisfies*

$$P_{\mathrm{rest}} \widetilde{Y}_t = P_{\mathrm{rest}} \widehat{Y}_t^0.$$

Proof. The selected variable sets are disjoint, so $P_{\mathrm{rest}} P_{\mathrm{corr}}^\top = 0$. Applying P_{rest} to Equation 13 removes the correction term. □

A.4 HISTORICAL AND FUTURE ERROR

Proof of Theorem 3. Substituting $R_{s_i}^{\text{obs}} = \Phi_{s_i} W_\star + E_{s_i}$ into C_k gives

$$C_k = \sum_{i=1}^{k} \gamma^{k-i} m_{s_i}^{-1} \Phi_{s_i}^\top (\Phi_{s_i} W_\star + E_{s_i}) \tag{33}$$

$$= G_k W_\star + D_k. \tag{34}$$

Equation 5 gives $(G_k + \lambda I_d) W_k = C_k$. Subtracting $(G_k + \lambda I_d) W_\star = G_k W_\star + \lambda W_\star$ proves the identity in Equation 16. The smallest singular value of $G_k + \lambda I_d$ is $\mu_k + \lambda$. The inverse spectral norm and triangle inequality give the stated bound. □

Proposition 2. *Define*

$$\mathcal{L}_k(W) = \sum_{i=1}^{k} \gamma^{k-i} m_{s_i}^{-1} \|R_{s_i}^{\text{obs}} - \Phi_{s_i} W\|_{\text{F}}^2.$$

Then

$$\mathcal{L}_k(0) - \mathcal{L}_k(W_k) \geq \lambda \|W_k\|_{\text{F}}^2 \geq 0.$$

Proof. Theorem 1 gives

$$\mathcal{L}_k(W_k) + \lambda \|W_k\|_{\text{F}}^2 \leq \mathcal{L}_k(0).$$

□

Proof of Theorem 4. Expanding the risk gives

$$\mathcal{R}(0) - \mathcal{R}(W) = 2 \operatorname{tr}(W^\top K) - \operatorname{tr}(W^\top \Sigma W). \tag{35}$$

The relation $K = (\Sigma + \lambda I_d) W_\lambda$ yields Equation 17. For the future distribution,

$$\mathcal{R}'(0) - \mathcal{R}'(W_\lambda) = \mathcal{R}(0) - \mathcal{R}(W_\lambda) + 2 \operatorname{tr}(W_\lambda^\top \Delta_K) - \operatorname{tr}(W_\lambda^\top \Delta_\Sigma W_\lambda). \tag{36}$$

The Frobenius trace inequality bounds the second term from below by $-2\|W_\lambda\|_{\text{F}} \|\Delta_K\|_{\text{F}}$. The spectral trace inequality bounds the third term from below by $-\|\Delta_\Sigma\|_2 \|W_\lambda\|_{\text{F}}^2$. □

A.5 FULL-FIELD IMPROVEMENT CERTIFICATE

Proof of Theorem 7. Let $H_t = \Sigma_t + \rho_t I_d$ and $\widehat{H}_t = \widehat{\Sigma}_t + \rho_t I_d$. Since $\Sigma_t, \widehat{\Sigma}_t \succeq 0$ and $\rho_t > 0$, both matrices are invertible. The resolvent identity gives

$$\widehat{W}_t - W_t^F = \widehat{H}_t^{-1} (\Delta_t^K + \Xi_t) + \widehat{H}_t^{-1} (\Sigma_t - \widehat{\Sigma}_t) H_t^{-1} K_t. \tag{37}$$

Since $\|H_t^{-1}\|_2 = 1/\mu_t^F$ and $\|\widehat{H}_t^{-1}\|_2 = 1/\widehat{\mu}_t$,

$$\|\widehat{W}_t - W_t^F\|_{\text{F}} \leq \eta_t. \tag{38}$$

For any W, the future risk satisfies

$$\mathcal{R}_{t,h}^+(0) - \mathcal{R}_{t,h}^+(W) = 2 \operatorname{tr}(W^\top K_{t,h}^+) - \operatorname{tr}(W^\top \Sigma_{t,h}^+ W). \tag{39}$$

Set $E_t = \widehat{W}_t - W_t^F$. Substituting $\widehat{W}_t = W_t^F + E_t$ and $K_t = (\Sigma_t + \rho_t I_d) W_t^F$ gives

$$\mathcal{R}_{t,h}^+(0) - \mathcal{R}_{t,h}^+(\widehat{W}_t) = \mathsf{S}_t + 2 \operatorname{tr}((W_t^F)^\top (K_{t,h}^+ - K_t)) - \operatorname{tr}((W_t^F)^\top (\Sigma_{t,h}^+ - \Sigma_t) W_t^F)$$
$$+ 2 \operatorname{tr}\Big(E_t^\top (K_{t,h}^+ - \Sigma_{t,h}^+ W_t^F)\Big) - \operatorname{tr}(E_t^\top \Sigma_{t,h}^+ E_t). \tag{40}$$

Applying the Frobenius and spectral trace inequalities to Equation 40, followed by $\|E_t\|_{\text{F}} \leq \eta_t$, yields Equation 29. □

Proof of Corollary 1. By the definitions of Σ_t and K_t and the triangle inequality,

$$\|\Sigma_{t,h}^{+} - \Sigma_t\|_2 \le A_t^{-1} \sum_{s \in \mathcal{M}_t} \alpha_{t,s} \|\Sigma_{t,h}^{+} - \Sigma^{(s)}\|_2 \le L_\Sigma \bar{a}_{t,h}, \tag{41}$$

$$\|K_{t,h}^{+} - K_t\|_{\mathrm{F}} \le A_t^{-1} \sum_{s \in \mathcal{M}_t} \alpha_{t,s} \|K_{t,h}^{+} - K^{(s)}\|_{\mathrm{F}} \le L_K \bar{a}_{t,h}. \tag{42}$$

□

Theorem 7 gives a deterministic bound in terms of the realized fields, sampling deviations, and measurement noise, covering both fixed and adaptive sensing layouts. Corollary 3 provides a joint probability bound under random sensing and conditionally sub-Gaussian noise.

