PHYSICAL ROBUSTNESS UNDER EQUATION CORRUPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robustness in machine learning usually means invariance, which is evaluated by perturbing an input while the correct label stays fixed. This evaluation setup fails for PDE surrogates: changing a coefficient or a boundary condition moves the physical solution, and changing the mesh or the sensor changes how the solution is measured. Physical robustness should instead be evaluated by how well the prediction tracks changes in the correct output, in both direction and magnitude. Nominal accuracy does not guarantee response fidelity: under a Burgers viscosity shift, an FNO with held-out field error below 2% still mispredicted the response by half the true response magnitude. To make physical robustness measurable and trainable, we decompose changed-target error into nominal error and sensitivity mismatch, introduce diagnostics that grade the response at held-out severities, and evaluate three training interventions on six specification-shift tasks. Our evaluation identifies effective training interventions for different perturbations, PDEs, and backbones: response alignment cut the response mismatch on public multi-viscosity Burgers data roughly sixfold and led on coefficient and boundary shifts, whereas stress augmentation cut evaluation-mesh and initial-condition error by 21.5% and 16.2% over a matched changed-label baseline. Adversarial training, a standard approach to robustness in machine learning, instead increased evaluation-mesh and initial-condition error by roughly 28%. Based on these results, we argue that a surrogate should be judged by how faithfully the prediction tracks the governing system, not by how little the prediction moves when the physics changes.

1 INTRODUCTION

Robustness in vision means invariance: the evaluation assumes that a corruption $x \mapsto x + \delta$ preserves the label, so a model's robustness is judged by how little the model's prediction moves. Scientific machine learning (SciML) applies the same evaluation template, but a physically meaningful perturbation changes the problem, and therefore the numerical solution used as the label (Karniadakis et al., 2021). Adding noise to a photograph of a cat still leaves a cat, but raising the viscosity of a flow leaves a different flow, and the solver returns a different field. A partial differential equation (PDE) surrogate replaces repeated numerical solves in simulation, design, and forecasting (Azizzadenesheli et al., 2024; Lam et al., 2023), and each such outer loop varies the specification and compares the resulting solutions. The loop asks the surrogate to predict the solution for settings not used for training, such as a new viscosity or boundary condition. As these settings change, the correct output changes too. Physical robustness should therefore be judged by how well the prediction matches the true change, not by how little the prediction moves: when the correct output moves, the prediction should match that change in direction and magnitude, a property we call response fidelity; an invariant prediction is wrong rather than robust. Figure 1 contrasts label-preserving with equation corruption and the response-fidelity criterion.

The specification consumed by a learned PDE surrogate is the governing equation together with the equation's coefficients, forcing, initial and boundary data, physical parameters, and the discretization and observation maps. Changing any one of these components changes the correct output. Changing the coefficients, forcing, initial or boundary data, or physical parameters moves the physical solution, as a new viscosity does; changing the discretization, the observation map, or the reported quantity of interest (QoI) moves how one and the same solution is measured, as a halved sensor spacing does.

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.

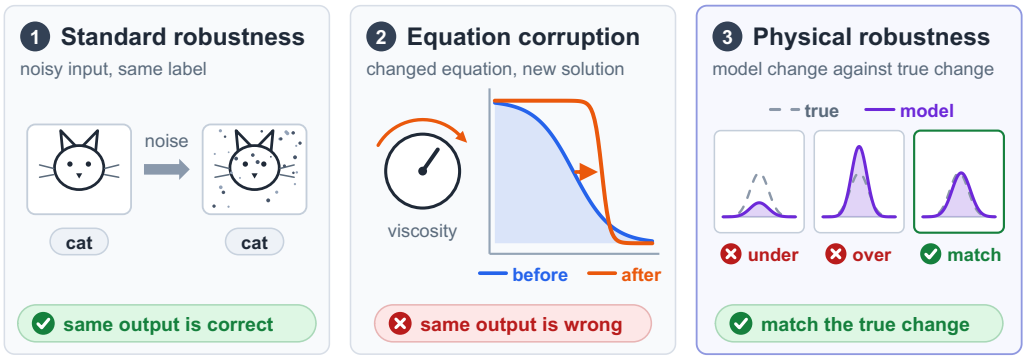

Figure 1: Physical robustness requires response fidelity rather than invariance. Under a standard label-preserving corruption, the correct output is unchanged (left). An admissible change to the governing specification instead moves the physical target (center), so the surrogate should match the direction and magnitude of the reference response; under-response and over-response are failures, and invariance is the extreme case of under-response (right).

To state how a surrogate should move when the correct output moves, write $\mathcal{S}$ for the true solution or observation map and $\mathcal{A}_\omega$ for the surrogate. For a physically admissible change $z \mapsto z_\delta$, we call a model physically robust when the model satisfies

$$\mathcal{A}_\omega(z_\delta) - \mathcal{A}_\omega(z) \approx \mathcal{S}(z_\delta) - \mathcal{S}(z),$$

while keeping the model's error at the nominal setting low. The vision template asks instead for $\mathcal{A}_\omega(z_\delta) - \mathcal{A}_\omega(z) \approx 0$, which is the right demand only when the target does not move. When the target changes, the invariance criterion favors a model that keeps the prediction fixed and misses the entire physical response. Clean-distribution generalization holds the specification fixed and never asks the prediction to move either. We therefore ask for a low nominal error and a calibrated response together, and we evaluate both at severities held out from training.

Neural PDE surrogates have conventionally been scored by the relative L^2 error of the predicted field. A low field error under that score does not imply that the prediction moves correctly. In our Burgers viscosity experiment, a Fourier neural operator (FNO) trained on the correct changed labels reached 1.06% nominal error and 1.81% held-out error, an accuracy that would ordinarily be reported as success. Despite this low field error, the FNO's mean held-out sensitivity-alignment error (SAE) was 0.50: the FNO's response error was 50% of the true response magnitude (Table 1). Figure 2 gives a sample-level view of this aggregate gap by comparing the parametric model with a model trained using physical sensitivity alignment (PhySA): two predictions that look equally accurate show markedly different response errors around the forming shock, and the excess amplification factor (EAF) reports the two models' relative response magnitudes.

Response fidelity is not a new demand. Classical model reduction already asks a parametrized surrogate to stay accurate at every parameter in a family, where the solution moves with the parameter rather than staying fixed (Benner et al., 2015; Hesthaven et al., 2016). Design, control, inversion, and global sensitivity analysis likewise consume differences or derivatives with respect to the specification (Lions, 1971; Saltelli et al., 2008). Recent work motivates a moving-target criterion without turning that criterion into an evaluation: changed-reference studies recompute solver targets after perturbing operator inputs (Adesoji & Chen, 2022; Sun, 2025), derivative-informed methods supervise solution sensitivities to PDE parameters (O'Leary-Roseberry et al., 2024; Behroozi et al., 2025), and shift benchmarks show that fixed-specification accuracy need not predict performance after a PDE or discretization change (Takamoto et al., 2022; Shikhman, 2026). However, none of these lines of work evaluates the response at a held-out severity, so a surrogate can satisfy all of these training and evaluation criteria and still miss the physical change: the FNO above reaches 1.81% held-out field error while the FNO's response error is 50% of the true response magnitude, and a field-error report cannot separate that failure from an overshoot of the same size. Evaluating response direction and magnitude at held-out severities makes the distinction measurable and lets us rank the training

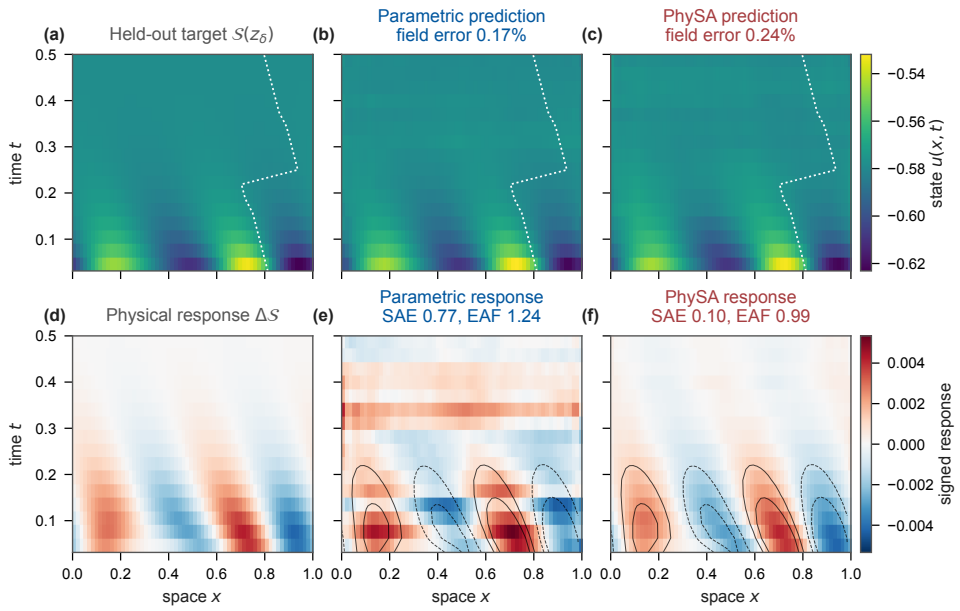

Figure 2: A small field error can conceal a large physical-response error. For a held-out change from $\nu = 0.1$ to 0.085, both FNO predictions have field error below 0.25%. However, the two models' response fields differ: the parametric model has SAE 0.77 and EAF 1.24, whereas PhySA has SAE 0.10 and EAF 0.99. A smaller SAE is a closer response match and EAF has a matched-response reference of one, so the parametric response misses the reference by 77% of the reference magnitude and overshoots that magnitude by 24%, while the PhySA response matches the reference in direction and in magnitude. The case is selected without consulting PhySA; App. G states the selection rule.

interventions below by the resulting responses rather than by field error alone. Table 3 in App. B summarizes the closest contrasts alongside the detailed positioning and perturbation taxonomy.

To make physical robustness measurable and trainable, we define admissible perturbation families and a moving-target risk, evaluate on seen and held-out severity grids, and score both changed-target error and response direction and magnitude. We connect these quantities through nominal-error and sensitivity-mismatch bounds, then test three interventions against matched changed-label baselines: physically admissible stress augmentation (PASA), PhySA, and intrinsic sensitivity factorization (ISF). The contributions are as follows.

1. We define physical robustness against an admissible specification-perturbation model that separates changes to the solution from changes to the representation or comparison map.
2. Our theoretical analysis decomposes changed-target error into nominal error and sensitivity mismatch, yields guarantees for finite-severity evaluation, aligned response anchors, and PDE-based residuals, and extends the pathwise bound to regularized inverse references with explicit stability and bias terms. At aligned anchors, the severity-discretization penalty improves from first to second order.
3. We introduce response diagnostics and evaluate three interventions on six specification-shift tasks and multiple backbones against matched changed-label baselines. The experiments identify the strengths and trade-offs of each intervention across perturbations, PDEs, and backbones, providing concrete guidance for choosing a training strategy based on both field accuracy and response fidelity.

2 Physical Robustness and Sensitivity Alignment

We turn response fidelity into an evaluation rule and a training goal. The aim is to show how nominal accuracy and response alignment control prediction error when the physical specification changes.

2.1 PHYSICAL ROBUSTNESS

We first define which changes enter the robustness test and how to measure prediction error at each changed setting.

Admissible specifications. Write the problem specification given to a learned PDE surrogate as

$$z := (\Omega, \mathcal{L}, a, f, u_0, g, \theta, \mathcal{M}, \mathcal{O}),$$

where the entries are the spatial domain and differential operator, the coefficient, forcing, initial and boundary data, the physical parameters, and the discretization and observation maps. Let

$$\mathcal{Z}_{\text{phys}} := \{z : C(z) = 0, G(z) \geq 0, \mathsf{W}(z) = 1\}$$

be the set of physically admissible PDE specifications. Here C encodes equality constraints such as conservation or compatibility, G encodes positivity or ellipticity constraints, and W requires a well-posed problem whose outputs can be compared reliably. These conditions define the benchmark's test domain: every perturbation must produce a physically valid problem with a reference that supports a reliable error measurement. For example, viscosity must remain positive, boundary values must agree at shared corners, and an evaluation mesh must be fine enough to resolve the field being compared. The physical problem and comparison map determine these conditions, independently of model predictions, and every model is evaluated on the same admissible changes. Conditions of the first two kinds are also used to constrain physical emulators through the emulator architecture or loss (Beucler et al., 2021). The set of admissible perturbations of z at budget ρ is

$$\Delta_{\text{phys}}(z; \rho) := \{\delta : z_\delta = T_\delta z, C(z_\delta) = 0, G(z_\delta) \geq 0, \mathsf{W}(z_\delta) = 1, D(z, z_\delta) \leq \rho\},$$

where D is a metric on specifications and ρ is the perturbation budget. Response-changing perturbations alter the continuous solution, as a viscosity change does. Representation-changing perturbations alter the comparison map, sensor operator, mesh, or QoI, changing how the solution is measured.

Physical robustness risk. Let $\mathcal{S}(z)$ be the true target in a common comparison space and $\mathcal{A}_\omega(z)$ be the learned output under the same observation or comparison map. We define the worst-case physical robustness risk

$$\mathcal{R}^{\text{wc}}_{\Delta,\rho}(\mathcal{A}_\omega) := \mathbb{E}_{z \sim P_Z}\left[\sup_{\delta \in \Delta_{\text{phys}}(z;\rho)} d_{\text{QoI}}(\mathcal{A}_\omega(z_\delta), \mathcal{S}(z_\delta))\right].$$

For each nominal specification, this risk takes the largest error over allowed changes and then averages over nominal specifications. Each prediction is compared with $\mathcal{S}(z_\delta)$, the target for the changed problem.

Severity-grid estimator. Evaluating every allowed change is usually infeasible for a neural PDE surrogate. In the experiments, we therefore use a finite common severity grid $\mathcal{G} := \{\delta_1, \ldots, \delta_K\}$ such that $z_\delta = T_\delta z$ is admissible for every evaluated z, with a split into seen and held-out severities. We report the split mean

$$\widehat{\mathcal{R}}_{\text{split}} := \frac{1}{|\mathcal{D}_{\text{test}}||\mathcal{G}_{\text{split}}|} \sum_{z \in \mathcal{D}_{\text{test}}} \sum_{\delta \in \mathcal{G}_{\text{split}}} d_{\text{QoI}}(\mathcal{A}_\omega(z_\delta), \mathcal{S}(z_\delta))$$

and severity envelopes

$$E_k := \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{z \in \mathcal{D}_{\text{test}}} d_{\text{QoI}}(\mathcal{A}_\omega(T_{\delta_k} z), \mathcal{S}(T_{\delta_k} z)).$$

The split mean measures average error over the selected severities, and E_k gives the error at each severity. A worst-case grid estimate takes each specification's largest grid error before averaging. The grid summaries directly measure how each model performs at the tested severities. The grid certificate in App. C.1 shows how to turn the worst-case grid estimate into an upper bound over the full continuous family: add a margin for possible error growth between tested severities and a margin for uncertainty from the finite test sample. The experiments report the measured grid errors; computing those extra margins requires bounds on how large the error can be and how fast the error can change with severity.

2.2 SENSITIVITY ALIGNMENT

We now separate error at the original setting from error caused by predicting the physical change incorrectly. In local coordinates given by a Banach chart around z, suppose $\mathcal{A}, \mathcal{S} : \mathcal{Z}_{\text{phys}} \to \mathcal{Y}$ are Fréchet differentiable at z along admissible increments. Then

$$\mathcal{A}(z+\delta) - \mathcal{S}(z+\delta) = \underbrace{\mathcal{A}(z) - \mathcal{S}(z)}_{\text{nominal}} + \underbrace{(D\mathcal{A}(z) - D\mathcal{S}(z))[\delta]}_{\text{sensitivity mismatch}} + o(\|\delta\|_{\mathcal{Z}}).$$

The first term is the model's existing error; the second is the first-order error in the model's predicted response to δ. If $D\mathcal{A}$ and $D\mathcal{S}$ are locally Lipschitz in this chart, the remainder is $O(\|\delta\|_{\mathcal{Z}}^2)$. A small first term does not ensure a small second term: nominal accuracy alone cannot determine whether the model responds in the right direction or by the right amount.

The next proposition turns this local decomposition into a bound for a finite change. The error at the changed setting is at most the starting error plus the sensitivity mismatch accumulated along an admissible path.

Proposition 1 (Local physical robustness bound)**.** *Let $\gamma : [0,1] \to \mathcal{Z}_{\text{phys}}$ be an admissible C^1 path with $\gamma(0) = z$ and $\gamma(1) = z_\delta$. If $\mathcal{A}$ and $\mathcal{S}$ are Fréchet differentiable along the path and the compositions of the model and target maps with γ are C^1, then*

$$\|\mathcal{A}(z_\delta) - \mathcal{S}(z_\delta)\|_{\mathcal{Y}} \le \|\mathcal{A}(z) - \mathcal{S}(z)\|_{\mathcal{Y}} + \int_0^1 \|(D\mathcal{A}(\gamma(t)) - D\mathcal{S}(\gamma(t)))[\dot\gamma(t)]\|_{\mathcal{Y}}\, dt.$$

If, in addition, the nominal error is at most ϵ_{nom}, the path length is at most ρ, and the operator-norm mismatch along the path is bounded by ϵ_{sens}, then the perturbed error is at most $\epsilon_{\text{nom}} + \epsilon_{\text{sens}}\rho$.

The bound separates two requirements: an accurate starting prediction and a response that follows the physical change. The integral accumulates the mismatch in response direction and magnitude along the path. If the sensitivities match along the entire path, even a large physical change leaves the prediction error equal to the nominal error.

2.3 ANCHORED RESPONSE FIDELITY

We now bound prediction error between the severity values used to align responses during training. These finitely many values are the response-alignment anchors. Proposition 1 uses sensitivities along the whole path; the following theorem uses nominal error, response mismatch at the anchors, and a bound on how sharply the error curve can bend.

Let $\gamma : [0, \rho] \to \mathcal{Z}_{\text{phys}}$ be an admissible path and set

$$r(t) \coloneqq \mathcal{A}(\gamma(t)) - \mathcal{S}(\gamma(t)) \in \mathcal{Y},$$

where $\mathcal{Y}$ is a Banach space. For response-alignment anchors $0 = t_0 < t_1 < \cdots < t_m = \rho$, define

$$h \coloneqq \max_{0 \le j < m} (t_{j+1} - t_j), \qquad \epsilon_{\text{nom}} \coloneqq \|r(0)\|_{\mathcal{Y}},$$
$$\epsilon_{\text{pair}} \coloneqq \max_{0 \le j \le m} \|[\mathcal{A}(\gamma(t_j)) - \mathcal{A}(\gamma(0))] - [\mathcal{S}(\gamma(t_j)) - \mathcal{S}(\gamma(0))]\|_{\mathcal{Y}}.$$

Theorem 1 (Uniform robustness between response-aligned anchors)**.** *In the anchor setup above, suppose that r is twice continuously differentiable and that $\|r''(t)\|_{\mathcal{Y}} \le M$ for every $t \in [0, \rho]$. Then the entire bracketed severity interval satisfies*

$$\sup_{0 \le t \le \rho} \|\mathcal{A}(\gamma(t)) - \mathcal{S}(\gamma(t))\|_{\mathcal{Y}} \le \epsilon_{\text{nom}} + \epsilon_{\text{pair}} + \frac{Mh^2}{8}. \tag{1}$$

Theorem 1 controls every severity between the anchors. The bound's three terms have distinct roles: ϵ_{nom} sets the starting error, ϵ_{pair} measures how well the model reproduces the anchor responses, and $Mh^2/8$ allows for error between anchors. The nominal field term remains necessary because matching response differences does not determine the starting error. For a fixed curvature bound M, halving the largest anchor gap h reduces the last term by a factor of four. The guarantee applies inside the outer anchors; extrapolation beyond the outer anchors and discontinuous perturbation paths require additional global regularity.

3 DIAGNOSTICS AND METHODS

Response diagnostics. For a pair (z, z_δ), the floor-stabilized physical-response scale is

$$\eta(z, z_\delta) \coloneqq \max\{\|\mathcal{S}(z_\delta) - \mathcal{S}(z)\|, \tau \max(\|\mathcal{S}(z)\|, \|\mathcal{S}(z_\delta)\|), \varepsilon_{\mathrm{diag}}\},$$

where $\tau \geq 0$ is the response-floor ratio and $\varepsilon_{\mathrm{diag}} > 0$ is a numerical floor in the output units. These floors prevent division by zero and limit inflation of the diagnostics when the true response is tiny. Every experiment reported here uses $\tau = 10^{-4}$ and $\varepsilon_{\mathrm{diag}} = 10^{-7}$. For QoI errors, fix the task's physical zero or reference value y_{ref} and define the corresponding target scale by

$$s_{\mathrm{QoI}}(z, z_\delta) \coloneqq \max\{d_{\mathrm{QoI}}(\mathcal{S}(z), y_{\mathrm{ref}}), d_{\mathrm{QoI}}(\mathcal{S}(z_\delta), y_{\mathrm{ref}})\}.$$

For a norm-induced QoI metric with $y_{\mathrm{ref}} = 0$, this target scale is the larger of the nominal and perturbed target norms. We report EAF, SAE, and the solution-normalized robust error (SNRE), defined by

$$\mathrm{EAF} \coloneqq \frac{\|\mathcal{A}(z_\delta) - \mathcal{A}(z)\|}{\eta(z, z_\delta)}, \quad \mathrm{SAE} \coloneqq \frac{\|[\mathcal{A}(z_\delta) - \mathcal{A}(z)] - [\mathcal{S}(z_\delta) - \mathcal{S}(z)]\|}{\eta(z, z_\delta)},$$
$$\mathrm{SNRE} \coloneqq \frac{d_{\mathrm{QoI}}(\mathcal{A}(z_\delta), \mathcal{S}(z_\delta))}{\max\{d_{\mathrm{QoI}}(\mathcal{S}(z_\delta), \mathcal{S}(z)), \tau s_{\mathrm{QoI}}(z, z_\delta), \varepsilon_{\mathrm{diag}}\}}.$$

EAF measures response magnitude, SAE measures response mismatch in direction and magnitude, and SNRE scales changed-target error by the physical response. EAF distinguishes under-response from over-response; SAE measures either as a mismatch.

Baselines. Clean training uses nominal examples; input-noise augmentation perturbs tensors but retains nominal labels. Parametric training uses examples from both the original setting and the perturbed settings selected for training. Each input is paired with the correct output for that setting: if the viscosity changes from $\nu = 0.1$ to 0.09, the training target is the solution at $\nu = 0.09$.

Finite-grid adversarial baseline. Finite-grid adversarial training evaluates the clean and all seen severities for each minibatch and backpropagates through the per-sample largest changed-target loss,

$$\mathcal{L}_{\mathrm{adv}} \coloneqq \frac{1}{B} \sum_{i=1}^{B} \max_{\delta \in \{0\} \cup \mathcal{G}_{\mathrm{seen}}} \ell(\mathcal{A}_\omega(T_\delta z_i), \mathcal{S}(T_\delta z_i)).$$

This finite-grid adversarial objective defines a matched robust-optimization baseline on the same finite grid.

PASA: augmentation with a PDE penalty. PASA trains on physically admissible perturbations, using the correct output for each perturbed setting. When a PDE residual can be computed, PASA adds a penalty for how much the predicted field violates the governing equation at that setting, following physics-informed operator learning (Li et al., 2024; Wang et al., 2021b). For Darcy, this penalty includes both the interior PDE residual and a boundary-mismatch term.

The weighted-parametric baseline samples original and perturbed settings in the same proportions as PASA and omits this physics penalty. Comparing this baseline with PASA isolates the contribution of the physics penalty.

PhySA: matching the physical response. PhySA compares the change in the prediction with the change in the reference output and penalizes response mismatches in direction and magnitude. PhySA adds the response-matching loss

$$\mathcal{L}_{\mathrm{SA}} \coloneqq \frac{\|[\mathcal{A}_\omega(z_\delta) - \mathcal{A}_\omega(z)] - [\mathcal{S}(z_\delta) - \mathcal{S}(z)]\|^2}{\max\{\|\mathcal{S}(z_\delta) - \mathcal{S}(z)\|^2, \varepsilon_{\mathrm{SA}}^2\}}.$$

Here $\varepsilon_{\mathrm{SA}} > 0$ is a training floor distinct from $\varepsilon_{\mathrm{diag}}$.

Corollary 1 (Observable PhySA specialization)**.** *Under the hypotheses of Theorem 1, let $\mathcal{L}_{\mathrm{SA}}$ be the PhySA response-matching loss defined in Sec. 3, with training floor $\varepsilon_{\mathrm{SA}}$, and let*

$$q \coloneqq \max_{0\le j\le m} \mathcal{L}_{\mathrm{SA}}(\gamma(0),\gamma(t_j)), \qquad R_{\mathrm{resp}} \coloneqq \max_{0\le j\le m} \max\{\|\mathcal{S}(\gamma(t_j)) - \mathcal{S}(\gamma(0))\|_{\mathcal{Y}}, \varepsilon_{\mathrm{SA}}\}.$$

Then $\epsilon_{\mathrm{pair}} \le R_{\mathrm{resp}}\sqrt{q}$, and hence the right-hand side of Eq. 1 can be replaced by $\epsilon_{\mathrm{nom}} + R_{\mathrm{resp}}\sqrt{q} + Mh^2/8$. If the composed maps $\mathcal{A}\circ\gamma$ and $\mathcal{S}\circ\gamma$ have second-derivative bounds M_A and M_S, respectively, one may take $M = M_A + M_S$.

Corollary 1 expresses the alignment term using the attained training loss q: $\epsilon_{\mathrm{pair}} \le R_{\mathrm{resp}}\sqrt{q}$. The response-alignment term $R_{\mathrm{resp}}\sqrt{q}$ in the robustness bound can be computed from the trained model and the reference responses at the anchors. A numerical bound for the whole interval also requires a valid curvature bound M.

ISF: penalizing the linearized PDE residual. For Burgers viscosity shifts, ISF checks the predicted change against the Burgers equation linearized around the model's nominal prediction. ISF penalizes the resulting response-scaled tangent residual

$$w_t + uw_x + wu_x - \nu w_{xx} - \Delta\nu u_{xx},$$

where $u \coloneqq \mathcal{A}_\omega(z)$, $w \coloneqq \mathcal{A}_\omega(z_\delta) - \mathcal{A}_\omega(z)$, and $\Delta\nu \coloneqq \widetilde{\nu} - \nu$ for nominal and perturbed viscosities ν and $\widetilde{\nu}$.

Among the interventions, PASA uses the matched mixture and adds a PDE residual when available, whereas PhySA and ISF use paired clean and perturbed responses at the perturbation levels used for training. In the Burgers experiments, ISF retains PhySA's clean and perturbed field losses, response-matching loss, and available PDE-residual terms, and adds $\lambda_{\tan}\mathcal{L}_{\tan}$ with $\lambda_{\tan} > 0$. The PhySA–ISF comparison therefore measures the effect of adding the tangent-residual penalty.

4 Experiments

We test whether low field error is enough to establish physical robustness and which training objectives help the model follow a change in the physical specification. Across six tasks, we measure both field accuracy and response fidelity at perturbation levels excluded from training, using matched changed-label baselines.

4.1 Setup and Reporting Conventions

Protocol. All results use five matched seeds. The Burgers tasks use PDEBench initial conditions with targets recomputed by a pseudo-spectral solver, or solution trajectories taken directly from the public multi-viscosity dataset (Takamoto et al., 2022). The Darcy tasks use targets computed by a finite-difference elliptic solver. Table 4 in App. G lists the seen and held-out severities of every task.

Backbones. The one-dimensional (1D) study includes FNO (Li et al., 2021), an operator transformer (Transformer) (Cao, 2021; Li et al., 2023a), a convolutional residual network (ConvResNet) in 1D (ConvResNet1D, or Conv1D for short) (Stachenfeld et al., 2022), and a coordinate-conditioned deep operator network (DeepONet)-style model (Chen & Chen, 1995; Lu et al., 2021a). For the Darcy tasks, we use the two-dimensional (2D) ConvResNet (ConvResNet2D, or Conv2D for short), with a padded 2D FNO (FNO2D) as an architecture control. Primary comparisons use the most accurate changed-label backbones for each task; less accurate or mismatched backbones serve as within-architecture replications. The mesh task changes only the comparison grid; the initial-condition task re-solves the changed PDE. Both reuse the checkpoints from the Burgers viscosity task, with no training on mesh or initial-condition shifts.

Gain and inference conventions. We compute the percentage reduction in held-out relative L^2 error for each matched seed and report the average of these percentages as gain. Confidence intervals (CIs) are Student-t intervals for those percentage gains; p values use two-sided paired-sample t tests on the raw matched errors. Benjamini–Hochberg (BH) q values are computed from unrounded p values over the pre-specified 36-comparison family to control the false discovery rate (FDR) (Benjamini & Hochberg, 1995).

Table 1: Condensed multi-seed summary for the primary backbones. Each seed is first averaged over the severities in the split; entries are means over five seeds. Tables 11 through 16 report all methods, backbones, and diagnostics.

