RESPONSE GEOMETRY FOR DECISIONS WITH NEURAL PDE SURROGATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural PDE surrogates can fit familiar solution fields while responding incorrectly to changes in physical parameters. We study whether this missing response information improves three decisions: choosing a forward predictor, choosing a surrogate for parameter recovery, and repairing an existing predictor. Response geometry compares parameter derivatives of model and reference fields. An exact pathwise error identity connects these derivatives to extrapolation; sensor projection explains local inverse sensitivity, and the rollout chain rule explains why one-step derivative fit can miss accumulated error. We evaluate these connected uses on the one-dimensional viscous Burgers equation, the two-dimensional steady Darcy flow equation with heterogeneous permeability, and the two-dimensional forced incompressible Navier–Stokes equations in vorticity form, separating ID-only probes, shift-calibrated scores, and target-assisted comparisons. A prespecified response selector reduces early-time Navier–Stokes OOD relative field error from 0.9934 to 0.3377, while its Burgers counterpart ranks population risk better but selects worse. Completed repair tests reduce Navier–Stokes field error from 0.6284 to 0.4899 using rollout-plus-response supervision, with a corrected paired-test gain and additional compute disclosed. The same repair worsens inverse parameter recovery. Burgers shows smaller path-repair gains; Darcy's reliable small gain is a direct-field control. These findings make response geometry useful as a task-specific decision variable: its information can improve selection and prediction, but field accuracy, inverse recovery, and long-time physical fidelity require separate tests.

1 INTRODUCTION

Imagine two fast fluid simulators that reproduce a known flow equally well. If viscosity is reduced slightly, one predicts a plausible increase in small-scale motion and the other barely changes. Their present-day field errors can look similar while their behavior under a physical intervention is different. Neural operators such as FNO and DeepONet learn maps from PDE inputs to solution fields (Li et al., 2021a; Lu et al., 2021). Their deployment decisions, however, are commonly made using field error measured inside the training parameter range. The central question of this paper is whether a model's *response* to a controlled parameter change provides information that field error misses.

We treat the true numerical solution operator as $\mathcal{S}(z, a)$ and a trained surrogate as $\mathcal{G}_\theta(z, a)$. Here z is a short vector of log-scaled physical parameters, and a contains the initial condition, coefficient field, or forcing that defines one parent case. "Response geometry" means how the entire solution field changes as z changes. It is local, directional, and observable through finite differences or solver tangents. This is not a new architecture or a claim that derivatives alone certify generalization. Its purpose is to organize three decisions around one measurable object: *which model to deploy forward, which model to invert through sensors, and which response errors to repair in a trained predictor*.

Derivative-informed operators already use Jacobian supervision for prediction and inversion (O'Leary-Roseberry et al., 2022; Behroozi et al., 2025; Yao et al., 2025). Our question concerns decisions: given frozen candidates or a fixed starting checkpoint, when does response information improve deployment under a protected physical shift? We test that question against field-only and physics-based controls.

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.

Common measurement: compare $\mathcal{G}_\theta(z, a)$ with $\mathcal{S}(z, a)$ and their directional responses $D_z\mathcal{G}_\theta v$ and $D_z\mathcal{S}v$ on authorized parent cases.

Forward selection	**Inverse selection**	**Predictor repair**
Accumulate response error along a parameter path; choose a frozen model.	Project response through sensors; avoid weakly identifiable directions.	Train against the faulty response while retaining familiar-case fit.

Figure 1: One physical response comparison feeds three distinct decisions. The comparison is only meaningful after its reference-solver and label budget are stated; the endpoint test cases remain protected.

Our contributions are (i) a pathwise error, sensor-conditioning, and rollout account that connects three uses of the same physical response; (ii) one controlled scaffold for forward selection (Pathway A1), inverse selection (Pathway A2), and predictor repair (Pathway B); and (iii) completed, protected comparisons on three PDEs. The results expose three distinctions: better population ranking need not select the best model, better forward repair need not improve inversion, and local response supervision need not control accumulated dynamics. These distinctions organize the framework and its limits.

The point of the framework is a decision that a practitioner can actually make. Suppose a scientist has several already trained operators and knows the physical range in which the next experiment will run, but does not know its solution field. She can spend a small reference-solver budget at familiar parameter settings to check both the value of each candidate and the direction in which its prediction moves. If a few labeled historical shifts are available, she can calibrate a response score on those shifts and then choose a frozen model for new protected cases in that shifted regime. If her task is parameter inference from sensors, she should choose again using the response *visible to those sensors*. If every candidate has a damaging response defect, she can spend additional training budget to correct a pre-trained model, provided familiar-case accuracy is preserved. These are three uses of the same measurement, with different target losses and information costs.

There is a genuine tension. Equation (3) makes response mismatch relevant to shifted error, yet ranking models by a few scalar response diagnostics can fail because the path is only sparsely sampled, errors can cancel, and the best model is an extreme order statistic. Our paper tests that tension explicitly. We retain the original source labels for auditability in Appendix B, but the main text uses descriptive method names: *path-response*, *spectral-response*, *joint-response*, and *sensor-geometry* scores. Their names describe the information each score measures rather than an unexplained version history.

2 Problem and the common response object

For a fixed parent a, let $u = \mathcal{S}(z, a)$ solve a parameterized PDE and $\widehat{u} = \mathcal{G}_\theta(z, a)$ be a neural surrogate. We use $z_j = \log(p_j/p_{j,\mathrm{ref}})$ for positive physical parameters p_j; one unit of z_j is a multiplicative, not additive, change in p_j. The parameter-response Jacobian is

$$\mathcal{J}_\star(z, a) = D_z\mathcal{S}(z, a), \qquad \mathcal{J}_\theta(z, a) = D_z\mathcal{G}_\theta(z, a), \qquad \Delta\mathcal{J}_\theta = \mathcal{J}_\theta - \mathcal{J}_\star. \tag{1}$$

Each column of $\mathcal{J}$ is a *tangent field*: the first-order change in every grid value when one physical knob moves. A directional probe v compares $\mathcal{J}_\theta v$ with $\mathcal{J}_\star v$ without constructing every column. The relative, weighted response-geometry error (RGE) is

$$\mathrm{RGE}_v(\theta; z, a) = \frac{\|W(\mathcal{J}_\theta - \mathcal{J}_\star)v\|_2}{\|W\mathcal{J}_\star v\|_2 + \epsilon}, \tag{2}$$

where W weights space, time, or a physical diagnostic and ϵ stabilizes near-zero reference responses. Absolute errors are also reported when a tiny denominator would make a relative score misleading. T-RGE denotes the time-resolved version; HP-RGE weights high spatial frequencies. These are distinct measurements of the same response discrepancy, not independent theories. Magnitude and direction both matter: for response vectors $q_\theta, q_\star$, $\|q_\theta - q_\star\|^2 = (\|q_\theta\| - \|q_\star\|)^2 + 2\,\|q_\theta\|\,\|q_\star\|\,(1 - \cos\phi)$.

