IDENTIFIABILITY BEFORE OPTIMIZATION: SPARSE OBSERVATION DESIGN FOR PDE SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

A vanishing physics residual can coexist with large solution error when the exposed partial differential equation (PDE) specification leaves the target ambiguous. We isolate this information failure by removing a public condition while keeping the hidden target fixed. For finite-dimensional linear operators, sparse observations restore identifiability exactly when they have full rank on the operator's nullspace; their conditioning controls reconstruction stability. We extend the analysis to estimated ambiguity spaces and local nonlinear problems. Motivated by these conditions, a target-independent portfolio selects observation locations before querying their values. Across seven controlled symbolic-search families, omitting initial data yields a perfect proxy optimum with relative L_2 error 1.000; nine public observations reduce mean error to 0.088, versus 0.074 with the full condition. On 16 PDE libraries, using ten observations with 5% representation mismatch and 1% Gaussian noise, the portfolio achieves mean error 0.0587, improves on uniform placement in 15/16 tasks, and meets an absolute 0.01 non-inferiority margin to A-optimal design. Independent neural-solver experiments reproduce the omission–repair gap: eight observations improve all 24 paired DeepXDE runs, although Allen–Cahn remains inaccurate. These results distinguish information sufficiency from optimization success and show how ambiguity analysis can guide sparse observation design.

1 INTRODUCTION

Agentic scientific search for partial differential equations (PDEs) increasingly optimizes public physics checks when a reference solution is unavailable. Agentic Symbolic Search (ASYS), for example, evolves differentiable symbolic programs, fits continuous parameters by limited-memory Broyden–Fletcher–Goldfarb– Shanno (L-BFGS) optimization, and selects survivors by residual, initial, boundary, and compatibility checks without consulting ground truth (Yu & Yang, 2026). The same premise underlies physics-informed learning (Raissi et al., 2019), symbolic regression (Brunton et al., 2016; Udrescu & Tegmark, 2020; Cranmer, 2023), and large language model (LLM)-guided program search (Romera-Paredes et al., 2024; Novikov et al., 2025; Shojaee et al., 2025): optimization is steered by a proxy for correctness rather than correctness itself.

Optimizing a proxy in place of the true objective is the setting of Goodhart's law and reward hacking, where over-optimization degrades the very quantity a proxy was meant to track (Skalse et al., 2022; Laidlaw et al., 2025; Gao et al., 2023; Pan et al., 2022; Manheim & Garrabrant, 2018; Amodei et al., 2016; Krakovna et al., 2020). Much of that literature is an offline story: a gold reward exists but is withheld until after optimization. Collocation methods exhibit a related scientific failure: a sampled residual can decouple from global error, motivating spectral-bias and adaptive-resampling studies (Rahaman et al., 2019; Daw et al., 2023). We ask a sharper question than whether this known gap exists: when does the public specification fail to identify the scientific target, and what is the smallest target-independent intervention that restores identifiability? This separates two causes that the same validation error can conceal. Better optimization can reduce fitting error within an identifiable discretized task. It cannot recover information that the task never exposed.

A public operator may leave a d-dimensional family of targets indistinguishable. Measurements must separate those directions before residual minimization can identify the target. We test this principle through matched omission and repair, then use the ambiguity space to design target-independent

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.

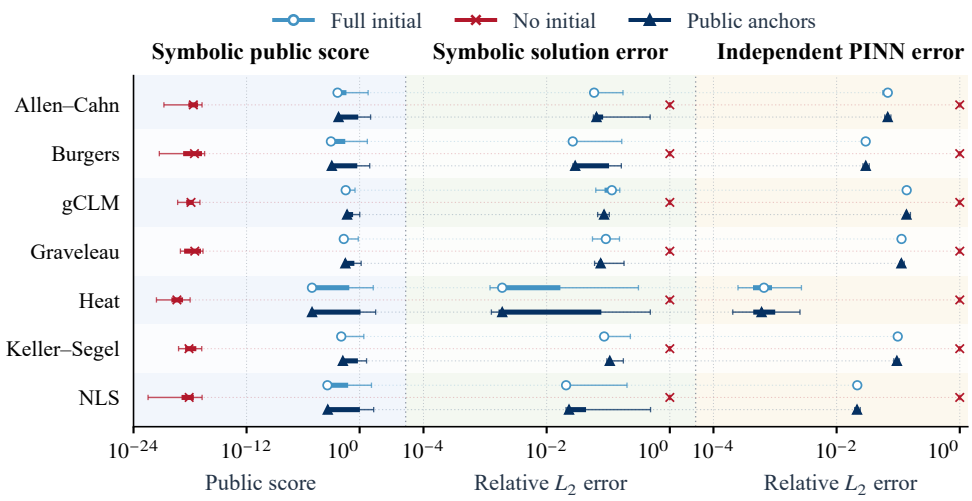

Figure 1: **Public information separates residual fit from identification.** Markers show medians; thick/thin bars show interquartile/full ranges. Dotted guides align separate log scales, not paired runs. Symbolic starts per case: 8 full, 16 omitted, 16 repaired. Neural pools: 12 seeds $\times$ 4 settings. Generalized Constantin–Lax–Majda (gCLM), nonlinear Schrödinger (NLS).

measurements. The analysis predicts both a rank threshold at $k = d$, where k is the number of scalar observations, and a stability penalty for poorly conditioned measurements.