Post-hoc certificate audit. We evaluated Equation 29 on 960 RealPDEBench states spanning all 20 cells, two trajectories per cell, three decision positions, two 1% sensing layouts, delays of one and eight blocks, and clean or Gaussian measurement noise with $\sigma = .05$. The computed lower bound did not exceed the realized improvement in any state and was strictly positive in 321/960 states, certifying lower future clean full-field risk than Frozen in those 321 states. The post-hoc audit was not used for routing, tuning, or model selection.

A.6 Numerical and autoregressive stability

Proposition 3. *For every update,*

$$\lambda_{\min}(G_k + \lambda I_d) \ge \lambda, \qquad \kappa_2(G_k + \lambda I_d) \le 1 + \frac{\lambda_{\max}(G_k)}{\lambda}, \qquad \|W_k\|_{\mathrm{F}} \le \frac{\|C_k\|_{\mathrm{F}}}{\lambda}.$$

Proof. The matrix G_k is positive semidefinite. Adding λI_d shifts every eigenvalue by λ, which gives the first two inequalities. The final inequality follows from $\|(G_k + \lambda I_d)^{-1}\|_2 \le 1/\lambda$ and Equation 5. □

Proposition 4. *Using the partition $W_k = [b_k^\top; A_k]$, define*

$$\mathcal{C}_k(z) = b_k + (I_q + A_k^\top) z.$$

Then

$$\|\mathcal{C}_k(z) - \mathcal{C}_k(z')\|_2 \le \|I_q + A_k^\top\|_2 \|z - z'\|_2.$$

Proof. The identity

$$\mathcal{C}_k(z) - \mathcal{C}_k(z') = (I_q + A_k^\top)(z - z')$$

and submultiplicativity of the induced 2-norm give the result. □

To extend this bound to the complete autoregressive state, let $\mathcal{C}_k^{\mathrm{full}}$ apply $\mathcal{C}_k$ to the selected variables and leave the remaining variables unchanged. Its linear part is

$$B_k = \mathrm{diag}(I_q + A_k^\top, I_{\mathrm{rest}}).$$

Let f be the frozen transition with Lipschitz constant L_f, and let H reconstruct the next surrogate input with Lipschitz constant L_H. On an invariant state set, the complete map $H \circ \mathcal{C}_k^{\mathrm{full}} \circ f$ is contractive under the sufficient condition

$$L_H L_f \|B_k\|_2 = L_H L_f \max\{\|I_q + A_k^\top\|_2, 1\} < 1. \tag{43}$$

A.7 Sparse-sensing identification

Regularization gives a unique correction matrix for every $m_s \ge 1$. Without regularization, a unique solution from one block requires

$$\mathrm{rank}(\Phi_s) = d = q + 1.$$

Across multiple blocks, the corresponding condition is $G_k \succ 0$, commonly referred to as persistent excitation.

Table 4: Feature dimensions under 1% spatial coverage.

System	Grid	T	q	d	Observed locations m_s	$m_s \geq d$?
Foil/Cylinder	64×128	20	40	41	82	Yes
Controlled Cylinder	64×128	10	20	21	82	Yes
FSI	64×64	20	40	41	41	Yes ($m_s = d$)

Proof of Theorem 5. Adding and subtracting $(\widehat{G} + \lambda I_d)^{-1} C^F$, followed by the resolvent identity, gives

$$\widehat{W} - W^F = (\widehat{G} + \lambda I_d)^{-1}(\widehat{C} - C^F) \tag{44}$$

$$+ \left[(\widehat{G} + \lambda I_d)^{-1} - (G^F + \lambda I_d)^{-1}\right] C^F, \tag{45}$$

$$(\widehat{G} + \lambda I_d)^{-1} - (G^F + \lambda I_d)^{-1} = (\widehat{G} + \lambda I_d)^{-1}(G^F - \widehat{G})(G^F + \lambda I_d)^{-1}. \tag{46}$$

Since $\widehat{G}, G^F \succeq 0$, both inverse spectral norms are at most $1/\lambda$. Submultiplicativity yields Equation 19. The full-field prediction bound follows from Cauchy–Schwarz. □

Corollary 2. *Fix decision t with $\mathcal{M}_t \neq \varnothing$ and condition on its complete clean fields and an exogenous arrival and update schedule, under which $\mathcal{M}_t$, $i(s)$, and m_s are fixed independently of sensing randomness and measurement noise. For every $s \in \mathcal{M}_t$, let S_s be a uniform sample without replacement of size m_s from Ω_s. Assume*

$$\|\varphi_{s,j}\|_2 \leq L_\varphi, \qquad \|r^\star_{s,j}\|_2 \leq L_r.$$

Conditional on the sampled locations and features, assume that the noise vectors are independent, mean zero, and σ-sub-Gaussian in every coordinate. Let $k_t = k(t)$, and define

$$g_s(\delta) = cL_\varphi^2 \left[\sqrt{\frac{\log(6dk_t/\delta)}{m_s}} + \frac{\log(6dk_t/\delta)}{m_s}\right], \tag{47}$$

$$c_s(\delta) = cL_\varphi L_r \sqrt{dq} \left[\sqrt{\frac{\log(6dqk_t/\delta)}{m_s}} + \frac{\log(6dqk_t/\delta)}{m_s}\right], \tag{48}$$

$$\epsilon_{\Sigma,t}(\delta) = A_t^{-1} \sum_{s \in \mathcal{M}_t} \alpha_{t,s} g_s(\delta), \tag{49}$$