Task	Method	Clean L^2	Held-out L^2	Held-out SAE
Burgers viscosity	FNO-parametric	0.0106	0.0181	0.5006
	FNO-PASA	0.0086	0.0177	0.4736
	Conv1D-parametric	0.0437	0.0474	0.9902
	Conv1D-PhySA	0.0373	0.0417	0.2477
Public Burgers	FNO-parametric	0.0238	0.2042	5.9812
	FNO-PhySA	0.0337	0.1393	1.0123
	FNO-ISF	0.0330	0.1379	1.0237
Evaluation mesh	FNO-parametric	0.0106	0.0125	0.0070
	FNO-PASA	0.0086	0.0098	0.0045
Initial condition	FNO-parametric	0.0106	0.0078	0.0217
	FNO-PASA	0.0086	0.0065	0.0200
Darcy forcing	Conv2D-parametric	0.0196	0.0266	0.1053
	Conv2D-PhySA	0.0182	0.0257	0.1043
Darcy boundary	Conv2D-parametric	0.0432	0.0634	0.0736
	Conv2D-PhySA	0.0328	0.0593	0.0689

4.2 RESULTS

Table 1 reports field and response errors for the primary backbones; Table 7 in App. I gives the paired gains and CIs. The paragraphs below describe the results and the implications for physical robustness.

Burgers viscosity shifts. The fixed-label convention of standard corruption robustness failed under viscosity shifts: the errors of the clean-trained and input-noise-trained FNO models rose from 0.0310 and 0.0413 at the nominal setting to 2.9671 and 2.9011 at a held-out severity (Table 9). Errors under changed-label recipes remained near the corresponding nominal errors. The clean-trained model's held-out EAF of 65.15 indicated over-response (Table 11). PASA reached 0.0177, a 1.9% gain over parametric FNO with $q = 0.7520$; the 6.3% PASA gain over the sampling-matched weighted baseline had unadjusted $p = 0.0339$ but $q = 0.0814$. After BH correction, neither PASA comparison met the 0.05 FDR threshold. Training on correct changed labels gives the parametric FNO a held-out field error of only 1.81%, but the parametric FNO's mean SAE of 0.50 means that the response error is still 50% of the true response magnitude (Table 11).

Response alignment: joint gains and trade-offs. On FNO, PhySA and ISF reduced SAE from PASA's 0.474 to 0.186 and 0.191 but increased field error. On ConvResNet1D, alignment improved both quantities: PhySA reduced held-out error from 0.0474 to 0.0417, a gain of 11.7% at $q = 0.0389$, and SAE from 0.990 to 0.248; ISF reached 0.0416 and 0.246. Evaluating field error and response fidelity together therefore gives a concrete guide to choosing a training objective: this joint evaluation identifies the joint gain on ConvResNet1D and makes the accuracy cost of tighter alignment on FNO explicit. Figure 4 shows how the responses vary across viscosity levels for one held-out case.

Public multi-viscosity Burgers. We use the public PDEBench solution trajectories directly as the changed targets in this stress test. On held-out $\nu \in \{0.02, 1.0\}$, PhySA and ISF lowered FNO error from 0.2042 to 0.1393 (29.0%) and 0.1379 (29.7%), with SAE 1.01 ± 0.32 and 1.02 ± 0.38 over the five seeds, rather than 5.98 ± 2.07 for the parametric baseline (Table 12). This reduction in SAE means that the models follow the public reference solutions much more faithfully when viscosity changes. The raw paired tests gave $p = 0.0287$ and $p = 0.0263$, while both BH-adjusted values were $q = 0.0814$; the field-error results were therefore suggestive but did not meet a 0.05 FDR threshold. The multiplicity family of Table 8 covers field error, so the response-mismatch gap is reported as a descriptive seedwise separation rather than as an adjusted test. The high-error case in Figure 5 makes the response difference visible: the true response is strongest early in the trajectory and fades over

time. ISF reproduces this pattern, whereas the parametric FNO predicts strong positive and negative bands even at later times.

Evaluation-mesh shifts. PASA lowered FNO evaluation-mesh error from 0.0125 to 0.0098, a 21.5% gain at $q = 0.0027$, whereas adversarial training reached 0.0160 (Table 13). Conv1D remained much less consistent across grids, at 0.2930 for parametric training and 0.2821 for PASA. This contrast agrees with the known importance of architecture for discretization consistency, that is, keeping the represented operator consistent as the grid changes (Bartolucci et al., 2023; Raonic et al., 2023). Here, PASA improves FNO's transfer to new grids, while the larger Conv1D errors point to a representation problem that the training objective alone does not fix.

Initial-condition shifts. For held-out smooth changes to u_0, PASA lowered FNO error from 0.0078 to 0.0065, a 16.2% gain at $q = 0.0024$, and improved 14.0% over the sampling-matched baseline at $q = 0.0192$ (Table 14). Finite-grid adversarial training instead increased FNO's held-out field error by 27.8% relative to parametric training, with $q = 0.0024$. PASA thus improves FNO on initial-condition changes even though training varied viscosity. PhySA reduced Conv1D error under viscosity shifts but did not reduce error under initial-condition shifts: the held-out errors were 0.0334 for Conv1D-PhySA and 0.0321 for Conv1D-parametric, with $q = 0.1440$ (Table 8).

Darcy shifts. PhySA lowered forcing-shift error from 0.0266 to 0.0257, a 3.4% gain at $q = 0.0814$, and boundary-shift error from 0.0634 to 0.0593, a 6.4% gain at $q = 0.0389$. Only the boundary comparison remained significant after BH correction. PASA did not improve either Darcy task (Table 8). Held-out SAE also moved little, from 0.1053 to 0.1043 under the forcing shift and from 0.0736 to 0.0689 under the boundary shift (Tables 15 and 16). Those Darcy baselines started much closer to the reference response than the Burgers baselines, so the alignment objective had a smaller gap to close. For Darcy, response alignment therefore refines an already accurate response, with the clearest field-error gain under changed boundary values.

Worst-severity, tail, and QoI summaries. Table 17 resolves the held-out averages into worst-severity and 90th-percentile severity-cell tail proxies. Table 18 gives the task-specific QoI summaries, and Tables 19 through 24 give the per-severity results. The positive mesh, initial-condition, and boundary field-error results persisted under the worst-severity and tail summaries. FNO-PASA on Burgers viscosity remained the main exception: the mean gain was small, and error at the worst held-out severity increased by 3.9% (Table 17). Checking only the mean field error over held-out severities would miss this deterioration at the hardest tested setting.

5 Conclusion

Our physical robustness criterion requires a surrogate to match the direction and magnitude of an admissible change in the PDE solution or comparison map. Our theoretical analysis proves that prediction error after an admissible specification change is bounded by nominal error plus accumulated sensitivity mismatch. Finite-grid, anchor, tangent, and equation-level guarantees connect this physical robustness criterion to practical diagnostics and training objectives. Our experiments provide concrete guidance for choosing training objectives across perturbations, PDEs, and backbones: PASA training improved FNO's robustness to initial-condition and evaluation-mesh shifts, while response alignment improved both response and field accuracy for ConvResNet1D viscosity shifts and ConvResNet2D boundary shifts. On public multi-viscosity Burgers data, response alignment cut the held-out response mismatch of a changed-label baseline roughly sixfold, from 5.98 to 1.01, while lowering field error by 29% at a BH-adjusted p value of $q = 0.08$. Low field error alone cannot establish physical robustness: under Burgers viscosity shifts, the parametric FNO achieved a held-out field error below 2%, which would ordinarily be reported as success, yet the parametric FNO's response error was half the magnitude of the true response. We should therefore evaluate SciML models by how faithfully the model predictions change with the governing system, while also requiring nominal accuracy.

AI USE STATEMENT

Generative AI tools were used as assistive tools for writing and editing, mathematical proofreading and consistency checking, and code development and debugging. Any AI-assisted outputs incorporated into this work were reviewed and validated by the authors as appropriate. The authors take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

Sec. 4.1 gives the evaluation and reporting protocol. App. G documents per-task data generation, severity grids, splits, backbones, and shared optimization and hyperparameters (Table 6); App. F each method's diagnostics and objective; App. H seeds and paired inference behind all gains, CIs, and q values; and App. I the per-severity, worst-cell, and QoI tables summarized in Sec. 4. App. E proves the statements in the main text, App. C, and App. D; App. L states experimental scope.

REFERENCES

Abolaji D. Adesoji and Pin-Yu Chen. Evaluating the Adversarial Robustness for Fourier Neural Operators. *CoRR*, abs/2204.04259, 2022.

Mark Ainsworth and J Tinsley Oden. A posteriori error estimation in finite element analysis. *Computer methods in applied mechanics and engineering*, 142(1-2):1–88, 1997.

Anastasios N. Angelopoulos and Stephen Bates. Conformal Prediction: A Gentle Introduction. *Found. Trends Mach. Learn.*, 16(4):494–591, 2023.

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. *CoRR*, abs/1907.02893, 2019.

Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019.

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.

A. B. Bakushinskii. Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion. *USSR Computational Mathematics and Mathematical Physics*, 24(4):181–182, 1984.

Mislav Balunovic, Maximilian Baader, Gagandeep Singh, Timon Gehr, and Martin T. Vechev. Certifying Geometric Robustness of Neural Networks. In *NeurIPS*, pp. 15287–15297, 2019.

Francesca Bartolucci, Emmanuel de Bézenac, Bogdan Raonic, Roberto Molinaro, Siddhartha Mishra, and Rima Alaifari. Representation Equivalent Neural Operators: a Framework for Alias-free Operator Learning. In *NeurIPS*, 2023.

Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic Differentiation in Machine Learning: a Survey. *J. Mach. Learn. Res.*, 18:153:1–153:43, 2017.

Abdolmehdi Behroozi, Chaopeng Shen, and Daniel Kifer. Sensitivity-Constrained Fourier Neural Operators for Forward and Inverse Problems in Parametric Differential Equations. 2025.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, 2010.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Peter Benner, Serkan Gugercin, and Karen Willcox. A Survey of Projection-Based Model Reduction Methods for Parametric Dynamical Systems. *SIAM Rev.*, 57(4):483–531, 2015.

Tom Beucler, Michael Pritchard, Stephan Rasp, Jordan Ott, Pierre Baldi, and Pierre Gentine. Enforcing analytic constraints in neural networks emulating physical systems. *Physical review letters*, 126(9):098302, 2021.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for Variance in Machine Learning Benchmarks. In *MLSys*, 2021.

Johannes Brandstetter, Max Welling, and Daniel E. Worrall. Lie Point Symmetry Data Augmentation for Neural PDE Solvers. In *ICML*, volume 162, pp. 2241–2256, 2022a.

Johannes Brandstetter, Daniel E. Worrall, and Max Welling. Message Passing Neural PDE Solvers. In *ICLR*, 2022b.

Martin Burger and Stanley Osher. Convergence rates of convex variational regularization. *Inverse Problems*, 20(5):1411–1421, 2004.

Shuhao Cao. Choose a Transformer: Fourier or Galerkin. In *NeurIPS*, pp. 24924–24940, 2021.

Tianping Chen and Hong Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Trans. Neural Networks*, 6(4):911–917, 1995.

Albert Cohen and Ronald A. DeVore. Approximation of high-dimensional parametric PDEs. *Acta Numer.*, 24:1–159, 2015.

Jeremy Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing. In *ICML*, volume 97, pp. 1310–1320, 2019.

Wojciech M. Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev Training for Neural Networks. In *NIPS*, pp. 4278–4287, 2017.

Heinz W. Engl, Karl Kunisch, and Andreas Neubauer. Convergence rates for Tikhonov regularisation of non-linear ill-posed problems. *Inverse Problems*, 5(4):523–540, 1989.

Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and Its Applications*. Kluwer Academic Publishers, Dordrecht, 1996.

Lawrence C Evans. *Partial differential equations*, volume 19. 2022.

Takashi Furuya, Davide Murari, and Carola-Bibiane Schönlieb. Approximation Theory for Lipschitz Continuous Transformers. *arXiv preprint arXiv:2602.15503*, 2026.

Vignesh Gopakumar, Ander Gray, Joel Oskarsson, Lorenzo Zanisi, Daniel Giles, Matt J Kusner, Stanislas Pamela, and Marc Peter Deisenroth. Uncertainty quantification of surrogate models using conformal prediction. *Machine Learning: Science and Technology*, 7(1):015025, 2026.

Somdatta Goswami, Katiana Kontolati, Michael D. Shields, and George Em Karniadakis. Deep transfer operator learning for partial differential equations under conditional shift. *Nat. Mac. Intell.*, 4(12):1155–1164, 2022.

Ishaan Gulrajani and David Lopez-Paz. In Search of Lost Domain Generalization. In *ICLR*, 2021.

Jayesh K. Gupta and Johannes Brandstetter. Towards Multi-spatiotemporal-scale Generalized PDE Modeling. *Trans. Mach. Learn. Res.*, 2023.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *ICLR*, 2019.

Maximilian Herde, Bogdan Raonic, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient Foundation Models for PDEs. In *NeurIPS*, 2024.

Jan S.. Hesthaven, Gianluigi Rozza, and Benjamin Stamm. *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*. 2016.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Bernd Hofmann, Barbara Kaltenbacher, Christiane Pöschl, and Otmar Scherzer. A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. *Inverse Problems*, 23(3):987–1010, 2007.

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.

Matthieu Kirchmeyer, Yuan Yin, Jérémie Donà, Nicolas Baskiotis, Alain Rakotomamonjy, and Patrick Gallinari. Generalizing to New Physical Systems via Context-Informed Dynamics Model. In *ICML*, volume 162, pp. 11283–11301, 2022.

Dmitrii Kochkov, Jamie A Smith, Ayya Alieva, Qing Wang, Michael P Brenner, and Stephan Hoyer. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21):e2101784118, 2021.

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 S. 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 *ICML*, volume 139, pp. 5637–5664, 2021.

Nikola B. Kovachki, Samuel Lanthaler, and Siddhartha Mishra. On Universal Approximation and Error Bounds for Fourier Neural Operators. *J. Mach. Learn. Res.*, 22:290:1–290:76, 2021.

Nikola B. Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Neural Operator: Learning Maps Between Function Spaces With Applications to PDEs. *J. Mach. Learn. Res.*, 24:89:1–89:97, 2023.

Aditi S. Krishnapriyan, Amir Gholami, Shandian Zhe, Robert M. Kirby, and Michael W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *NeurIPS*, pp. 26548–26560, 2021.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Samuel Lanthaler, Siddhartha Mishra, and George E Karniadakis. Error estimates for deeponets: A deep learning framework in infinite dimensions. *Transactions of Mathematics and its Applications*, 6(1):tnac001, 2022.

Zijie Li, Kazem Meidani, and Amir Barati Farimani. Transformer for Partial Differential Equations' Operator Learning. *Trans. Mach. Learn. Res.*, 2023a.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Fourier Neural Operator for Parametric Partial Differential Equations. In *ICLR*, 2021.

Zongyi Li, Daniel Zhengyu Huang, Burigede Liu, and Anima Anandkumar. Fourier Neural Operator with Learned Deformations for PDEs on General Geometries. *J. Mach. Learn. Res.*, 24:388:1–388:26, 2023b.

Zongyi Li, Nikola B. Kovachki, Christopher B. Choy, Boyi Li, Jean Kossaifi, Shourya Prakash Otta, Mohammad Amin Nabian, Maximilian Stadler, Christian Hundt, Kamyar Azizzadenesheli, and Animashree Anandkumar. Geometry-Informed Neural Operator for Large-Scale 3D PDEs. In *NeurIPS*, 2023c.

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.

Jacques-Louis Lions. Optimal control of systems governed by partial differential equations. *(No Title)*, 1971.

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. *Nat. Mach. Intell.*, 3(3):218–229, 2021a.

Lu Lu, Raphaël Pestourie, Wenjie Yao, Zhicheng Wang, Francesc Verdugo, and Steven G. Johnson. Physics-Informed Neural Networks with Hard Constraints for Inverse Design. *SIAM J. Sci. Comput.*, 43(6):B1105–B1132, 2021b.

Lu Lu, Xuhui Meng, Shengze Cai, Zhiping Mao, Somdatta Goswami, Zhongqiang Zhang, and George Em Karniadakis. A comprehensive and fair comparison of two neural operators (with practical extensions) based on fair data. *Computer Methods in Applied Mechanics and Engineering*, 393:114778, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*, 2018.

Michael McCabe, Bruno Régaldo-Saint Blancard, Liam Holden Parker, Ruben Ohana, Miles D. Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Géraud Krawezik, François Lanusse, Mariel Pettee, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. Multiple Physics Pretraining for Spatiotemporal Surrogate Models. In *NeurIPS*, 2024.

Siddhartha Mishra and Roberto Molinaro. Estimates on the generalization error of physics-informed neural networks for approximating PDEs. *IMA Journal of Numerical Analysis*, 43(1):1–43, 2023.

Jose G. Moreno-Torres, Troy Raeder, Rocío Alaíz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognit.*, 45(1):521–530, 2012.

Davide Murari, Takashi Furuya, and Carola-Bibiane Schönlieb. Approximation Theory for 1-Lipschitz ResNets. *arXiv preprint arXiv:2505.12003*, 2025.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina Julia 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 Holden Parker, Miles D. Cranmer, and Shirley Ho. The Well: a Large-Scale Collection of Diverse Physics Simulations for Machine Learning. In *NeurIPS*, 2024.

Thomas O'Leary-Roseberry, Peng Chen, Umberto Villa, and Omar Ghattas. Derivative-Informed Neural Operator: An efficient framework for high-dimensional parametric derivative learning. *J. Comput. Phys.*, 496:112555, 2024.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. FiLM: Visual Reasoning with a General Conditioning Layer. In *AAAI*, 2018.

Apostolos F. Psaros, Xuhui Meng, Zongren Zou, Ling Guo, and George Em Karniadakis. Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *J. Comput. Phys.*, 477:111902, 2023.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.*, 378:686–707, 2019.

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 *NeurIPS*, 2023.

Andrew Slavin Ross and Finale Doshi-Velez. Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing Their Input Gradients. In *AAAI*, pp. 1660–1669, 2018.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks. In *ICLR*, 2020.

Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global sensitivity analysis: the primer*. 2008.

Otmar Scherzer, Markus Grasmair, Harald Grossauer, Markus Haltmeier, and Frank Lenzen. *Variational Methods in Imaging*, volume 167 of *Applied Mathematical Sciences*. Springer, New York, 2009.

Lennon Shikhman. Diagnosing Failure Modes of Neural Operators Across Diverse PDE Families. *Trans. Mach. Learn. Res.*, 2026.

Suraj Srinivas and François Fleuret. Knowledge Transfer with Jacobian Matching. In *ICML*, volume 80, pp. 4730–4738, 2018.

Kimberly L. Stachenfeld, Drummond Buschman Fielding, Dmitrii Kochkov, Miles D. Cranmer, Tobias Pfaff, Jonathan Godwin, Can Cui, Shirley Ho, Peter W. Battaglia, and Alvaro Sanchez-Gonzalez. Learned Simulators for Turbulence. In *ICLR*, 2022.

Shashank Subramanian, Peter Harrington, Kurt Keutzer, Wahid Bhimji, Dmitriy Morozov, Michael W. Mahoney, and Amir Gholami. Towards Foundation Models for Scientific Machine Learning: Characterizing Scaling and Transfer Behavior. In *NeurIPS*, 2023.

Yifei Sun. Towards Universal Solvers: Using PGD Attack in Active Learning to Increase Generalizability of Neural Operators as Knowledge Distillation from Numerical PDE Solvers. *CoRR*, abs/2510.18989, 2025.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In *ICML*, volume 119, pp. 9229–9248, 2020.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An Extensive Benchmark for Scientific Machine Learning. In *NeurIPS*, 2022.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IROS*, pp. 23–30, 2017.

Kiwon Um, Robert Brand, Yun (Raymond) Fei, Philipp Holl, and Nils Thuerey. Solver-in-the-Loop: Learning from Differentiable Physics to Interact with Iterative PDE-Solvers. In *NeurIPS*, 2020.

Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and Mitigating Gradient Flow Pathologies in Physics-Informed Neural Networks. *SIAM J. Sci. Comput.*, 43(5):A3055–A3081, 2021a.

Sifan Wang, Hanwen Wang, and Paris Perdikaris. Learning the solution operator of parametric partial differential equations with physics-informed DeepONets. *Science advances*, 7(40):eabi8605, 2021b.

Sifan Wang, Shyam Sankaran, Hanwen Wang, and Paris Perdikaris. An Expert's Guide to Training Physics-informed Neural Networks. *CoRR*, abs/2308.08468, 2023.

Liu Yang, Siting Liu, Tingwei Meng, and Stanley J Osher. In-context operator learning with data prompts for differential equation problems. *Proceedings of the National Academy of Sciences*, 120 (39):e2310142120, 2023.

Yuan Yin, Matthieu Kirchmeyer, Jean-Yves Franceschi, Alain Rakotomamonjy, and Patrick Gallinari. Continuous PDE Dynamics Forecasting with Implicit Neural Representations. In *ICLR*, 2023.

Jeremy Yu, Lu Lu, Xuhui Meng, and George Em Karniadakis. Gradient-enhanced physics-informed neural networks for forward and inverse PDE problems. *Computer Methods in Applied Mechanics and Engineering*, 393:114823, 2022.

Kai Zhang, Wangmeng Zuo, and Lei Zhang. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018.

Min Zhu, Handi Zhang, Anran Jiao, George Em Karniadakis, and Lu Lu. Reliable extrapolation of deep neural operators informed by physics or sparse observations. *Computer Methods in Applied Mechanics and Engineering*, 412:116064, 2023.

GUIDE TO THE APPENDICES

Further results and details. The supporting material is collected in the appendices. App. G records the admissibility checks and severity grids used in the experiments, and App. A lists the symbols. App. C specifies the local coordinate assumptions and gives the additional certificates. Corollary 1 in Sec. 3 expresses the anchor-alignment term through the attained PhySA loss. The finite-grid result (Theorem C.1) accounts for untested severities and test-sample uncertainty. The tangent results (Theorem C.2 and Proposition C.1) turn residuals evaluated at the model state into changed-target error bounds, without differentiating through a solver. The Burgers and Darcy results give equation-level response and residual bounds, and the finite-difference result (Proposition C.4) relates the diagnostics to sensitivity mismatch. App. D extends the pathwise bound to regularized inverse references and separates fidelity to that reference from inverse stability and bias (Corollary D.1). All proofs are collected in App. E.

Further details. App. F explains near-null responses and the finite-difference interpretation. The appendix also connects ISF to the tangent certificates, states the initial-value constraint, and gives the combined objective, loss weights, and evaluation protocol. App. C.3 states the conditions under which PDE residuals bound changed-target error.

Further experimental details. App. G specifies data generation, admissibility checks, severity grids, data splits, architectures, losses, and optimization settings. App. H gives the seedwise inference definitions and limitations. App. I collects the full results, and App. K reports the within-architecture replications. App. C.3 states the assumptions behind the equation-level certificates and explains how the certificates apply to the experiments.

A LIST OF NOTATION

Table 2 lists the symbols that recur in the paper: the specification and admissible-perturbation model, the target maps and risk functionals, the local, tangent, and regularized-inverse certificates, the Burgers and Darcy instantiations, and the response diagnostics and training objectives. Symbols used only inside a single result or proof are defined at the point of use.

Table 2: List of notation for the symbols used repeatedly in the paper.

Symbol	Meaning
Specification and admissible perturbations	
z	PDE problem specification
$\Omega, \mathcal{L}$	Spatial domain and differential operator
a, f, u_0, g, θ	Coefficient, forcing, initial, boundary data, parameters
$\mathcal{M}, \mathcal{O}$	Discretization and observation maps
$\mathcal{Z}_{\mathrm{phys}}$	Set of physically admissible specifications
C, G, W	Equality, inequality, and well-posedness constraints
T_δ, δ	Perturbation map and perturbation, $z_\delta = T_\delta z$
$\rho, \Delta_{\mathrm{phys}}(z;\rho)$	Perturbation budget and admissible perturbation set
Target maps, risk, and evaluation	
$\mathcal{S}$	True target map: changed solution or observation
$\mathcal{A}_\omega$	Learned surrogate with parameters ω
$\mathcal{Z}_{\mathrm{inv}}, \mathcal{X}$	Inverse-specification set and reconstruction space
$x_t^\dagger$	Selected exact solution for the exact specification $\gamma^\dagger(t)$
$\mathcal{S}_{\alpha_{\mathrm{reg}}}, \mathcal{A}_{\alpha_{\mathrm{reg}}}$	Regularized inverse reference and learned surrogate
$C_{\alpha_{\mathrm{reg}}}, b_{\alpha_{\mathrm{reg}}}, \sigma$	Inverse stability, bias, and specification-discrepancy terms
$\beta_{\alpha_{\mathrm{reg}}}(\sigma)$	Reference-to-exact discrepancy of the regularized inverse
$\mathcal{Y}$	Output space of the reported field or QoI norm
d_{QoI}	Metric in the common comparison space
$\mathcal{R}^{\mathrm{wc}}_{\Delta,\rho}$	Worst-case physical robustness risk
$\mathcal{G}, \mathcal{G}_{\mathrm{seen}}, \mathcal{G}_{\mathrm{held}}$	Severity grid with seen and held-out splits
$\widehat{\mathcal{R}}_{\mathrm{split}}, E_k$	Split-mean estimator and severity envelope
Θ, d_Θ	Compact severity set and severity metric
$\ell_z(\theta)$	Changed-target loss of z at severity θ
Local decomposition and tangent certificates	
$\mathcal{P}$,$\mathcal{U}, \mathcal{V}$	Parameter, state, and residual Banach spaces
$D\mathcal{A}, D\mathcal{S}$	Fréchet derivatives (sensitivities) of the two maps
$\gamma(t)$	Admissible C^1 path in a specification chart
$\epsilon_{\mathrm{nom}}, \epsilon_{\mathrm{sens}}$	Nominal error and sensitivity-mismatch bound
ϵ_{pair}	Paired response mismatch at the alignment anchors
$F, D_u F, D_p F$	PDE residual operator with state and parameter derivatives
p_t, v	Parameter path $p_0 + tv$ and unit perturbation direction
a_t, u_t, e_t	Model output, true solution, and error $a_t - u_t$
$\widetilde{\tau}(t), \widetilde{q}_h$	Model-state tangent residual and response-scaled residual
L_t, C_{stab}	Linearized PDE operator and inverse-stability bound
Burgers and Darcy instantiations	
$\nu, \widetilde{\nu}, \Delta\nu$	Nominal viscosity, perturbed viscosity, and viscosity gap
q_b	Primal Burgers residual of an approximate trajectory b
w	Model response $\mathcal{A}_\omega(z_\delta) - \mathcal{A}_\omega(z)$
λ, Λ	Uniform lower and upper Darcy ellipticity bounds
V, V^*	Energy space $H_0^1(\Omega)$ and dual space
$\mathfrak{r}_{\widehat{u}}$	Darcy variational residual of a surrogate field $\widehat{u}$

Table 2: List of notation (continued).

Symbol	Meaning
Response diagnostics and training objectives	
$\eta(z, z_\delta)$	Floor-stabilized physical-response scale
$\tau, \varepsilon_{\mathrm{diag}}$	Response-floor ratio and diagnostic floor
$s_{\mathrm{QoI}}, y_{\mathrm{ref}}$	QoI target scale and physical reference value
EAF	Model-response magnitude relative to η
SAE	Normalized response mismatch against $\mathcal{S}$
SNRE	Changed-target error over the physical-response scale
$\mathcal{L}_{\mathrm{SA}}, \varepsilon_{\mathrm{SA}}$	PhySA response-matching loss and training floor
$\mathcal{L}_{\mathrm{PDE}}, \mathcal{L}_{\tan}$	PDE-residual and tangent-residual penalties
$\mathcal{L}_{\mathrm{adv}}$	Finite-grid worst-severity adversarial loss
$\alpha, \lambda_{\mathrm{SA}}, \lambda_{\mathrm{PDE}}, \lambda_{\tan}$	Weights on the perturbed-target and penalty terms

Table 3: Positioning against the closest lines of work. Moving target means a reference recomputed at the perturbed setting. Admissible means that physical constraints decide which perturbations are allowed. Split means dividing severities into training and held-out levels. Certificate names the guaranteed quantity. Response records whether the model response is compared with the reference in direction and magnitude; supervised marks matching during training without evaluation at held-out severities. A tinted ✓, a ✗, and – mark a property that is present, absent, or not applicable.