Contributions. This paper makes three contributions. First, a stress test that withholds reference solutions separates missing information from approximation and optimization error. Second, nullspace analysis characterizes how many observations are needed, where they should be placed, and how conditioning affects reconstruction; the analysis covers estimated ambiguity spaces and gives a local nonlinear result. Third, a public portfolio converts these ideas into an automatic observation policy and is tested across symbolic search, random-feature and finite-difference solvers, physics-informed neural networks (PINNs), and standard DeepXDE problems. Together, these results shift the question from whether a residual is small to whether the public specification identifies the target. Figure 1 contrasts public scores with returned errors and tests repair across two solver stacks.

2 Problem Setup

Scope. The testbed combines proxy-only search with separate reference-based evaluation. Its seven one-dimensional reaction–diffusion surrogates are controlled problems chosen to isolate specification effects. The basis-limited search uses population 8, 12 generations, and 96 collocation points. Poisson–Neumann and two-dimensional heat examples broaden the types of ambiguity considered. The surrogate suite comprises Allen–Cahn, Burgers, generalized Constantin–Lax–Majda (gCLM), Graveleau, heat, Keller–Segel, and nonlinear Schrödinger (NLS) cases. A separate DeepXDE experiment uses the standard heat, viscous Burgers, and Allen–Cahn equations.

Cases and the hidden true error. The workload is a fixed suite of M cases, $\mathcal{C} = \{c_1, \dots, c_M\}$, each with a hidden reference $u_c^\star$, held fixed across all methods. A candidate θ in the search space Θ induces a prediction $\hat{u}_c(\theta)$, with per-case and aggregate hidden error

$$E_c(\theta) = \frac{\|\hat{u}_c(\theta) - u_c^\star\|_2}{\|u_c^\star\|_2}, \qquad \bar{E}(\theta) = \frac{1}{M}\sum_{c=1}^{M} E_c(\theta), \tag{1}$$

where $\|\cdot\|_2$ is the discrete L_2 norm on the evaluation grid and $\|u_c^\star\|_2 > 0$. Search selects $\hat{\theta}$ from the accepted final-generation candidates by minimizing $s(\mathbf{p}(\theta))$. Here $\mathbf{p} \in \mathbb{R}^4$ (or $\mathbb{R}^5$ with sparse public anchors) contains the public score components, and s is a fixed scalarization. Hidden errors are evaluated only after selection.

Reference access. Search observes only the public score $\theta \mapsto \mathbf{p}(\theta)$ and the prescribed data. Neither reference trajectories nor hidden errors enter fitting, scalarization, or selection. After predictions are fixed, a separate evaluation computes returned errors and candidate-pool correlations.

3 Specification Stress Test and Repair

The protocol first fixes which information search can observe, then removes one public condition while retaining the same hidden target. If the remaining specification is ambiguous, measurement locations are chosen from the public operator's nullspace and committed before their values are queried. Residual optimization controls visible directions. The measurements must control the directions that the residual cannot detect.

A candidate u_θ is a differentiable symbolic program: a sum of parametric atoms over $z = (x, t)$ whose atom-type multiset (the structure) is chosen by the outer evolutionary loop and whose parameters θ are fitted by L-BFGS. Each case exposes only public operators (a residual $\mathcal{R}[u]$, an initial residual $\mathcal{I}[u]$, a boundary residual $\mathcal{B}[u]$) and samplers for interior, initial, boundary, and optional anchor points. Let C, I, B denote interior, initial, and boundary point sets, and $\widehat{\mathbb{E}}_D$ the empirical average over any point set D. The default proxy is

$$\begin{aligned} S(u) = &\widehat{\mathbb{E}}_C[\mathcal{R}[u]^2] + 2\,\widehat{\mathbb{E}}_I[\mathcal{I}[u]^2] + 2\,\widehat{\mathbb{E}}_B[\mathcal{B}[u]^2] \\ &+ 1.5\,\widehat{\mathbb{E}}_I[\mathcal{R}[u]^2]. \end{aligned} \tag{2}$$

The last term is a compatibility check: the residual is re-evaluated on the initial slice. When a stress variant withholds dense initial information, its initial term is omitted. The repair instead adds $20\,\widehat{\mathbb{E}}_A[(u-a)^2]$ for fixed public measurements a on the anchor set A. We test one to nine anchors, up to 5% deterministic amplitude noise, off-slice observations, and 25 two-dimensional observations. The inner loop minimizes only the active public proxy. The hidden error equation 1 is computed on a separate validation grid after search is complete and never enters fitting or selection.

Sparse target values are public measurements, distinct from the withheld dense trajectory. The default anchor coefficient is 20; equal-cost comparisons use eight values at weight 2, identical optimization limits, and paired seeds. Search-pressure experiments use 3, 6, 12, and 24 generations.