$$\epsilon_{K,t}(\delta) = A_t^{-1} \sum_{s \in \mathcal{M}_t} \alpha_{t,s} c_s(\delta), \tag{50}$$

$$V_t = A_t^{-2} \sum_{s \in \mathcal{M}_t} \frac{\alpha_{t,s}^2}{m_s}, \tag{51}$$

$$\epsilon_{\Xi,t}(\delta) = \sigma L_\varphi \sqrt{2dqV_t \log(6dq/\delta)}, \tag{52}$$

where c is a universal constant. With probability at least $1 - \delta$, jointly over all matured blocks,

$$\|\Delta_t^\Sigma\|_2 \leq \epsilon_{\Sigma,t}(\delta), \qquad \|\Delta_t^K\|_{\mathrm{F}} \leq \epsilon_{K,t}(\delta), \qquad \|\Xi_t\|_{\mathrm{F}} \leq \epsilon_{\Xi,t}(\delta). \tag{53}$$

Sampling masks need not be independent across blocks.

Proof. For each block, matrix Bernstein bounds the centered feature matrices, while scalar Bernstein and a union bound over the dq entries bound the clean feature–residual products (Tropp, 2012). The corresponding bounds for sampling without replacement follow from the matrix comparison principle (Gross & Nesme, 2010). A union bound over the k_t matured blocks, followed by the weighted triangle inequality, gives $\epsilon_{\Sigma,t}$ and $\epsilon_{K,t}$.

The exogenous schedule fixes the weights $\alpha_{t,s}/A_t$ and sample sizes m_s with respect to measurement noise. Each entry of Ξ_t therefore has a sub-Gaussian variance proxy bounded by $\sigma^2 L_\varphi^2 V_t$. A union bound over its dq entries and $\|\Xi_t\|_{\mathrm{F}} \leq \sqrt{dq}\|\Xi_t\|_{\max}$ gives $\epsilon_{\Xi,t}$. A final union bound gives probability at least $1 - \delta$. □

Corollary 3. *Under the conditions of Corollary 2, define*

$$\underline{\mu}_t(\delta) = \rho_t + \max\{\lambda_{\min}(\Sigma_t) - \epsilon_{\Sigma,t}(\delta), 0\}, \tag{54}$$

$$\overline{\eta}_t(\delta) = \frac{\epsilon_{K,t}(\delta) + \epsilon_{\Xi,t}(\delta)}{\underline{\mu}_t(\delta)} + \frac{\epsilon_{\Sigma,t}(\delta) L_\varphi L_r}{\underline{\mu}_t(\delta)\mu_t^F}. \tag{55}$$

With probability at least $1-\delta$*,*

$$\begin{aligned}\mathcal{R}^+_{t,h}(0) - \mathcal{R}^+_{t,h}(\widehat{W}_t) \geq \mathsf{S}_t - 2\|W_t^F\|_{\mathrm{F}} d_{K,t,h} - \|W_t^F\|_{\mathrm{F}}^2 d_{\Sigma,t,h} \\ - 2\overline{\eta}_t(\delta)\zeta_{t,h} - \|\Sigma^+_{t,h}\|_2 \overline{\eta}_t(\delta)^2.\end{aligned} \tag{56}$$

A positive right-hand side therefore guarantees future clean full-field improvement with probability at least $1-\delta$*, conditional on the realized complete fields.*

Proof. On the event in Equation 53, Weyl's inequality gives $\widehat{\mu}_t \geq \underline{\mu}_t(\delta)$. The boundedness assumptions imply $\|K_t\|_{\mathrm{F}} \leq L_\varphi L_r$. Substituting these bounds into Equation 37 gives $\|\widehat{W}_t - W_t^F\|_{\mathrm{F}} \leq \overline{\eta}_t(\delta)$. The exact decomposition in Equation 40 then gives Equation 56. □

For noise outside the sub-Gaussian class, Theorem 7 applies with the realized $\|\Xi_t\|_{\mathrm{F}}$, while Corollary 3 requires a corresponding concentration bound.

A.8 Measurement-noise perturbation

Proof of Theorem 6. Measurement noise changes only the feature–residual statistic:

$$C_k^{\mathrm{obs}} = C_k^\star + D_k^\varepsilon.$$

Hence

$$W_k^{\mathrm{obs}} - W_k^\star = (G_k + \lambda I_d)^{-1} D_k^\varepsilon.$$

The inverse spectral norm is $1/(\lambda + \lambda_{\min}(G_k))$, which gives the first bound in Equation 20. The second follows from $\lambda_{\min}(G_k) \geq 0$. □

B Experimental Details

B.1 Benchmarks and Model Tuning

RealPDEBench. Table 5 summarizes the surrogate checkpoints and trajectory partitions used for each RealPDEBench system. *num.* denotes a numerical checkpoint, and *FT* denotes a fine-tuned checkpoint.

Table 5: RealPDEBench partitions.

System	Surrogate/Checkpoint	Physical Trajectories		Model Records	
		Development	Evaluation	Development	Evaluation
Foil	CNO/num.	10	10	10	10
	CNO/FT			10	10
	Transolver/num.			10	10
	Transolver/FT			10	10
Cylinder	FNO/num.	25	26	25	26
	U-Net/num.			25	26
	Transolver/num.			25	26
	Transolver/FT			25	26
Controlled Cylinder	FNO/num.	27	28	27	28
	U-Net/num.			27	28
	CNO/num.			27	28
	CNO/FT			27	28
	Transolver/num.			27	28
	Transolver/FT			27	28
FSI	FNO/num.	6	6	6	6
	U-Net/num.			6	6
	CNO/num.			6	6
	CNO/FT			6	6
	Transolver/num.			6	6
	Transolver/FT			6	6
Total	20 cells	68	70	338	348

The 20 cells comprise the compatible combinations within the selected four-system, four-surrogate panel. Compatibility requires matched forecast and measurement variables, the shared delayed observation–prediction protocol, and trajectory-held-out evaluation. CNO is excluded for Cylinder because its available assets use 16 input channels rather than three. FNO and U-Net are excluded for Foil because their artifacts follow a different adaptation protocol. Matched fine-tuned checkpoints are available only for CNO and Transolver.