Line of work	Moving target	Admissible	Split	Certificate	Response
Corruption certification (Cohen et al., 2019; Balunovic et al., 2019)	✗	✗	✗	invariance	✗
Changed-reference attacks (Adesoji & Chen, 2022; Sun, 2025)	✓	✗	✗	none	✗
Sensitivity supervision (O'Leary-Roseberry et al., 2024; Behroozi et al., 2025)	–	✗	✗	none	✓ supervised
Symmetry augmentation (Brandstetter et al., 2022a)	✓	✓ exact	✗	none	✗
Shift benchmarks (Takamoto et al., 2022; Shikhman, 2026)	✓	✗	✗	none	✗
This paper	✓	✓	✓	fidelity	✓

B Related Work and Positioning

This appendix positions the paper against the literatures the paper draws on and departs from, comparing the evaluation criterion used in each such literature with the physical robustness criterion of Sec. 2.1. Figure 3 summarizes the perturbation and evidence taxonomy, Table 3 places the closest lines of work side by side, and the paragraphs below treat each line in full.

Changed-reference and sensitivity-regularized precedents. Adesoji & Chen (2022) and Sun (2025) recompute a solver reference at perturbed operator inputs, while Behroozi et al. (2025) regularize solution sensitivities to equation parameters. Our setting first constrains perturbations by physical admissibility, then grades the direction and magnitude of the response at held-out severities and supplies certificates for that moving target.

Regularized inverse maps and Lipschitz models. Ill-posed inverse relations motivate replacing the discontinuous exact inverse by a family of stable regularized maps (Engl et al., 1996). Architecture-side Lipschitz constraints provide complementary control: Murari et al. (2025) prove approximation guarantees for 1-Lipschitz residual networks, and Furuya et al. (2026) establish an analogous Lipschitz-constrained approximation theory for transformers. Such constraints control model amplification, whereas physical response fidelity asks the model response to match the chosen reference response; a small model Lipschitz constant alone does not establish that alignment. Corollary D.1 makes the complementary inverse-specific stability and regularization-bias terms explicit.

Operator learning and parametrized PDEs. Neural operators learn maps from function-space inputs to solution fields. Representative architectures include the FNO (Li et al., 2021), DeepONet (Lu et al., 2021a), geometry-aware Fourier and graph operators (Li et al., 2023b;c), and attention-based operator learners (Cao, 2021; Li et al., 2023a); Kovachki et al. (2023) give a broader formulation. These models make coefficients, forcing terms, grids, or geometries part of the supervised input. DeepONet's branch–trunk factorization also instantiates the classical universal approximation result for nonlinear operators (Chen & Chen, 1995).

Operator learning is the learned counterpart of the classical parametrized-PDE problem. Approximation theory characterizes parameter dependence (Cohen & DeVore, 2015), while projection-based and certified reduced-basis methods build surrogates for an entire parameter range (Benner et al., 2015;

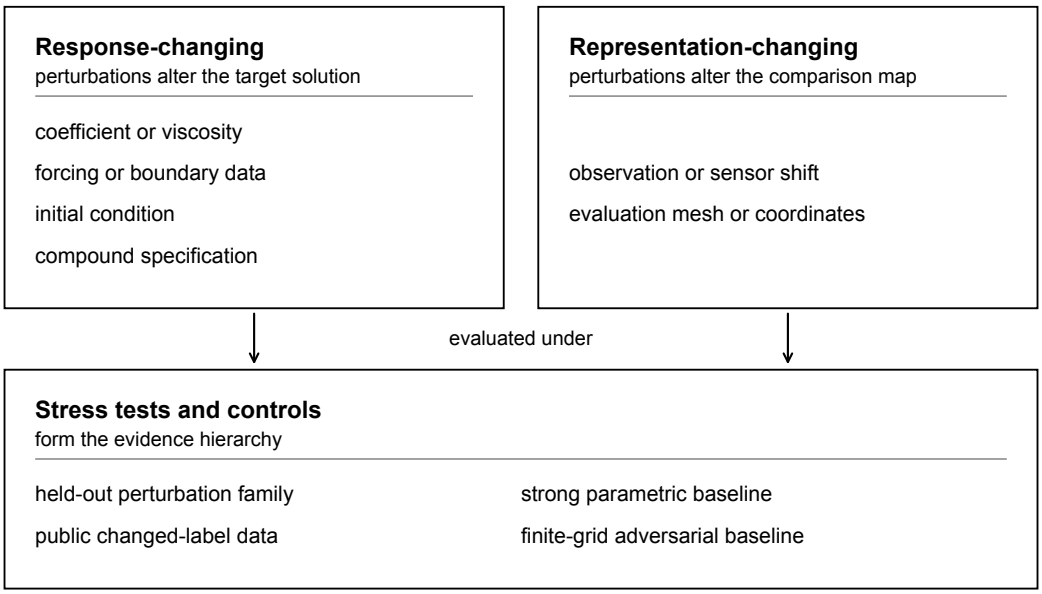

Figure 3: Perturbation and evidence taxonomy. Response-changing shifts alter the solution; representation-changing shifts alter the observation, mesh, or comparison map. Primary comparisons use accurate changed-label baselines, while secondary backbones diagnose architecture dependence.

Hesthaven et al., 2016). Neural-operator theory instead provides universality and approximation bounds for a fixed target operator (Kovachki et al., 2021; Lanthaler et al., 2022).

Physics-informed and hybrid variants add residual, rollout, or graph-structured biases, including physics-informed neural networks (PINNs) (Raissi et al., 2019), physics-informed neural operators and DeepONets (Li et al., 2024; Wang et al., 2021b), and message-passing PDE solvers (Brandstetter et al., 2022b). Residual objectives can introduce optimization failures (Krishnapriyan et al., 2021), so we treat the physics residual as an intervention to ablate rather than a default remedy. Discretization consistency is a separate concern that can be addressed by controlling aliasing (Bartolucci et al., 2023), through convolutional operator design (Raonic et al., 2023), or with grid-free representations (Yin et al., 2023). Our evaluation-mesh task probes discretization consistency. Unlike the usual fixed-operator test criterion, however, physical robustness compares the model response with the response of a solution or observation operator under a specified admissible change.

Transfer and adaptation across specifications. Several studies evaluate transfer after changing a physical specification. These transfer studies cover coefficients and parameters (Subramanian et al., 2023), pretrained multi-physics operators (Herde et al., 2024), joint pretraining across physical systems (McCabe et al., 2024), and learned corrections tested at new forcings and Reynolds numbers (Kochkov et al., 2021). Each study scores accuracy at the new specification. We instead split one admissible perturbation family into seen and held-out severities and score the response along that family. This severity split is extrapolative in the sense studied by Zhu et al. (2023), but the resulting evaluation measures the response to a controlled change rather than accuracy at an isolated unseen input.

Adaptation methods consume information from the new specification. Examples include task-specific fine-tuning (Goswami et al., 2022), environment context vectors (Kirchmeyer et al., 2022), demonstrations for in-context operator learning (Yang et al., 2023), and parameter updates at test time (Sun et al., 2020). Our protocol instead evaluates one fixed model at held-out severities, without adaptation, demonstrations, or per-severity context estimation.

Benchmarks and data protocols. PDEBench standardizes supervised evaluation across equations and solvers (Takamoto et al., 2022); PDEArena emphasizes multiscale dynamics and architecture comparisons (Gupta & Brandstetter, 2023); and The Well broadens the coverage to a large multi-

system archive (Ohana et al., 2024). These resources provide nominal specifications and, in some cases, several parameter values. These benchmark resources alone do not define an admissible perturbation family with seen and held-out severities.

Shikhman (2026) directly stress-tests operator architectures under changes in parameters, conditions, resolution, and rollout horizon. The study finds that in-distribution accuracy does not predict robustness and that architecture rankings depend on the equation class, consistent with the regime dependence in Sec. 4.2. The architecture sweep measures degradation through field error; our evaluation additionally fixes admissibility and severity splits and grades response direction and magnitude against strong changed-label baselines. We reuse PDEBench Burgers initial conditions and public multi-viscosity trajectories. For the other tasks, the same split-and-solver protocol supplies a changed target for every perturbed specification.

Robustness, augmentation, and distribution shift. Classical adversarial and corruption robustness typically assumes that a perturbation preserves the label (Madry et al., 2018; Hendrycks & Dietterich, 2019). Randomized smoothing certifies constancy in a norm ball (Cohen et al., 2019), and transformation certificates extend this constancy guarantee to parametrized families (Balunovic et al., 2019). Our physical robustness certificates require fidelity to a target that moves, so invariance can be a failure.

Changed-reference work in operator learning is especially close. Under an ℓ_∞ input attack, Adesoji & Chen (2022) recompute the solver reference at the perturbed input. Sun (2025) puts that solver inside the adversarial training loop. Both studies therefore use moving targets. Our protocol adds physical admissibility, seen and held-out severity levels, and explicit grading of response direction and magnitude. These additions distinguish under-response from over-response even when worst-case changed-target errors are similar.

Domain randomization (Tobin et al., 2017) is close to our parametric baseline, while our finite-grid adversarial baseline parallels a worst-group objective (Sagawa et al., 2020). Here, however, every severity has a corresponding target $\mathcal{S}(z_\delta)$. Lie point symmetry augmentation also moves inputs and targets together (Brandstetter et al., 2022a), but only within the exact symmetry group of a fixed PDE. An admissible specification change can leave that orbit and therefore requires a recomputed or streamed target.

Distributionally robust and domain-generalization benchmarks usually retain the target attached to each example under distribution shift (Koh et al., 2021; Gulrajani & Lopez-Paz, 2021). Classical bounds for these distribution-shift settings presume that one labeling rule can remain approximately correct across domains (Ben-David et al., 2010). Invariant risk minimization makes that premise explicit by seeking a common optimal predictor (Arjovsky et al., 2019). Our setting is closer to controlled concept shift than covariate shift (Moreno-Torres et al., 2012): the correct predictor changes with a known physical severity. This dependence on physical severity motivates a split of severities into seen and held-out levels rather than a detect-and-adapt protocol.

Sensitivity matching and derivative-informed learning. Adjoint and implicit-function methods differentiate PDE solutions with respect to parameters (Lions, 1971; Baydin et al., 2017); global sensitivity analysis studies the same dependence through variance decompositions (Saltelli et al., 2008). Those traditions characterize one simulator. Our diagnostics instead compare the sensitivity of a learned surrogate with the sensitivity of the solution or observation operator the surrogate replaces. Input-gradient penalties (Ross & Doshi-Velez, 2018) and teacher–student Jacobian matching (Srinivas & Fleuret, 2018) concern a fixed task's input tensor. Here the differentiated object is the specification that defines the task.

Sobolev training supervises target derivatives as well as values (Czarnecki et al., 2017). In our setting, the derivative lies along an admissible specification path and is observed through finite differences at severity anchors. Related SciML methods differentiate PDE residuals (Yu et al., 2022), supervise the solution operator's parametric Jacobian (O'Leary-Roseberry et al., 2024), or train against a differentiable solver's trajectories (Um et al., 2020). Our objectives require only solved specifications at finitely many severities and are evaluated at held-out severities. The shared principle is that a learned solver's directional derivative should align with the physical reference's directional derivative.

Most directly, Behroozi et al. (2025) constrain an FNO's sensitivity to equation parameters. The sensitivity-constrained FNO targets forward and inverse parameter estimation, whereas our evaluation targets robustness over an admissible perturbation set with severity splits and certificates. The sensitivity-constrained FNO uses parametric derivative supervision; our objective uses finite response differences at solved anchors. Propositions C.4 and C.1 certify this finite-difference construction. Theorem C.2 allows linearization at the model state without a differentiable simulator, and Theorem 1 converts alignment at finitely many anchors into control between the anchors.

C Further Theoretical Results

This appendix states the finite-grid, tangent-residual, equation-level, and finite-difference results summarized in Sec. 2.1 and in the appendix guide. The anchored-response result is stated in Sec. 2.3, and its observable PhySA specialization (Corollary 1) is stated in Sec. 3. Proofs are given in App. E.

For the forward local statements of Secs. 2.1 and 2.2 and for the PDE results below, we use z as a coordinate in a Banach chart of $\mathcal{Z}_{\mathrm{phys}}$ and take $\mathcal{Y}$ to be the Hilbert or Banach output space induced by the reported field, energy, observation, or QoI norm. Proposition 1 and Theorem 1 use only the chart and path regularity, so App. D applies the same argument in a chart of $\mathcal{Z}_{\mathrm{inv}}$. The Banach-chart formulation does not require every specification component to be additive in the component's native representation: geometry deformations, sensor masks, mesh rules, and parameter changes enter through admissible paths in the corresponding specification chart.

If $D\mathcal{A}$ and $D\mathcal{S}$ are locally Lipschitz in that chart, the remainder term of the local decomposition of Sec. 2.2 sharpens from $o(\|\delta\|_{\mathcal{Z}})$ to $O(\|\delta\|_{\mathcal{Z}}^2)$. This sharpening holds in particular when the two maps are twice continuously differentiable with locally bounded second derivatives. The required regularity of $\mathcal{S}$ is not an extra idealization for the families studied here: smooth, and often holomorphic, dependence of the solution on coefficient, forcing, and boundary parameters is classical for parametrized elliptic problems (Cohen & DeVore, 2015), and Theorems C.3 and C.4 below give the corresponding quantitative statements for the two PDE families used in the experiments.

C.1 Finite-Grid Certificate for Continuous Physical Risk

The following result identifies conditions under which a finite admissible grid gives a quantitative certificate for the continuous perturbation family, separating deterministic severity-coverage error from finite test-set uncertainty.

Let (Θ, d_Θ) be a compact metric severity set that parametrizes a common admissible perturbation family, so that for every $\theta \in \Theta$ and for P_Z-almost every z, $T_\theta z = T_{\delta(\theta,z)} z$ for some $\delta(\theta, z) \in \Delta_{\mathrm{phys}}(z; \rho)$. For a fixed model $\mathcal{A}$, define

$$\ell_z(\theta) \coloneqq d_{\mathrm{QoI}}(\mathcal{A}(T_\theta z), \mathcal{S}(T_\theta z)),$$

and, for any finite $\mathcal{G} \subset \Theta$, write

$$\mathcal{R}_\Theta^{\mathrm{wc}} \coloneqq \mathbb{E}_z \left[\sup_{\theta \in \Theta} \ell_z(\theta) \right], \qquad \mathcal{R}_{\mathcal{G}}^{\mathrm{wc}} \coloneqq \mathbb{E}_z \left[\max_{g \in \mathcal{G}} \ell_z(g) \right]$$

for the continuous-family and finite-grid risks, respectively.

Theorem C.1 (Finite-sample grid certificate for continuous physical risk)**.** *Suppose that $(z, \theta) \mapsto \ell_z(\theta)$ is jointly measurable and that constants $B, L < \infty$ satisfy*

$$0 \le \ell_z(\theta) \le B, \qquad |\ell_z(\theta) - \ell_z(\theta')| \le L d_\Theta(\theta, \theta')$$

for P_Z-almost every z and all $\theta, \theta' \in \Theta$. Let $\mathcal{G} \subset \Theta$ be a finite h-net: for every $\theta \in \Theta$, there is a $g \in \mathcal{G}$ with $d_\Theta(\theta, g) \le h$. Fix $\mathcal{A}$ before drawing the test sample, or else condition on independent training data used to select $\mathcal{A}$. Let $z_1, \dots, z_n$ be independent draws from P_Z, and define

$$\widehat{\mathcal{R}}_{\mathcal{G},n}^{\mathrm{wc}} \coloneqq \frac{1}{n} \sum_{i=1}^{n} \max_{g \in \mathcal{G}} \ell_{z_i}(g).$$

For every $\alpha \in (0, 1)$, with probability at least $1 - \alpha$ over the test sample,

$$\mathcal{R}_\Theta^{\mathrm{wc}} \le \widehat{\mathcal{R}}_{\mathcal{G},n}^{\mathrm{wc}} + Lh + B\sqrt{\frac{\log(1/\alpha)}{2n}}. \tag{2}$$

The Lh term in Eq. 2 is deterministic discretization error, whereas the final term is sampling error. The theorem also fixes the relevant order of operations for the parametrized risk: the grid maximum is taken per specification before averaging. The maximum of the severity envelopes is a useful descriptive lower proxy because $\max_g \mathbb{E}[\ell_z(g)] \le \mathbb{E}[\max_g \ell_z(g)]$, but that envelope maximum alone is not the estimator in Eq. 2.

The same ordering separates the two finite-grid summaries reported in Sec. 2.1. The split mean $\widehat{\mathcal{R}}_{\mathrm{split}}$ estimates average-over-severity error and is not the finite-grid estimator of the worst-case risk. That worst-case estimator takes the grid maximum separately for each specification and then averages, as in Theorem C.1. The order of the two operations is not a convention. Because the maximum of the severity envelopes E_k is at most the average of the per-specification grid maxima, a robustness number formed by taking the worst envelope E_k understates the worst-case risk, so a summary reported in that order is a lower proxy and not a conservative one.

The experiments use the fixed severity grid for direct model comparisons. To turn the worst-case grid estimate into a numerical upper bound for the continuous family, Eq. 2 requires bounds B and L on the loss and the loss variation with severity; the experiments do not estimate these bounds. Classical certified reduced-basis methods compute residual-based error estimates at queried parameters (Hesthaven et al., 2016), while the grid certificate quantifies the additional margins for severity coverage and test-sample uncertainty.

Order of the severity-discretization penalty. Theorem C.1 and Theorem 1 both pay for evaluating finitely many severities, and the two prices have different orders in the spacing h. The grid certificate pays Lh, where L is a Lipschitz constant of the loss in the severity coordinate, and that term is first order because a bounded first difference is all that a Lipschitz hypothesis supplies. The anchor theorem pays $Mh^2/8$, where M bounds the second derivative of the error path r, and the constant $1/8$ is the sharp one for a twice-differentiable function that is controlled at both endpoints of a subinterval. The improvement is bought by the alignment condition rather than by a weaker assumption: the anchor bound requires ϵ_{pair} to be small at every anchor, which is exactly what the response-matching objective optimizes, and that bound requires curvature control that the Lipschitz hypothesis does not. The two statements therefore do not strictly order one bound by the other. For a benchmark that can afford a fixed number of solved severities, however, the comparison is the quantitative reason to spend that budget on aligned anchors rather than on grid evaluation alone, since halving h reduces the coverage term by a factor of two in the first case and by a factor of four in the second.

C.2 Model-State Tangent Residuals

The anchor result justifies response matching along a covered 1D severity path. The next two results give the PDE-based justification for the tangent-residual objectives used in the experiments. For a physical PDE $F(p, u_p) = 0$, the implicit-function sensitivity is

$$Du_p[\delta_p] = -(D_u F(p, u_p))^{-1} D_p F(p, u_p)[\delta_p],$$

whenever the linearized operator is invertible on the constraint-compatible subspace. This implicit-function identity motivates response-matching losses and tangent residuals.

The following theorem allows the tangent residual to be evaluated at the model state, as in a trainable objective, rather than assuming access to a linearization only at the true solution. The subsequent proposition gives a finite-response version for the secants used by ISF.

Let $\mathcal{P}$, $\mathcal{U}$, and $\mathcal{V}$ be Banach spaces, let $p_t := p_0 + tv$ be an admissible path for $t \in [0, \rho]$ with $\|v\|_{\mathcal{P}} = 1$, and let $F : \mathcal{P} \times \mathcal{U} \to \mathcal{V}$ be continuously Fréchet differentiable on a neighborhood of the path. Set $u_t := \mathcal{S}(p_t)$, $a_t := \mathcal{A}(p_t)$, and $e_t := a_t - u_t$, and define

$$\widetilde{\tau}(t) := D_u F(p_t, a_t)[\dot{a}_t] + D_p F(p_t, a_t)[v] \in \mathcal{V},$$

with $\widehat{L}_t := D_u F(p_t, a_t)$, $b_t := D_p F(p_t, u_t)[v]$, and $\widehat{b}_t := D_p F(p_t, a_t)[v]$.

Theorem C.2 (PDE-stability certificate for model-state tangent residuals)**.** *In the tangent-path setup above, suppose that $F(p_t, u_t) = 0$ and that u_t and a_t are continuously differentiable in t. After a fixed lifting of any inhomogeneous essential constraints, assume that $\dot{u}_t, \dot{a}_t$ belong to a constraint-compatible Banach space $\mathcal{U}_0$ and that*

$$L_t := D_u F(p_t, u_t)|_{\mathcal{U}_0} : \mathcal{U}_0 \to \mathcal{V} \quad \textit{is bijective}, \qquad \sup_{0 \le t \le \rho} \|L_t^{-1}\|_{\mathrm{op}} \le C_{\mathrm{stab}}.$$

Assume also that constants $K_u, K_p, G < \infty$ satisfy, for every t,

$$\|D_u F(p_t, a_t) - D_u F(p_t, u_t)\|_{\mathrm{op}} \le K_u \|e_t\|_{\mathcal{U}},$$
$$\|D_p F(p_t, a_t) - D_p F(p_t, u_t)\|_{\mathrm{op}} \le K_p \|e_t\|_{\mathcal{U}}, \qquad \|\dot{a}_t\|_{\mathcal{U}} \le G.$$

With $K \coloneqq C_{\mathrm{stab}}(K_u G + K_p)$, one has

$$\sup_{0 \le t \le \rho} \|\mathcal{A}(p_t) - \mathcal{S}(p_t)\|_{\mathcal{U}} \le e^{K\rho} \left(\|\mathcal{A}(p_0) - \mathcal{S}(p_0)\|_{\mathcal{U}} + C_{\mathrm{stab}} \int_0^\rho \|\widetilde{\tau}(t)\|_{\mathcal{V}} \, dt \right). \tag{3}$$

Proposition C.1 (Finite-response residual certificate)**.** *Under the hypotheses and notation of Theorem C.2, fix $h \in (0, \rho]$. Suppose that $t \mapsto u_t$ is twice continuously differentiable on $[0, h]$ with $\|\ddot{u}_t\|_{\mathcal{U}} \le M_S$, and define*

$$\widehat{w}_h \coloneqq \frac{a_h - a_0}{h} \in \mathcal{U}_0, \qquad \widetilde{\tau}_h \coloneqq D_u F(p_0, a_0)[\widehat{w}_h] + D_p F(p_0, a_0)[v].$$

Equivalently, define the response-scaled residual

$$\widetilde{q}_h \coloneqq h\widetilde{\tau}_h = D_u F(p_0, a_0)[a_h - a_0] + D_p F(p_0, a_0)[hv].$$

If $\|\widehat{w}_h\|_{\mathcal{U}} \le G_h$, then

$$\|a_h - u_h\|_{\mathcal{U}} \le [1 + hC_{\mathrm{stab}}(K_u G_h + K_p)] \|e_0\|_{\mathcal{U}} + C_{\mathrm{stab}} \|\widetilde{q}_h\|_{\mathcal{V}} + \frac{M_S h^2}{2}. \tag{4}$$

Theorem C.2 and Proposition C.1 show that a small tangent residual is useful only together with nominal accuracy and stability of the linearized PDE. The factor C_{stab} exposes sensitivity near ill-posed or nearly singular regimes; $e^{K\rho}$ accounts for evaluating the linearization at the model state; $\widetilde{q}_h = h\widetilde{\tau}_h$ is the full finite response-scaled residual, whose interior Burgers-PDE component is used by ISF; and $M_S h^2/2$ is the finite-response consistency cost.

The hypothesis that the linearization may be taken at the model state a_t is what makes the objective implementable without a differentiable solver. A certificate stated only at the true solution u_t would require $D_u F(p_t, u_t)$ and $D_p F(p_t, u_t)$, that is, the Jacobians of the simulator, which is the access assumed by derivative-informed operator learning (O'Leary-Roseberry et al., 2024), by sensitivity-constrained operator training (Behroozi et al., 2025), and by solver-in-the-loop correction (Um et al., 2020). Theorem C.2 substitutes the model state for the true solution and charges $e^{K\rho}$ with $K = C_{\mathrm{stab}}(K_u G + K_p)$ for the substitution, so the drift is explicit rather than assumed away. Proposition C.1 completes the correspondence between what is proved and what is run, because the quantity certified there is the finite secant $(a_h - a_0)/h$ that the training loss evaluates rather than a derivative that no finite implementation forms.

C.3 Equation-Level Certificates

The preceding results are operator-level statements. The next two equation-level energy estimates and the accompanying residual propositions instantiate the framework for the two primary PDE families in the experiments, using the standard energy method for parabolic and elliptic problems (Evans, 2022). These equation-level results simultaneously control compound changes in the physical specification and convert the primal PDE residual used by PASA into a changed-target error bound.

Burgers setting. The Burgers certificates below are stated on the 1D periodic torus.

Theorem C.3 (Energy certificate for compound viscous-Burgers shifts)**.** *Let $T > 0$, let $\mathbb{T}$ be the 1D periodic torus, and let $\nu, \widetilde{\nu} > 0$. Suppose that u and $\widetilde{u}$ are strong periodic solutions on $[0, T]$ of*

$$u_t + uu_x - \nu u_{xx} = 0, \qquad \widetilde{u}_t + \widetilde{u}\widetilde{u}_x - \widetilde{\nu}\widetilde{u}_{xx} = 0,$$

with initial data u_0 and $\widetilde{u}_0$, respectively. Assume that $u_x, \widetilde{u}_x \in L^1(0, T; L^\infty(\mathbb{T}))$ and $u_x \in L^2(0, T; L^2(\mathbb{T}))$, and put $A_u(t) \coloneqq \int_0^t \|u_x(s)\|_{L^\infty} \, ds$. Then, for every $t \in [0, T]$,

$$\|\widetilde{u}(t) - u(t)\|_{L^2}^2 \le e^{A_u(t)} \left[\|\widetilde{u}_0 - u_0\|_{L^2}^2 + \frac{|\widetilde{\nu} - \nu|^2}{\widetilde{\nu}} \int_0^t \|u_x(s)\|_{L^2}^2 \, ds \right]. \tag{5}$$

Proposition C.2 (Burgers primal-residual certificate)**.** *Under the hypotheses of Theorem C.3, let v denote either exact solution above, let μ be the viscosity of that solution, and let b be any sufficiently regular periodic approximate trajectory whose primal residual $q_b := b_t + bb_x - \mu b_{xx}$ belongs to $L^2(0,T;L^2(\mathbb{T}))$. For every $\kappa > 0$,*

$$\|b(t) - v(t)\|_{L^2}^2 \le e^{A_v(t)+\kappa t}\left[\|b(0) - v(0)\|_{L^2}^2 + \kappa^{-1}\int_0^t \|q_b(s)\|_{L^2}^2\, ds\right], \tag{6}$$

where $A_v(t) := \int_0^t \|v_x(s)\|_{L^\infty}\, ds$.

Theorem C.3 is a global-in-severity estimate on any fixed time horizon on which the stated strong solutions exist: the estimate covers the simultaneous initial-condition-plus-viscosity perturbations represented by the stated hypotheses. Proposition C.2 separates initial-state error from the time-integrated primal residual. The exponential factor makes the long-horizon and low-viscosity limitation explicit rather than hiding that limitation in an unspecified solution-operator Lipschitz constant.

Darcy setting. Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain, $V := H_0^1(\Omega)$ with $\|\phi\|_V := \|\nabla\phi\|_{L^2}$, and let $\mathsf{E} : H^{1/2}(\partial\Omega) \to H^1(\Omega)$ be a bounded right inverse of the trace operator. Suppose that $a, \widetilde{a} \in L^\infty(\Omega)$ satisfy the common ellipticity bounds

$$0 < \lambda \le a(x), \widetilde{a}(x) \le \Lambda < \infty \quad \text{for almost every } x \in \Omega.$$

Let $f, \widetilde{f} \in V^*$ and $g, \widetilde{g} \in H^{1/2}(\partial\Omega)$, and let $u, \widetilde{u} \in H^1(\Omega)$ be the weak solutions with traces $g, \widetilde{g}$ of

$$-\nabla\cdot(a\nabla u) = f, \qquad -\nabla\cdot(\widetilde{a}\nabla\widetilde{u}) = \widetilde{f},$$

which exist and are unique under the stated ellipticity bounds by the Lax–Milgram theorem (Evans, 2022).

Theorem C.4 (Energy certificate for joint Darcy specification shifts)**.** *Under the preceding Darcy setting, let $H := \mathsf{E}(\widetilde{g} - g)$. The response to simultaneous coefficient, forcing, and boundary changes satisfies*

$$\|\nabla(\widetilde{u} - u)\|_{L^2} \le \lambda^{-1}\|\widetilde{f} - f\|_{V^*} + \lambda^{-1}\|\widetilde{a} - a\|_{L^\infty}\|\nabla u\|_{L^2} + \left(1 + \frac{\Lambda}{\lambda}\right)\|\nabla H\|_{L^2}. \tag{7}$$

Proposition C.3 (Darcy variational-residual certificate)**.** *Under the hypotheses of Theorem C.4, let $\widehat{u} \in H^1(\Omega)$ be any surrogate field for the perturbed specification and define the corresponding variational residual $\mathfrak{r}_{\widehat{u}} \in V^*$ by*

$$\langle \mathfrak{r}_{\widehat{u}}, \phi\rangle := \langle \widetilde{f}, \phi\rangle - \int_\Omega \widetilde{a}\nabla\widehat{u}\cdot\nabla\phi\, dx, \qquad \phi \in V.$$

Writing $\widehat{H} := \mathsf{E}(\operatorname{Tr}\widehat{u} - \widetilde{g})$, one has

$$\|\nabla(\widehat{u} - \widetilde{u})\|_{L^2} \le \lambda^{-1}\|\mathfrak{r}_{\widehat{u}}\|_{V^*} + \left(1 + \frac{\Lambda}{\lambda}\right)\|\nabla\widehat{H}\|_{L^2}. \tag{8}$$

Corollary C.1 (Boundary-compatible Darcy residual certificate)**.** *Under the hypotheses and notation of Proposition C.3, if $\operatorname{Tr}\widehat{u} = \widetilde{g}$, then*

$$\|\widehat{u} - \widetilde{u}\|_{L^2} \le C_P\lambda^{-1}\|\mathfrak{r}_{\widehat{u}}\|_{V^*},$$

where C_P is the Poincaré constant of Ω.