2.1 FROM LOCAL RESPONSE TO SHIFTED-FIELD ERROR

Let $z(t)$, $0 \le t \le 1$, connect an in-distribution anchor z_0 to a target z_1, and set $e_\theta(z) = \mathcal{G}_\theta(z, a) - \mathcal{S}(z, a)$. If both solution maps are differentiable along the same path, the fundamental theorem of calculus gives the exact identity

$$e_\theta(z_1) = e_\theta(z_0) + \int_0^1 \Delta\mathcal{J}_\theta(z(t), a)\, \dot{z}(t)\, dt, \quad \|e_\theta(z_1)\| \le \|e_\theta(z_0)\| + \int_0^1 \|\Delta\mathcal{J}_\theta(z(t), a)\dot{z}(t)\|\, dt. \tag{3}$$

The first term is the starting field error. The integral is accumulated response error: if the surrogate tilts differently from the true solution as the physical knobs move, its endpoint can drift. This bound does *not* imply that any finite-probe score is a certified OOD-error bound. With a straight step hv and a bound M on the second directional derivative of e_θ along that segment, Taylor's theorem yields

$$\|e_\theta(z_0 + hv)\| \le \|e_\theta(z_0)\| + |h|\, \|\Delta\mathcal{J}_\theta(z_0)v\| + \tfrac{1}{2}Mh^2. \tag{4}$$

The $Mh^2/2$ remainder is the area under a linearly growing slope uncertainty. Small anchor error and small local response error matter only over a distance for which curvature is controlled. Strong shocks, parameter interactions, and long chaotic horizons can violate a useful small-M approximation.

With two knobs, a mixed curvature term $h_1 h_2 \partial_{z_1 z_2} e_\theta$ captures interactions that axis-only tests miss. Our 3×3 bank measures mixed second differences separately from joint first derivatives, whose linearity adds no curvature information (Appendix A.1).

2.2 THE SAME JACOBIAN THROUGH SENSORS

Inverse use starts with observations $y = O\mathcal{S}(z_\star, a) + \eta$, where O samples a sparse sensor set and η has positive-definite covariance Σ. For a frozen surrogate, we estimate z by minimizing $\|O\mathcal{G}_\theta(z, a) - y\|^2_{\Sigma^{-1}}$ within physical bounds. Let $J = O\mathcal{J}_\theta(z_\star, a)$ and $b = O[\mathcal{G}_\theta(z_\star, a) - \mathcal{S}(z_\star, a)]$. The unconstrained local least-squares approximation gives, when $J^\top\Sigma^{-1}J$ is invertible,

$$\widehat{z} - z_\star \;\approx\; (J^\top\Sigma^{-1}J)^{-1}J^\top\Sigma^{-1}(\eta - b). \tag{5}$$

Thus a small whole-field L2 does not guarantee a good inverse result: the projected bias $J^\top\Sigma^{-1}b$ can align with a weak parameter direction. For Gaussian noise, $F = J^\top\Sigma^{-1}J$ is local Fisher information (how strongly the sensors distinguish nearby parameters). A small eigenvalue means two parameter changes can produce nearly indistinguishable observations. Active bounds or a different optimizer basin can invalidate the local approximation. Equation (5) motivates sensor-aware RGE and identifiability features; it does not guarantee global optimization or remove non-identifiability. The estimator is numerical optimization through the selected *forward* model, not a separately trained inverse network.

3 FROM RESPONSE MEASUREMENT TO THREE DECISIONS

The mathematical object in Equation (1) is shared, but a useful operational score must match the task. Forward deployment cares about a future field, inverse deployment about a parameter inferred through sensors, and repair about changing the model without sacrificing known behavior. This distinction prevents a large feature inventory from becoming an unexplained collection of methods.

3.1 PATHWAY A1: CHOOSING A FORWARD SURROGATE

For each frozen candidate, we first measure relative field error on held-out ID parents. Response diagnostics use an authorized ID reference bank: axis and joint finite secants, mixed second differences, *path-response error* along a predeclared path inside the ID rectangle, and *spectral-response error* with boundary-compatible frequency filters. Time-resolved measurements distinguish stages of a dynamical trajectory. The evaluator computes an ID-to-OOD path separately; its target-adjacent labels never enter the strict-ID feature bank. Every node consumes reference solves, which a fair cost comparison must count against direct field measurements. Labeled near-target nodes belong to the target-assisted regime.

In the calibrated regime, each candidate m has ID response features f_m, anchor field error, and a capacity measure. The fitted score estimates log shifted-field risk:

$$s_m = \beta_0 + \beta_{\mathrm{ID}}\widetilde{L}_{\mathrm{ID}}(m) + \beta_C\widetilde{C}_m + \sum_k \beta_k \widetilde{f}_{mk}, \quad \beta_{\mathrm{ID}}, \beta_C, \beta_k \geq 0. \tag{6}$$

Tildes denote the frozen transforms fitted on development data: log scaling, residualization of response features against field error and capacity, standardization, and sign alignment. A nonnegative coefficient therefore acts on an *aligned transformed* feature; it does not make the raw physical response error monotone by theorem. Ridge-stabilized nonnegative least squares (NNLS) fits the coefficients against labeled shifted *development* outcomes, with its penalty selected by leave-one-family-out development regret. The archived architecture groups distinguish FNO and SC-FNO training recipes even though they share network structure. The same 36 baseline candidates per PDE appear at protected evaluation, but new independent parent cases do. This tests transfer across cases within a fixed zoo, not cold-start transfer to unseen architectures. Raw ID response rules are reported separately from this shift-calibrated score; sparse positive Lasso and robust/rank fits are comparisons, not new physical principles.

The rule is frozen before protected outcomes are opened. Its practical output is simply one model identity and a predicted risk score. Only afterward do we compute the selected model's OOD field L2. Direct OOD L2 over all candidates gives a hindsight oracle ceiling. It cannot be used by a practitioner without paying for the very target labels that selection is meant to avoid.

3.2 Pathway A2: choosing a surrogate for inverse recovery

Here the output is again a *model identity*, but the downstream task differs. We do not train an inverse network. For every selected forward model, the same bounded multi-start optimizer solves

$$\widehat{z}_\theta(y) = \arg\min_{z\in\mathcal{Z}} \left\| \Sigma^{-1/2}[O\mathcal{G}_\theta(z,a) - y] \right\|_2^2. \tag{7}$$

To anticipate failure using ID cases, the *sensor-geometry score* measures three concepts implied by Equation (5): observation-space field bias $\Sigma^{-1/2}b$, observation-Jacobian mismatch $\Sigma^{-1/2}O\Delta\mathcal{J}_\theta$, and weak directions of F. The compact score tests whether the model and reference agree on the most observable parameter direction; the augmented score adds a penalty for near-collapse of identifiability. These are empirical features fitted to inverse outcomes from development cases, not a valid confidence interval for a single recovered parameter. Where a Navier–Stokes-only enstrophy/palinstrophy channel is defined, we keep it NS-specific instead of inventing a same-named Darcy or Burgers measurement.

We report the mean of per-parent log-parameter RMSEs, so proportional errors have comparable meaning across positive physical parameters. Each parent contributes $[\frac{1}{2}\sum_{j=1}^{2}(\widehat{z}_j - z_{\star,j})^2]^{1/2}$. Failure means this error exceeds the frozen 0.5 threshold or every restart fails numerically; it is not just an optimizer-crash rate. We also test the reconstructed field $\mathcal{G}_\theta(\widehat{z}, a)$ against the reference field. A model can match sensor data with a wrong z when the inverse is poorly identified; reconstructed-field and parameter errors must therefore remain separate. The direct OOD inverse-RMSE selector is an oracle comparator, not a deployed choice. Inverse score weights use labeled shifted *development* inverse cases; protected inverse outcomes are opened only after freezing.