Combustion is excluded because its experimental stream provides a single OH^* chemiluminescence-intensity channel, whereas its numerical stream contains 15 reacting-flow variables. Its inclusion would require a cross-modal observation operator, changing the task from same-variable output calibration to inverse observation.

PDEBench. Table 6 summarizes the source parameters and the constructed mild and strong shifts for each PDEBench system.

Table 6: Physical shifts in PDEBench.

System	Source	Mild Shift	Strong Shift
Diffusion–Reaction	$(D_u, D_v, k) = (.001, .005, .005)$	joint $1.25\times$	joint $1.50\times$
Diffusion–Sorption	$D = 5 \times 10^{-4}$	$D = 6.25 \times 10^{-4}$	$D = 7.5 \times 10^{-4}$
Compressible NS	$\Gamma_{\text{gas}} = 5/3$	$\Gamma_{\text{gas}} = 1.5$	$\Gamma_{\text{gas}} = 1.4$

Table 7 summarizes the Development and Evaluation partitions. Each physical trajectory is generated under one declared shift and produces two model records, one for each surrogate.

Table 7: PDEBench partitions.

System	Surrogate/Shift	Physical Trajectories		Model Records	
		Development	Evaluation	Development	Evaluation
Diffusion–Reaction	FNO/mild	51	45	29	19
	FNO/strong			22	26
	U-Net/mild			29	19
	U-Net/strong			22	26
Diffusion–Sorption	FNO/mild	52	44	27	21
	FNO/strong			25	23
	U-Net/mild			27	21
	U-Net/strong			25	23
Compressible NS	FNO/mild	49	47	27	21
	FNO/strong			22	26
	U-Net/mild			27	21
	U-Net/strong			22	26
Total	12 cells	152	136	304	272

We use the officially released FNO and U-Net checkpoints. We exclude the PDEBench PINN baselines (Raissi et al., 2019) because they require condition-specific optimization rather than inference with a frozen surrogate.

Baseline tuning. Table 8 reports the search space and selected configuration for each baseline. CMA is parameter-free. Each tunable baseline uses one global configuration selected on the RealPDEBench Development set and used on PDEBench. Here, s, η, r, K, h, g_0, and $\rho_{\max}$ denote the number of adaptation steps, learning rate, rank, context length, hidden dimension, gate initialization, and spectral-radius cap, respectively.

Table 8: Baseline tuning on the RealPDEBench Development set.

Method	Development search space	Selected configuration
CMA	No tunable hyperparameter	Parameter-free
EWMA	$\beta \in \{.50, .90, .99\}$	$\beta = .50$
RLS	$(\lambda, \gamma) \in \{(.001, 1), (.01, 1), (.1, 1), (.01, .99)\}$	$\lambda = .01,\ \gamma = .99$
EnKF	32 members; $(\sigma_{\text{proc}}, \sigma_{\text{obs}}) \in \{(0, .10), (.01, .05), (.01, .10), (.01, .20), (.05, .10)\}$	32 members; $(\sigma_{\text{proc}}, \sigma_{\text{obs}}) = (.01, .05)$
DMD	$(\lambda, \gamma, \rho_{\max}) \in \{(.001, 1, .995), (.01, .99, .995), (.01, 1, .95), (.01, 1, .995), (.01, 1, 1), (.1, 1, .995)\}$	$\lambda = .01,\ \gamma = 1,\ \rho_{\max} = .95$
Always-8	$(r, \eta) \in \{(4, .10), (8, .03), (8, .10), (8, .30), (16, .10)\}$, $s = 8$	$r = 8,\ \eta = .30,\ s = 8$
LoRA	$(r, \eta) \in \{(4, .10), (8, .03), (8, .10), (8, .30), (16, .10)\}$, $\alpha = r,\ s = 8$	$r = \alpha = 16,\ \eta = .10,\ s = 8$
COSA	$K \in \{5, 10\},\ s \in \{1, 3, 5\},$ $\eta \in \{3{\times}10^{-4}, 10^{-3}\}$	$K = 10,\ s = 5,\ \eta = 10^{-3}$
ADAPT-Z	$h \in \{32, 64\},\ s \in \{1, 3, 5\},\ \eta \in \{10^{-4}, 3{\times}10^{-4}\}$	$h = 64,\ s = 5,\ \eta = 3{\times}10^{-4}$
TAFAS	$s \in \{1, 3\},\ \eta \in \{3{\times}10^{-4}, 10^{-3}\},$ $g_0 \in \{0, .01, .05\}$	$s = 3,\ \eta = 10^{-3},\ g_0 = .05$

BROC **hyperparameter sensitivity.** BROC uses its default hyperparameters $(\lambda, \gamma) = (10^{-2}, 1)$ throughout Evaluation. Table 9 reports the post hoc sensitivity analysis on the RealPDEBench Evaluation set. Improved Cells counts cells with lower MSE than the default configuration.

Table 9: BROC hyperparameter sensitivity on the RealPDEBench Evaluation set. Negative values indicate lower MSE; positive values indicate higher MSE.