Theorem C.4 treats the coefficient, forcing, and Dirichlet data as one joint specification, rather than proving three unrelated single-factor statements. Eq. 7 controls the true physical response, while Proposition C.3 gives the surrogate-field estimate in Eq. 8, and Corollary C.1 gives the boundary-compatible L^2 specialization of that estimate. Proposition C.3 is the changed-specification analogue of classical residual-based a posteriori error estimation for elliptic problems (Ainsworth & Oden, 1997), and of the residual-plus-stability route by which Mishra & Molinaro (2023) bound the generalization error of a PINN at a fixed specification, with the perturbed coefficient, forcing, and boundary data in place of the nominal ones.

The general estimate in Eq. 8 also shows why a Darcy residual penalty alone is insufficient when the learned field violates the changed boundary condition: the boundary-lift term is separately necessary, and that term vanishes exactly when the changed Dirichlet data are hard-enforced rather than penalized, as in the constrained formulations of Lu et al. (2021b). The backbones evaluated here impose the boundary data softly, so the experiments report both the PDE residual and the boundary mismatch.

What the Burgers and Darcy bounds explain. The equation-level certificates in App. C.3 are upper bounds derived from the Burgers and Darcy equations. Under the stated assumptions, the energy bounds limit how much the exact solution changes after a specification shift, while the residual bounds limit prediction error using the PDE residual and initial or boundary mismatch. The viscosity term in Eq. 5 carries a factor $|\widetilde{\nu} - \nu|^2/\widetilde{\nu}$, so the bound allows a larger physical response as the perturbed viscosity decreases below the nominal value. The changed-label FNO errors in Table 9 grow fastest in the two lowest-viscosity columns. The largest measured FNO errors thus occur where the bound allows larger physical changes. The bound does not predict the size of those model errors: Eq. 5 compares two exact solutions and contains no learned prediction.

For Darcy, Eq. 8 bounds the gradient error by two terms: the PDE residual inside the domain and a boundary-lift term for disagreement with the prescribed boundary values. Even an exact solution with the wrong boundary values can have zero interior residual, so checking that residual alone does not identify the correct solution. The boundary-lift term vanishes when the predicted boundary values match the prescribed values exactly. To address both sources of error, our Darcy PASA loss includes penalties for the interior residual and boundary mismatch. These residual and boundary penalties impose soft constraints: training can leave boundary error, which must still be accounted for in Eq. 8. A small interior residual therefore gives a small upper bound only when the boundary contribution is also controlled. By contrast, PhySA directly penalizes the mismatch between the predicted change and the change in the reference solution.

Finally, the coefficient term in Eq. 7 scales with the nominal solution gradient $\|\nabla u\|_{L^2}$, so the certificate flags specifications with steep nominal solutions as potentially sensitive to coefficient changes before evaluation.

Scope of the equation-level certificates. The equation-level certificates add global energy estimates on the stated domains and horizons but require positive viscosity and integrable Burgers solution gradients, or uniform Darcy ellipticity and control of the variational residual in H^{-1}. Near-shock low-viscosity regimes, loss of ellipticity, and collocation residuals that do not control the dual norm fall outside those guarantees. The experiments do not estimate these constants, so none of the residual metrics is a numerical a posteriori certificate.

C.4 Finite-Difference Diagnostic

The following result is not equation-specific: the proposition below certifies that the response-mismatch difference quotient behind the diagnostics of Sec. 3 is a consistent estimate of the directional sensitivity mismatch of Sec. 2.2.

Proposition C.4 (Finite-difference diagnostic consistency)**.** *Let $\gamma : [0, \rho_0] \to \mathcal{Z}_{\mathrm{phys}}$ be an admissible 1D perturbation path in a Banach chart, with $\gamma(0) = z$, $\dot{\gamma}(0) = v$, and $\|v\|_{\mathcal{Z}} = 1$. Write $z_\rho := \gamma(\rho)$ for $0 < \rho \le \rho_0$. Suppose that $\mathcal{A}$ and $\mathcal{S}$ are Fréchet differentiable at z, and that $g_A := \mathcal{A} \circ \gamma$ and $g_S := \mathcal{S} \circ \gamma$ are twice continuously differentiable and satisfy $\|g_A''(t)\|_{\mathcal{Y}} \le M_A$ and $\|g_S''(t)\|_{\mathcal{Y}} \le M_S$ on $[0, \rho_0]$. Then*

$$\frac{[\mathcal{A}(z_\rho) - \mathcal{A}(z)] - [\mathcal{S}(z_\rho) - \mathcal{S}(z)]}{\rho} = (D\mathcal{A}(z) - D\mathcal{S}(z))[v] + R_\rho, \qquad \|R_\rho\|_{\mathcal{Y}} \le \frac{\rho}{2}(M_A + M_S).$$

Consequently, the response-mismatch difference quotient estimates the directional sensitivity mismatch with first-order finite-difference error.

D Extension to Regularized Inverse Problems

An ill-posed inverse problem can have a set-valued inverse relation, and even a fixed single-valued selection need not be continuous. The differentiability-based results of Sec. 2.2 and App. C therefore cannot in general be applied directly to an exact inverse. We instead use a fixed, single-valued regularized inverse as a stable reference and state separately the price of comparing that reference with a selected exact solution (Engl et al., 1996). Proofs are given in App. E.

Let $\mathcal{X}$ be a Banach reconstruction space and let $\mathcal{Z}_{\mathrm{inv}}$ be an admissible inverse-specification set whose paths lie in one Banach chart with norm $\|\cdot\|_{\mathcal{Z}}$. An inverse specification may contain measured data,

a forward operator, and an acquisition model. For a fixed regularization level $\alpha_{\text{reg}} > 0$, write

$$\mathcal{S}_{\alpha_{\text{reg}}}, \mathcal{A}_{\alpha_{\text{reg}}} : \mathcal{Z}_{\text{inv}} \to \mathcal{X}$$

for the regularized inverse reference and a learned surrogate for that fixed reference. Thus $\mathcal{S}_{\alpha_{\text{reg}}}$ is a reconstruction rule, not the exact inverse target. The subscript α_{reg} distinguishes the regularization level from the coefficient α of the generic perturbed-target loss in Sec. 3. A canonical reconstruction rule is variational regularization: for a forward operator $\mathcal{F}$ from a domain $\mathcal{C} \subset \mathcal{X}$ into an observation space $\mathcal{Y}_{\text{obs}}$ and a penalty R,

$$\mathcal{S}_{\alpha_{\text{reg}}}(y) \in \arg\min_{x \in \mathcal{C}} \{\|\mathcal{F}(x) - y\|^2_{\mathcal{Y}_{\text{obs}}} + \alpha_{\text{reg}} R(x)\}, \tag{9}$$

with a fixed selection whenever the minimizer is not unique. Here $z = y$ when the forward and acquisition models are fixed; a perturbed forward or acquisition model enters z alongside the data.

Let $\gamma^\dagger, \gamma^\sigma : [0, \rho] \to \mathcal{Z}_{\text{inv}}$ be paired exact and observed specification paths satisfying

$$\sup_{0 \le t \le \rho} \|\gamma^\sigma(t) - \gamma^\dagger(t)\|_{\mathcal{Z}} \le \sigma. \tag{10}$$

Here σ is a specification-discrepancy bound; in the fixed-operator specializations below, σ is the observation-noise level. For every t, fix a selected exact solution $x_t^\dagger \in \mathcal{X}$ associated with $\gamma^\dagger(t)$. If the inverse is non-unique, this selection is part of the estimand and is fixed in advance. When only the observation noise changes for one fixed latent object, $x_t^\dagger$ is constant and the exact response is zero.

Suppose that, for some $C_{\alpha_{\text{reg}}} < \infty$,

$$\|\mathcal{S}_{\alpha_{\text{reg}}}(\gamma^\sigma(t)) - \mathcal{S}_{\alpha_{\text{reg}}}(\gamma^\dagger(t))\|_{\mathcal{X}} \le C_{\alpha_{\text{reg}}} \|\gamma^\sigma(t) - \gamma^\dagger(t)\|_{\mathcal{Z}} \tag{11}$$

for $0 \le t \le \rho$, and define the uniform regularization bias

$$b_{\alpha_{\text{reg}}} \coloneqq \sup_{0 \le t \le \rho} \|\mathcal{S}_{\alpha_{\text{reg}}}(\gamma^\dagger(t)) - x_t^\dagger\|_{\mathcal{X}} < \infty.$$

Corollary D.1 (Pathwise bounds via a regularized inverse reference)**.** *Suppose that γ^σ is C^1, that $\mathcal{A}_{\alpha_{\text{reg}}}$ and $\mathcal{S}_{\alpha_{\text{reg}}}$ are Fréchet differentiable in the chosen chart along γ^σ, and that the compositions of the surrogate and regularized inverse reference with γ^σ are C^1. Define*

$$\begin{aligned}
\epsilon_{\text{nom},\alpha_{\text{reg}}} &\coloneqq \|\mathcal{A}_{\alpha_{\text{reg}}}(\gamma^\sigma(0)) - \mathcal{S}_{\alpha_{\text{reg}}}(\gamma^\sigma(0))\|_{\mathcal{X}}, \\
E_{\alpha_{\text{reg}}} &\coloneqq \int_0^\rho \|(D\mathcal{A}_{\alpha_{\text{reg}}}(\gamma^\sigma(t)) - D\mathcal{S}_{\alpha_{\text{reg}}}(\gamma^\sigma(t)))[\dot\gamma^\sigma(t)]\|_{\mathcal{X}}\, dt, \\
L_\gamma &\coloneqq \int_0^\rho \|\dot\gamma^\sigma(t)\|_{\mathcal{Z}}\, dt.
\end{aligned}$$

Then

$$\sup_{0 \le t \le \rho} \|\mathcal{A}_{\alpha_{\text{reg}}}(\gamma^\sigma(t)) - \mathcal{S}_{\alpha_{\text{reg}}}(\gamma^\sigma(t))\|_{\mathcal{X}} \le \epsilon_{\text{nom},\alpha_{\text{reg}}} + E_{\alpha_{\text{reg}}}, \tag{12}$$

$$\sup_{0 \le t \le \rho} \|\mathcal{A}_{\alpha_{\text{reg}}}(\gamma^\sigma(t)) - x_t^\dagger\|_{\mathcal{X}} \le \epsilon_{\text{nom},\alpha_{\text{reg}}} + E_{\alpha_{\text{reg}}} + C_{\alpha_{\text{reg}}}\sigma + b_{\alpha_{\text{reg}}}. \tag{13}$$

If, in addition,

$$\|D\mathcal{A}_{\alpha_{\text{reg}}}(\gamma^\sigma(t)) - D\mathcal{S}_{\alpha_{\text{reg}}}(\gamma^\sigma(t))\|_{\text{op}} \le \epsilon_{\text{sens},\alpha_{\text{reg}}}$$

for all t, then $E_{\alpha_{\text{reg}}} \le \epsilon_{\text{sens},\alpha_{\text{reg}}} L_\gamma$.

The stability and bias terms enter the proof of Eq. 13 only through the reference-to-exact discrepancy

$$\beta_{\alpha_{\text{reg}}}(\sigma) \coloneqq \sup_{0 \le t \le \rho} \|\mathcal{S}_{\alpha_{\text{reg}}}(\gamma^\sigma(t)) - x_t^\dagger\|_{\mathcal{X}} \le C_{\alpha_{\text{reg}}}\sigma + b_{\alpha_{\text{reg}}}, \tag{14}$$

so Eq. 13 remains true with $C_{\alpha_{\text{reg}}}\sigma + b_{\alpha_{\text{reg}}}$ replaced by $\beta_{\alpha_{\text{reg}}}(\sigma)$. Classical regularization theory bounds $\beta_{\alpha_{\text{reg}}}(\sigma)$ directly, and the nonlinear case of Proposition D.1 below uses that route without a Lipschitz constant as in Eq. 11.

Normalization and the trade-off between stability and bias. For $s, t \in [0, \rho]$, let $N_{\alpha_{\mathrm{reg}}}(s,t)$ be the response mismatch against the regularized reference and let $N^{\dagger}(s,t)$ be the mismatch against the selected exact response:

$$N_{\alpha_{\mathrm{reg}}}(s,t) \coloneqq \|[\mathcal{A}_{\alpha_{\mathrm{reg}}}(\gamma^{\sigma}(t)) - \mathcal{A}_{\alpha_{\mathrm{reg}}}(\gamma^{\sigma}(s))] - [\mathcal{S}_{\alpha_{\mathrm{reg}}}(\gamma^{\sigma}(t)) - \mathcal{S}_{\alpha_{\mathrm{reg}}}(\gamma^{\sigma}(s))]\|_{\mathcal{X}},$$
$$N^{\dagger}(s,t) \coloneqq \|[\mathcal{A}_{\alpha_{\mathrm{reg}}}(\gamma^{\sigma}(t)) - \mathcal{A}_{\alpha_{\mathrm{reg}}}(\gamma^{\sigma}(s))] - [x_t^{\dagger} - x_s^{\dagger}]\|_{\mathcal{X}}.$$

Define $\eta_{\alpha_{\mathrm{reg}}}(s,t)$ by the diagnostic scale of Sec. 3, with $\mathcal{S}$ replaced by $\mathcal{S}_{\alpha_{\mathrm{reg}}}$, and set

$$\mathrm{SAE}_{\alpha_{\mathrm{reg}}}(s,t) \coloneqq \frac{N_{\alpha_{\mathrm{reg}}}(s,t)}{\eta_{\alpha_{\mathrm{reg}}}(s,t)}.$$

Let $\mathcal{L}_{\mathrm{SA},\alpha_{\mathrm{reg}}}(s,t)$ be the corresponding PhySA loss and define

$$d_{\alpha_{\mathrm{reg}}}(s,t) \coloneqq \max\{\|\mathcal{S}_{\alpha_{\mathrm{reg}}}(\gamma^{\sigma}(t)) - \mathcal{S}_{\alpha_{\mathrm{reg}}}(\gamma^{\sigma}(s))\|_{\mathcal{X}}, \varepsilon_{\mathrm{SA}}\}.$$

The definitions and the triangle inequality give

$$N_{\alpha_{\mathrm{reg}}}(s,t) = \eta_{\alpha_{\mathrm{reg}}}(s,t)\mathrm{SAE}_{\alpha_{\mathrm{reg}}}(s,t) = d_{\alpha_{\mathrm{reg}}}(s,t)\sqrt{\mathcal{L}_{\mathrm{SA},\alpha_{\mathrm{reg}}}(s,t)}, \tag{15}$$
$$N^{\dagger}(s,t) \le \eta_{\alpha_{\mathrm{reg}}}(s,t)\mathrm{SAE}_{\alpha_{\mathrm{reg}}}(s,t) + 2\beta_{\alpha_{\mathrm{reg}}}(\sigma). \tag{16}$$

Thus a small normalized loss or diagnostic can coexist with a large absolute mismatch when the regularized response scale is large. The regularization level also controls the response scale: under the hypotheses of Corollary D.1, the physical-response term in $\eta_{\alpha_{\mathrm{reg}}}(s,t)$ is at most the operator norm of $D\mathcal{S}_{\alpha_{\mathrm{reg}}}$ along the path times the path length between s and t, and that operator norm is at most $1/(2\sqrt{\alpha_{\mathrm{reg}}})$ in the linear specialization below, so the denominator of $\mathrm{SAE}_{\alpha_{\mathrm{reg}}}$ cannot grow without bound as the response of an exact inverse can. The exact-response term $2\beta_{\alpha_{\mathrm{reg}}}(\sigma) \le 2(C_{\alpha_{\mathrm{reg}}}\sigma + b_{\alpha_{\mathrm{reg}}})$ falls with the level through the stability term and rises with the level through the bias term. The numerator, denominator, specification-discrepancy level, and regularization level should therefore be reported together; when absolute response error is primary, an unnormalized response term can also be retained in the training objective.

Tikhonov specializations. Both specializations take $\mathcal{Z}_{\mathrm{inv}} = \mathcal{Y}_{\mathrm{obs}}$ with Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}_{\mathrm{obs}}$, $\gamma^{\dagger}(t) = y_t^{\dagger}$, $\gamma^{\sigma}(t) = y_t^{\sigma}$, and $\|y_t^{\sigma} - y_t^{\dagger}\|_{\mathcal{Y}_{\mathrm{obs}}} \le \sigma$, so that σ is the observation-noise level. In the linear case, let $\mathcal{K} : \mathcal{X} \to \mathcal{Y}_{\mathrm{obs}}$ be a fixed bounded linear operator, take $\mathcal{F} = \mathcal{K}$, $\mathcal{C} = \mathcal{X}$, and $R(x) = \|x\|_{\mathcal{X}}^2$ in Eq. 9, and suppose that $y_t^{\dagger} = \mathcal{K}x_t^{\dagger}$ and that the selected minimum-norm solutions satisfy the uniform source condition

$$x_t^{\dagger} = (\mathcal{K}^*\mathcal{K})^{\mu}\psi_t, \qquad 0 < \mu \le 1, \qquad B_{\mu} \coloneqq \sup_t \|\psi_t\|_{\mathcal{X}} < \infty.$$

In the nonlinear case, let $\mathcal{F} : \mathcal{C} \to \mathcal{Y}_{\mathrm{obs}}$ be a weakly sequentially closed forward operator on a closed convex set $\mathcal{C} \subset \mathcal{X}$ that extends to a twice continuously Fréchet differentiable map on an open set containing $\mathcal{C}$ with $\|\mathcal{F}''(x)\| \le L_{\mathcal{F}}$ there, take $R(x) = \|x - x_0\|_{\mathcal{X}}^2$ for a fixed prior guess $x_0 \in \mathcal{X}$, so that Eq. 9 is Tikhonov regularization of a nonlinear problem (Engl et al., 1989; 1996), and suppose that $y_t^{\dagger} = \mathcal{F}(x_t^{\dagger})$ with $x_t^{\dagger} \in \mathcal{C}$ and that the selected exact solutions satisfy the source condition

$$x_t^{\dagger} - x_0 = \mathcal{F}'(x_t^{\dagger})^*\psi_t, \qquad B \coloneqq \sup_t \|\psi_t\|_{\mathcal{Y}_{\mathrm{obs}}} < \infty, \qquad L_{\mathcal{F}}B < 1,$$

the nonlinear counterpart of the linear source condition with $\mu = 1/2$. Minimizers over $\mathcal{C}$ exist for every y but need not be unique, so $\mathcal{S}_{\alpha_{\mathrm{reg}}}$ is a fixed selection, exactly as $x_t^{\dagger}$ is. Write $x_t \coloneqq \mathcal{S}_{\alpha_{\mathrm{reg}}}(y_t^{\sigma})$, let $J_{\alpha_{\mathrm{reg}},y}(x) \coloneqq \|\mathcal{F}(x) - y\|_{\mathcal{Y}_{\mathrm{obs}}}^2 + \alpha_{\mathrm{reg}}\|x - x_0\|_{\mathcal{X}}^2$ be the Tikhonov functional, and let $\varrho_{\alpha_{\mathrm{reg}}} \coloneqq \sup_{0 \le t \le \rho} \|\mathcal{F}(x_t) - y_t^{\sigma}\|_{\mathcal{Y}_{\mathrm{obs}}}$ be the largest residual of the selected reference along the path.

Proposition D.1 (Tikhonov specializations)**.** *In the linear case, the minimizer in Eq. 9 is unique and*

$$\mathcal{S}_{\alpha_{\mathrm{reg}}}(y) = (\mathcal{K}^*\mathcal{K} + \alpha_{\mathrm{reg}}I)^{-1}\mathcal{K}^*y, \qquad C_{\alpha_{\mathrm{reg}}} \le \frac{1}{2\sqrt{\alpha_{\mathrm{reg}}}},$$
$$b_{\alpha_{\mathrm{reg}}} \le c_{\mu}B_{\mu}\alpha_{\mathrm{reg}}^{\mu}, \qquad c_{\mu} \coloneqq \sup_{r \ge 0} \frac{r^{\mu}}{1 + r} \le 1,$$

so that

$$\beta_{\alpha_{\mathrm{reg}}}(\sigma) \le \frac{\sigma}{2\sqrt{\alpha_{\mathrm{reg}}}} + c_\mu B_\mu \alpha_{\mathrm{reg}}^{\mu}. \tag{17}$$

In the nonlinear case, for every t and every global minimizer x of $J_{\alpha_{\mathrm{reg}},y_t^\sigma}$ over $\mathcal{C}$, in particular for $x = x_t$,

$$\|x - x_t^\dagger\|_{\mathcal{X}} \le \frac{\sigma + \alpha_{\mathrm{reg}} B}{\sqrt{\alpha_{\mathrm{reg}}(1 - L_{\mathcal{F}} B)}}, \qquad \|\mathcal{F}(x) - y_t^\sigma\|_{\mathcal{Y}_{\mathrm{obs}}} \le \sigma + 2\alpha_{\mathrm{reg}} B, \tag{18}$$

so that

$$\beta_{\alpha_{\mathrm{reg}}}(\sigma) \le \frac{\sigma}{\sqrt{\alpha_{\mathrm{reg}}(1 - L_{\mathcal{F}} B)}} + B\sqrt{\frac{\alpha_{\mathrm{reg}}}{1 - L_{\mathcal{F}} B}}. \tag{19}$$

If, in addition, the selected minimizers are interior points of $\mathcal{C}$, $\alpha_{\mathrm{reg}} > L_{\mathcal{F}} \varrho_{\alpha_{\mathrm{reg}}}$, and the minimizer of $J_{\alpha_{\mathrm{reg}},y}$ is unique for every y in a neighborhood of the observed path in $\mathcal{Y}_{\mathrm{obs}}$, then $\mathcal{S}_{\alpha_{\mathrm{reg}}}$ is continuously Fréchet differentiable on a neighborhood of that path and

$$\|D\mathcal{S}_{\alpha_{\mathrm{reg}}}(y_t^\sigma)\|_{\mathrm{op}} \le \frac{1}{2\sqrt{\alpha_{\mathrm{reg}} - L_{\mathcal{F}} \varrho_{\alpha_{\mathrm{reg}}}}}. \tag{20}$$

Balancing the two terms of Eq. 17 gives $\alpha_{\mathrm{reg}} \asymp (\sigma/B_\mu)^{2/(2\mu+1)}$ and an inverse-specific error of order $\sigma^{2\mu/(2\mu+1)}$ for fixed B_μ. The nonlinear bound in Eq. 19 has the same $\alpha_{\mathrm{reg}}^{-1/2}$ noise amplification and $\alpha_{\mathrm{reg}}^{1/2}$ bias as the linear case with $\mu = 1/2$, so balancing gives $\alpha_{\mathrm{reg}} \asymp \sigma/B$ and an inverse-specific error of order $\sqrt{\sigma}$, and the nonlinearity enters only through the product $L_{\mathcal{F}} B$, which must be below one. Neither balance includes $\epsilon_{\mathrm{nom},\alpha_{\mathrm{reg}}}$, $E_{\alpha_{\mathrm{reg}}}$, or the attained absolute alignment term, which can also depend on α_{reg}.

The linear bounds of Proposition D.1 follow from spectral calculus (Engl et al., 1996), and Eq. 18 is the estimate behind the convergence-rate theorem of Engl et al. (1989), written in the present notation; the estimate uses only the Lipschitz continuity of $\mathcal{F}'$ with constant $L_{\mathcal{F}}$ that the bound on $\mathcal{F}''$ implies, and the estimate does not use that $x_t^\dagger$ minimizes $\|x - x_0\|_{\mathcal{X}}$ among the solutions of $\mathcal{F}(x) = y_t^\dagger$, so the selected solutions need only satisfy the source condition. The differentiability of $\mathcal{S}_{\alpha_{\mathrm{reg}}}$ along the observed path that Corollary D.1 assumes is not implied by the minimization in Eq. 9; the last part of Proposition D.1 establishes that differentiability together with the C^1 regularity of $\mathcal{S}_{\alpha_{\mathrm{reg}}} \circ \gamma^\sigma$ that the corollary requires, under a curvature condition. The condition $\alpha_{\mathrm{reg}} > L_{\mathcal{F}} \varrho_{\alpha_{\mathrm{reg}}}$ holds whenever $2L_{\mathcal{F}} B < 1$ and $\alpha_{\mathrm{reg}} > L_{\mathcal{F}}\sigma/(1 - 2L_{\mathcal{F}} B)$, because $\varrho_{\alpha_{\mathrm{reg}}} \le \sigma + 2\alpha_{\mathrm{reg}} B$ by Eq. 18, and in particular for the balanced level $\alpha_{\mathrm{reg}} = \sigma/B$ when $L_{\mathcal{F}} B < 1/3$.

The bound in Eq. 20 reduces to the linear constant $1/(2\sqrt{\alpha_{\mathrm{reg}}})$ when $L_{\mathcal{F}} = 0$, so the bound on the sensitivity of the nonlinear reference, and hence on the reference response that $\mathcal{A}_{\alpha_{\mathrm{reg}}}$ must track, again scales as $\alpha_{\mathrm{reg}}^{-1/2}$ as the regularization is weakened, with the residual-weighted curvature $L_{\mathcal{F}} \varrho_{\alpha_{\mathrm{reg}}}$ reducing the effective level. For convex non-quadratic penalties such as total variation or sparsity-promoting norms, the corresponding stability and convergence-rate results are formulated in Bregman distances rather than in the norm of $\mathcal{X}$ (Burger & Osher, 2004; Hofmann et al., 2007; Scherzer et al., 2009), and the resulting references are in general not differentiable in the data, so such penalties fall outside the norm-based and differentiable statement of Corollary D.1 and would require a separate treatment.

Implementation with a fixed level. For fixed α_{reg}, the paired batching logic carries over when $\mathcal{A}_{\alpha_{\mathrm{reg}}}$ is intended to emulate $\mathcal{S}_{\alpha_{\mathrm{reg}}}$: $\mathcal{S}_{\alpha_{\mathrm{reg}}}$ must replace $\mathcal{S}$ in both pointwise target terms and the response target. If the model is instead trained directly against $x^\dagger$, then $\mathcal{S}_{\alpha_{\mathrm{reg}}}$ is an analytical comparator rather than the training target. Target construction must respect the forward and acquisition models, and a noise-only pair for one fixed object has zero exact response even when the object's regularized reconstruction moves.

Adaptive regularization levels. A fixed level is restrictive from the standpoint of regularization theory, where α_{reg} is chosen as a function of the noise level and possibly of the data (Engl et al., 1996). Two extensions, both left to future work, keep an adaptive level inside the framework.

First, the level can become an input. One model $\mathcal{A}(z, \alpha_{\mathrm{reg}})$ that receives the level as a conditioning input (Zhang et al., 2018; Perez et al., 2018) can emulate the joint map $(z, \alpha_{\mathrm{reg}}) \mapsto \mathcal{S}_{\alpha_{\mathrm{reg}}}(z)$, and

Corollary D.1 applies to the joint maps on the enlarged specification set $\mathcal{Z}_{\mathrm{inv}} \times [\alpha_{\min}, \alpha_{\max}]$ with $\alpha_{\min} > 0$, with $\log \alpha_{\mathrm{reg}}$ as the additional chart coordinate. A path may then move the data, the level, or both, and the chain rule splits the reference response into a data term and a level term,

$$\frac{d}{dt}\mathcal{S}_{\alpha_{\mathrm{reg}}(t)}(\gamma(t)) = D_z\mathcal{S}_{\alpha_{\mathrm{reg}}(t)}(\gamma(t))[\dot\gamma(t)] + \alpha_{\mathrm{reg}}(t)\partial_{\alpha_{\mathrm{reg}}}\mathcal{S}_{\alpha_{\mathrm{reg}}(t)}(\gamma(t))\frac{d}{dt}\log\alpha_{\mathrm{reg}}(t),$$

with D_z the partial derivative in the data coordinate, and the model must match both terms. For quadratic Tikhonov regularization, $\partial_{\alpha_{\mathrm{reg}}}\mathcal{S}_{\alpha_{\mathrm{reg}}}(y) = -(\mathcal{K}^*\mathcal{K} + \alpha_{\mathrm{reg}}I)^{-1}\mathcal{S}_{\alpha_{\mathrm{reg}}}(y)$, so in the logarithmic coordinate the level term is bounded by the norm of the regularized reconstruction, uniformly in α_{reg}, whereas the data term carries the factor $1/(2\sqrt{\alpha_{\mathrm{reg}}})$, and the constants of the corollary are uniform over $[\alpha_{\min}, \alpha_{\max}]$ only. A level-only pair has zero exact response, like a noise-only pair, and the exact-solution bound along a path with level $\alpha_{\mathrm{reg}}(t)$ carries $\sup_t \beta_{\alpha_{\mathrm{reg}}(t)}(\sigma)$.