3.3 Pathway B: improving an existing predictor

Selection is limited by the available candidates. Repair starts every competing intervention from the *same* pre-trained FNO-FiLM checkpoint and changes its weights. FiLM conditions the network using learned feature scales and offsets driven by the physical parameters. The central comparison is the familiar field objective augmented by physics and response supervision:

$$L_{\mathrm{core}}(\theta) = L_{\mathrm{field}}^{\mathrm{ID}} + \lambda_r L_{\mathrm{physics}} + \lambda_s L_{\mathrm{tangent}}, \tag{8}$$

$$L_{\mathrm{response}} = \mathbb{E}_{a,z,v} \left\| W[\mathcal{J}_\theta^{\mathrm{step}} - \mathcal{J}_\star^{\mathrm{step}}]v \right\|_2^2. \tag{9}$$

L_{physics} compares model and reference residuals of the implemented discrete equation. L_{tangent} compares the log-parameter derivatives of those residuals, using the reference tangent to differentiate

its local field. These are the code's `physics` and `tangent` terms. They differ from L_{response}, which directly matches field derivatives. Local dynamical losses use one-step Jacobians $\mathcal{J}^{\text{step}}$ at a fixed input state. Initial-state geometry and rollout losses additionally differentiate the parameter-dependent input and accumulated trajectory. Burgers amplitude changes the initial state, so its fixed-state amplitude tangent is zero even when its total response is nonzero. Physics-plus-tangent, direct full-response, and combined controls therefore test distinct objectives. Darcy flux-response loss includes conductivity and pressure derivatives. NS losses match nonzero reference defects rather than impose a false zero-residual target. Budgets, rollout, and spatial-scale restrictions are separate ablations.

Each of the 231 new arms has an *active ID-retention mechanism*: a dual penalty on replayed ID field loss is adjusted during training, and a checkpoint is accepted only if rollout error on eight fixed tune parents stays within 1.05 times its starting value. If none qualifies, deployment falls back to the starting checkpoint. Fresh ID cases then test retention; the tune gate does not guarantee it. Repair variants ask whether joint directions, path information, high-frequency/fractional weights, or targeted acquisition improve over matched controls. Direct-state comparisons match reference-node counts, not automatically runtime. Field-only, physics-plus-tangent, and direct full-response controls share starting checkpoints and update counts; GPU time is audited separately. Historical B6 controls do not have the new active-retention mechanism. The completed campaign evaluates all 231 new arms, 135 historical controls, and 45 measured-compute checkpoints on the same new parent cohort.

The link between local repair and deployment is the rollout chain rule. Write g_θ for one surrogate step, $\widehat{u}_{n+1} = g_\theta(\widehat{u}_n, z)$, and $Q_n = D_z\widehat{u}_n$. Then

$$Q_{n+1} = D_u g_\theta(\widehat{u}_n, z)Q_n + D_z g_\theta(\widehat{u}_n, z). \tag{10}$$

The first term propagates the response already carried by the state; the second adds the direct effect of moving a physical knob. Matching only the second term at familiar states cannot guarantee the total trajectory response. The NS rollout intervention therefore matches both fields and total derivatives at saved steps $2, 4, 8$ ($t = 0.2, 0.4, 0.8$). With U and Q denoting these stacked fields and responses, its extra loss is

$$L_{\text{roll+resp}} = \tfrac{1}{2}\rho(U_\theta - U_\star, U_\star) + \tfrac{1}{2}\rho(Q_\theta - Q_\star, Q_\star), \qquad \rho(E, T) = \frac{\langle |E|^2 \rangle}{\max(\langle |T|^2 \rangle, \epsilon)}. \tag{11}$$

Brackets average the represented entries; the denominator compares errors relative to reference size. This term augments ID field training and replay retention. Its field-only rollout control removes the second term. Spatial-scale, path/mixed-curvature, budget, sensor-bias, and acquisition variants remain component tests, not separate claims of a universally optimal repair recipe.

4 Experimental design and reliability

4.1 Three PDEs and independent parent cases

Our testbed comprises the one-dimensional viscous Burgers equation (advection and diffusion), the two-dimensional steady Darcy flow equation with heterogeneous permeability (elliptic pressure), and the two-dimensional forced incompressible Navier–Stokes equations in vorticity form (nonlinear vortex evolution). We use Burgers, Darcy, and NS as shorthand for these three PDEs. The two varied physical coordinates per PDE create axis, same-sign, opposite-sign, mild, and severe joint shifts; parent initial conditions/coefficient fields are independently seeded. The reference generator produced 5,120 cases, including 512 train, 64 tune, and 64 audit cases per PDE, two protected 64-parent cohorts paired across seven shift strata, and 512 extra fixed-viscosity NS cases. Spatial grid cells are never treated as independent statistical replicates; a parent case is the unit. Each case carries field values, solver tangents, frozen sensors, noise draws, and integrity hashes.

A *parent* is one latent initial condition or coefficient realization held fixed while the two physical knobs are moved. Pairing the resulting ID and shifted cases answers a controlled question: what did the parameter shift do to this same physical specimen? It avoids confusing a difficult random initial condition with a difficult shift. Parents, seeds, and reference roles are disjoint across training, tuning, selector calibration, and protected confirmation. All normalizations and checkpoint choices

Table 1: Physical shifts in the frozen reference generator. A is initial-amplitude for Burgers and source/forcing amplitude otherwise. "Severe" shows the joint endpoint; axis and milder shifts are also tested.

PDE: varied knobs	Training range	Severe joint endpoint
Burgers: ν, initial A	$\nu \in [0.05, 0.10]$, $A \in [0.5, 1.5]$	$(\nu, A) = (0.01, 2.0)$
Darcy: σ, source A	$\sigma \in [0.5, 1.2]$, $A \in [0.5, 1.5]$	$(\sigma, A) = (1.6, 2.0)$
Navier–Stokes: ν, forcing A	$\nu \in [0.0008, 0.0012]$, $A \in [0.8, 1.2]$	$(\nu, A) = (0.0002, 2.0)$

are fitted from authorized roles only. In the source-restoration evaluation, 12 fresh development and 12 fresh confirmation parents per PDE are paired across predeclared path and severe-shift nodes; the latter are generated only after the selection protocol is frozen.

Burgers' severe-shift reference was refined to a 2048-point ETDRK4 solution after its source 128-point grid failed a shock-resolution gate; Darcy's 64^2 reference passed a 128^2 comparison after 32^2 response error was too large. The original source grids remain available for replication. Navier–Stokes retains the planned 64^2 operational source map, but its $T = 10$ trajectory failed pointwise grid convergence. Chaotic decorrelation complicates that comparison but does not by itself explain the failure or establish statistical convergence. We therefore separate early-time field comparisons from long-time statistical, spectral, and budget diagnostics and carry this numerical uncertainty into interpretation.

Both fields and solver tangents were checked against mesh or finite-difference ladders; Appendix A records the numerical qualifications.