Setting	Overall MSE	Cell Range	Improved Cells	Lowest-MSE Cells
$\lambda = 10^{-4}$	$-.035340$	$(-.128050, +.003890)$	17/20	15/20
$\lambda = 10^{-3}$	$-.025590$	$(-.077190, +.001260)$	19/20	4/20
$\lambda = 10^{-2}$		Reference	–	1/20
$\lambda = 10^{-1}$	$+.053010$	$(+.001060, +.138560)$	0/20	0/20
$\lambda = 1$	$+.133000$	$(+.000950, +.305600)$	0/20	0/20
$\gamma = .90$	$+.012490$	$(-.000110, +.039260)$	1/20	0/20
$\gamma = .95$	$+.006270$	$(-.000230, +.019670)$	2/20	2/20
$\gamma = .99$	$+.001210$	$(-.000180, +.003780)$	4/20	2/20
$\gamma = 1$		Reference	–	16/20

The results show that the default hyperparameters are not optimal: smaller λ values achieve lower MSE. Nevertheless, we retain the default parameters in all reported comparisons.

B.2 Accuracy and Efficiency

Accuracy. Tables 10 and 11 report the complete MSE for every method and cell on RealPDEBench and PDEBench, respectively. Figures A1 and A2 visualize the corresponding MSE ratios relative to Frozen as heatmaps.

Table 10: Complete MSE results on RealPDEBench; lower is better.

System	Surrogate/Checkpoint	Frozen	CMA	EWMA	RLS	Always-8	LoRA	EnKF	DMD	COSA	ADAPT-Z	TAFAS	**BROC**
Foil	CNO/num.	.040603	.040082	.040090	.037321	.039072	.037195	.040177	.060473	.038640	.035173	.036373	.028606
	CNO/FT	.024359	.024183	.024183	.022692	.023081	.021669	.024215	.030735	.022782	.021481	.021302	.019449
	Transolver/num.	.103107	.101733	.101737	.100351	.098267	.085331	.101948	.107596	.100050	.086537	.097711	.051242
	Transolver/FT	.102210	.099629	.099629	.099433	.101878	.096795	.100041	.104225	.101123	.096850	.097840	.081878
Cylinder	FNO/num.	.144634	.137645	.137679	.125360	.142706	.132206	.138564	.148402	.136516	.126641	.133962	.094635
	U-Net/num.	.088789	.083897	.083972	.066244	.084087	.074820	.084585	.125660	.079303	.070622	.075313	.054471
	Transolver/num.	.254704	.207387	.207678	.198544	.222007	.199693	.213877	.253659	.238303	.203548	.223123	.148310
	Transolver/FT	.064274	.062325	.062368	.062312	.063754	.061575	.062644	.060992	.063000	.060252	.061709	.050605
Controlled Cylinder	FNO/num.	.074754	.064215	.064222	.062642	.070112	.063700	.064410	.093426	.070347	.064031	.067297	.048586
	U-Net/num.	.034852	.034448	.034458	.033828	.034684	.033902	.034464	.054420	.034311	.033700	.033473	.029987
	CNO/num.	.026090	.025908	.025919	.025708	.025992	.025731	.025921	.042179	.025829	.025558	.024995	.023447
	CNO/FT	.006526	.006523	.006527	.006526	.006523	.006524	.006535	.011538	.006523	.006532	.006536	.006519
	Transolver/num.	.077153	.069326	.069623	.066185	.075969	.066245	.069417	.062676	.074339	.066890	.069057	.040636
	Transolver/FT	.026813	.026823	.026832	.026843	.026784	.026440	.026829	.025675	.026604	.025891	.026284	.019631
FSI	FNO/num.	.107008	.106612	.106610	.099997	.106284	.100405	.106661	.071201	.101792	.094320	.097344	.070605
	U-Net/num.	.029278	.029151	.029152	.028708	.029187	.028825	.029192	.035143	.028847	.028198	.026809	.024844
	CNO/num.	.035674	.035481	.035482	.035209	.035562	.035197	.035526	.040526	.035257	.034555	.032714	.031378
	CNO/FT	.006193	.006191	.006193	.006191	.006192	.006186	.006211	.008614	.006189	.006176	.006196	.005881
	Transolver/num.	.051744	.051607	.051596	.051279	.051587	.050330	.051643	.044229	.050754	.047758	.048400	.034646
	Transolver/FT	.028448	.028450	.028451	.028443	.028413	.027860	.028471	.025766	.027654	.025114	.026461	.016507

Table 11: Complete MSE results on PDEBench; lower is better.

System	Surrogate/Checkpoint	Frozen	CMA	EWMA	RLS	Always-8	LoRA	EnKF	DMD	COSA	ADAPT-Z	TAFAS	**BROC**
Diffusion–Reaction	FNO/mild	.001223	.001218	.001216	.001699	.001200	.001134	.001209	.001192	.001219	.001098	.001183	.001051
	FNO/strong	.001910	.001788	.001789	.002597	.001857	.001689	.001795	.001755	.001902	.001588	.001842	.001175
	U-Net/mild	.019595	.009696	.009360	.031913	.019263	.014673	.010170	.014058	.019561	.016271	.019316	.003211
	U-Net/strong	.022518	.010149	.009654	.032414	.022001	.016688	.010837	.015800	.022486	.018743	.022241	.003400
Diffusion–Sorption	FNO/mild	.000309	.000192	.000181	.000324	.000274	.000189	.000167	.000201	.000302	.000169	.000274	.000019
	FNO/strong	.000963	.000600	.000554	.001062	.000866	.000578	.000490	.000655	.000950	.000594	.000888	.000048
	U-Net/mild	.004290	.011843	.010654	.029705	.004316	.004223	.006859	.007363	.004293	.003508	.004305	.002972
	U-Net/strong	.004026	.009982	.009079	.032104	.003962	.003588	.005281	.006059	.004032	.002880	.004056	.002876
Compressible NS	FNO/mild	.684919	.265479	.263157	.248066	.416497	.250009	.278699	.684919	.684616	.676007	.684584	.223822
	FNO/strong	.752109	.399080	.396621	.367647	.543547	.279608	.411057	.752109	.751827	.743283	.751732	.364093
	U-Net/mild	3.075761	.321712	.317531	.312278	1.806857	.093365	.346013	3.075761	3.074783	3.039799	3.073482	.500454
	U-Net/strong	3.300940	.446881	.444528	.427128	2.161821	.642895	.488137	3.300940	3.299813	3.264975	3.298654	.558000