Second, the level can be tied to the noise level. Classical a priori rules take $\alpha_{\mathrm{reg}} = \alpha_{\mathrm{reg}}(\sigma)$, for instance $\alpha_{\mathrm{reg}}(\sigma) = (\sigma/B_\mu)^{2/(2\mu+1)}$ from the balance above, and a posteriori rules such as the discrepancy principle take $\alpha_{\mathrm{reg}} = \alpha_{\mathrm{reg}}(\sigma, z)$ (Engl et al., 1996); for a linear problem with an unbounded exact inverse, no rule that depends on the data alone yields a convergent regularization method (Bakushinskii, 1984), so the noise level is a necessary input rather than a nuisance parameter. The learned counterpart is a model $\mathcal{A}(z, \hat\sigma)$ whose second input $\hat\sigma$ is the noise level supplied at deployment and which emulates the composite reference $\mathcal{S}_{\mathrm{ad}}(z, \hat\sigma) \coloneqq \mathcal{S}_{\alpha_{\mathrm{reg}}(\hat\sigma)}(z)$, in the spirit of learned reconstruction (Arridge et al., 2019) and noise-conditioned denoising (Zhang et al., 2018). With the a priori rule, the linear source condition, and $\hat\sigma = \sigma$, Eq. 17 gives $\beta_{\alpha_{\mathrm{reg}}(\sigma)}(\sigma) \le (1/2 + c_\mu)B_\mu^{1/(2\mu+1)}\sigma^{2\mu/(2\mu+1)}$, and the chain rule gives $\|\hat\sigma\partial_{\hat\sigma}\mathcal{S}_{\mathrm{ad}}(z, \hat\sigma)\|_{\mathcal{X}} \le (2/(2\mu+1))\|\mathcal{S}_{\alpha_{\mathrm{reg}}(\hat\sigma)}(z)\|_{\mathcal{X}}$, so $\log\hat\sigma$ is the natural coordinate for the noise input and the guarantee is uniform on $[\sigma_{\min}, \sigma_{\max}]$ with $\sigma_{\min} > 0$. A data-dependent rule such as the discrepancy principle keeps the composite reference differentiable only where the implicit function theorem applies to the discrepancy equation.

The extension here is theoretical; the experiments in this paper evaluate forward solution and observation maps rather than inverse reconstructions.

E PROOFS

This appendix collects the proofs of every theoretical statement in the main text, in App. C, and in App. D.

Proof of Theorem C.1. For each $\theta \in \Theta$, choose $g_\theta \in \mathcal{G}$ with $d_\Theta(\theta, g_\theta) \le h$. The Lipschitz assumption gives, for every fixed z,

$$\ell_z(\theta) \le \ell_z(g_\theta) + Lh \le \max_{g\in\mathcal{G}} \ell_z(g) + Lh.$$

Taking the supremum over θ and expectation over z yields $\mathcal{R}^{\mathrm{wc}}_\Theta \le \mathcal{R}^{\mathrm{wc}}_\mathcal{G} + Lh$.

Set $X_i \coloneqq \max_{g\in\mathcal{G}} \ell_{z_i}(g)$. The variables X_i are independent and identically distributed, and each variable lies in $[0, B]$. Hoeffding's inequality (Hoeffding, 1963) therefore yields

$$\mathbb{P}\left(\mathbb{E}[X_1] - \frac{1}{n}\sum_{i=1}^n X_i > t\right) \le \exp\left(-\frac{2nt^2}{B^2}\right).$$

With $t = B\sqrt{\log(1/\alpha)/(2n)}$, the probability on the left is at most α. Combining this one-sided concentration bound with the deterministic coverage bound proves Eq. 2. □

Proof of Proposition 1. By the fundamental theorem of calculus in Banach spaces,

$$(\mathcal{A} - \mathcal{S})(\gamma(1)) = (\mathcal{A} - \mathcal{S})(\gamma(0)) + \int_0^1 D(\mathcal{A} - \mathcal{S})(\gamma(t))[\dot\gamma(t)]\,dt.$$

Taking the $\mathcal{Y}$ norm and applying the triangle inequality gives the integral bound. The final statement follows from $\|(D\mathcal{A} - D\mathcal{S})(\gamma(t))[\dot\gamma(t)]\| \le \epsilon_{\mathrm{sens}}\|\dot\gamma(t)\|_\mathcal{Z}$ and $\int_0^1 \|\dot\gamma(t)\|_\mathcal{Z}\,dt \le \rho$. □

Proof of Theorem 1. Fix $t \in [t_j, t_{j+1}]$ and abbreviate $a \coloneqq t_j$, $b \coloneqq t_{j+1}$, and $\lambda \coloneqq (t-a)/(b-a)$. Let $I(t) \coloneqq (1-\lambda)r(a) + \lambda r(b)$ be the linear interpolant of the two endpoint errors. The fundamental theorem of calculus for Banach-valued continuously differentiable functions, applied twice, gives the exact interpolation remainder

$$r(t) - I(t) = -\frac{b-t}{b-a}\int_a^t (s-a)r''(s)\,ds - \frac{t-a}{b-a}\int_t^b (b-s)r''(s)\,ds.$$

Consequently,

$$\begin{aligned}\|r(t) - I(t)\|_{\mathcal{Y}} &\le \frac{M}{2(b-a)}[(b-t)(t-a)^2 + (t-a)(b-t)^2] \\ &= \frac{M}{2}(t-a)(b-t) \le \frac{M(b-a)^2}{8} \le \frac{Mh^2}{8}. \end{aligned} \tag{21}$$

For each anchor, define $m_j \coloneqq r(t_j) - r(0)$. Expanding the definition of r shows that $\max_j \|m_j\|_{\mathcal{Y}} = \epsilon_{\text{pair}}$. Moreover,

$$I(t) = r(0) + (1-\lambda)m_j + \lambda m_{j+1},$$

so convexity of the norm and the triangle inequality imply

$$\|I(t)\|_{\mathcal{Y}} \le \epsilon_{\text{nom}} + (1-\lambda)\|m_j\|_{\mathcal{Y}} + \lambda\|m_{j+1}\|_{\mathcal{Y}} \le \epsilon_{\text{nom}} + \epsilon_{\text{pair}}. \tag{22}$$

Combining Eq. 21 and Eq. 22 and then taking the supremum over every grid interval proves Eq. 1. □

Proof of Corollary 1. The numerator of $\mathcal{L}_{\text{SA}}(\gamma(0), \gamma(t_j))$ is exactly $\|m_j\|_{\mathcal{Y}}^2$. Therefore

$$\|m_j\|_{\mathcal{Y}} \le \sqrt{q}\max\{\|\mathcal{S}(\gamma(t_j)) - \mathcal{S}(\gamma(0))\|_{\mathcal{Y}}, \varepsilon_{\text{SA}}\} \le \sqrt{q}R_{\text{resp}}.$$

Taking the maximum over j proves the stated PhySA bound. Finally, $r''(t) = (\mathcal{A}\circ\gamma)''(t) - (\mathcal{S}\circ\gamma)''(t)$, so the triangle inequality gives $\|r''(t)\|_{\mathcal{Y}} \le M_A + M_S$ whenever the separate curvature bounds hold. □

Proof of Theorem C.2. Differentiating $F(p_t, u_t) = 0$ by the Banach-space chain rule gives the true tangent equation

$$L_t\dot{u}_t + D_pF(p_t, u_t)[v] = 0. \tag{23}$$

Since $\widetilde{\tau}(t) = \widehat{L}_t\dot{a}_t + \widehat{b}_t$ and $e_t = a_t - u_t$, subtracting Eq. 23 from the model tangent gives

$$L_t\dot{e}_t = \widetilde{\tau}(t) + (L_t - \widehat{L}_t)\dot{a}_t + (b_t - \widehat{b}_t). \tag{24}$$

Applying L_t^{-1} to Eq. 24, followed by the assumed bounds of the theorem and $\|v\|_{\mathcal{P}} = 1$, yields

$$\|\dot{e}_t\|_{\mathcal{U}} \le C_{\text{stab}}\|\widetilde{\tau}(t)\|_{\mathcal{V}} + K\|e_t\|_{\mathcal{U}}. \tag{25}$$

The fundamental theorem of calculus and Eq. 25 then yield, for $0 \le t \le \rho$,

$$\|e_t\|_{\mathcal{U}} \le \|e_0\|_{\mathcal{U}} + C_{\text{stab}}\int_0^t \|\widetilde{\tau}(s)\|_{\mathcal{V}}\,ds + K\int_0^t \|e_s\|_{\mathcal{U}}\,ds.$$

The integral form of Gronwall's inequality (Evans, 2022) gives

$$\|e_t\|_{\mathcal{U}} \le e^{Kt}\left(\|e_0\|_{\mathcal{U}} + C_{\text{stab}}\int_0^t \|\widetilde{\tau}(s)\|_{\mathcal{V}}\,ds\right).$$

Taking the supremum and enlarging the integral to $[0, \rho]$ proves Eq. 3. □

Proof of Proposition C.1. Write $w_0 \coloneqq \dot{u}_0$ and introduce the same secant residual at the true base state,

$$\tau_h^\star \coloneqq L_0\widehat{w}_h + D_pF(p_0, u_0)[v].$$

The Lipschitz assumptions at $t = 0$ give

$$\|\tau_h^\star\|_{\mathcal{V}} \le \|\widetilde{\tau}_h\|_{\mathcal{V}} + (K_uG_h + K_p)\|e_0\|_{\mathcal{U}}.$$

Eq. 23 at $t = 0$ implies $L_0(\widehat{w}_h - w_0) = \tau_h^\star$. Consequently,

$$\|\widehat{w}_h - w_0\|_{\mathcal{U}} \le C_{\mathrm{stab}}[\|\widetilde{\tau}_h\|_{\mathcal{V}} + (K_u G_h + K_p)\|e_0\|_{\mathcal{U}}]. \tag{26}$$

Finally, Taylor's theorem with integral remainder gives

$$u_h - u_0 = h w_0 + R_h^S, \qquad \|R_h^S\|_{\mathcal{U}} \le \frac{M_S h^2}{2}.$$

Using $a_h - a_0 = h\widehat{w}_h$, we have the exact decomposition

$$e_h = e_0 + h(\widehat{w}_h - w_0) - R_h^S.$$

Substitution of Eq. 26, the identity $h\|\widetilde{\tau}_h\|_{\mathcal{V}} = \|\widetilde{q}_h\|_{\mathcal{V}}$, and the triangle inequality prove Eq. 4. □

Proof of Theorem C.3. Set $w \coloneqq \widetilde{u} - u$. Subtracting the two Burgers equations and using $\widetilde{u} = u + w$ gives

$$w_t + \widetilde{u} w_x + w u_x - \widetilde{\nu} w_{xx} = (\widetilde{\nu} - \nu) u_{xx}.$$

Taking the $L^2(\mathbb{T})$ inner product with w is justified by the stated strong-solution regularity. Periodicity gives

$$\int_{\mathbb{T}} (\widetilde{u} w_x + w u_x) w \, dx = -\frac{1}{2}\int_{\mathbb{T}} \widetilde{u}_x w^2 \, dx + \int_{\mathbb{T}} u_x w^2 \, dx = \frac{1}{2}\int_{\mathbb{T}} u_x w^2 \, dx,$$

because $\widetilde{u}_x = u_x + w_x$ and $\int_{\mathbb{T}} w_x w^2 \, dx = (1/3)\int_{\mathbb{T}} (w^3)_x \, dx = 0$. After one further periodic integration by parts on the right-hand side,

$$\frac{1}{2}\frac{d}{dt}\|w\|_{L^2}^2 + \widetilde{\nu}\|w_x\|_{L^2}^2 + \frac{1}{2}\int_{\mathbb{T}} u_x w^2 \, dx = -(\widetilde{\nu} - \nu)\int_{\mathbb{T}} u_x w_x \, dx.$$

The L^∞ bound on u_x, Cauchy–Schwarz, and Young's inequality imply

$$\frac{d}{dt}\|w\|_{L^2}^2 + \widetilde{\nu}\|w_x\|_{L^2}^2 \le \|u_x\|_{L^\infty}\|w\|_{L^2}^2 + \frac{|\widetilde{\nu} - \nu|^2}{\widetilde{\nu}}\|u_x\|_{L^2}^2.$$

Dropping the nonnegative gradient term and applying the integral form of Gronwall's inequality proves Eq. 5. □

Proof of Proposition C.2. Set $e \coloneqq b - v$. Subtracting the exact equation for v from the residual equation for b gives

$$e_t + b e_x + e v_x - \mu e_{xx} = q_b.$$

Exactly as above, periodicity and $b = v + e$ imply

$$\int_{\mathbb{T}} (b e_x + e v_x) e \, dx = \frac{1}{2}\int_{\mathbb{T}} v_x e^2 \, dx.$$

Taking the L^2 inner product with e, discarding the nonnegative diffusion term, and using $2\langle q_b, e\rangle \le \kappa^{-1}\|q_b\|_{L^2}^2 + \kappa\|e\|_{L^2}^2$ gives

$$\frac{d}{dt}\|e\|_{L^2}^2 \le (\|v_x\|_{L^\infty} + \kappa)\|e\|_{L^2}^2 + \kappa^{-1}\|q_b\|_{L^2}^2.$$

Gronwall's inequality proves Eq. 6. □

Proof of Theorem C.4. Let $w \coloneqq \widetilde{u} - u$ and $q \coloneqq w - H \in V$. Subtracting the two weak equations gives, for every $\phi \in V$,

$$\int_\Omega \widetilde{a}\nabla q \cdot \nabla\phi \, dx = \langle \widetilde{f} - f, \phi\rangle + \int_\Omega (a - \widetilde{a})\nabla u \cdot \nabla\phi \, dx - \int_\Omega \widetilde{a}\nabla H \cdot \nabla\phi \, dx.$$

Taking $\phi = q$ and using coercivity, Hölder's inequality, and the definition of the V^* norm yields

$$\lambda\|\nabla q\|_{L^2} \le \|\widetilde{f} - f\|_{V^*} + \|\widetilde{a} - a\|_{L^\infty}\|\nabla u\|_{L^2} + \Lambda\|\nabla H\|_{L^2},$$

with the zero case immediate. Since $w = q + H$, the triangle inequality proves Eq. 7. □

Proof of Proposition C.3. Let $e := \widehat{u} - \widetilde{u}$ and $p := e - \widehat{H} \in V$. By the definition of the residual and the weak equation for $\widetilde{u}$,

$$\int_\Omega \widetilde{a}\nabla p \cdot \nabla\phi \, dx = -\langle \mathfrak{r}_{\widehat{u}}, \phi\rangle - \int_\Omega \widetilde{a}\nabla\widehat{H} \cdot \nabla\phi \, dx.$$

Testing with p and repeating the coercivity argument gives

$$\|\nabla p\|_{L^2} \le \lambda^{-1}\|\mathfrak{r}_{\widehat{u}}\|_{V^*} + \frac{\Lambda}{\lambda}\|\nabla\widehat{H}\|_{L^2}.$$

Because $e = p + \widehat{H}$, this bound proves Eq. 8. □

Proof of Corollary C.1. Exact boundary compatibility gives $\widehat{H} = 0$ and $e = \widehat{u} - \widetilde{u} \in V$. Apply Eq. 8, followed by Poincaré's inequality (Evans, 2022). □

Proof of Proposition C.4. Taylor's theorem with integral remainder gives

$$g_A(\rho) - g_A(0) = \rho g_A'(0) + \int_0^\rho (\rho - t) g_A''(t) \, dt,$$

and the analogous identity holds for g_S. Subtracting the identities and dividing by ρ gives the stated formula because the chain rule yields $g_A'(0) - g_S'(0) = (D\mathcal{A}(z) - D\mathcal{S}(z))[v]$. The triangle inequality and the two second-derivative bounds give

$$\|R_\rho\|_{\mathcal{Y}} \le \frac{1}{\rho}\int_0^\rho (\rho - t)(M_A + M_S) \, dt = \frac{\rho}{2}(M_A + M_S).$$

□

Proof of Corollary D.1. The proof of Proposition 1 uses only the Banach-chart structure of the specification set, so the proposition applies verbatim in the chart of $\mathcal{Z}_{\mathrm{inv}}$. For each $t \in (0, \rho]$, apply Proposition 1 to the restriction of γ^σ to $[0, t]$, reparametrized to $[0, 1]$, with $(\mathcal{A}, \mathcal{S}) = (\mathcal{A}_{\alpha_{\mathrm{reg}}}, \mathcal{S}_{\alpha_{\mathrm{reg}}})$. Enlarging the derivative integral to $[0, \rho]$ and taking the supremum over t proves Eq. 12; the case $t = 0$ is immediate. For every t, the triangle inequality gives

$$\|\mathcal{A}_{\alpha_{\mathrm{reg}}}(\gamma^\sigma(t)) - x_t^\dagger\|_{\mathcal{X}} \le \|\mathcal{A}_{\alpha_{\mathrm{reg}}}(\gamma^\sigma(t)) - \mathcal{S}_{\alpha_{\mathrm{reg}}}(\gamma^\sigma(t))\|_{\mathcal{X}} + \|\mathcal{S}_{\alpha_{\mathrm{reg}}}(\gamma^\sigma(t)) - x_t^\dagger\|_{\mathcal{X}},$$

and the last term is at most $\beta_{\alpha_{\mathrm{reg}}}(\sigma)$ of Eq. 14, which is at most $C_{\alpha_{\mathrm{reg}}}\sigma + b_{\alpha_{\mathrm{reg}}}$ by a second application of the triangle inequality at $\mathcal{S}_{\alpha_{\mathrm{reg}}}(\gamma^\dagger(t))$ together with Eq. 10 and Eq. 11. Combining this estimate with Eq. 12 proves Eq. 13, both as stated and with $C_{\alpha_{\mathrm{reg}}}\sigma + b_{\alpha_{\mathrm{reg}}}$ replaced by $\beta_{\alpha_{\mathrm{reg}}}(\sigma)$. The final assertion follows from the operator-norm bound and the definition of L_γ. □

Proof of Proposition D.1. In the linear case, $J_{\alpha_{\mathrm{reg}}, y}$ is strictly convex, so the minimizer is unique and solves the normal equation $(\mathcal{K}^*\mathcal{K} + \alpha_{\mathrm{reg}} I)x = \mathcal{K}^* y$. Spectral calculus for $\mathcal{K}^*\mathcal{K}$ gives $\|\mathcal{S}_{\alpha_{\mathrm{reg}}}\|_{\mathrm{op}} \le \sup_{s \ge 0} s/(s^2 + \alpha_{\mathrm{reg}}) = 1/(2\sqrt{\alpha_{\mathrm{reg}}})$, which bounds $C_{\alpha_{\mathrm{reg}}}$ because $\mathcal{S}_{\alpha_{\mathrm{reg}}}$ is linear, and

$$\mathcal{S}_{\alpha_{\mathrm{reg}}}(y_t^\dagger) - x_t^\dagger = -\alpha_{\mathrm{reg}}(\mathcal{K}^*\mathcal{K} + \alpha_{\mathrm{reg}} I)^{-1}(\mathcal{K}^*\mathcal{K})^\mu \psi_t,$$

whose norm is at most $\sup_{s \ge 0} \alpha_{\mathrm{reg}} s^{2\mu}(s^2 + \alpha_{\mathrm{reg}})^{-1}\|\psi_t\|_{\mathcal{X}} = c_\mu \alpha_{\mathrm{reg}}^\mu \|\psi_t\|_{\mathcal{X}}$ by the substitution $s^2 = \alpha_{\mathrm{reg}} r$; the bound $c_\mu \le 1$ follows from $r^\mu \le 1 + r$. Taking the supremum over t bounds $b_{\alpha_{\mathrm{reg}}}$, and Eq. 14 gives Eq. 17.

In the nonlinear case, minimizers of $J_{\alpha_{\mathrm{reg}}, y}$ over $\mathcal{C}$ exist for every y because $\mathcal{F}$ is continuous and weakly sequentially closed and $\mathcal{C}$ is weakly closed (Engl et al., 1996). Fix t and a global minimizer x of $J_{\alpha_{\mathrm{reg}}, y_t^\sigma}$ over $\mathcal{C}$, and argue as in Engl et al. (1989). Comparing the value of $J_{\alpha_{\mathrm{reg}}, y_t^\sigma}$ at x with the value at $x_t^\dagger$, expanding $\|x - x_0\|_{\mathcal{X}}^2 = \|x - x_t^\dagger\|_{\mathcal{X}}^2 + 2\langle x - x_t^\dagger, x_t^\dagger - x_0\rangle_{\mathcal{X}} + \|x_t^\dagger - x_0\|_{\mathcal{X}}^2$, and using $\|\mathcal{F}(x_t^\dagger) - y_t^\sigma\|_{\mathcal{Y}_{\mathrm{obs}}} \le \sigma$ gives

$$\|\mathcal{F}(x) - y_t^\sigma\|_{\mathcal{Y}_{\mathrm{obs}}}^2 + \alpha_{\mathrm{reg}}\|x - x_t^\dagger\|_{\mathcal{X}}^2 \le \sigma^2 + 2\alpha_{\mathrm{reg}}\langle x_t^\dagger - x, x_t^\dagger - x_0\rangle_{\mathcal{X}}.$$

The source condition, the identity $\mathcal{F}'(x_t^\dagger)(x_t^\dagger - x) = y_t^\dagger - \mathcal{F}(x) + \zeta_t$ with the Taylor remainder ζ_t of $\mathcal{F}$ at $x_t^\dagger$, which satisfies $\|\zeta_t\|_{\mathcal{Y}_{\text{obs}}} \le (L_{\mathcal{F}}/2)\|x - x_t^\dagger\|_{\mathcal{X}}^2$ because the segment between $x_t^\dagger$ and x lies in the convex set $\mathcal{C}$, and the bound $\|y_t^\dagger - \mathcal{F}(x)\|_{\mathcal{Y}_{\text{obs}}} \le \sigma + \|\mathcal{F}(x) - y_t^\sigma\|_{\mathcal{Y}_{\text{obs}}}$ give

$$\langle x_t^\dagger - x, x_t^\dagger - x_0\rangle_{\mathcal{X}} \le B\left(\sigma + \|\mathcal{F}(x) - y_t^\sigma\|_{\mathcal{Y}_{\text{obs}}} + \frac{L_{\mathcal{F}}}{2}\|x - x_t^\dagger\|_{\mathcal{X}}^2\right).$$

Inserting this bound, moving the term in $\|x - x_t^\dagger\|_{\mathcal{X}}^2$ to the left, and completing the square in the residual yields

$$(\|\mathcal{F}(x) - y_t^\sigma\|_{\mathcal{Y}_{\text{obs}}} - \alpha_{\text{reg}}B)^2 + \alpha_{\text{reg}}(1 - L_{\mathcal{F}}B)\|x - x_t^\dagger\|_{\mathcal{X}}^2 \le (\sigma + \alpha_{\text{reg}}B)^2,$$

which proves Eq. 18; applying the first bound of Eq. 18 to $x = x_t$ and taking the supremum over t proves Eq. 19.

For the last part, write the first-order optimality condition at an interior minimizer as $\Xi(x, y) = 0$ with $\Xi(x, y) \coloneqq \mathcal{F}'(x)^*(\mathcal{F}(x) - y) + \alpha_{\text{reg}}(x - x_0)$. The map Ξ is continuously differentiable because $\mathcal{F}$ is twice continuously differentiable, and the partial derivative of Ξ in x is half the Hessian of $J_{\alpha_{\text{reg}},y}$, which satisfies

$$\begin{aligned}\langle J''_{\alpha_{\text{reg}},y_t^\sigma}(x_t)h, h\rangle_{\mathcal{X}} &= 2\|\mathcal{F}'(x_t)h\|_{\mathcal{Y}_{\text{obs}}}^2 + 2\langle \mathcal{F}(x_t) - y_t^\sigma, \mathcal{F}''(x_t)[h,h]\rangle_{\mathcal{Y}_{\text{obs}}} + 2\alpha_{\text{reg}}\|h\|_{\mathcal{X}}^2 \\ &\ge 2\|\mathcal{F}'(x_t)h\|_{\mathcal{Y}_{\text{obs}}}^2 + 2(\alpha_{\text{reg}} - L_{\mathcal{F}}\varrho_{\alpha_{\text{reg}}})\|h\|_{\mathcal{X}}^2\end{aligned}$$

for all $h \in \mathcal{X}$ by the bound on $\mathcal{F}''$ and the definition of $\varrho_{\alpha_{\text{reg}}}$. The Hessian is therefore coercive and invertible, and the implicit function theorem provides an open ball $\mathcal{B}$ around x_t and a C^1 map $y \mapsto x(y)$, defined for y near y_t^σ, such that $x(y)$ is the unique zero of $\Xi(\cdot, y)$ in $\mathcal{B}$, with $Dx(y) = 2J''_{\alpha_{\text{reg}},y}(x(y))^{-1}\mathcal{F}'(x(y))^*$. Minimizers of $J_{\alpha_{\text{reg}},y}$ depend stably on the data: if $y_k \to y$, every sequence of minimizers for y_k has a subsequence converging to a minimizer for y (Engl et al., 1996). Since the minimizer is unique for every y near the path, the minimizer $\mathcal{S}_{\alpha_{\text{reg}}}(y)$ converges to x_t as $y \to y_t^\sigma$, so for y close to y_t^σ the minimizer $\mathcal{S}_{\alpha_{\text{reg}}}(y)$ is an interior point of $\mathcal{C}$ lying in $\mathcal{B}$ and hence equals $x(y)$. Thus $\mathcal{S}_{\alpha_{\text{reg}}}$ coincides with the C^1 map $x(\cdot)$ near y_t^σ, and the union over t of these neighborhoods is a neighborhood of the observed path on which $\mathcal{S}_{\alpha_{\text{reg}}}$ is continuously Fréchet differentiable. For Eq. 20, fix $v \in \mathcal{Y}_{\text{obs}}$ and let $u \coloneqq 2J''_{\alpha_{\text{reg}},y_t^\sigma}(x_t)^{-1}\mathcal{F}'(x_t)^*v$, so that $\langle J''_{\alpha_{\text{reg}},y_t^\sigma}(x_t)u, u\rangle_{\mathcal{X}} = 2\langle v, \mathcal{F}'(x_t)u\rangle_{\mathcal{Y}_{\text{obs}}}$. The coercivity bound then gives

$$(\alpha_{\text{reg}} - L_{\mathcal{F}}\varrho_{\alpha_{\text{reg}}})\|u\|_{\mathcal{X}}^2 \le \|v\|_{\mathcal{Y}_{\text{obs}}}\|\mathcal{F}'(x_t)u\|_{\mathcal{Y}_{\text{obs}}} - \|\mathcal{F}'(x_t)u\|_{\mathcal{Y}_{\text{obs}}}^2 \le \frac{\|v\|_{\mathcal{Y}_{\text{obs}}}^2}{4},$$

which proves Eq. 20. □

F Further Diagnostic and Method Details

This appendix collects the floor conventions behind the diagnostics of Sec. 3, the reading rules for those diagnostics, the tangent-residual derivation behind ISF, the combined training objective, and the training and evaluation protocol summarized by Procedure 1.

Floor-stabilized response scales. The three definitions of Sec. 3 share the floor-stabilized denominator η defined there because a normalized response diagnostic is otherwise undefined exactly where a benchmark most needs one. When an admissible perturbation produces a near-null physical response, the ratio of a response mismatch to that response diverges for reasons that have nothing to do with the model. The construction of η keeps the denominator above a relative and an absolute floor, and the benchmark logs which of the three terms was active, so a ratio set by a floor can be told apart from a ratio set by the physics.

For a QoI metric, fix the task's physical zero or reference value y_{ref} and define the corresponding target scale by

$$s_{\text{QoI}}(z, z_\delta) \coloneqq \max\{d_{\text{QoI}}(\mathcal{S}(z), y_{\text{ref}}), d_{\text{QoI}}(\mathcal{S}(z_\delta), y_{\text{ref}})\}.$$

For a norm-induced QoI metric with $y_{\text{ref}} = 0$, this target scale is simply the maximum of the nominal and perturbed target norms. The τ terms are used only to stabilize near-null physical responses.

Reading the diagnostics. EAF measures model-response magnitude relative to the stabilized denominator η of Sec. 3. When the physical-response term determines η, values above or below one indicate over- or under-amplification. When a relative or numerical floor determines η, the matched-response reference value is $\|\mathcal{S}(z_\delta) - \mathcal{S}(z)\|/\eta$, which is at most one rather than exactly one. SAE measures response direction and magnitude mismatch, and SNRE normalizes robust error by the floor-stabilized physical-response scale. For tasks with response-scale metadata, the evaluation logs raw root-mean-square (RMS) response values, the absolute SAE and SNRE numerators, the response floor, and a near-null indicator.