4.2 Models, baselines, and protected protocol

The source zoo contains 156 checkpoints: 108 core combinations of three PDEs, four families (FNO, FNO-FiLM, DeepONet, and a physical-parameter-response-supervised SC-FNO adaptation), three capacities, and three seeds; 27 negative controls; nine Burgers spectral-resolution controls; and twelve fixed-viscosity NS controls. The headline selectors choose among the *36 baseline candidates per PDE*; controls probe whether the measurements detect deliberate failures. SC-FNO adapts a prior derivative-supervised training approach to our FNO implementation; it is not an additional distinct network architecture or our original invention. Training and checkpoint selection read only train/tune data. Candidate response features use authorized ID reference banks; shifted development outcomes fit the source-composite weights. Forward rules are frozen before confirmation references are generated. Inverse rules are frozen before inverse recovery outcomes are computed on the same B8 confirmation parents; those reference fields were already used in forward testing. Repair starts from nine matched FNO-FiLM checkpoints and uses independent repair confirmation parents. The older B3 population and later B8 source-restoration population differ, so their top-one values are not pooled.

We report four information regimes: *strict ID* (only ID reference labels and responses), *shift-calibrated* (labeled shifted development parents, protected target parents untouched), *target-assisted* (a limited number of labeled near-target fields), and *oracle* (all target labels needed for hindsight selection). B8 source NNLS and B9 inverse composites are shift-calibrated because their weights see shifted development outcomes, even though their candidate features are measured on ID cases. This distinction is essential: OOD at evaluation means a protected parameter regime absent from model training, even when separate development labels were used to fit a selector. A target-assisted method must be compared with direct field L2 on the *same number* of new solver calls; otherwise response probes get an unfair information advantage.

Our primary forward endpoint is mean relative OOD field L2 of the *one selected model*; top-one regret is its gap from the direct-OOD oracle. Spearman correlation over the 36 candidates is secondary: it measures overall ranking association but can improve even when the minimum-score model is poor. For inverse, primary error is mean per-parent log-parameter RMSE, supplemented by threshold/numerical failure rate and reconstructed-field L2. Repair compares severe-shift field L2 against an equal-update active-ID field-only arm across three matched training seeds and 12 fresh parents; its inverse subset has eight parents. Hierarchical bootstrap intervals resample seeds and

Table 2: Protected forward selection on 12 fresh confirmation parents per PDE. Lower OOD relative field L2 is better; NS uses $t = 1, 2$ rather than the long-horizon $T = 10$ trajectory. ρ is candidate-level Spearman correlation with OOD error. The response column uses the prespecified shift-calibrated rule, not the best rule chosen retrospectively. Oracle uses protected labels and is unavailable in deployment.

PDE	ID L2	Response score	Oracle	ρ_{ID}	ρ_{response}
Burgers	0.2148	0.2993	0.1982	0.5676	0.7315
Darcy	0.5272	0.5272	0.4334	0.4314	0.4178
Navier–Stokes	0.9934	0.3377	0.3288	-0.3176	0.8996

parents. For the sign-flip test, seed replicates are averaged within each parent before enumerating signs; this assesses a paired symmetric-null comparison, not pixel-level replication. Holm correction covers all new repair arms within each PDE. Fresh-ID error and measured GPU seconds qualify the gains. The small number of parents limits precision. Secondary probe budgets, finite-difference steps, sensor/noise choices, and grid resolution test sensitivity without being recast as new primaries after inspection.

5 Results: response information depends on the decision

Forward decision and the ranking paradox. Table 2 shows why a global metric correlation is not enough. On Burgers, the prespecified path-response score correlates *more* strongly with OOD error than ID L2 over 36 candidates, yet selects a model with worse field error: 0.2993 versus 0.2148. Its top-ranked tail was wrong. Darcy ties the ID-L2 model despite oracle headroom of 0.0938. By contrast, the NS ID-L2 rule selects a DeepONet with 0.9934 error and negative candidate-level rank correlation; the prespecified *joint-response score* selects an SC-FNO with 0.3377 error. The paired parent-level sign-flip test survives the conservative within-PDE Holm adjustment ($p \simeq 0.0146$). This is a substantial, specific NS gain, not evidence that the same score works universally. The selected architecture changes from DeepONet to FNO. Family-held-out development tuning does not isolate a within-family benefit or prove which model component caused the gain. The exact source mappings of the three prespecified scores appear in Appendix B.

Several raw ID response probes selected better Burgers or Darcy models in exploratory inspection. They are useful hypotheses for a future frozen test, not replacements for the prespecified rule in this table. Target-assisted direct near-anchor L2 was often strong, which reinforces the need to count target reference solves fairly. The displayed-master reconstruction of *OS-RGE v4* and the earlier operationalized proxy are separately audited: original executable source was unavailable, and the displayed equations do not provide a valid certified OOD-error bound. Neither is relabeled as our strict-ID method.

Inverse decision. Table 3 tests whether changing the *selection task* changes the selected model's inverse behavior. Under the prespecified NS sensor-geometry score, inverse log-parameter RMSE is 0.6911 versus 0.8607 for the ID-L2-chosen model, and failures fall from 0.667 to 0.500. Its paired inverse improvement does not survive multiplicity correction. Burgers is worse and Darcy ties under their own frozen primary rules. The exploratory compact sensor-geometry score reaches 0.0624 and zero failures on NS confirmation parents, with Holm-adjusted $p \simeq 0.0622$; this is a candidate for a *new independent* confirmation cohort, not a confirmatory headline. It selects a different FNO than the forward response rule, illustrating the sensor task's distinct geometry. The completed secondary tests retain all 16 frozen shortlisted models under two noise levels, half the sensors, and twenty prior restarts. NS's primary improves point estimates in all four conditions, but all paired intervals span zero. The compact rule stays at 0.061–0.064 with no threshold/numerical failures; these are robustness views of the same 12 parents, not independent confirmation.

Reconstructed-field error and inverse parameter error are reported separately because a visually plausible field can arise from a wrong parameter. The NS primary selected model's reconstructed-field L2 is 0.5899, compared with 0.9934 for the ID-L2-selected model. Even this field improvement does not establish that the parameter improvement is statistically secure. Burgers' high inverse errors, in-

Table 3: Protected inverse selection. The score column is each PDE's prespecified rule, fitted on development inverse outcomes. Entries are mean per-parent log-parameter RMSE (lower is better); the final column shows NS threshold/numerical failure fractions only.

PDE	ID-L2 model	Prespecified score	Oracle	Failure: ID → rule
Burgers	1.1560	1.2069	1.0368	—
Darcy	0.1151	0.1151	0.1031	—
Navier–Stokes	0.8607	0.6911	0.0624	$0.667 \rightarrow 0.500$

Table 4: Completed predictor repair on three seeds and 12 new field parents; inverse errors use the fixed eight-parent subset. "Field control" means active-ID field-only training, not the selector in Table 2. Lower errors are better. p_H is the within-PDE Holm-adjusted field test. Main-text names map to exact source labels in Appendix C.

PDE	Intervention	Field control	Repaired L2	Inverse RMSE	p_H
Burgers	Three-node path	0.4180	0.4074	1.1742	0.0117
Darcy	Endpoint fields	0.3007	0.2960	0.1104	0.0488
NS	Full response	0.6284	0.5726	0.5289	0.0137
NS	Short rollout	0.6284	0.5263	1.0941	0.0859
NS	Rollout + response	0.6284	0.4899	1.0560	0.0137

cluding the oracle's 1.0368, also warn that the sensor/parameter configuration may be intrinsically difficult; a better forward selector cannot create information absent from the observations.