	Surrogate/Checkpoint	Frozen	CMA	EWMA	RLS	Always-8	LoRA	EnKF	DMD	COSA	ADAPT-Z	TAFAS	**BROC (Ours)**
Foil	CNO / num.	1.000	0.987	0.987	0.919	0.962	0.916	0.990	1.489	0.952	0.866	0.896	**0.705**
	CNO / FT	1.000	0.993	0.993	0.932	0.948	0.890	0.994	1.262	0.935	0.882	0.875	**0.798**
	Transolver / num.	1.000	0.987	0.987	0.973	0.953	0.828	0.989	1.044	0.970	0.839	0.948	**0.497**
	Transolver / FT	1.000	0.975	0.975	0.973	0.997	0.947	0.979	1.020	0.989	0.948	0.957	**0.801**
Cylinder	FNO / num.	1.000	0.952	0.952	0.867	0.987	0.914	0.958	1.026	0.944	0.876	0.926	**0.654**
	U-Net / num.	1.000	0.945	0.946	0.746	0.947	0.843	0.953	1.415	0.893	0.795	0.848	**0.613**
	Transolver / num.	1.000	0.814	0.815	0.780	0.872	0.784	0.840	0.996	0.936	0.799	0.876	**0.582**
	Transolver / FT	1.000	0.970	0.970	0.969	0.992	0.958	0.975	0.949	0.980	0.937	0.960	**0.787**
Controlled Cylinder	FNO / num.	1.000	0.859	0.859	0.838	0.938	0.852	0.862	1.250	0.941	0.857	0.900	**0.650**
	U-Net / num.	1.000	0.988	0.989	0.971	0.995	0.973	0.989	1.561	0.984	0.967	0.960	**0.860**
	CNO / num.	1.000	0.993	0.993	0.985	0.996	0.986	0.994	1.617	0.990	0.980	0.958	**0.899**
	CNO / FT	1.000	1.000	1.000	1.000	1.000	1.000	1.001	1.768	1.000	1.001	1.002	**0.999**
	Transolver / num.	1.000	0.899	0.902	0.858	0.985	0.859	0.900	0.812	0.964	0.867	0.895	**0.527**
	Transolver / FT	1.000	1.000	1.001	1.001	0.999	0.986	1.001	0.958	0.992	0.966	0.980	**0.732**
FSI	FNO / num.	1.000	0.996	0.996	0.934	0.993	0.938	0.997	0.665	0.951	0.881	0.910	**0.660**
	U-Net / num.	1.000	0.996	0.996	0.981	0.997	0.985	0.997	1.200	0.985	0.963	0.916	**0.849**
	CNO / num.	1.000	0.995	0.995	0.987	0.997	0.987	0.996	1.136	0.988	0.969	0.917	**0.880**
	CNO / FT	1.000	1.000	1.000	1.000	1.000	0.999	1.003	1.391	0.999	0.997	1.001	**0.950**
	Transolver / num.	1.000	0.997	0.997	0.991	0.997	0.973	0.998	0.855	0.981	0.923	0.935	**0.670**
	Transolver / FT	1.000	1.000	1.000	1.000	0.999	0.979	1.001	0.906	0.972	0.883	0.930	**0.580**
Overall		1.000	0.948	0.949	0.912	0.967	0.902	0.953	1.059	0.961	0.885	0.919	**0.674**

Figure A1: MSE ratio heatmap on RealPDEBench; lower is better.

	Surrogate/Checkpoint	Frozen	CMA	EWMA	RLS	Always-8	LoRA	EnKF	DMD	COSA	ADAPT-Z	TAFAS	**BROC (Ours)**
Diffusion-Reaction	FNO mild	1.000	0.996	0.994	1.390	0.982	0.928	0.989	0.975	0.997	0.898	0.968	**0.859**
	FNO strong	1.000	0.936	0.937	1.360	0.973	0.884	0.940	0.919	0.996	0.831	0.964	**0.615**
	U-Net mild	1.000	0.495	0.478	1.629	0.983	0.749	0.519	0.717	0.998	0.830	0.986	**0.164**
	U-Net strong	1.000	0.451	0.429	1.440	0.977	0.741	0.481	0.702	0.999	0.832	0.988	**0.151**
Diffusion-Sorption	FNO mild	1.000	0.622	0.587	1.047	0.886	0.611	0.542	0.649	0.978	0.548	0.886	**0.061**
	FNO strong	1.000	0.622	0.575	1.103	0.899	0.600	0.508	0.680	0.986	0.616	0.922	**0.050**
	U-Net mild	1.000	2.761	2.484	6.924	1.006	0.984	1.599	1.716	1.001	0.818	1.004	**0.693**
	U-Net strong	1.000	2.479	2.255	7.974	0.984	0.891	1.312	1.505	1.001	0.715	1.008	**0.714**
Compressible NS	FNO mild	1.000	0.388	0.384	0.362	0.608	0.365	0.407	1.000	1.000	0.987	1.000	**0.327**
	FNO strong	1.000	0.531	0.527	0.489	0.723	**0.372**	0.547	1.000	1.000	0.988	0.999	0.484
	U-Net mild	1.000	0.105	0.103	0.102	0.587	**0.030**	0.112	1.000	1.000	0.988	0.999	0.163
	U-Net strong	1.000	0.135	0.135	**0.129**	0.655	0.195	0.148	1.000	1.000	0.989	0.999	0.169
overall		1.000	1.016	0.934	2.760	0.864	0.604	0.687	1.071	0.999	0.856	0.993	**0.341**

Figure A2: MSE ratio heatmap on PDEBench; lower is better.