Relation to predictive uncertainty. All three quantities are deterministic summaries of a trained checkpoint's response and are therefore complementary to, rather than substitutes for, predictive uncertainty estimates, whose methods and metrics for SciML are surveyed by Psaros et al. (2023) and whose distribution-free variants calibrate coverage for PDE surrogates by wrapping the surrogate in conformal prediction (Angelopoulos & Bates, 2023; Gopakumar et al., 2026). Those methods place a calibrated interval around the prediction at one specification; the diagnostics here instead ask whether the prediction moves correctly when the specification changes.

Finite-difference interpretation. Proposition C.4 shows that, when the physical-response term sets the SAE denominator, dividing the response-mismatch numerator and denominator by ρ makes SAE an estimate of the response-normalized directional sensitivity mismatch. When $\|\mathcal{S}(z_\rho) - \mathcal{S}(z)\|$ is near zero, normalized diagnostics such as SAE and SNRE should be read together with the response norm and the absolute field or QoI error.

Reference implementation. Listing 1 is the implementation of the three diagnostics and fixes the evaluation conventions that the definitions above do not determine. Each norm is taken per sample over the non-batch axes and averaged over the test set only afterwards, so the reported diagnostic is a mean of per-specification ratios rather than a ratio of means; one floor-stabilized denominator η serves all three ratios; and the unnormalized quantities named in the reading rules above are returned beside the three ratios, so a ratio whose denominator is set by a floor can be recognized from the logged record alone.

For the field diagnostics reported in this paper, d_{QoI} is the Euclidean norm and $y_{\mathrm{ref}} = 0$, so s_{QoI} is the larger of the two target norms, and the SNRE denominator coincides with η.

```python
EPS = 1e-7                      # numerical floor eps_diag
RESPONSE_FLOOR_RATIO = 1e-4     # relative response floor tau


def norm(t):
    """Per-sample Euclidean norm over all non-batch axes."""
    return torch.linalg.vector_norm(t.flatten(1), dim=1)


def response_metrics(pred_clean, pred_pert, y_clean, y_pert):
    """EAF, SAE, and SNRE for one perturbation pair (z, z_delta).

    pred_clean = A(z), pred_pert = A(z_delta), y_clean = S(z),
    y_pert = S(z_delta). Every quantity below is computed per
    sample; averaging over the test set happens afterwards.
    """
    model_resp = pred_pert - pred_clean      # A(z_delta) - A(z)
    true_resp = y_pert - y_clean             # S(z_delta) - S(z)
    true_norm = norm(true_resp)
    model_norm = norm(model_resp)
    align = norm(model_resp - true_resp)
    pert_err = norm(pred_pert - y_pert)

    # s_QoI for a norm-induced QoI metric with y_ref = 0.
    target_scale = torch.maximum(
        norm(y_clean), norm(y_pert)
    ).clamp_min(EPS)
    response_floor = RESPONSE_FLOOR_RATIO * target_scale

    # eta = max(||S(z_delta) - S(z)||, tau * s_QoI, eps_diag),
    # the one denominator shared by all three ratios.
    denom = torch.maximum(true_norm, response_floor).clamp_min(EPS)

    sqrt_dim = true_resp.flatten(1).shape[1] ** 0.5
    near_null = (true_norm < response_floor).to(true_norm.dtype)
    return {
        "eaf": model_norm / denom,
        "sae": align / denom,
        "snre": pert_err / denom,
        # Unnormalized quantities logged beside the ratios.
        "true_resp_rms": true_norm / sqrt_dim,
        "model_resp_rms": model_norm / sqrt_dim,
        "sae_abs": align / sqrt_dim,
        "snre_abs": pert_err / sqrt_dim,
        "response_floor_rms": response_floor / sqrt_dim,
        "near_null_response": near_null,
    }
```

Listing 1: Reference implementation of the EAF, SAE, and SNRE diagnostics of Sec. 3, at the τ and $\varepsilon_{\text{diag}}$ used in every reported experiment. The per-sample Euclidean norm is factored into the helper `norm`. All returned values are batched, and a reported diagnostic is the mean of those values over the test set.

Tangent-residual derivation for ISF. Theorem C.2 and Proposition C.1 give the conditional justification for the ISF objective of Sec. 3: for a nonzero scalar step, take $h = |\Delta\nu|$ and the unit parameter direction $v = \operatorname{sgn}(\Delta\nu)$, so the ISF tangent residual is precisely the interior PDE component of the $\widetilde{q}_h = h\widetilde{\tau}_h$ form in the proposition.

If the initial-value constraint is included in $F(\nu, u) \coloneqq (u_t + uu_x - \nu u_{xx}, u(0) - u_0)$, then the full response-scaled residual is

$$\widetilde{q}_h^{\text{full}} = (w_t + uw_x + wu_x - \nu w_{xx} - \Delta\nu u_{xx}, w(0)).$$

Thus, when the finite model response is constraint-compatible, and in particular when the common initial condition is hard-enforced so that $w(0) = 0$, the full residual is the pair consisting of the interior PDE residual above and a zero initial-condition component. Controlling that interior term then controls the full product-space residual under the natural product norm. Otherwise, Proposition C.1 does not apply to the interior term alone: the initial-condition component must also be controlled.

Under either reading, a stable linearized PDE converts a small ISF residual and small nominal error into a finite changed-target error bound, with explicit penalties for model-state linearization drift and finite-step curvature.

Procedure 1: finite-grid physical robustness evaluation. The protocol first specifies a physically admissible perturbation family T_δ together with a severity grid $\mathcal{G} = \mathcal{G}_{\text{seen}} \cup \mathcal{G}_{\text{held}}$, and then computes or streams the changed targets $\mathcal{S}(T_\delta z_i)$ for each nominal specification z_i in the common comparison space. The clean and input-noise baselines are trained on the nominal target alone, whereas the parametric, weighted-parametric, finite-grid adversarial, PASA, PhySA, and ISF recipes are trained on the changed targets at $\mathcal{G}_{\text{seen}}$. Every checkpoint is then evaluated on clean, seen, and held-out severities under the changed-target semantics, never against the nominal label of a changed specification. The reported quantities are field and QoI errors, the finite-grid worst and 90th-percentile ($q_{0.90}$) severity-cell tail proxies, and the EAF, SAE, and SNRE response diagnostics over matched seeds.

Training protocol. For each PDE family, we first sample nominal specifications z_i and admissible perturbation parameters δ_{ik} from the severity grid. We then solve or stream the corresponding changed targets $\mathcal{S}(T_{\delta_{ik}} z_i)$ and split severities into seen and held-out subsets.

Clean training uses only the nominal target $\mathcal{S}(z)$, while the input-noise baseline applies tensor noise without changing that target; all other method families receive the changed target $\mathcal{S}(T_\delta z)$. Parametric training samples seen changed labels without an extra constraint, and weighted-parametric training uses the same mixture of clean and perturbed samples as PASA. Finite-grid adversarial training minimizes the worst loss over the seen severity grid through a finite inner maximum over the clean and seen changed labels.

How we combine the loss terms. Every method except the input-noise baseline, which keeps the nominal target, can be written as an empirical risk over pairs (z, z_δ):

$$\min_{\omega} \mathbb{E}[\ell(\mathcal{A}_\omega(z), \mathcal{S}(z)) + \alpha \ell(\mathcal{A}_\omega(z_\delta), \mathcal{S}(z_\delta)) + \lambda_{\text{SA}} \mathcal{L}_{\text{SA}} + \lambda_{\text{PDE}} \mathcal{L}_{\text{PDE}} + \lambda_{\tan} \mathcal{L}_{\tan}],$$

with method-specific coefficients and sampling. Parametric training retains the changed-label term.

Because these objectives add residual and alignment penalties to a data term, the relative weights matter: an imbalance of gradient magnitudes across the terms of a composite physics-informed objective is a documented optimization failure mode (Wang et al., 2021a), and loss weighting is one of the training choices that separate a working physics-informed model from a diverging one (Wang et al., 2023). The alignment loss is therefore normalized by the physical-response scale, and every method family within a task is compared under a common backbone, data split, and seed protocol rather than under a per-method tuning budget. The certificates of App. C state when these residual and alignment penalties control changed-target field error.

Evaluation protocol. At evaluation time, every method is tested on clean, seen, and held-out severities with the same changed-target semantics. The benchmark does not require invariance to admissible changes. The benchmark's strong baselines are supervised directly on the changed targets at the perturbation levels used for training.

G Experimental Setup Details

This appendix lists the severity grids in Table 4, the admissibility conditions that bound each grid, the per-task protocol summarized in Sec. 4, and the optimization and hardware settings referenced by the reproducibility statement.

Table 4: Severity grids of the six tasks. The initial-condition row reports relative RMS amplitudes. Every evaluation-mesh and initial-condition shift is held out.

Task	Component	Nominal	Seen severities	Held-out severities
Burgers viscosity	ν	0.1	$0.09, 0.075, 0.06$	$0.085, 0.07, 0.055, 0.05$
Public Burgers	ν	0.1	$0.04, 0.2, 0.4$	$0.02, 1.0$
Evaluation mesh	$\mathcal{M}$ (grid points)	64	–	$32, 48, 96, 128$
Initial condition	u_0 (relative RMS)	0	–	$0.05, 0.10, 0.15, 0.20$
Darcy forcing	f (scale)	0	$0.15, 0.35, 0.55$	$0.25, 0.45, 0.65$
Darcy boundary	g (amplitude)	0	$0.04, 0.08, 0.12$	$0.06, 0.10, 0.14$

Table 5: Admissibility condition of Sec. 2.1 that bounds each severity grid, where C is the compatibility condition, G the positivity and ellipticity condition, and W the well-posedness and comparability condition. The severity grids are listed in Table 4 and in the per-task paragraphs of this appendix.

Task	Component	Condition	Bound set by
Burgers viscosity	ν	G, W	$\nu > 0$; shock resolved at 64 points
Public multi-viscosity	ν	W	viscosities released with the labels
Initial condition	u_0	C, W	periodicity kept; band-limited
Evaluation mesh	$\mathcal{M}$	W	sampling rate the field needs
Darcy forcing	f	G, W	$f > 0$ and smooth at 32×32
Darcy boundary	g	C, W	corner compatibility; bounded lift

Admissibility of the severity grids. Three kinds of failure decide admissibility in practice, and the constraints of Sec. 2.1 separate those kinds. A specification can be physically impossible, as when a viscosity or a diffusion coefficient is driven to zero or below, which is what G excludes. A specification can stay well posed in the continuum and still have no trustworthy target on the finite grid a benchmark can afford, as when the viscosity is lowered until the shock is thinner than the grid spacing, so that a score computed at that severity measures the label rather than the model. A specification can also have a target that no common comparison space reaches, as when an evaluation mesh falls below the sampling rate the reported field needs, which is the failure that the comparability half of W excludes and the reason a representation-changing family is a bounded set of meshes rather than an arbitrary regridding. The severity grids in Table 4 are bounded by these three conditions, and Table 5 records which condition sets each bound.

Burgers viscosity task. Burgers uses PDEBench initial conditions for 1D viscous Burgers when available and a pseudo-spectral fourth-order Runge–Kutta solver to generate labels for $\nu = 0.1 \cdot \{1, 0.9, 0.85, 0.75, 0.7, 0.6, 0.55, 0.5\}$. Training uses the seen severities $\{0.9, 0.75, 0.6\}$ and holds out $\{0.85, 0.7, 0.55, 0.5\}$.

The four 1D backbones named in Sec. 4.1 are FNO; Transformer, a softmax-attention encoder in the attention-based operator family (Cao, 2021; Li et al., 2023a); ConvResNet1D, a residual-convolution backbone of the general-purpose learned-simulator type (Stachenfeld et al., 2022); and a coordinate-conditioned DeepONet-style backbone with branch and trunk networks (Chen & Chen, 1995; Lu et al., 2021a). Each backbone is trained for 150 epochs on a 64-point grid, with 1024 training, 256 validation, and 256 test initial conditions.

Public multi-viscosity Burgers task. The raw public-label Burgers stress test uses PDEBench multi-viscosity trajectories stored in Hierarchical Data Format version 5 (HDF5) files. For each sampled initial condition, the target trajectories are streamed directly from the public files for $\nu \in \{0.02, 0.04, 0.1, 0.2, 0.4, 1.0\}$ and downsampled from 1024 to 64 grid points. The backbones are then trained for 150 epochs with 1024 training, 256 validation, and 256 test samples. The clean viscosity is $\nu = 0.1$, seen changed-label viscosities are $\{0.04, 0.2, 0.4\}$, and held-out public labels are $\{0.02, 1.0\}$.

Table 6: Shared training settings used in the experiments. All runs use the five seeds from 2026 through 2030, training batch size 32, evaluation batch size 64, width 48, four backbone blocks, and 16 Fourier modes for FNO. Weighted-parametric and PASA sample a perturbed specification with probability 0.7.

Task family	Epochs	Train, val, test	Nonzero loss weights
Burgers tasks	150	1024, 256, 256	$\lambda_{\mathrm{SA}} = 5 \times 10^{-3}$, $\lambda_{\mathrm{PDE}} = 10^{-4}$, $\lambda_{\tan} = 10^{-5}$
Darcy tasks	200	512, 128, 128	$\lambda_{\mathrm{SA}} = 10^{-3}$, $\lambda_{\mathrm{PDE}} = 10^{-5}$

Burgers mesh and initial-condition tasks. The Burgers mesh and initial-condition tasks evaluate the same recipe- and seed-matched 64-point checkpoints; each task's clean entry is re-evaluated within that task, and all reported gains compare methods within one task. For Burgers mesh shifts, we evaluate clean-viscosity trajectories on 32, 48, 96, and 128 grid points after periodic interpolation from the PDEBench-derived 64-point field.

For Burgers initial-condition shifts, we evaluate smooth low-frequency periodic perturbations with target RMS amplitudes $\{0.05, 0.10, 0.15, 0.20\}$ relative to each initial condition's standard deviation; each perturbed trajectory is re-solved with the pseudo-spectral solver at $\nu = 0.1$.

Darcy forcing and boundary tasks. Forcing-perturbed Darcy fixes a and solves $-\nabla \cdot (a\nabla u) = f_s$ for smooth positive forcing fields f_s with perturbation scales $\{0.15, 0.25, 0.35, 0.45, 0.55, 0.65\}$, using $\{0.15, 0.35, 0.55\}$ during training and holding out $\{0.25, 0.45, 0.65\}$. Boundary-perturbed Darcy fixes a and $f = 1$ but changes the Dirichlet data g_s through smooth low-frequency boundary fields with amplitudes $\{0.04, 0.06, 0.08, 0.10, 0.12, 0.14\}$, using $\{0.04, 0.08, 0.12\}$ during training and holding out $\{0.06, 0.10, 0.14\}$. The model receives a, g_s, and the coordinates, and the target is the solution satisfying the changed boundary condition.

ConvResNet2D is trained for 200 epochs with 512 training, 128 validation, and 128 test samples for each of the forcing and boundary tasks, in both cases on 32×32 grids. Both Darcy protocols are run with an FNO2D backbone using nonperiodic spatial padding as an architecture-control experiment. Even with this nonperiodic padding, FNO2D was far less accurate than ConvResNet2D on the Darcy tasks, so ConvResNet2D is the strong baseline for the Darcy comparisons in the main text.

Optimization, selection, and hardware. Every training and evaluation job appends one record per task, method, seed, and severity cell to a single canonical result table, and all numerical tables reported here are generated from that table by a deterministic reporting step. Training runs are process-parallel with one graphics processing unit (GPU) per job on 12 Intel Data Center GPU Max 1550 accelerators accessed through PyTorch's `xpu` backend.

Unless stated otherwise, training uses adaptive moment estimation with decoupled weight decay, with learning rate $2 \cdot 10^{-3}$, weight decay 10^{-4}, gradient clipping at 1.0, cosine annealing over the full epoch budget, and model selection by the mean relative L^2 over seen validation severities. The primary baselines and interventions share the same split into training, validation, and test data, the same seeds, the same backbone width and layer settings, and the same changed-label targets for a given task; App. H states the seed protocol and the paired inference convention built on that protocol.

Table 6 summarizes the shared configuration.

Qualitative case selection. Every qualitative figure fixes the case shown before the compared methods are examined, by ranking held-out cases on one pre-specified statistic of a baseline model and never consulting the intervention. Figure 2 uses the empirical $q_{0.90}$ SAE of the parametric FNO, Figure 4 the median held-out SAE of the same model, and Figure 5 the $q_{0.90}$ held-out field error of the same model. The statistic differs across the three figures because each figure isolates a different failure. The two SAE rankings grade response alignment at a trained and at a swept severity, whereas the field-error ranking grades accuracy under label extrapolation. Aggregate results in every case come from the five seeds of Table 6.

H SEED PROTOCOL AND PAIRED SEED-LEVEL INFERENCE CONVENTION

This appendix collects the seed-level details behind every gain, CI, and p value reported in the main text.

Seed protocol. Every task is run with the same list of $R = 5$ seeds, and each result reported in the paper is averaged over those seeds. A seed controls parameter initialization, batch ordering, and the sampling of perturbation parameters from the severity grid, so baselines and interventions within a task share not only those seeds but also the split into training, validation, and test data, the backbone width and layer settings, and the changed-label targets. A comparison at a fixed seed therefore differs only in the training recipe, which is what makes the comparisons below paired rather than independent.

Every training and evaluation job writes one record per task, method, seed, and severity cell to the canonical result table, and the multi-seed summary tables first average each seed over the severities in a split and then report the mean $\pm$ standard deviation of those per-seed values.

Paired held-out gain. Let $\mathcal{G}_{\text{held}}$ be the held-out severity cells and let R be the number of matched seeds. For method m and seed r, first define

$$\overline{L}_{m,r} \coloneqq \frac{1}{|\mathcal{G}_{\text{held}}|} \sum_{\delta \in \mathcal{G}_{\text{held}}} L_{m,r,\delta}.$$

Unless stated otherwise, the reported paired held-out field-error gain is

$$\text{Gain}(m) \coloneqq \frac{100}{R} \sum_{r=1}^{R} \frac{\overline{L}_{\text{base},r} - \overline{L}_{m,r}}{\overline{L}_{\text{base},r}}.$$

Positive gain therefore means lower relative L^2 error for the method; negative gain means degradation. The same sign convention applies to the $q_{0.90}$, worst-severity, QoI, and response-diagnostic summaries.

Uncertainty and multiplicity. The reported 95% CI is the Student-t interval for these R seedwise percentage gains, whereas the paired p value is computed by a two-sided paired-sample t test on the raw matched pairs $(\overline{L}_{\text{base},r}, \overline{L}_{m,r})$. BH-adjusted q values (Benjamini & Hochberg, 1995) are reported only within the explicitly named 36-comparison family of Table 8; comparisons outside that family retain unadjusted paired p values. The tabulated q values are computed from the unrounded p values before display rounding.

Seed variation is only one of several sources of benchmark variance (Bouthillier et al., 2021), and with $R = 5$ seeds these CIs describe that source alone. Every comparison in the paper is therefore reported as a paired seedwise contrast under a shared data split rather than as an unpaired difference of means.

Descriptive ratios. For secondary comparisons outside the paired-ablation table, a percentage explicitly identified as a descriptive ratio is instead computed from the underlying unrounded held-out means whose rounded values appear in the summary table,

$$100 \frac{\overline{L}_{\text{base}} - \overline{L}_{\text{method}}}{\overline{L}_{\text{base}}},$$

each mean being taken over seeds and held-out severity cells.

All tabulated gains are likewise computed from unrounded values, whereas the tabulated entries are rounded to four decimal places. An accompanying paired p value still uses the raw matched seed pairs. A descriptive ratio and the corresponding paired seedwise gain need not be numerically identical because averaging seedwise ratios does not commute with taking a ratio of averages. The trade-off audit in App. J uses this descriptive convention throughout.

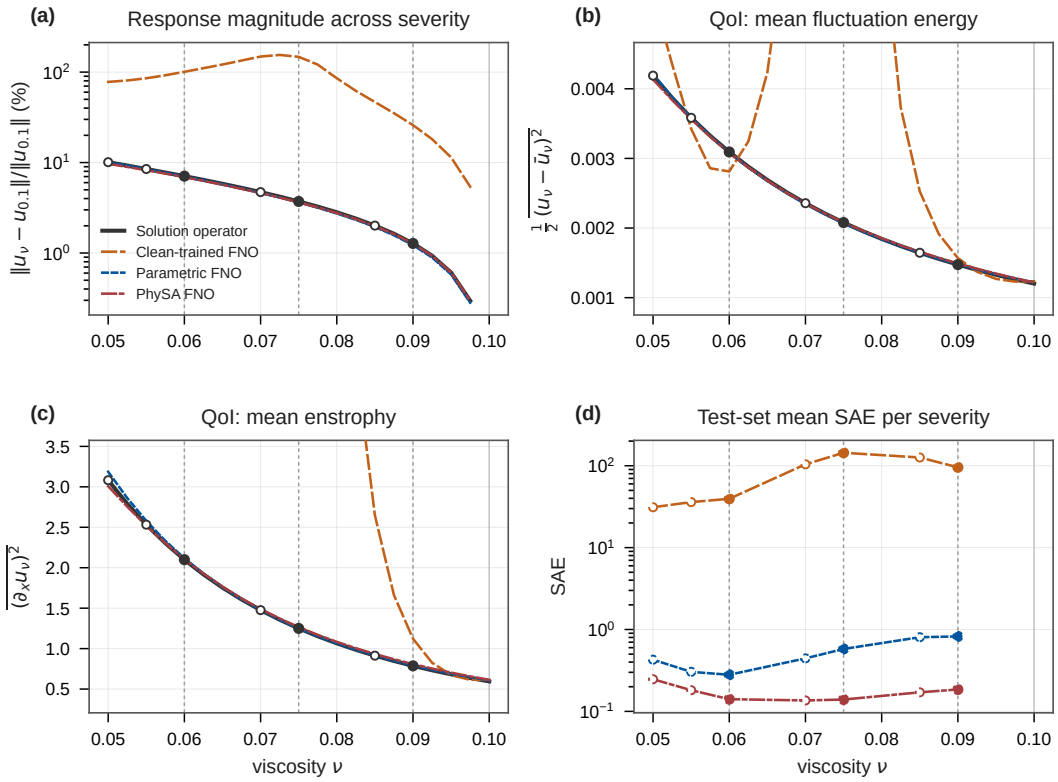

Figure 4: Dense response sweep between Burgers viscosity anchors. For a case selected as stated in App. G, curves cover 21 viscosities in $[0.05, 0.1]$ and dots mark benchmark severities; line styles distinguish models in addition to color. The clean model over-responds by one to two orders of magnitude. Parametric and PhySA track response magnitude and physical QoIs similarly, but PhySA reduces mean SAE from 0.52 to 0.17 across the grid. This response sweep provides a qualitative consistency check, not an estimate of the anchor certificate's rate.

I Full Result Tables

This appendix collects, in four groups, the tables referenced in Sec. 4.

The evidence hierarchy of Sec. 4.1 separates two levels of empirical evidence. Primary evidence comes from strong-baseline comparisons: FNO or ConvResNet backbones whose parametric changed-label recipes already reach low held-out field error, plus paired tests against parametric, weighted-parametric, and finite-grid adversarial baselines. A physical robustness framework should explain when a response-oriented objective helps and when a simpler domain-randomized baseline is already enough. Secondary evidence comes from weaker or mismatched backbones, such as DeepONet-style 1D runs and FNO2D Darcy runs; these secondary runs show architecture dependence and diagnostic behavior rather than state-of-the-art accuracy, and App. K reports those runs as within-architecture replications rather than as evidence about an operator family. Table 10 records each task's evidence source, target-generation route, and result-row count.

Table 7 summarizes the primary held-out comparisons. Table 8 is the primary inferential comparison family in full, and Table 9 resolves the Burgers viscosity task per severity for the FNO backbone. Table 1 in the main text gives the condensed multi-seed summary, and Table 10 records the evidence source and target-generation route of each task. Tables 11 through 16 give the complete multi-seed summaries, one task per table, each with a continuation that reports the response diagnostics. Tables 17 and 18 give the worst-severity and tail proxies together with the QoI and physics diagnostics. Tables 19 through 24 give the per-severity robustness envelopes for every task, in the task order used in Sec. 4 and under the column and entry conventions of Table 9. The compact comparison and envelope tables abbreviate parametric training as param and adversarial training as adv.

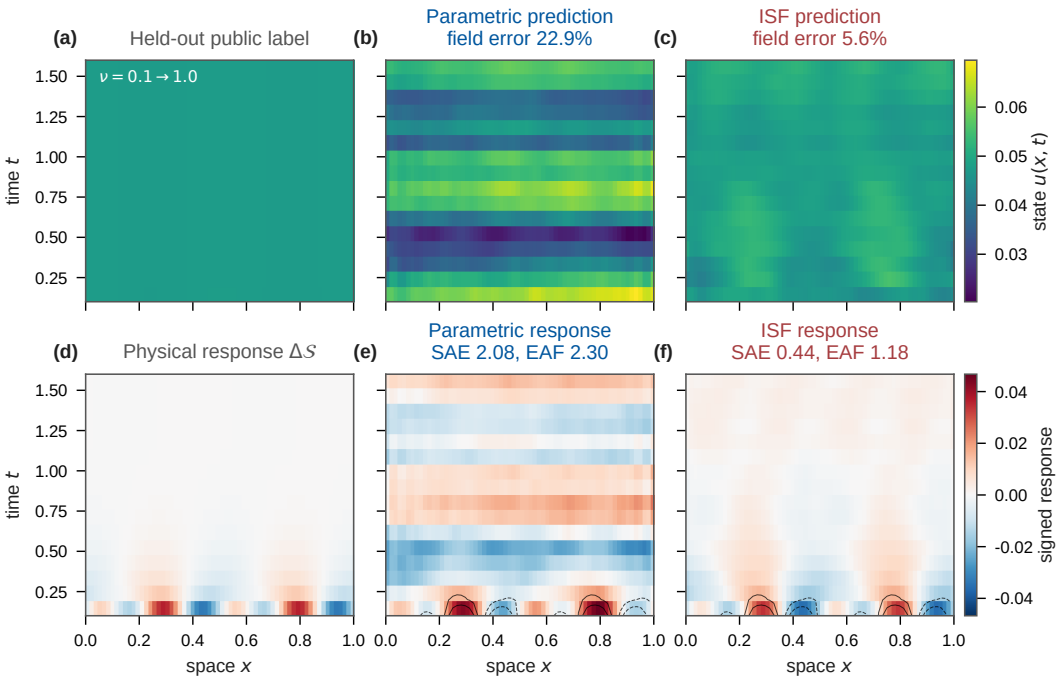

Figure 5: Response alignment under public-label extrapolation. At held-out $\nu = 1.0$, $2.5\times$ the largest seen viscosity, the target is streamed from PDEBench. The case is selected without consulting ISF (App. G). The reference response is strongest early and fades over time. Parametric FNO retains strong bands at later times, with 22.9% field error and SAE 2.08, whereas ISF more closely follows the early reference response and the subsequent decay, with 5.6% field error and SAE 0.44. Table 7 gives the five-seed comparison.

Table 7: Primary held-out comparisons. Each intervention is compared with that task's changed-label parametric baseline, and gain is the paired reduction in field error, so positive is better. Intervals are 95% CIs for the seedwise gains, and BH q is adjusted over the 36 comparisons in Table 8. The compact backbone labels are defined in Sec. 4.1.

Task	Intervention	Gain (%)	95% CI	BH q
Burgers viscosity	FNO-PASA	1.9	$[-11.4, 15.2]$	0.7520
Burgers viscosity	Conv1D-PhySA	11.7	$[5.1, 18.3]$	0.0389
Public Burgers viscosity	FNO-ISF	29.7	$[12.5, 46.9]$	0.0814
Evaluation mesh	FNO-PASA	21.5	$[18.1, 24.8]$	0.0027
Initial condition	FNO-PASA	16.2	$[13.4, 19.0]$	0.0024
Darcy forcing	Conv2D-PhySA	3.4	$[0.5, 6.2]$	0.0814
Darcy boundary	Conv2D-PhySA	6.4	$[2.5, 10.4]$	0.0389

Table 8: Paired held-out ablations for the pre-specified primary inferential comparison family, grouped by task in the italic rows. Positive gain means the method on the right reduces held-out relative L^2 compared with the method on the left under the same seed. CIs are Student-t intervals for the seedwise percentage gains defined in App. H; q is BH adjusted over the 36 comparisons using unrounded p values.