Repair decision: fields improve, inverse recovery can worsen. The protected repair campaign and final provenance audit are complete. Table 4 reports deployed, ID-gated checkpoints; all methods share the same starting checkpoints within a PDE/seed. Burgers' three-node path and mixed-Taylor variants improve over active-ID field training by about 2.5%, with corrected $p = 0.0117$, but remain worse than the historical field-only control (0.3989). Its lowest observed new-arm error, physics-plus-tangent at 0.3612, is not reliable after correction ($p = 0.3988$). Darcy's small corrected gain comes from direct endpoint-field supervision; the path arm's lower point estimate (0.2937) does not survive correction ($p = 0.4940$). These controls prevent us from attributing every gain to response geometry.

NS provides the clearest repair result: rollout-plus-response reduces early-time OOD field L2 by 22.0% (0.6284 to 0.4899; paired difference -0.1385, hierarchical 95% interval $[-0.1969, -0.0721]$, corrected $p = 0.0137$). Fresh ID field L2 also falls from 0.2498 to 0.1567; each repaired seed is within the fresh-ID retention threshold. Local full-response training gives a smaller 8.9% OOD gain. Signed-joint, fixed-band, fractional-response, and random-reference arms also pass the corrected field test; the best observed arm was identified after evaluation, so its status as a winner remains descriptive. Short rollout alone has a favorable interval but corrected $p = 0.0859$. The separate rollout-versus-rollout-plus-response contrast was not multiplicity-adjusted; it cannot alone establish a uniquely causal contribution of the derivative term.

The same NS rollout-plus-response intervention worsens inverse RMSE from the active-ID field control's 0.3937 to 1.0560 and increases failure fraction from 0.3333 to 0.8750. The signed-joint arm reaches 0.2550 descriptively, near the unchanged checkpoint's 0.2591. Thus forward improvement cannot be treated as inverse repair. Equation (5) supplies a plausible local explanation, but we have not identified a unique causal mechanism. Late-time spectral fidelity is also poor; early-time success does not establish a faithful $T = 10$ trajectory or spectrum.

Gate effects matter. Several NS budget, susceptibility, and sensor-bias variants reject all trained seeds and return the starting checkpoint; their common 0.6143 field error is a fallback, not successful repair. Direct path-state controls are worse than their response counterparts on NS but also worse than field-only training. Wrong-direction and shuffled-allocation controls nearly match the Burgers response variants. Targeted acquisition and discrepancy correction do not establish consistent gains. Appendix C reports every registered arm, including these failures.

The NS winner costs additional training. Its nearest available B6 field-only checkpoints consume only 11.3–13.9% of its training GPU time; the finite compute bank saturates below the repair budget. We therefore establish a gain over the fixed-update field recipe, not superiority at equal compute. Alternate-grid diagnostics remain secondary: NS full-response gives 0.6171 versus 0.6693 for field-only, with an interval spanning zero; Burgers patch-tail gives 0.4190 versus full-response 0.4270 and field-only 0.4200. They do not certify continuum accuracy.

6 Relation to earlier operator learning

FNO and DeepONet established powerful ways to approximate maps between functions (Li et al., 2021a; Lu et al., 2021). Physics-informed neural operators add PDE residuals to field training (Li et al., 2021b). Derivative-informed operators go further: DINO uses compressed derivative information for high-dimensional parameter maps, and SC-FNO and DIFNO supervise response derivatives to improve prediction and inverse tasks (O'Leary-Roseberry et al., 2022; Behroozi et al., 2025; Yao et al., 2025). Recent SC-NO work also studies sampled solver Jacobians across multiple systems (Behroozi et al., 2026). We do not claim that derivative supervision itself is new. Our focus is the *choice among already trained models*, the distinction between whole-field and sensor-conditioned choice, and a controlled repair intervention under explicit label access. SC-FNO is consequently a strong candidate/baseline in our zoo, not a method invented here. The elementary path identity supplies a common explanation for the three decisions; it is not a new approximation theorem.

7 Discussion, limitations, and conclusion

The framework reduces many measurements to a short physical chain: a physical perturbation produces a field response; its mismatch accumulates into forward error along a path, is filtered and potentially amplified by sensors during inversion, and provides a target for repair. This account explains why field fit, derivative fit, physics consistency, and inverse identifiability are complementary rather than interchangeable. It also changes what a useful result looks like. The Burgers rank/top-one disagreement is not a failed table to conceal; it shows that a score can diagnose population-wide risk and still choose badly at the extreme. The NS forward gain shows the opposite regime, where ID field validation fails to recognize a response-faithful candidate. The inverse comparison asks a different question again: which model presents a trustworthy local instrument to the sensor optimizer?

Several limits restrict the claim. Equation (3) is exact, but estimating its integral from a few probes requires assumptions about unsampled curvature and path coverage. With strong shocks or far extrapolation those assumptions can fail. A physics residual only measures the chosen discretized equation and can favor a wrong coarse-grid solution. Parameter Jacobians, additional reference solves, and multi-start inversion cost time; these costs must be weighed against simply acquiring direct target labels. A shift-calibrated selector already uses labeled historical OOD cases and should not be marketed as strict-ID generalization. The same fixed 36-model candidate zoo appears in development and protected confirmation, so a separate zoo would be needed to claim transport to novel model families. Twelve parent-level confirmation units give limited power, especially after multiplicity correction. Finally, chaotic NS pointwise trajectories at $T = 10$ are not grid-converged DNS; early-time fields and late-time statistics answer different physical questions.

The completed evidence supports a coherent, conditional result: response information can improve NS forward choice and early-time predictor repair, while neither global ranking nor improved field fit guarantees the downstream decision we care about. The inverse lead remains exploratory and needs an independent confirmation cohort. Our recommended workflow is to state the task and reference budget, freeze a corresponding response rule or intervention, and evaluate its actual deployment loss. Response geometry is useful because it exposes distinctions that a single field-error table hides, including negative effects.

REPRODUCIBILITY STATEMENT

The main text specifies the physical shifts, data roles, candidate families, information regimes, frozen decision points, and parent-level endpoints. Appendix A records the governing and differentiated PDEs and the resolution qualification; Appendix B maps presentation names to archived source labels. The frozen release manifests, seeds, hashes, and evaluation scripts provide a basis for an anonymized code/data supplement. The completed repair audit checks all 231 new checkpoint/training records, 135 historical controls, 45 compute checkpoints, and all retained parent outcomes. Primary row counts, frozen digests, finite selected errors, reported means, and Holm adjustments were verified after download. One raw inverse case has failed-restart diagnostics; the frozen adverse-error rule retains it, with unavailable diagnostics disclosed rather than dropping the parent. An anonymized executable supplement remains necessary for complete artifact release.

AI USE STATEMENT

Generative AI was used to assist with writing and polishing and with synthetic-data generation workflows. The authors take full responsibility for the final submission.

REFERENCES

Abdolmehdi Behroozi, Chaopeng Shen, and Daniel Kifer. Sensitivity-constrained fourier neural operators for forward and inverse problems in parametric differential equations. In *International Conference on Learning Representations (ICLR)*, 2025.

Abdolmehdi Behroozi, Chaopeng Shen, Daniel Kifer, and Kathryn Lawson. Sensitivity-constrained neural operators for data-efficient forward and inverse modeling of partial differential equation systems. *arXiv preprint arXiv:2608.29888*, 2026. URL https://arxiv.org/abs/2608.29888.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *International Conference on Learning Representations (ICLR)*, 2021a.

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

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

Thomas O'Leary-Roseberry, Peng Chen, Umberto Villa, and Omar Ghattas. Derivative-informed neural operator: An efficient framework for high-dimensional parametric derivative learning. *arXiv preprint arXiv:2206.10745*, 2022. URL https://arxiv.org/abs/2206.10745.