Efficiency. Table 12 reports the latency of every method on RealPDEBench and PDEBench.

Table 12: The latency of each method on RealPDEBench and PDEBench; lower is better.

Model	RealPDEBench Latency (ms)	PDEBench Latency (ms)
Frozen	0.121 [0.109, 0.129]	0.116 [0.110, 0.124]
CMA	2.810 [2.177, 3.496]	2.373 [2.147, 3.158]
EWMA	2.801 [2.165, 3.478]	2.359 [2.133, 3.141]
BROC	5.317 [4.406, 6.673]	4.348 [4.298, 4.519]
EnKF	5.542 [5.517, 5.572]	4.666 [4.535, 4.736]
RLS	6.104 [5.190, 7.745]	5.191 [5.034, 5.359]
DMD	43.076 [34.010, 48.057]	19.766 [14.331, 21.798]
COSA	30.152 [24.188, 33.684]	33.670 [31.452, 34.504]
LoRA	26.675 [21.766, 38.611]	38.032 [36.976, 38.496]
Always-8	47.186 [44.858, 47.408]	45.604 [45.439, 45.867]
ADAPT-Z	63.651 [39.153, 74.201]	74.250 [73.617, 74.533]
TAFAS	398.260 [348.381, 620.413]	116.291 [113.213, 117.891]

B.3 Sparse Sensing

Table 13 reports the complete full-field BROC/Frozen MSE ratios under dense and sparse sensing configurations. All ratios are below one.

Table 13: Sparse sensing on RealPDEBench and PDEBench; lower MSE ratios are better.

System	Dense	Uniform 5%	Random 5%	Uniform 1%	Random 1%
RealPDEBench					
Foil	0.545263	0.551557	0.557622	0.599518	0.616879
Cylinder	0.642016	0.642485	0.648710	0.654922	0.677994
Controlled Cylinder	0.630474	0.631914	0.633032	0.639754	0.645773
FSI	0.693950	0.698792	0.702365	0.714742	0.740916
Overall	0.627926	0.631187	0.635432	0.652234	0.670391
PDEBench					
Diffusion–Reaction	0.188282	0.188286	0.189263	0.193822	0.190540
Diffusion–Sorption	0.312781	0.312934	0.341524	0.318143	0.784227
Compressible NS	0.151499	0.159202	0.181251	0.303764	0.251495
Overall	0.217521	0.220141	0.237346	0.271910	0.408754

B.4 Noise

Measurement noise is evaluated with fixed random 1% probes to represent corruption under sparse sensing. Clean and noisy runs use the same trajectories, probe locations, and placement seeds. Surrogate-input noise instead uses dense, clean observations to isolate input perturbations from sensing sparsity. We perturb the dynamic input variables as

$$\bar{X}_n^{(\sigma)} = \bar{X}_n + \sigma\xi_n, \qquad \xi_n \sim \mathcal{N}(0, I),$$

while leaving the grid and physical parameters unchanged. Tables 14 and 15 report the complete system results. We report BROC/Frozen ratios at $\sigma = .10$, the highest tested Gaussian noise level.

Table 14: Measurement noise on RealPDEBench and PDEBench; lower MSE values and ratios are better.

System	Clean	Gaussian .05	Gaussian .10	Student-t_3 .05	Gaussian*	Ratio (.10)
RealPDEBench						
Foil	.171915	.171938	.172023	.171942	.171958	.614
Cylinder	.093987	.094015	.094106	.094018	.094044	.681
Controlled Cylinder	.030524	.030545	.030604	.030543	.030563	.648
FSI	.050829	.050876	.051000	.050875	.050915	.742
PDEBench						
Diffusion–Reaction	.002251	.002310	.002402	.002312	.002376	.222
Diffusion–Sorption	.006644	.006859	.007164	.006863	.007009	.339
Compressible NS	.540614	.543579	.556463	.543655	.548305	.299

Table 15: Surrogate-input noise on RealPDEBench and PDEBench; lower MSE values and ratios are better.

System	Clean	Gaussian .05	Gaussian .10	Ratio (.10)
RealPDEBench				
Foil	.152841	.153678	.155051	.551
Cylinder	.089031	.089609	.090904	.657
Controlled Cylinder	.029787	.030677	.030981	.532
FSI	.047826	.048057	.048669	.704
PDEBench				
Diffusion–Reaction	.002206	.002293	.002335	.213
Diffusion–Sorption	.001564	.001641	.001689	.435
Compressible NS	.392284	.391326	.392207	.211

B.5 AUTOREGRESSIVE EVALUATION

At each prespecified rollout origin, we freeze the committed BROC state and disable all subsequent observation access and correction-state updates. We compare two recursive inputs. *Corrected* feeds the BROC-corrected variables into the next surrogate evaluation, while variables outside the correction retain their Frozen predictions. *Reference* instead feeds the complete Frozen prediction into the next evaluation; the BROC correction is applied only to the output being scored.

For each horizon h, we average the MSE over the first h autoregressive blocks and normalize it by the Frozen autoregressive MSE over the same blocks. Table 16 reports the system results. The benchmark summaries in Table 3 use the *Corrected* input.

Table 16: Autoregressive evaluation on RealPDEBench and PDEBench; lower MSE ratios are better. The longest horizon is $h = 7$ for Compressible NS and $h = 8$ otherwise.