Comparison	Gain (%)	95% CI	p	BH q
Burgers viscosity				
FNO-param→FNO-weighted	−4.6093	$[-16.1090, 6.8904]$	0.3221	0.4848
FNO-param→FNO-adv	−1.6814	$[-14.3476, 10.9847]$	0.7580	0.8026
FNO-adv→FNO-PASA	2.9388	$[-12.6144, 18.4920]$	0.5599	0.7096
FNO-weighted→FNO-PASA	6.2825	$[1.0286, 11.5364]$	0.0339	0.0814
FNO-param→FNO-PASA	1.8987	$[-11.3861, 15.1836]$	0.6651	0.7520
Conv1D-param→Conv1D-weighted	0.7888	$[-4.9554, 6.5329]$	0.6684	0.7520
Conv1D-weighted→Conv1D-PASA	4.7336	$[3.8506, 5.6166]$	0.0002	0.0024
Conv1D-param→Conv1D-PASA	5.5008	$[0.4931, 10.5085]$	0.0455	0.0946
Conv1D-param→Conv1D-PhySA	11.7359	$[5.1242, 18.3476]$	0.0103	0.0389
Conv1D-param→Conv1D-ISF	11.9056	$[5.1378, 18.6734]$	0.0105	0.0389
Public Burgers viscosity				
FNO-param→FNO-weighted	1.0647	$[-10.0201, 12.1495]$	0.5716	0.7096
FNO-param→FNO-adv	4.8903	$[-5.6493, 15.4300]$	0.3225	0.4848
FNO-param→FNO-PASA	2.8006	$[-6.1252, 11.7265]$	0.3232	0.4848
FNO-param→FNO-PhySA	28.9713	$[11.2878, 46.6549]$	0.0287	0.0814
FNO-param→FNO-ISF	29.7306	$[12.5411, 46.9201]$	0.0263	0.0814
FNO-adv→FNO-ISF	25.9527	$[8.5749, 43.3305]$	0.0339	0.0814

Table 8: Paired held-out ablations (continued). Columns and multiplicity family follow the first part.

Comparison	Gain (%)	95% CI	p	BH q
Burgers evaluation mesh				
FNO-param→FNO-PASA	21.4521	$[18.0597, 24.8445]$	0.0003	0.0027
FNO-weighted→FNO-PASA	19.1256	$[11.5402, 26.7110]$	0.0081	0.0389
Conv1D-param→Conv1D-PASA	3.7343	$[2.3105, 5.1582]$	0.0020	0.0144
Burgers initial condition				
FNO-param→FNO-weighted	2.3903	$[-4.2202, 9.0008]$	0.4127	0.5943
FNO-param→FNO-adv	−27.8057	$[-32.7810, -22.8305]$	0.0002	0.0024
FNO-weighted→FNO-PASA	14.0288	$[9.6718, 18.3858]$	0.0032	0.0192
FNO-param→FNO-PASA	16.2130	$[13.4401, 18.9859]$	0.0001	0.0024
Conv1D-param→Conv1D-PhySA	−3.9854	$[-8.6379, 0.6671]$	0.0760	0.1440
Darcy forcing				
Conv2D-param→Conv2D-adv	2.3909	$[0.0233, 4.7586]$	0.0468	0.0946
Conv2D-adv→Conv2D-PhySA	0.9783	$[-2.8806, 4.8372]$	0.5252	0.7003
Conv2D-param→Conv2D-weighted	−0.4552	$[-4.0398, 3.1294]$	0.7320	0.7985
Conv2D-weighted→Conv2D-PASA	−0.0454	$[-2.7279, 2.6372]$	0.9800	0.9800
Conv2D-param→Conv2D-PASA	−0.4617	$[-2.6879, 1.7645]$	0.6102	0.7322
Conv2D-param→Conv2D-PhySA	3.3771	$[0.5222, 6.2319]$	0.0301	0.0814
Darcy boundary				
Conv2D-param→Conv2D-adv	0.3194	$[-4.2849, 4.9238]$	0.8469	0.8711
Conv2D-adv→Conv2D-PhySA	6.0318	$[0.2494, 11.8141]$	0.0473	0.0946
Conv2D-param→Conv2D-weighted	0.3786	$[-0.8614, 1.6187]$	0.4560	0.6314
Conv2D-weighted→Conv2D-PASA	−0.8714	$[-2.2957, 0.5528]$	0.1654	0.2977
Conv2D-param→Conv2D-PASA	−0.4834	$[-1.5947, 0.6279]$	0.2899	0.4848
Conv2D-param→Conv2D-PhySA	6.4322	$[2.4696, 10.3948]$	0.0108	0.0389

Table 9: Burgers viscosity robustness envelope for the FNO backbone, excerpted from Table 19. Viscosity decreases monotonically from $\nu = 0.1$ at the clean column C to $\nu = 0.05$ at S7*, and asterisks mark held-out severities that no method trains on. Entries are mean relative L^2 over five seeds, so lower is better. The label-preserving recipes degrade by close to two orders of magnitude, whereas every changed-label recipe stays within a factor of four of that recipe's own clean error.

Method	C	S1	S2*	S3	S4*	S5	S6*	S7*
ν	0.100	0.090	0.085	0.075	0.070	0.060	0.055	0.050
FNO-clean	0.0310	0.3885	0.6233	1.3698	1.9527	2.7243	2.9671	2.8921
FNO-noise	0.0413	0.3532	0.5808	1.2988	1.9438	2.7486	2.9011	2.8293
FNO-param	0.0106	0.0100	0.0102	0.0110	0.0116	0.0140	0.0190	0.0314
FNO-adv	0.0136	0.0132	0.0133	0.0132	0.0129	0.0126	0.0172	0.0300
FNO-PASA	0.0086	0.0084	0.0090	0.0102	0.0107	0.0125	0.0184	0.0327

Table 10: Data and target provenance. Burgers nominal specifications come from PDEBench; changed targets are streamed when public labels exist and otherwise generated by the stated solver. Rows count evaluation records.

Task	Specification shift	Evidence source	Target generation	Rows
Burgers viscosity	viscosity	public initial data	pseudo-spectral solves	1200
Public Burgers	viscosity	public changed labels	streamed trajectories	240
Evaluation mesh	mesh	public initial data	periodic interpolation	750
Initial condition	initial field	public initial data	pseudo-spectral solves	750
Darcy forcing	forcing field	controlled benchmark	elliptic solves	490
Darcy boundary	boundary data	controlled benchmark	elliptic solves	490

Table 11: Full multi-seed summary for Burgers viscosity: field errors. Each seed is first averaged over the severities in the indicated split; entries report mean ± standard deviation across five seeds. A dash denotes a split that is not defined for the task.

Method	Clean L^2	Seen L^2	Held-out L^2
FNO-clean	0.0310 ± 0.0286	1.4942 ± 0.5165	2.1088 ± 0.7940
FNO-noise	0.0413 ± 0.0241	1.4669 ± 0.5411	2.0638 ± 0.7642
FNO-parametric	0.0106 ± 0.0006	0.0117 ± 0.0007	0.0181 ± 0.0008
FNO-weighted	0.0103 ± 0.0011	0.0118 ± 0.0010	0.0189 ± 0.0013
FNO-adversarial	0.0136 ± 0.0010	0.0130 ± 0.0008	0.0184 ± 0.0017
FNO-PASA	0.0086 ± 0.0006	0.0104 ± 0.0005	0.0177 ± 0.0014
FNO-PhySA	0.0241 ± 0.0028	0.0251 ± 0.0026	0.0292 ± 0.0020
FNO-ISF	0.0250 ± 0.0029	0.0258 ± 0.0026	0.0299 ± 0.0022
Transformer-clean	0.3207 ± 0.1361	1.3363 ± 0.3102	1.5728 ± 0.3785
Transformer-noise	0.2303 ± 0.1511	1.2846 ± 0.2112	1.4690 ± 0.2199
Transformer-parametric	0.0859 ± 0.0086	0.0911 ± 0.0101	0.1067 ± 0.0101
Transformer-weighted	0.0958 ± 0.0138	0.0987 ± 0.0166	0.1122 ± 0.0176
Transformer-adversarial	0.0949 ± 0.0080	0.0898 ± 0.0083	0.1099 ± 0.0175
Transformer-PASA	0.0929 ± 0.0175	0.0963 ± 0.0203	0.1110 ± 0.0210
Transformer-PhySA	0.2262 ± 0.0119	0.2301 ± 0.0109	0.2313 ± 0.0096
Transformer-ISF	0.2266 ± 0.0110	0.2307 ± 0.0107	0.2322 ± 0.0096
Conv1D-clean	0.0631 ± 0.0276	1.5985 ± 0.3969	1.8850 ± 0.5183
Conv1D-noise	0.0516 ± 0.0201	1.6223 ± 0.5273	1.9954 ± 0.6988
Conv1D-parametric	0.0437 ± 0.0010	0.0435 ± 0.0023	0.0474 ± 0.0029
Conv1D-weighted	0.0417 ± 0.0019	0.0428 ± 0.0015	0.0469 ± 0.0013
Conv1D-adversarial	0.0455 ± 0.0020	0.0441 ± 0.0018	0.0465 ± 0.0024
Conv1D-PASA	0.0383 ± 0.0015	0.0403 ± 0.0013	0.0447 ± 0.0012
Conv1D-PhySA	0.0373 ± 0.0009	0.0389 ± 0.0012	0.0417 ± 0.0014
Conv1D-ISF	0.0371 ± 0.0014	0.0388 ± 0.0016	0.0416 ± 0.0016
DeepONet-parametric	0.3615 ± 0.0309	0.3962 ± 0.0309	0.4085 ± 0.0280
DeepONet-weighted	0.3441 ± 0.0198	0.3766 ± 0.0200	0.3900 ± 0.0184
DeepONet-adversarial	0.3491 ± 0.0193	0.3824 ± 0.0181	0.3963 ± 0.0161
DeepONet-PASA	0.3622 ± 0.0061	0.3940 ± 0.0035	0.4054 ± 0.0029
DeepONet-PhySA	0.3233 ± 0.0118	0.3575 ± 0.0104	0.3728 ± 0.0091
DeepONet-ISF	0.3285 ± 0.0178	0.3620 ± 0.0165	0.3765 ± 0.0147

Table 11: Full multi-seed summary for Burgers viscosity: response diagnostics (continued). EAF has a matched-response reference of one when the physical-response term sets the denominator; a dash denotes a diagnostic that is not defined for the task.

Method	Held-out SAE	Held-out EAF	Held-out SNRE
FNO-clean	65.2646 ± 17.7603	65.1529 ± 17.7412	65.8893 ± 18.5399
FNO-noise	64.4006 ± 19.0575	64.2963 ± 19.0487	63.6126 ± 19.6436
FNO-parametric	0.5006 ± 0.0419	1.3055 ± 0.0357	0.5857 ± 0.0483
FNO-weighted	0.5052 ± 0.0282	1.3072 ± 0.0244	0.5864 ± 0.0434
FNO-adversarial	0.5404 ± 0.0402	1.3453 ± 0.0330	0.6683 ± 0.0518
FNO-PASA	0.4736 ± 0.0199	1.2790 ± 0.0185	0.5165 ± 0.0300
FNO-PhySA	0.1860 ± 0.0069	0.9956 ± 0.0042	1.2574 ± 0.1715
FNO-ISF	0.1911 ± 0.0069	0.9983 ± 0.0077	1.2893 ± 0.1635
Transformer-clean	63.7310 ± 14.9441	63.6483 ± 14.9214	65.7504 ± 17.3976
Transformer-noise	64.2321 ± 13.6728	64.0798 ± 13.6506	67.4580 ± 10.3938
Transformer-parametric	2.2266 ± 0.1754	2.5841 ± 0.1367	4.5464 ± 0.3566
Transformer-weighted	2.2994 ± 0.2189	2.6315 ± 0.1897	4.6325 ± 0.2807
Transformer-adversarial	2.4444 ± 0.1535	2.8276 ± 0.1256	4.6139 ± 0.3000
Transformer-PASA	2.2536 ± 0.2374	2.5788 ± 0.1965	4.4973 ± 0.2897
Transformer-PhySA	0.8397 ± 0.0219	0.6861 ± 0.0278	7.6106 ± 0.4030
Transformer-ISF	0.8311 ± 0.0197	0.6850 ± 0.0330	7.5457 ± 0.4097
Conv1D-clean	60.1753 ± 6.0904	60.0269 ± 6.1267	60.1594 ± 6.2127
Conv1D-noise	59.3476 ± 8.5557	59.1845 ± 8.5742	59.2982 ± 9.1826
Conv1D-parametric	0.9902 ± 0.0508	1.6506 ± 0.0367	1.8200 ± 0.0905
Conv1D-weighted	0.9925 ± 0.0634	1.6553 ± 0.0566	1.7807 ± 0.0821
Conv1D-adversarial	1.0857 ± 0.0761	1.7531 ± 0.0641	1.9260 ± 0.0544
Conv1D-PASA	0.9291 ± 0.0479	1.5936 ± 0.0426	1.6208 ± 0.0854
Conv1D-PhySA	0.2477 ± 0.0084	0.9739 ± 0.0026	1.3985 ± 0.1033
Conv1D-ISF	0.2461 ± 0.0087	0.9747 ± 0.0038	1.4077 ± 0.0969
DeepONet-parametric	1.6647 ± 0.1668	0.9728 ± 0.2106	8.9047 ± 1.2067
DeepONet-weighted	1.6452 ± 0.0869	0.9388 ± 0.1272	8.5177 ± 1.0022
DeepONet-adversarial	2.0421 ± 0.2510	1.4327 ± 0.3036	9.3878 ± 1.2552
DeepONet-PASA	1.6151 ± 0.1331	0.8890 ± 0.1853	8.7879 ± 0.7824
DeepONet-PhySA	1.0448 ± 0.0164	0.1420 ± 0.0272	6.4992 ± 0.4309
DeepONet-ISF	1.0374 ± 0.0111	0.1259 ± 0.0265	6.5352 ± 0.4310

Table 12: Full multi-seed summary for Public Burgers viscosity: field errors. Each seed is first averaged over the severities in the indicated split; entries report mean $\pm$ standard deviation across five seeds. A dash denotes a split that is not defined for the task.

Method	Clean L^2	Seen L^2	Held-out L^2
FNO-clean	0.0208 ± 0.0043	0.8311 ± 0.2750	1.7500 ± 0.5849
FNO-noise	0.0274 ± 0.0140	0.8112 ± 0.2810	1.7020 ± 0.6006
FNO-parametric	0.0238 ± 0.0022	0.0286 ± 0.0033	0.2042 ± 0.0518
FNO-weighted	0.0226 ± 0.0029	0.0287 ± 0.0028	0.1990 ± 0.0386
FNO-adversarial	0.0318 ± 0.0032	0.0362 ± 0.0036	0.1931 ± 0.0465
FNO-PASA	0.0213 ± 0.0026	0.0271 ± 0.0026	0.1962 ± 0.0408
FNO-PhySA	0.0337 ± 0.0010	0.0537 ± 0.0014	0.1393 ± 0.0093
FNO-ISF	0.0330 ± 0.0016	0.0530 ± 0.0014	0.1379 ± 0.0092

Table 12: Full multi-seed summary for Public Burgers viscosity: response diagnostics (continued). EAF has a matched-response reference of one when the physical-response term sets the denominator; a dash denotes a diagnostic that is not defined for the task.

Method	Held-out SAE	Held-out EAF	Held-out SNRE
FNO-clean	98.0631 ± 15.4842	97.8992 ± 15.4753	98.0500 ± 15.5642
FNO-noise	94.6438 ± 15.2990	94.4781 ± 15.2874	94.6185 ± 15.2701
FNO-parametric	5.9812 ± 2.0675	6.3784 ± 2.0619	6.1750 ± 2.1060
FNO-weighted	6.4283 ± 2.3017	6.8307 ± 2.3159	6.7018 ± 2.4676
FNO-adversarial	5.2457 ± 1.3290	5.6607 ± 1.3741	5.4255 ± 1.3871
FNO-PASA	6.2212 ± 2.0985	6.6252 ± 2.1168	6.4238 ± 2.2148
FNO-PhySA	1.0123 ± 0.3150	1.3958 ± 0.3093	2.5221 ± 1.6173
FNO-ISF	1.0237 ± 0.3786	1.4120 ± 0.3729	2.4639 ± 1.5706

Table 13: Full multi-seed summary for Burgers evaluation mesh: field errors. Each seed is first averaged over the severities in the indicated split; entries report mean $\pm$ standard deviation across five seeds. A dash denotes a split that is not defined for the task.

Method	Clean L^2	Seen L^2	Held-out L^2
FNO-clean	0.0450 ± 0.0320	–	0.0471 ± 0.0319
FNO-noise	0.0413 ± 0.0241	–	0.0415 ± 0.0237
FNO-parametric	0.0106 ± 0.0006	–	0.0125 ± 0.0012
FNO-weighted	0.0103 ± 0.0011	–	0.0123 ± 0.0019
FNO-adversarial	0.0136 ± 0.0010	–	0.0160 ± 0.0018
FNO-PASA	0.0086 ± 0.0006	–	0.0098 ± 0.0009
FNO-PhySA	0.0241 ± 0.0028	–	0.0313 ± 0.0033
FNO-ISF	0.0250 ± 0.0029	–	0.0319 ± 0.0035
Transformer-clean	0.3693 ± 0.0911	–	0.4941 ± 0.0997
Transformer-noise	0.2303 ± 0.1511	–	0.3287 ± 0.1822
Transformer-parametric	0.0859 ± 0.0086	–	0.1726 ± 0.0144
Transformer-weighted	0.0958 ± 0.0138	–	0.1882 ± 0.0314
Transformer-adversarial	0.0949 ± 0.0080	–	0.1842 ± 0.0107
Transformer-PASA	0.0929 ± 0.0175	–	0.1818 ± 0.0360
Transformer-PhySA	0.2262 ± 0.0119	–	0.3766 ± 0.0075
Transformer-ISF	0.2266 ± 0.0110	–	0.3731 ± 0.0082
Conv1D-clean	0.0590 ± 0.0150	–	0.2977 ± 0.0079
Conv1D-noise	0.0516 ± 0.0201	–	0.2906 ± 0.0052
Conv1D-parametric	0.0437 ± 0.0010	–	0.2930 ± 0.0038
Conv1D-weighted	0.0417 ± 0.0019	–	0.2909 ± 0.0047
Conv1D-adversarial	0.0455 ± 0.0020	–	0.2934 ± 0.0039
Conv1D-PASA	0.0383 ± 0.0015	–	0.2821 ± 0.0036
Conv1D-PhySA	0.0373 ± 0.0009	–	0.2832 ± 0.0038
Conv1D-ISF	0.0371 ± 0.0014	–	0.2828 ± 0.0035
DeepONet-parametric	0.3615 ± 0.0309	–	0.3598 ± 0.0195
DeepONet-weighted	0.3441 ± 0.0198	–	0.3639 ± 0.0158
DeepONet-adversarial	0.3491 ± 0.0193	–	0.3809 ± 0.0249
DeepONet-PASA	0.3622 ± 0.0061	–	0.3645 ± 0.0049
DeepONet-PhySA	0.3233 ± 0.0118	–	0.3460 ± 0.0120
DeepONet-ISF	0.3285 ± 0.0178	–	0.3564 ± 0.0210

Table 13: Full multi-seed summary for Burgers evaluation mesh: response diagnostics (continued). EAF has a matched-response reference of one when the physical-response term sets the denominator; a dash denotes a diagnostic that is not defined for the task.

Method	Held-out SAE	Held-out EAF	Held-out SNRE
FNO-clean	0.0136 ± 0.0058	–	–
FNO-noise	0.0056 ± 0.0021	–	–
FNO-parametric	0.0070 ± 0.0014	–	–
FNO-weighted	0.0069 ± 0.0018	–	–
FNO-adversarial	0.0088 ± 0.0017	–	–
FNO-PASA	0.0045 ± 0.0010	–	–
FNO-PhySA	0.0205 ± 0.0028	–	–
FNO-ISF	0.0208 ± 0.0028	–	–
Transformer-clean	0.3583 ± 0.0782	–	–
Transformer-noise	0.2528 ± 0.1318	–	–
Transformer-parametric	0.1582 ± 0.0127	–	–
Transformer-weighted	0.1716 ± 0.0293	–	–
Transformer-adversarial	0.1648 ± 0.0085	–	–
Transformer-PASA	0.1636 ± 0.0325	–	–
Transformer-PhySA	0.3173 ± 0.0087	–	–
Transformer-ISF	0.3128 ± 0.0058	–	–
Conv1D-clean	0.2914 ± 0.0094	–	–
Conv1D-noise	0.2837 ± 0.0056	–	–
Conv1D-parametric	0.2913 ± 0.0030	–	–
Conv1D-weighted	0.2890 ± 0.0045	–	–
Conv1D-adversarial	0.2906 ± 0.0035	–	–
Conv1D-PASA	0.2806 ± 0.0036	–	–
Conv1D-PhySA	0.2799 ± 0.0033	–	–
Conv1D-ISF	0.2797 ± 0.0031	–	–
DeepONet-parametric	0.1421 ± 0.0325	–	–
DeepONet-weighted	0.1329 ± 0.0235	–	–
DeepONet-adversarial	0.1606 ± 0.0565	–	–
DeepONet-PASA	0.1306 ± 0.0200	–	–
DeepONet-PhySA	0.1057 ± 0.0205	–	–
DeepONet-ISF	0.1183 ± 0.0312	–	–

Table 14: Full multi-seed summary for Burgers initial condition: field errors. Each seed is first averaged over the severities in the indicated split; entries report mean ± standard deviation across five seeds. A dash denotes a split that is not defined for the task.

Method	Clean L^2	Seen L^2	Held-out L^2
FNO-clean	0.0450 ± 0.0320	–	0.0287 ± 0.0193
FNO-noise	0.0413 ± 0.0241	–	0.0279 ± 0.0161
FNO-parametric	0.0106 ± 0.0006	–	0.0078 ± 0.0004
FNO-weighted	0.0103 ± 0.0011	–	0.0076 ± 0.0008
FNO-adversarial	0.0136 ± 0.0010	–	0.0100 ± 0.0007
FNO-PASA	0.0086 ± 0.0006	–	0.0065 ± 0.0004
FNO-PhySA	0.0241 ± 0.0028	–	0.0168 ± 0.0013
FNO-ISF	0.0250 ± 0.0029	–	0.0173 ± 0.0016
Transformer-clean	0.3693 ± 0.0911	–	0.2352 ± 0.0369
Transformer-noise	0.2303 ± 0.1511	–	0.1519 ± 0.0886
Transformer-parametric	0.0859 ± 0.0086	–	0.0620 ± 0.0068
Transformer-weighted	0.0958 ± 0.0138	–	0.0703 ± 0.0172
Transformer-adversarial	0.0949 ± 0.0080	–	0.0692 ± 0.0079
Transformer-PASA	0.0929 ± 0.0175	–	0.0689 ± 0.0194
Transformer-PhySA	0.2262 ± 0.0119	–	0.1864 ± 0.0032
Transformer-ISF	0.2266 ± 0.0110	–	0.1854 ± 0.0026
Conv1D-clean	0.0590 ± 0.0150	–	0.0427 ± 0.0077
Conv1D-noise	0.0516 ± 0.0201	–	0.0386 ± 0.0108
Conv1D-parametric	0.0437 ± 0.0010	–	0.0321 ± 0.0002
Conv1D-weighted	0.0417 ± 0.0019	–	0.0314 ± 0.0003
Conv1D-adversarial	0.0455 ± 0.0020	–	0.0350 ± 0.0012
Conv1D-PASA	0.0383 ± 0.0015	–	0.0308 ± 0.0005
Conv1D-PhySA	0.0373 ± 0.0009	–	0.0334 ± 0.0012
Conv1D-ISF	0.0371 ± 0.0014	–	0.0334 ± 0.0013
DeepONet-parametric	0.3615 ± 0.0309	–	0.3228 ± 0.0063
DeepONet-weighted	0.3441 ± 0.0198	–	0.3226 ± 0.0048
DeepONet-adversarial	0.3491 ± 0.0193	–	0.3288 ± 0.0057
DeepONet-PASA	0.3622 ± 0.0061	–	0.3248 ± 0.0046
DeepONet-PhySA	0.3233 ± 0.0118	–	0.3172 ± 0.0033
DeepONet-ISF	0.3285 ± 0.0178	–	0.3154 ± 0.0027

Table 14: Full multi-seed summary for Burgers initial condition: response diagnostics (continued). EAF has a matched-response reference of one when the physical-response term sets the denominator; a dash denotes a diagnostic that is not defined for the task.

Method	Held-out SAE	Held-out EAF	Held-out SNRE
FNO-clean	0.0433 ± 0.0161	1.0078 ± 0.0061	0.6127 ± 0.4222
FNO-noise	0.0406 ± 0.0135	0.9989 ± 0.0133	0.6580 ± 0.4155
FNO-parametric	0.0217 ± 0.0008	1.0005 ± 0.0004	0.1586 ± 0.0081
FNO-weighted	0.0213 ± 0.0017	1.0001 ± 0.0005	0.1549 ± 0.0129
FNO-adversarial	0.0273 ± 0.0014	0.9991 ± 0.0006	0.2018 ± 0.0105
FNO-PASA	0.0200 ± 0.0014	1.0002 ± 0.0006	0.1374 ± 0.0077
FNO-PhySA	0.0560 ± 0.0038	0.9951 ± 0.0016	0.3245 ± 0.0260
FNO-ISF	0.0568 ± 0.0044	0.9955 ± 0.0018	0.3323 ± 0.0309
Transformer-clean	0.5517 ± 0.0152	0.6833 ± 0.0500	4.2419 ± 0.6878
Transformer-noise	0.4141 ± 0.2067	0.7874 ± 0.1514	2.6785 ± 1.4926
Transformer-parametric	0.1895 ± 0.0333	0.9505 ± 0.0194	1.1879 ± 0.1394
Transformer-weighted	0.2366 ± 0.1175	0.9193 ± 0.0716	1.3133 ± 0.3094
Transformer-adversarial	0.2128 ± 0.0393	0.9196 ± 0.0283	1.3497 ± 0.1713
Transformer-PASA	0.2368 ± 0.1241	0.9202 ± 0.0764	1.2907 ± 0.3413
Transformer-PhySA	0.6014 ± 0.0101	0.7123 ± 0.0166	3.2977 ± 0.0856
Transformer-ISF	0.5971 ± 0.0079	0.7123 ± 0.0149	3.2548 ± 0.0418
Conv1D-clean	0.1904 ± 0.0112	0.9438 ± 0.0151	0.8772 ± 0.1682
Conv1D-noise	0.1963 ± 0.0107	0.9341 ± 0.0067	0.7959 ± 0.2444
Conv1D-parametric	0.1798 ± 0.0141	0.9480 ± 0.0071	0.6459 ± 0.0191
Conv1D-weighted	0.1788 ± 0.0140	0.9519 ± 0.0072	0.6370 ± 0.0126
Conv1D-adversarial	0.1946 ± 0.0183	0.9426 ± 0.0103	0.7017 ± 0.0090
Conv1D-PASA	0.1845 ± 0.0164	0.9476 ± 0.0081	0.6256 ± 0.0132
Conv1D-PhySA	0.2110 ± 0.0103	0.9383 ± 0.0059	0.6284 ± 0.0244
Conv1D-ISF	0.2133 ± 0.0099	0.9348 ± 0.0064	0.6279 ± 0.0245
DeepONet-parametric	1.1183 ± 0.0680	0.4631 ± 0.1366	5.8669 ± 0.2370
DeepONet-weighted	1.0884 ± 0.0232	0.4080 ± 0.0514	5.7964 ± 0.1941
DeepONet-adversarial	1.1653 ± 0.1365	0.5097 ± 0.2192	6.0002 ± 0.2198
DeepONet-PASA	1.0906 ± 0.0317	0.3935 ± 0.0437	5.8192 ± 0.1630
DeepONet-PhySA	1.1069 ± 0.0926	0.4161 ± 0.1594	5.4558 ± 0.0669
DeepONet-ISF	1.0954 ± 0.0498	0.4192 ± 0.0882	5.4623 ± 0.0888

Table 15: Full multi-seed summary for Darcy forcing: field errors. Each seed is first averaged over the severities in the indicated split; entries report mean $\pm$ standard deviation across five seeds. A dash denotes a split that is not defined for the task.

Method	Clean L^2	Seen L^2	Held-out L^2
Conv2D-clean	0.0226 ± 0.0039	0.5189 ± 0.1100	0.6091 ± 0.1144
Conv2D-noise	0.0183 ± 0.0005	0.1897 ± 0.0279	0.2510 ± 0.0393
Conv2D-parametric	0.0196 ± 0.0007	0.0239 ± 0.0006	0.0266 ± 0.0007
Conv2D-weighted	0.0195 ± 0.0007	0.0239 ± 0.0010	0.0268 ± 0.0012
Conv2D-adversarial	0.0208 ± 0.0003	0.0240 ± 0.0004	0.0260 ± 0.0006
Conv2D-PASA	0.0195 ± 0.0005	0.0239 ± 0.0006	0.0267 ± 0.0007
Conv2D-PhySA	0.0182 ± 0.0008	0.0229 ± 0.0010	0.0257 ± 0.0011
FNO2D-clean	0.7406 ± 0.0379	0.8151 ± 0.0994	0.8351 ± 0.1171
FNO2D-noise	0.6930 ± 0.0778	0.7121 ± 0.0715	0.7235 ± 0.0691
FNO2D-parametric	0.6838 ± 0.0815	0.6975 ± 0.0811	0.7059 ± 0.0815
FNO2D-weighted	0.6941 ± 0.0680	0.7070 ± 0.0666	0.7148 ± 0.0661
FNO2D-adversarial	0.7012 ± 0.0743	0.7172 ± 0.0745	0.7278 ± 0.0752
FNO2D-PASA	0.6895 ± 0.0657	0.7029 ± 0.0641	0.7109 ± 0.0633
FNO2D-PhySA	0.6981 ± 0.0775	0.7140 ± 0.0802	0.7231 ± 0.0821

Table 15: Full multi-seed summary for Darcy forcing: response diagnostics (continued). EAF has a matched-response reference of one when the physical-response term sets the denominator; a dash denotes a diagnostic that is not defined for the task.