Boyuan Yao, Dingcheng Luo, Lianghao Cao, Nikola Kovachki, Thomas O'Leary-Roseberry, and Omar Ghattas. Derivative-informed fourier neural operator: Universal approximation and applications to PDE-constrained optimization. *arXiv preprint arXiv:2512.14086*, 2025. URL https://arxiv.org/abs/2512.14086.

A PDES, PHYSICAL RESPONSES, AND NUMERICAL CAVEATS

One-dimensional viscous Burgers equation. On the periodic interval $[0, 2\pi]$, $u_t + uu_x = \nu u_{xx}$ and $u(x, 0) = Au_{0,a}(x)$ for a fixed parent shape $u_{0,a}$. For the total log-viscosity tangent $q_\nu = \partial_{\log \nu} u$, differentiating the PDE gives

$$(q_\nu)_t + u(q_\nu)_x + q_\nu u_x = \nu(q_\nu)_{xx} + \nu u_{xx}, \qquad q_\nu(x, 0) = 0. \tag{12}$$

The amplitude tangent has the same homogeneous evolution with $q_A(x,0) = u(x,0)$. Thus a model can match present fields while underestimating how shocks steepen as viscosity falls. The source 128-point grid failed the severe-shock resolution gate. The retained 2048-point ETDRK4 reference agreed with a 4096-point check within the frozen field and response thresholds; original lower-resolution outputs remain available for source replication.

Two-dimensional steady Darcy flow equation with heterogeneous permeability. On the unit square with homogeneous Dirichlet boundary, $-\nabla \cdot (K(x;\sigma)\nabla u) = Af(x)$ and $K = \exp(\sigma a(x))$ for a fixed parent field a. If $q = \partial_{\log\sigma} u$ and $K' = \partial_{\log\sigma} K = \sigma a K$, differentiation yields

$$-\nabla \cdot (K\nabla q) = \nabla \cdot (K'\nabla u), \qquad q|_{\partial\Omega} = 0. \tag{13}$$

For flux $j = -K\nabla u$, $\partial_{\log\sigma} j = -K'\nabla u - K\nabla q$: both terms are required. For the log-source-amplitude tangent q_A, the right-hand side is Af. The 32^2 source response was insufficiently resolved, while the 64^2 reference passed a 128^2 comparison under the response threshold. Boundary-compatible transforms are used for Darcy spectral features; a periodic FFT would change the physical question.

Two-dimensional forced incompressible Navier–Stokes equations in vorticity form. On $[0, 2\pi]^2$, the solver evolves periodic vorticity ω, not a stored velocity output: $\omega_t + u\cdot\nabla\omega = \nu\Delta\omega + Ag$ with divergence-free velocity $u = \mathcal{B}\omega$ recovered through the periodic Biot–Savart map $\mathcal{B}$. For any log-parameter tangent q, its velocity perturbation is $\mathcal{B}q$, and its equation contains both $u \cdot \nabla q$ and $(\mathcal{B}q) \cdot \nabla\omega$. The log-viscosity source is $\nu\Delta\omega$; the log-forcing source is Ag. Omitting the induced velocity response would be a physically incorrect tangent. The 64^2 operational source was compared with 128^2, 256^2, and 512^2 trajectories. At $T = 10$, pointwise differences did not converge; no DNS or continuum pointwise claim is made there. Earlier field windows and later spectra/aggregate budgets are interpreted separately.

A.1 Interpreting the common response

This decomposition tells us what a response probe detects. A model may react by the right *amount* but in the wrong spatial locations, making the angular term large. Another may place the change correctly but underreact everywhere, making the magnitude term large. High-pass weighting asks whether either mistake is concentrated in fine spatial scales; time-resolved weighting asks whether it appears only after the dynamics develop. Neither weighting is universally appropriate: high-frequency noise in a coarse reference can otherwise dominate a useful field predictor. We therefore report unweighted field error and resolution checks alongside weighted response scores.

For a locally twice-differentiable error map with two physical knobs, the second-order term is especially revealing. If $h = (h_1, h_2)$, then

$$e_\theta(z_0 + h) \approx e_\theta(z_0) + \Delta\mathcal{J}_\theta(z_0)h + \tfrac{1}{2}h_1^2\Delta H_{11} + h_1h_2\Delta H_{12} + \tfrac{1}{2}h_2^2\Delta H_{22}, \tag{14}$$

where $\Delta H_{ij} = \partial_{z_i}\partial_{z_j} e_\theta$ is a field-valued curvature defect. The mixed term $h_1h_2\Delta H_{12}$ is invisible to tests confined to the two coordinate axes. Changing viscosity and forcing together may alter a flow more than the sum of their separate effects. Our 3×3 parameter bank measures this curvature with mixed second differences and separately compares axis and joint finite secants. These are different diagnostics: at one point, $\mathcal{J}(v_1 + v_2) = \mathcal{J}v_1 + \mathcal{J}v_2$ provides no new second-order information, and symmetric joint secants cancel even-order terms. Equation (14) motivates the mixed-difference test, not a claim that joint first derivatives alone identify interactions.

The equation explains why the forward and inverse winners may differ. Whole-field L2 averages error across the domain, while the inverse optimizer sees only $O\mathcal{G}_\theta$. A small error in a sensor-visible pattern that looks like a parameter change can create a large bias after multiplication by F^{-1}. Conversely, a model with somewhat larger field error may have a trustworthy sensor response and a better-conditioned inverse landscape. We compare model and reference observation Jacobians on known ID calibration cases, and measure the alignment of their strongly observable parameter directions. This does not reveal the unknown test parameter; it asks whether the local measuring instrument provided by the surrogate is credible before using it to invert data.

Physics and tangent losses. For the implemented discrete PDE residual $\mathcal{R}$, the compared defects are

$$r_\theta = \mathcal{R}(\widehat{u}, z) - \mathcal{R}(u_\star, z), \qquad s_\theta = D_z\mathcal{R}(\widehat{u}, z) - D_z\mathcal{R}(u_\star, z). \tag{15}$$

The first asks whether the two fields satisfy the discrete equation equally well. The second asks whether that agreement changes correctly when a parameter moves; differentiation includes both explicit PDE coefficients and the corresponding field tangent. Relative squared versions supply L_{physics} and L_{tangent}. The reference residual is retained because a coarse stored trajectory need not have zero residual under the training discretization. Direct response loss instead compares the field tangents themselves. These three targets are scientifically different, even when they share a reference solver.

B Presentation names and archived method labels

The paper removes unexplained source version numbers from its primary exposition without changing the tested methods. Table 5 gives exact archived strings so results can be traced to frozen protocols and to the team's earlier reports.

Table 5: Descriptive manuscript names and exact archived labels. The final three inverse primaries are PDE-specific; the compact NS result is exploratory.