System	Recursive input	$h = 2$	$h = 4$	Longest horizon
RealPDEBench				
Foil	Corrected	.768 [.746, .787]	.824 [.798, .851]	.552 [.520, .585]
	Reference	.843 [.807, .872]	.913 [.889, .939]	.731 [.694, .769]
Cylinder	Corrected	.681 [.632, .731]	.789 [.756, .823]	.486 [.450, .523]
	Reference	.889 [.833, .946]	.874 [.844, .905]	.667 [.628, .708]
Controlled Cylinder	Corrected	.737 [.706, .770]	.812 [.786, .839]	.524 [.495, .555]
	Reference	.813 [.792, .836]	.892 [.868, .917]	.704 [.672, .737]
FSI	Corrected	.764 [.749, .778]	.479 [.458, .601]	.238 [.109, .468]
	Reference	.803 [.796, .812]	.836 [.817, .857]	.610 [.576, .646]
PDEBench				
Diffusion–Reaction	Corrected	.657 [.646, .668]	.662 [.641, .685]	.698 [.652, .748]
	Reference	.714 [.708, .720]	.825 [.821, .828]	.918 [.916, .920]
Diffusion–Sorption	Corrected	.439 [.433, .446]	.395 [.387, .405]	.456 [.443, .472]
	Reference	.577 [.572, .582]	.568 [.563, .573]	.599 [.593, .605]
Compressible NS	Corrected	.136 [.104, .172]	.168 [.128, .210]	.289 [.179, .449]
	Reference	.252 [.209, .293]	.333 [.277, .386]	.377 [.316, .435]

Corrected produces a lower MSE ratio than *Reference* across all systems and reported horizons. Thus, feeding corrected variables into subsequent surrogate evaluations provides an additional benefit beyond correcting only the scored outputs.

B.6 Physical Fidelity

RealPDEBench. We evaluate the velocity-divergence discrepancy of BROC on RealPDEBench. It is calculated as the root mean square (RMS) difference between the predicted and observed velocity-divergence fields. Results are averaged over the two numerical-checkpoint cells and four future blocks for each system. Table 17 reports the discrepancies of BROC and Frozen and the corresponding reductions.

Table 17: Velocity-divergence discrepancy on RealPDEBench; lower discrepancies and larger reductions are better.

System	Frozen	BROC	Reduction
Foil	.007343	.003848	47.597%
Cylinder	.011138	.008565	23.101%
Controlled Cylinder	.008367	.004638	44.567%
FSI	.014900	.010163	31.793%

PDEBench. We evaluate a broader set of physical diagnostics on PDEBench. For Diffusion–Reaction, we evaluate the coupled PDE residual, errors in the global activator and inhibitor quantities, the boundary normal-gradient error, and the radial-spectrum discrepancy. For Diffusion–Sorption, we evaluate the interior PDE residual, total-concentration error, negative-concentration violations, and front-location error. For Compressible NS, we evaluate the Euler conservation residual, errors in global mass, momentum, and energy, non-positive-density violations, and density total-variation error. Each entry averages the four shifted FNO and U-Net cells for the corresponding system. Table 18 reports the results for Frozen and BROC together with the corresponding BROC/Frozen ratios. Ratios below one favor BROC.

Table 18: Physical-fidelity diagnostics on PDEBench; lower values and ratios are better.

Fidelity	System	Object	Frozen	BROC	Ratio
Residual	Diffusion–Reaction	PDE Residual	.619358	.453714	.733
	Diffusion–Sorption		1.081313	.864511	.799
	Compressible NS		5.034278	2.477343	.492
Property	Diffusion–Reaction	Activator	.008667	.002640	.305
		Inhibitor	.007253	.002572	.355
	Diffusion–Sorption	Concentration	.023666	.013936	.589
	Compressible NS	Mass	.545808	.177244	.325
		Momentum	1.356861	.496743	.366
		Energy	.593477	.074275	.125
Violation	Diffusion–Reaction	Boundary Gradient	.519774	.440689	.848
	Diffusion–Sorption	Concentration	.007216	.000914	.127
	Compressible NS	Density	.003588	.001848	.515
Specific	Diffusion–Reaction	Spatial Spectrum	.027247	.005694	.209
	Diffusion–Sorption	Front Location	.007423	.003659	.493
	Compressible NS	Density Variation	34.885760	25.467650	.730

B.7 Mechanism Evidence

Table 19 reports the complete mechanism results. Residual-energy entries correspond to $r = 1, 2, 4$ for RealPDEBench and $r = 1, 2, 3$ for PDEBench. We use records from the Development and Evaluation sets for the temporal-persistence analysis as a post hoc diagnostic.

Table 19: Mechanism evidence on RealPDEBench and PDEBench.

System	Cumulative Residual Energy	Spearman Correlation	Positive Future Gain
RealPDEBench			
Foil	56.35 / 84.35 / 95.50%	.980 [.953, .991]	100.00%
Cylinder	54.34 / 86.21 / 97.09%	.973 [.945, .987]	100.00%
Controlled Cylinder	53.56 / 85.84 / 97.08%	.974 [.950, .982]	97.88%
FSI	59.39 / 89.49 / 98.31%	.945 [.873, .969]	100.00%
Overall	54.88 / 85.96 / 97.18%	.980 [.962, .988]	98.98%
PDEBench			
Diffusion–Reaction	77.60 / 94.00 / 96.90%	.896 [.853, .931]	97.40%
Diffusion–Sorption	98.20 / 99.50 / 99.90%	.934 [.903, .950]	91.67%
Compressible NS	93.00 / 99.30 / 100.00%	.918 [.879, .939]	100.00%
Overall	91.22 / 99.09 / 99.93%	.911 [.887, .924]	96.35%

All four RealPDEBench systems show strong temporal persistence, with correlations of .945–.980; positive future gains occur in 97.88–100% of their model records. On PDEBench, the correlations range from .896 to .934, with positive future gains in 91.67–100% of the model records.