Method	Held-out SAE	Held-out EAF	Held-out SNRE
Conv2D-clean	4.0016 ± 0.8043	3.7232 ± 0.6908	4.0066 ± 0.7990
Conv2D-noise	1.4612 ± 0.2382	1.0686 ± 0.2693	1.4674 ± 0.2370
Conv2D-parametric	0.1053 ± 0.0054	1.0027 ± 0.0036	0.1804 ± 0.0029
Conv2D-weighted	0.1066 ± 0.0035	1.0001 ± 0.0050	0.1801 ± 0.0033
Conv2D-adversarial	0.0946 ± 0.0028	1.0013 ± 0.0039	0.1801 ± 0.0064
Conv2D-PASA	0.1065 ± 0.0024	1.0007 ± 0.0048	0.1800 ± 0.0013
Conv2D-PhySA	0.1043 ± 0.0029	1.0025 ± 0.0036	0.1728 ± 0.0058
FNO2D-clean	3.9263 ± 1.1500	3.8178 ± 1.3113	5.9849 ± 0.7190
FNO2D-noise	1.4765 ± 0.2722	1.1014 ± 0.2044	5.2408 ± 0.4473
FNO2D-parametric	1.2438 ± 0.2441	0.7265 ± 0.3940	5.1220 ± 0.5343
FNO2D-weighted	1.3085 ± 0.3061	0.8059 ± 0.3980	5.1958 ± 0.4050
FNO2D-adversarial	1.6832 ± 0.5535	1.3240 ± 0.6794	5.2611 ± 0.5107
FNO2D-PASA	1.3564 ± 0.2087	0.8842 ± 0.3122	5.1655 ± 0.3900
FNO2D-PhySA	1.2587 ± 0.2294	0.7565 ± 0.3758	5.2448 ± 0.4988

Table 16: Full multi-seed summary for Darcy boundary: field errors. Each seed is first averaged over the severities in the indicated split; entries report mean $\pm$ standard deviation across five seeds. A dash denotes a split that is not defined for the task.

Method	Clean L^2	Seen L^2	Held-out L^2
Conv2D-clean	0.0216 ± 0.0013	0.7780 ± 0.0069	0.8503 ± 0.0059
Conv2D-noise	0.0242 ± 0.0016	0.7751 ± 0.0054	0.8489 ± 0.0053
Conv2D-parametric	0.0432 ± 0.0005	0.0594 ± 0.0009	0.0634 ± 0.0011
Conv2D-weighted	0.0410 ± 0.0011	0.0591 ± 0.0014	0.0632 ± 0.0015
Conv2D-adversarial	0.0651 ± 0.0018	0.0618 ± 0.0021	0.0632 ± 0.0020
Conv2D-PASA	0.0416 ± 0.0010	0.0597 ± 0.0011	0.0637 ± 0.0013
Conv2D-PhySA	0.0328 ± 0.0007	0.0549 ± 0.0020	0.0593 ± 0.0020
FNO2D-clean	0.8131 ± 0.1196	1.0077 ± 0.0748	1.0242 ± 0.0720
FNO2D-noise	0.9667 ± 0.2599	0.9921 ± 0.0981	0.9956 ± 0.0700
FNO2D-parametric	1.5950 ± 0.4365	1.2065 ± 0.1083	1.1455 ± 0.0774
FNO2D-weighted	1.1869 ± 0.2812	1.1253 ± 0.1657	1.0950 ± 0.1266
FNO2D-adversarial	1.4342 ± 0.1531	1.2015 ± 0.0518	1.1468 ± 0.0487
FNO2D-PASA	1.3169 ± 0.5447	1.1660 ± 0.2705	1.1298 ± 0.2124
FNO2D-PhySA	1.2987 ± 0.3413	1.1523 ± 0.1982	1.1218 ± 0.1713

Table 16: Full multi-seed summary for Darcy boundary: response diagnostics (continued). EAF has a matched-response reference of one when the physical-response term sets the denominator; a dash denotes a diagnostic that is not defined for the task.

Method	Held-out SAE	Held-out EAF	Held-out SNRE
Conv2D-clean	1.0019 ± 0.0118	0.1260 ± 0.0180	1.0021 ± 0.0119
Conv2D-noise	1.0001 ± 0.0002	0.0022 ± 0.0004	1.0003 ± 0.0002
Conv2D-parametric	0.0736 ± 0.0011	0.9925 ± 0.0006	0.0750 ± 0.0016
Conv2D-weighted	0.0736 ± 0.0020	0.9921 ± 0.0009	0.0747 ± 0.0023
Conv2D-adversarial	0.0765 ± 0.0011	0.9904 ± 0.0017	0.0752 ± 0.0020
Conv2D-PASA	0.0739 ± 0.0018	0.9921 ± 0.0009	0.0753 ± 0.0017
Conv2D-PhySA	0.0689 ± 0.0023	0.9925 ± 0.0007	0.0700 ± 0.0022
FNO2D-clean	1.1365 ± 0.0941	0.5549 ± 0.1885	1.2203 ± 0.0894
FNO2D-noise	1.0018 ± 0.0048	0.0212 ± 0.0146	1.1871 ± 0.0941
FNO2D-parametric	1.4701 ± 0.3072	1.0703 ± 0.4074	1.3756 ± 0.1030
FNO2D-weighted	1.3109 ± 0.2380	0.8279 ± 0.3274	1.3110 ± 0.1649
FNO2D-adversarial	1.3603 ± 0.0768	0.9821 ± 0.1350	1.3784 ± 0.0656
FNO2D-PASA	1.3737 ± 0.4006	0.9047 ± 0.5025	1.3543 ± 0.2715
FNO2D-PhySA	1.2726 ± 0.1489	0.7713 ± 0.2112	1.3429 ± 0.2083

Table 17: Worst held-out severity and severity-cell tail proxies. The worst severity is the largest held-out-severity mean relative L^2 over seeds, and merged baseline cells mark shared recipes. Tail cells are test-set means for each pair of seed and severity, and $q_{0.90}$ is the 90th percentile of those cells, not a certified sample-level tail bound.

Comparison	Worst severity			Severity-cell tail				
				$q_{0.90}$		max		$q_{0.90}$ gain
	Base	Method	Gain (%)	Base	Method	Base	Method	(%)
Burgers viscosity								
FNO-param→FNO-PASA	0.0314	0.0327	−3.9	0.0323	0.0322	0.0329	0.0410	0.3
FNO-param→FNO-adv		0.0300	4.5		0.0281		0.0397	13.0
Conv1D-param→Conv1D-PhySA	0.0553	0.0472	14.7	0.0545	0.0473	0.0639	0.0493	13.2
Public Burgers viscosity								
FNO-param→FNO-PhySA	0.2735	0.1633	40.3	0.3393	0.1655	0.3610	0.1821	51.2
FNO-param→FNO-ISF		0.1598	41.6		0.1621		0.1730	52.2
FNO-param→FNO-adv		0.2352	14.0		0.2724		0.3164	19.7
Burgers evaluation mesh								
FNO-param→FNO-PASA	0.0138	0.0107	22.7	0.0147	0.0113	0.0162	0.0135	23.2
FNO-param→FNO-adv		0.0178	−29.0		0.0183		0.0217	−24.9
Burgers initial condition								
FNO-param→FNO-PASA	0.0081	0.0067	16.6	0.0082	0.0071	0.0086	0.0074	13.6
FNO-param→FNO-adv		0.0104	−28.1		0.0109		0.0113	−32.2
Conv1D-param→Conv1D-PhySA	0.0332	0.0362	−9.0	0.0332	0.0370	0.0340	0.0378	−11.5
Darcy forcing								
Conv2D-param→Conv2D-PhySA	0.0335	0.0328	2.1	0.0340	0.0329	0.0351	0.0351	3.2
Conv2D-param→Conv2D-adv		0.0310	7.3		0.0311		0.0330	8.6
Darcy boundary								
Conv2D-param→Conv2D-PhySA	0.0714	0.0681	4.6	0.0717	0.0685	0.0733	0.0709	4.5
Conv2D-param→Conv2D-adv		0.0676	5.3		0.0676		0.0710	5.7

Table 18: QoI and physics-diagnostic held-out summary for the primary comparisons. Each italic row also names the compared pair, so a block reports the three diagnostics of one comparison, in the task order used throughout Sec. 4. Entries are means over held-out seed-and-severity cells, with root-mean-square error (RMSE) reported for the PDE and boundary diagnostics. Gains are computed from the unrounded means, so a row whose two entries differ only in the fourth decimal place, such as the Darcy forcing center-pressure error, carries a much smaller gain than the rounded entries alone would suggest.

QoI or diagnostic	Baseline	Method	Gain (%)
Burgers viscosity: FNO-parametric→FNO-PASA			
Relative energy error	0.0454	0.0360	20.7
Peak absolute error	0.1057	0.1041	1.5
PDE RMSE	0.7735	0.4388	43.3
Burgers viscosity: Conv1D-parametric→Conv1D-PhySA			
Relative energy error	0.1596	0.0739	53.7
Peak absolute error	0.1389	0.1330	4.2
PDE RMSE	1.9581	1.3064	33.3
Public Burgers viscosity: FNO-parametric→FNO-ISF			
Relative energy error	0.9590	0.1036	89.2
Peak absolute error	0.2291	0.2213	3.4
PDE RMSE	9.8757	10.3240	−4.5
Burgers initial condition: FNO-parametric→FNO-PASA			
Relative energy error	0.0083	0.0087	−4.7
Peak absolute error	0.0400	0.0373	6.8
PDE RMSE	1.0229	0.4995	51.2
Darcy forcing: Conv2D-parametric→Conv2D-PhySA			
Relative maximum error	0.0176	0.0170	3.2
Center absolute error	0.0011	0.0012	−0.3
PDE RMSE	0.9338	0.8688	7.0
Darcy boundary: Conv2D-parametric→Conv2D-PhySA			
Relative mean error	0.1231	0.1120	9.0
Relative maximum error	0.0163	0.0149	8.4
Boundary RMSE	0.0036	0.0033	7.1

Table 19: Burgers robustness envelope for every backbone and method. Table 9 excerpts the FNO rows.

Method	C	S1	S2*	S3	S4*	S5	S6*	S7*
ν	0.100	0.090	0.085	0.075	0.070	0.060	0.055	0.050
FNO-clean	0.0310	0.3885	0.6233	1.3698	1.9527	2.7243	2.9671	2.8921
FNO-noise	0.0413	0.3532	0.5808	1.2988	1.9438	2.7486	2.9011	2.8293
FNO-param	0.0106	0.0100	0.0102	0.0110	0.0116	0.0140	0.0190	0.0314
FNO-weighted	0.0103	0.0099	0.0103	0.0115	0.0120	0.0140	0.0195	0.0336
FNO-adv	0.0136	0.0132	0.0133	0.0132	0.0129	0.0126	0.0172	0.0300
FNO-PASA	0.0086	0.0084	0.0090	0.0102	0.0107	0.0125	0.0184	0.0327
FNO-PhySA	0.0241	0.0240	0.0239	0.0241	0.0244	0.0271	0.0309	0.0376
FNO-ISF	0.0250	0.0248	0.0248	0.0249	0.0252	0.0279	0.0315	0.0381
Transformer-clean	0.3207	0.5178	0.7363	1.6681	2.3615	1.8231	1.6764	1.5170
Transformer-noise	0.2303	0.5676	0.8072	1.4241	1.8244	1.8622	1.7056	1.5389
Transformer-param	0.0859	0.0860	0.0888	0.0925	0.0927	0.0947	0.1053	0.1399
Transformer-weighted	0.0958	0.0944	0.0964	0.0993	0.0994	0.1025	0.1118	0.1413
Transformer-adv	0.0949	0.0916	0.0917	0.0905	0.0898	0.0873	0.1020	0.1561
Transformer-PASA	0.0929	0.0924	0.0944	0.0972	0.0971	0.0994	0.1102	0.1424
Transformer-PhySA	0.2262	0.2272	0.2280	0.2300	0.2312	0.2331	0.2331	0.2327
Transformer-ISF	0.2266	0.2276	0.2284	0.2305	0.2319	0.2341	0.2343	0.2342
Conv1D-clean	0.0631	0.5669	0.9727	1.9767	2.2101	2.2517	2.2134	2.1439
Conv1D-noise	0.0516	0.5468	0.9355	1.8847	2.2743	2.4352	2.4190	2.3527
Conv1D-param	0.0437	0.0422	0.0423	0.0430	0.0435	0.0452	0.0484	0.0553
Conv1D-weighted	0.0417	0.0406	0.0409	0.0422	0.0433	0.0457	0.0486	0.0549
Conv1D-adv	0.0455	0.0445	0.0446	0.0444	0.0442	0.0434	0.0452	0.0521
Conv1D-PASA	0.0383	0.0377	0.0382	0.0398	0.0409	0.0436	0.0466	0.0531
Conv1D-PhySA	0.0373	0.0372	0.0373	0.0381	0.0388	0.0413	0.0436	0.0472
Conv1D-ISF	0.0371	0.0370	0.0372	0.0380	0.0387	0.0412	0.0435	0.0472
DeepONet-param	0.3615	0.3745	0.3816	0.3966	0.4041	0.4175	0.4223	0.4262
DeepONet-weighted	0.3441	0.3556	0.3620	0.3761	0.3836	0.3980	0.4043	0.4102
DeepONet-adv	0.3491	0.3608	0.3673	0.3819	0.3898	0.4044	0.4106	0.4173
DeepONet-PASA	0.3622	0.3739	0.3804	0.3942	0.4013	0.4139	0.4183	0.4216
DeepONet-PhySA	0.3233	0.3354	0.3421	0.3566	0.3645	0.3805	0.3883	0.3964
DeepONet-ISF	0.3285	0.3404	0.3470	0.3612	0.3689	0.3843	0.3915	0.3986

Table 20: Public PDEBench Burgers multi-viscosity envelope, whose severity axis is the public viscosity ν. Targets are streamed from public HDF5 trajectories with shared initial conditions.

Method	C	S1*	S2	S3	S4	S5*
ν	0.10	0.02	0.04	0.20	0.40	1.00
FNO-clean	0.0208	0.8586	0.5537	0.5702	1.3694	2.6414
FNO-noise	0.0274	0.8478	0.5457	0.5588	1.3290	2.5562
FNO-param	0.0238	0.2735	0.0437	0.0197	0.0222	0.1349
FNO-weighted	0.0226	0.2537	0.0437	0.0203	0.0221	0.1444
FNO-adv	0.0318	0.2352	0.0319	0.0348	0.0418	0.1510
FNO-PASA	0.0213	0.2527	0.0430	0.0187	0.0196	0.1398
FNO-PhySA	0.0337	0.1633	0.0804	0.0343	0.0464	0.1153
FNO-ISF	0.0330	0.1598	0.0790	0.0339	0.0460	0.1160

Table 21: Burgers evaluation-mesh envelope. Columns give the evaluation grids in points. The 64-point column is the clean grid used for training and evaluation, and the remaining columns evaluate the same clean-viscosity PDE state on 32, 48, 96, and 128 points, every one of which is held out from training.

Method	64	32	48	96	128
FNO-clean	0.0450	0.0463	0.0461	0.0473	0.0486
FNO-noise	0.0413	0.0404	0.0410	0.0420	0.0425
FNO-param	0.0106	0.0118	0.0117	0.0128	0.0138
FNO-weighted	0.0103	0.0118	0.0115	0.0124	0.0134
FNO-adv	0.0136	0.0147	0.0150	0.0166	0.0178
FNO-PASA	0.0086	0.0107	0.0096	0.0093	0.0098
FNO-PhySA	0.0241	0.0239	0.0277	0.0346	0.0389
FNO-ISF	0.0250	0.0243	0.0284	0.0353	0.0397
Transformer-clean	0.3693	0.7449	0.4873	0.3422	0.4019
Transformer-noise	0.2303	0.5054	0.3146	0.2198	0.2749
Transformer-param	0.0859	0.2153	0.1328	0.1400	0.2025
Transformer-weighted	0.0958	0.2461	0.1492	0.1480	0.2095
Transformer-adv	0.0949	0.2347	0.1444	0.1473	0.2104
Transformer-PASA	0.0929	0.2389	0.1412	0.1433	0.2036
Transformer-PhySA	0.2262	0.4228	0.2856	0.3285	0.4695
Transformer-ISF	0.2266	0.4232	0.2795	0.3272	0.4624
Conv1D-clean	0.0590	0.3171	0.1793	0.2551	0.4392
Conv1D-noise	0.0516	0.3010	0.1707	0.2530	0.4375
Conv1D-param	0.0437	0.3209	0.1695	0.2452	0.4366
Conv1D-weighted	0.0417	0.3147	0.1670	0.2456	0.4361
Conv1D-adv	0.0455	0.3184	0.1745	0.2447	0.4361
Conv1D-PASA	0.0383	0.2903	0.1584	0.2440	0.4356
Conv1D-PhySA	0.0373	0.3050	0.1528	0.2396	0.4354
Conv1D-ISF	0.0371	0.3029	0.1531	0.2397	0.4356
DeepONet-param	0.3615	0.3625	0.3591	0.3532	0.3644
DeepONet-weighted	0.3441	0.3816	0.3607	0.3514	0.3618
DeepONet-adv	0.3491	0.3941	0.3734	0.3647	0.3914
DeepONet-PASA	0.3622	0.3836	0.3635	0.3499	0.3612
DeepONet-PhySA	0.3233	0.3773	0.3477	0.3262	0.3329
DeepONet-ISF	0.3285	0.3814	0.3468	0.3411	0.3562

Table 22: Burgers smooth initial-condition shift envelope. Columns give the target RMS perturbation amplitudes relative to each initial condition's standard deviation. The zero-amplitude column uses the unperturbed PDEBench-derived initial condition, and the remaining columns add smooth periodic low-frequency perturbations and re-solve Burgers at the clean viscosity. Every nonzero amplitude is held out from training.

Method	0	0.05	0.10	0.15	0.20
FNO-clean	0.0450	0.0309	0.0293	0.0277	0.0269
FNO-noise	0.0413	0.0299	0.0282	0.0269	0.0264
FNO-param	0.0106	0.0081	0.0078	0.0076	0.0076
FNO-weighted	0.0103	0.0079	0.0076	0.0075	0.0075
FNO-adv	0.0136	0.0104	0.0100	0.0098	0.0097
FNO-PASA	0.0086	0.0067	0.0065	0.0064	0.0064
FNO-PhySA	0.0241	0.0174	0.0168	0.0165	0.0166
FNO-ISF	0.0250	0.0179	0.0173	0.0169	0.0170
Transformer-clean	0.3693	0.2510	0.2367	0.2277	0.2256
Transformer-noise	0.2303	0.1594	0.1521	0.1478	0.1484
Transformer-param	0.0859	0.0645	0.0622	0.0609	0.0604
Transformer-weighted	0.0958	0.0719	0.0701	0.0696	0.0697
Transformer-adv	0.0949	0.0715	0.0693	0.0682	0.0680
Transformer-PASA	0.0929	0.0703	0.0686	0.0682	0.0685
Transformer-PhySA	0.2262	0.1862	0.1850	0.1848	0.1895
Transformer-ISF	0.2266	0.1852	0.1843	0.1838	0.1885
Conv1D-clean	0.0590	0.0431	0.0424	0.0423	0.0429
Conv1D-noise	0.0516	0.0382	0.0380	0.0386	0.0396
Conv1D-param	0.0437	0.0316	0.0315	0.0321	0.0332
Conv1D-weighted	0.0417	0.0308	0.0308	0.0314	0.0326
Conv1D-adv	0.0455	0.0341	0.0343	0.0352	0.0366
Conv1D-PASA	0.0383	0.0296	0.0302	0.0311	0.0324
Conv1D-PhySA	0.0373	0.0307	0.0323	0.0342	0.0362
Conv1D-ISF	0.0371	0.0306	0.0323	0.0345	0.0364
DeepONet-param	0.3615	0.3153	0.3178	0.3240	0.3341
DeepONet-weighted	0.3441	0.3148	0.3165	0.3247	0.3345
DeepONet-adv	0.3491	0.3183	0.3204	0.3321	0.3444
DeepONet-PASA	0.3622	0.3173	0.3197	0.3268	0.3353
DeepONet-PhySA	0.3233	0.3059	0.3122	0.3202	0.3306
DeepONet-ISF	0.3285	0.3046	0.3108	0.3182	0.3278

Table 23: Darcy forcing robustness envelope. The Scale row gives the perturbation scale of the smooth forcing field.

Method	C	S1	S2*	S3	S4*	S5	S6*
Scale	0	0.15	0.25	0.35	0.45	0.55	0.65
Conv2D-clean	0.0226	0.3139	0.4526	0.5518	0.6280	0.6909	0.7466
Conv2D-noise	0.0183	0.0694	0.1239	0.1851	0.2492	0.3146	0.3798
Conv2D-param	0.0196	0.0202	0.0212	0.0228	0.0252	0.0286	0.0335
Conv2D-weighted	0.0195	0.0201	0.0212	0.0228	0.0253	0.0288	0.0338
Conv2D-adv	0.0208	0.0212	0.0220	0.0232	0.0249	0.0274	0.0310
Conv2D-PASA	0.0195	0.0201	0.0212	0.0229	0.0253	0.0288	0.0337
Conv2D-PhySA	0.0182	0.0189	0.0201	0.0218	0.0243	0.0279	0.0328
FNO2D-clean	0.7406	0.7739	0.7968	0.8172	0.8361	0.8541	0.8722
FNO2D-noise	0.6930	0.6955	0.7003	0.7077	0.7184	0.7330	0.7518
FNO2D-param	0.6838	0.6853	0.6887	0.6943	0.7023	0.7129	0.7267
FNO2D-weighted	0.6941	0.6957	0.6990	0.7041	0.7114	0.7213	0.7342
FNO2D-adv	0.7012	0.7023	0.7063	0.7129	0.7229	0.7365	0.7541
FNO2D-PASA	0.6895	0.6912	0.6945	0.6998	0.7074	0.7176	0.7309
FNO2D-PhySA	0.6981	0.7005	0.7046	0.7108	0.7195	0.7308	0.7454

Table 24: Darcy boundary-condition robustness envelope. The Amplitude row gives the amplitude of the smooth low-frequency Dirichlet perturbation.

Method	C	S1	S2*	S3	S4*	S5	S6*
Amplitude	0	0.04	0.06	0.08	0.10	0.12	0.14
Conv2D-clean	0.0216	0.5970	0.7439	0.8281	0.8779	0.9089	0.9290
Conv2D-noise	0.0242	0.5910	0.7388	0.8250	0.8768	0.9094	0.9310
Conv2D-param	0.0432	0.0545	0.0575	0.0592	0.0613	0.0645	0.0714
Conv2D-weighted	0.0410	0.0544	0.0572	0.0588	0.0609	0.0640	0.0714
Conv2D-adv	0.0651	0.0668	0.0638	0.0602	0.0581	0.0583	0.0676
Conv2D-PASA	0.0416	0.0550	0.0578	0.0594	0.0615	0.0647	0.0718
Conv2D-PhySA	0.0328	0.0491	0.0526	0.0548	0.0572	0.0607	0.0681
FNO2D-clean	0.8131	0.9629	1.0089	1.0274	1.0328	1.0328	1.0310
FNO2D-noise	0.9667	0.9833	0.9907	0.9948	0.9970	0.9983	0.9990
FNO2D-param	1.5950	1.3534	1.2490	1.1770	1.1266	1.0892	1.0608
FNO2D-weighted	1.1869	1.1959	1.1537	1.1169	1.0864	1.0630	1.0448
FNO2D-adv	1.4342	1.3296	1.2626	1.1925	1.1308	1.0823	1.0468
FNO2D-PASA	1.3169	1.2499	1.2036	1.1583	1.1199	1.0898	1.0659
FNO2D-PhySA	1.2987	1.2234	1.1785	1.1423	1.1136	1.0911	1.0733

Table 25: Descriptive field-error and diagnostic-error trade-off audit for the 15-comparison worst-severity and tail subset of the primary strong-baseline family. Each italic row also names the response diagnostic entering the Diagnostic column. The final column is a thresholded reading that treats gains within ± 1.0 percentage point as near-neutral, so a comparison is read as a field-only gain when the response diagnostic is near-neutral and as a diagnostic regression when the field gain is near-neutral.

	Gain (%)		
Comparison	Field	Diagnostic	Interpretation
Burgers viscosity; diagnostic: SAE			
FNO-param→FNO-PASA	2.2	5.4	field and response improve
FNO-param→FNO-adv	−1.5	−7.9	both worsen
Conv1D-param→Conv1D-PhySA	11.9	75.0	field and response improve
Public Burgers viscosity; diagnostic: SAE			
FNO-param→FNO-PhySA	31.8	83.1	field and response improve
FNO-param→FNO-ISF	32.5	82.9	field and response improve
FNO-param→FNO-adv	5.4	12.3	field and response improve
Burgers evaluation mesh; diagnostic: grid SAE			
FNO-param→FNO-PASA	21.5	36.2	field and response improve
FNO-param→FNO-adv	−27.8	−25.4	both worsen
Burgers initial condition; diagnostic: SAE			
FNO-param→FNO-PASA	16.2	7.8	field and response improve
FNO-param→FNO-adv	−27.9	−26.0	both worsen
Conv1D-param→Conv1D-PhySA	−4.0	−17.4	both worsen
Darcy forcing; diagnostic: SAE			
Conv2D-param→Conv2D-PhySA	3.4	0.9	field-only gain
Conv2D-param→Conv2D-adv	2.4	10.1	field and response improve
Darcy boundary; diagnostic: SAE			
Conv2D-param→Conv2D-PhySA	6.4	6.4	field and response improve
Conv2D-param→Conv2D-adv	0.3	−4.0	diagnostic worsens

J Primary Field-Error and Diagnostic-Error Trade-Off Audit

Table 25 reports the case-by-case trade-off between field error and diagnostic error for the 15 comparisons carried through the worst-severity and tail analyses of Table 17. All but one of these 15 comparisons are also rows of the 36-comparison primary inferential family of Table 8; the mesh adversarial contrast is reported as a descriptive ratio only. The inferential claims in the main text rest on the paired effects in that table. Each gain in this appendix is a descriptive ratio rather than the paired seedwise gain used for inference (App. H).

K Secondary-Tier Backbone Replications

This appendix reports the secondary-tier replications referred to in the appendix guide. The backbones used in these runs are much less accurate at nominal settings than the primary-tier FNO and ConvResNet baselines, so each run replicates a contrast within one architecture rather than supplying evidence about an operator family. The DeepONet-style runs in particular use a plain branch and trunk configuration without the practical extensions and architecture-specific tuning that Lu et al. (2022) show are needed before two operator families can be compared fairly, so these runs are not evidence about DeepONet as a family. Each percentage below is a descriptive ratio in the sense of App. H.

Burgers viscosity shifts. The DeepONet Burgers parametric baseline has held-out error 0.4085 in Table 11, far above the held-out errors of the FNO and ConvResNet baselines under the same data split. Within the weaker DeepONet-style architecture, however, PhySA lowered held-out error to

0.3728 and ISF lowered the same error to 0.3765; the descriptive ratios of the underlying unrounded means are 8.7% and 7.8%, respectively, while the paired tests give $p = 0.0582$ and $p = 0.1012$.

Evaluation-mesh shifts. The same contrast appears at weaker accuracy in the DeepONet-style mesh result. The descriptive ratio of the underlying unrounded held-out means for DeepONet-PhySA relative to DeepONet-parametric is 3.8%, with paired $p = 0.1655$, and the DeepONet-PhySA held-out error remains 0.3460 in Table 13, well above the FNO mesh errors.

L Scope of the Experiments

The experiments cover six specification-shift tasks: response-changing Burgers viscosity and smooth initial-condition perturbations, raw public PDEBench multi-viscosity Burgers trajectories, Darcy forcing and boundary perturbations, and representation-changing evaluation-mesh shifts, each with five seeds, held-out severities, and strong changed-label parametric, weighted-parametric, and finite-grid adversarial baselines. The backbones are FNO, ConvResNet1D, transformer-style, and DeepONet-style Burgers models together with ConvResNet2D and FNO2D Darcy models. That coverage is a slice rather than a survey of the PDE families, datasets, and operator architectures now in use, which is why the comparisons above are reported as perturbation-, PDE-, and backbone-dependent outcomes on these tasks rather than as a ranking that transfers to every specification shift or every surrogate. The regularized-inverse extension of App. D is theoretical: the experiments do not evaluate inverse reconstructions or data-dependent regularization choices.