Manuscript name	Exact archived label
Path-response score (Burgers forward primary)	`V3-1 path NNLS source`
Spectral-response score (Darcy forward primary)	`HP-RGE spectral NNLS source`
Joint-response score (NS forward primary)	`Multi-Parameter NNLS source`
Sensor-geometry score (NS inverse primary)	`SA-RGE/IP v6 + HR-IS hinge/ratio`
Information-volume risk (Burgers inverse primary)	`I3c_logdet_risk_aligned single-channel`
RGE inverse score (Darcy inverse primary)	`G_RGE single-channel`
Compact sensor-geometry score (exploratory)	`HR-IS core_5`
Target-assisted reconstructed comparator	`OS-RGE v4 displayed-master reconstruction`

The first three scores are fitted nonnegative composites over different feature families, with ID L2 and capacity retained in the design matrix. "Path" combines uniform/interior response and field-plus-response features; "spectral" combines high-pass and fractional filters; "joint" combines parameter-axis and same/opposite-sign directions. The NS inverse score combines sensor-visible response and identifiability features. Its active-curvature channel uses a declared Gauss–Newton directional reconstruction because the chat did not specify an estimator. The Burgers information-volume score is the risk-aligned negative log determinant of a regularized sensor-information matrix, not the PDE's mixed Hessian. The archived target-assisted reconstruction is not a source-identical implementation of a missing executable and is not a mathematical error certificate.

C Repair interventions and complete protected results

The protected comparison includes every registered repair arm, not just the winning NS intervention. In the tables below, "ID" is field error on fresh ID parents; "OOD" is severe-shift field error of the gated checkpoint; "raw" uses the checkpoint selected without the field-retention gate; "Inv." is mean per-parent two-coordinate log-parameter RMSE; "accept" is the fraction of the three trained seeds accepted on the fixed tune gate, not a guarantee on fresh ID parents; p_H adjusts field tests over all new arms within a PDE. The inverse subset has eight parents, and field endpoints have twelve. Module winners are descriptive.

The normalized names in Table 4 correspond to `M2_three_node_path`, `M2_direct_states_endpoints`, `M0_full_response_active_ID`, `M5_rollout_2_4_8_NS`, and `M5_rollout_plus_response_NS`. Prefixes identify archived intervention modules, not successive versions of the paper's method.

Intervention families. M0 compares field, physics-plus-tangent, and direct response losses with active ID retention. M1 changes parameter-direction supervision and compares signed/wrong joint directions, a physical prior, positive-Lasso allocation, and shuffled NNLS allocation. M2 tests finite chords, three-node quadrature, and mixed Taylor error against equal-reference-node direct-field controls. M3 uses spectral H^1/H^2 weights, fractional $H^{1.5}$ weights, a fixed frequency band, and localized patches, with random-band/patch controls; these are implemented discrete filters, not proofs about continuum Sobolev norms. M4 tests NS transient energy/enstrophy defects and parameter responses of energy/enstrophy/spectral-tail quantities, matching the nonzero discrete reference defects. M5 adds the total-response rollout loss in Equation (11). M6 tests sensor-visible bias and Darcy flux. M8 compares targeted versus random ID reference acquisition and a discrepancy-correction wrapper versus continued training. Historical B6 controls additionally test NNLS-attributed, largest-defect, random/all directions, gradient balancing, sensor Jacobian/Fisher losses, and short unrolled inversion. No pruning intervention is included.

Reading physical and statistical limits. All repair arms are frozen before the new confirmation cohort is generated. B9 robustness instead reuses its original inverse parents; it is not an independent replication of its exploratory compact winner. B8/B9 calibrated score weights use shifted development outcomes. Raw ID-only rules and target-assisted comparisons remain distinct. The displayed-master reconstruction of OS-RGE v4 is not source-identical and is not an OOD-error certificate. At $T = 10$, the NS 64^2 operational source failed pointwise refinement; later spectra and aggregate budgets are diagnostics, not a substitute convergence proof.

The primary all-arm Holm family does not cover every secondary matched contrast. For example, Burgers path-plus-mixed-Taylor minus its equal-node direct-field control is -0.0146 with unadjusted hierarchical interval $[-0.0331, -0.0026]$; NS finite-chord minus endpoint fields is -0.4141 with interval $[-0.6368, -0.2380]$, while neither arm beats the NS active-ID field control. A favorable within-module contrast can therefore coexist with a poor absolute deployment choice. The time bank is finite, and the NS rollout-plus-response winner exceeds its range; the reported nearest-time control is not equal-compute.

C.1 BURGERS: ALL NEW REPAIR ARMS

Table 6: Burgers: complete new-arm endpoints. Exact source names are retained.

Arm	ID	OOD	Raw	Inv.	Accept	p_H
M0_field_only_active_ID	0.0122	0.4180	0.4180	1.2282	1.0000	1.0000
M0_full_response_active_ID	0.0127	0.4256	0.4256	1.2164	1.0000	1.0000
M0_physics_tangent_active_ID	0.0263	0.3612	0.3612	1.3537	1.0000	0.3988
M1_physics_prior_allocation	0.0264	0.3615	0.3615	1.3555	1.0000	0.3988
M1_positive_Lasso_subset	0.0267	0.3617	0.3617	1.3573	1.0000	0.3988
M1_shuffled_NNLS_allocation	0.0263	0.3619	0.3619	1.3525	1.0000	0.3988
M1_signed_joint_response	0.0264	0.3615	0.3615	1.3545	1.0000	0.3988
M1_wrong_joint_response	0.0261	0.3615	0.3615	1.3506	1.0000	0.3988
M2_direct_states_endpoints	0.0119	0.4111	0.4111	1.1783	1.0000	0.0117
M2_direct_states_path_nodes	0.0121	0.4167	0.4167	1.2261	1.0000	1.0000
M2_direct_states_path_plus_grid	0.0124	0.4221	0.4221	1.2295	1.0000	1.0000
M2_finite_chord	0.0123	0.4136	0.4136	1.1820	1.0000	0.0635
M2_path_plus_mixed_Taylor	0.0125	0.4075	0.4075	1.1700	1.0000	0.0117
M2_three_node_path	0.0119	0.4074	0.4074	1.1742	1.0000	0.0117
M3_H1_H2_response	0.0123	0.4221	0.4221	1.1981	1.0000	1.0000
M3_fixed_Fourier_band	0.0122	0.4224	0.4224	1.2011	1.0000	1.0000
M3_fractional_response	0.0124	0.4213	0.4213	1.2295	1.0000	1.0000
M3_patch_tail	0.0130	0.4171	0.4171	1.1913	1.0000	1.0000
M3_random_band_patch	0.0123	0.4216	0.4216	1.1900	1.0000	1.0000
M6_sensor_visible_bias	0.0127	0.4117	0.4117	1.2158	1.0000	0.0359
M8_discrepancy_correction	0.0115	0.4233	0.4233	1.2736	1.0000	1.0000
M8_matched_continued_training	0.0124	0.4258	0.4258	1.2194	1.0000	1.0000
M8_random_ID_reference	0.0118	0.4155	0.4155	1.2353	1.0000	0.3988
M8_targeted_ID_reference	0.0115	0.4217	0.4217	1.2300	1.0000	1.0000

C.2 DARCY: ALL NEW REPAIR ARMS

Table 7: Darcy: complete new-arm endpoints. Exact source names are retained.

Arm	ID	OOD	Raw	Inv.	Accept	p_H
M0_field_only_active_ID	0.0708	0.3007	0.3041	0.1124	1.0000	1.0000
M0_full_response_active_ID	0.0711	0.3007	0.3022	0.1107	1.0000	1.0000
M0_physics_tangent_active_ID	0.0709	0.3013	0.4903	0.1184	0.0000	1.0000
M1_physics_prior_allocation	0.0709	0.3013	0.4761	0.1184	0.0000	1.0000
M1_positive_Lasso_subset	0.0709	0.3013	0.4710	0.1184	0.0000	1.0000
M1_shuffled_NNLS_allocation	0.0709	0.3013	0.4761	0.1184	0.0000	1.0000
M1_signed_joint_response	0.0709	0.3013	0.4655	0.1184	0.0000	1.0000
M1_wrong_joint_response	0.0709	0.3013	0.4870	0.1184	0.0000	1.0000
M2_direct_states_endpoints	0.0709	0.2960	0.2960	0.1104	1.0000	0.0488
M2_direct_states_path_nodes	0.0709	0.2960	0.3009	0.1116	1.0000	0.0937
M2_direct_states_path_plus_grid	0.0696	0.2983	0.3017	0.1154	1.0000	1.0000
M2_finite_chord	0.0709	0.2984	0.2984	0.1110	1.0000	1.0000
M2_path_plus_mixed_Taylor	0.0700	0.2979	0.2993	0.1132	1.0000	1.0000
M2_three_node_path	0.0703	0.2937	0.2937	0.1116	1.0000	0.4940
M3_H1_H2_response	0.0712	0.3070	0.3086	0.1118	1.0000	1.0000
M3_fixed_Fourier_band	0.0708	0.3021	0.3025	0.1136	1.0000	1.0000
M3_fractional_response	0.0707	0.3051	0.3072	0.1127	1.0000	1.0000
M3_patch_tail	0.0709	0.3010	0.3024	0.1122	1.0000	1.0000
M3_random_band_patch	0.0697	0.2995	0.2995	0.1101	1.0000	1.0000
M6_Darcy_flux	0.0702	0.3029	0.3029	0.1172	1.0000	1.0000
M6_sensor_visible_bias	0.0726	0.3015	0.3015	0.1212	1.0000	1.0000
M8_discrepancy_correction	0.0699	0.3006	0.3006	0.1181	1.0000	1.0000
M8_matched_continued_training	0.0698	0.3030	0.3030	0.1167	1.0000	1.0000
M8_random_ID_reference	0.0702	0.3065	0.3076	0.1138	1.0000	1.0000
M8_targeted_ID_reference	0.0701	0.3054	0.3054	0.1143	1.0000	1.0000

C.3 NAVIER–STOKES: ALL NEW REPAIR ARMS

Table 8: Navier–Stokes: complete new-arm endpoints. Exact source names are retained.

Arm	ID	OOD	Raw	Inv.	Accept	p_H
M0_field_only_active_ID	0.2498	0.6284	0.6569	0.3937	0.6667	1.0000
M0_full_response_active_ID	0.2412	0.5726	0.5184	0.5289	0.3333	0.0137
M0_physics_tangent_active_ID	0.2357	0.6143	0.6380	0.2591	0.0000	0.3127
M1_physics_prior_allocation	0.2357	0.6143	0.6513	0.2591	0.0000	0.3127
M1_positive_Lasso_subset	0.2357	0.6143	0.6002	0.2591	0.0000	0.3127
M1_shuffled_NNLS_allocation	0.2357	0.6143	0.6518	0.2591	0.0000	0.3127
M1_signed_joint_response	0.2499	0.5822	0.5951	0.2550	0.3333	0.0137
M1_wrong_joint_response	0.2357	0.6143	0.5901	0.2591	0.0000	0.3127
M2_direct_states_endpoints	0.1888	1.0348	1.0348	0.8505	1.0000	1.0000
M2_direct_states_path_nodes	0.1882	1.0625	1.0625	0.8123	1.0000	1.0000
M2_direct_states_path_plus_grid	0.1909	1.0052	1.0052	0.6105	1.0000	1.0000
M2_finite_chord	0.2409	0.6207	0.8589	0.4286	0.3333	1.0000
M2_path_plus_mixed_Taylor	0.2256	0.8725	1.0701	0.8248	0.6667	1.0000
M2_three_node_path	0.2404	0.8602	1.0538	0.7944	0.6667	1.0000
M3_H1_H2_response	0.2370	0.6291	0.5685	0.3664	0.3333	1.0000
M3_fixed_Fourier_band	0.2486	0.5783	0.5186	0.5480	0.3333	0.0137
M3_fractional_response	0.2433	0.5952	0.5737	0.3036	0.3333	0.0137
M3_patch_tail	0.2357	0.6143	0.5542	0.2591	0.0000	0.3127
M3_random_band_patch	0.2357	0.6143	0.5513	0.2591	0.0000	0.3127
M4_susceptibility_tail_NS	0.2357	0.6143	0.6752	0.2591	0.0000	0.3127
M4_transient_budget_NS	0.2357	0.6143	0.8428	0.2591	0.0000	0.3127
M5_rollout_2_4_8_NS	0.1293	0.5263	0.5263	1.0941	1.0000	0.0859
M5_rollout_plus_response_NS	0.1567	0.4899	0.4899	1.0560	1.0000	0.0137
M6_sensor_visible_bias	0.2357	0.6143	0.7450	0.2591	0.0000	0.3127
M8_discrepancy_correction	0.2423	0.6256	0.6343	0.2203	1.0000	1.0000
M8_matched_continued_training	0.2236	0.6506	0.6386	0.5548	1.0000	1.0000
M8_random_ID_reference	0.2464	0.5893	0.5382	0.5426	0.3333	0.0137
M8_targeted_ID_reference	0.2357	0.6143	0.5387	0.2591	0.0000	0.3127

C.4 HISTORICAL CONTROLS ON THE SAME PROTECTED PARENTS

Each cell below shows OOD field L2 / inverse RMSE, averaged across three seeds. These controls do not have the new active-ID dual mechanism. Their exact names allow direct comparison with earlier reports; the parent cohort is the new repair cohort, not the original historical evaluation cohort.

Table 9: All 15 B6 control methods re-evaluated on the B10 parent cases.

Control	Burgers	Darcy	Navier–Stokes
NNLS_attributed_direction_repair	0.3853 / 1.3669	0.3013 / 0.1184	0.6143 / 0.2591
active_direction_Jacobian	0.3932 / 1.1898	0.2884 / 0.1043	0.5524 / 0.5574
all_direction_repair	0.3747 / 1.3086	0.3013 / 0.1184	0.5708 / 0.3249
field_only	0.3989 / 1.1975	0.2978 / 0.1149	0.6276 / 0.3158
full_response	0.4005 / 1.2006	0.2875 / 0.1046	0.5556 / 0.6655
gradient_balanced_full	0.3779 / 1.3056	0.3013 / 0.1184	0.6143 / 0.2591
largest_direct_defect_repair	0.3989 / 1.1899	0.2978 / 0.1149	0.6276 / 0.3158
observation_Jacobian_plus_Fisher_shape	0.3953 / 1.1933	0.2864 / 0.0972	0.6143 / 0.2591
physics_only	0.3982 / 1.1932	0.2978 / 0.1149	0.6424 / 0.4118
physics_plus_tangent	0.3738 / 1.3100	0.3013 / 0.1184	0.5858 / 0.3873
random_direction_repair	0.4050 / 1.2665	0.3013 / 0.1184	0.6143 / 0.2591
response_plus_observation_inverse	0.4277 / 1.2362	0.2934 / 0.1084	0.6143 / 0.2591
short_unrolled_inverse	0.3988 / 1.1959	0.2978 / 0.1147	0.6143 / 0.2591
tangent_only	0.3754 / 1.3144	0.3013 / 0.1184	0.5757 / 0.4096
unchanged_base	0.4238 / 1.2748	0.3013 / 0.1184	0.6143 / 0.2591