The algorithm-independent lower bounds in Appendix A show both that indistinguishable targets force half their separation as worst-case error and that a deterministic adaptive policy using fewer than d continuous linear scalar queries cannot eliminate a d-dimensional ambiguity. More computation cannot restore information that was never queried.

Corollary 1 (Residual-only adaptation cannot repair omitted information)**.** *Suppose two admissible targets u_0, u_1 induce the same public operator, source, boundary data, and residual response at every queried coordinate. Any randomized solver whose training updates and adaptive collocation points depend only on this public residual transcript has the same output distribution on the two tasks. For any fixed prediction norm $\|\cdot\|$, with expectation over solver randomness,*

$$\max_{j\in\{0,1\}} \mathbb{E}\|\widehat{u} - u_j\| \geq \tfrac{1}{2}\|u_0 - u_1\|. \tag{3}$$

If a linear public measurement H satisfies $H(u_0 - u_1) \neq 0$, the indistinguishability premise is broken and the lower bound no longer follows.

The lower bound in equation 3 covers residual-prioritized resampling, including retain–resample–release (R3): changing where an exposed operator is checked can improve discretization, but cannot reveal a condition absent from the public task. Appendix A gives the coupling proof. The independent PINN/R3 experiment tests this prediction with a separate solver and seed set.

Specification stress variants. Each stress case retains the default hidden reference and residual/boundary checks but removes dense initial data, admitting the zero function. The repair arm adds nine fixed public values on the initial slice while leaving the dense initial term absent. This tests whether sparse observations rule out the zero solution.

3.1 Public Anchor Design from the Ambiguity Space

The stress tests suggest a constructive repair: design measurements against what the public operator cannot see. Let $A \in \mathbb{R}^{m\times n}$ be the discretized public physics operator, with $d = \dim\ker A > 0$, and let $N \in \mathbb{R}^{n\times d}$ span $\ker A$. The k-row observation operator $H \in \mathbb{R}^{k\times n}$ is chosen using this subspace, without target values or validation error.

Theorem 2 (Nullspace-complete anchors)**.** *Let N be an orthonormal basis for $\ker A$. The augmented specification (A, H) uniquely identifies each solution if and only if $\operatorname{rank}(HN) = d$. Under this rank condition, any candidate error $e = \widehat{u} - u^\star$ satisfies*

$$\|e\|_2 \le \|A^\dagger\|_2\|Ae\|_2 + \sigma_{\min}(HN)^{-1}\big(\|He\|_2 + \|H\|_2\|A^\dagger\|_2\|Ae\|_2\big). \tag{4}$$

Here $A^\dagger$ is the Moore–Penrose inverse. Vector and matrix norms are Euclidean and spectral, respectively. For a matrix with d columns, $\sigma_{\min}$ includes zero singular values when its rank is below d. The bound separates two failure modes: residual control acts on $(\ker A)^\perp$, whereas public anchors control the invisible component. We therefore choose sensors greedily to maximize $\log\det(N^\top H^\top HN + \epsilon I)$, where I is the $d\times d$ identity and $\epsilon > 0$ is a regularization parameter. This is a target-free D-optimal criterion. The controlled extension compares this rule with random and uniform sensors on additive, amplitude, translation, phase, two-interface, and four-mode ambiguities, across sensor counts, observation noise, and independent trials. Appendix A gives the proof and the fixed decision rule.

For automatic repair, we linearize a public constraint vector $F(\theta)$ at a reference-free pilot and estimate a weak right-singular subspace $\widehat{N}$. Candidate locations are scored only through public rows $H\widehat{N}$ and committed before values are queried. The comparison includes random, uniform, leverage, column-pivoted orthogonal–triangular (QR) factorization, and A-, D-, E-, and D–E-optimal designs. Dense references remain behind the firewall.

Theorem 3 (Robustness to an estimated ambiguity space)**.** *Let N and $\widehat{N}$ have $d > 0$ orthonormal columns, and let $\Theta(N, \widehat{N})$ denote their principal angles. Set $\delta = \|\sin\Theta(N, \widehat{N})\|_2$, define $\eta(\delta) = \sqrt{2 - 2\sqrt{1-\delta^2}}$, and suppose $\sigma_{\min}(H\widehat{N}) = \gamma > \|H\|_2\eta(\delta)$. Then*

$$\sigma_{\min}(HN) \ge \gamma - \|H\|_2\eta(\delta). \tag{5}$$

Consequently, the bound in Theorem 2 remains valid with the factor $\sigma_{\min}(HN)^{-1}$ replaced by $(\gamma - \|H\|_2\eta(\delta))^{-1}$.

The bound follows from the Procrustes characterization of principal angles and Weyl's singular-value inequality. It motivates reporting both subspace recovery and measurement conditioning, and comparing E-optimality's stability criterion with D-optimality's volume criterion.

The theory links estimated nullity, subspace error, and $\sigma_{\min}(H\widehat{N})$ to rank recovery and reconstruction stability. The public portfolio compares conditioned designs using a column-normalized library X on candidate points. Here d denotes the library dimension. Let X_g contain its values at n_g public reconstruction points and X_S its rows at selected locations S. Define $M_S = X_S^\top X_S + 10^{-10}I_d$ and $G = X_g^\top X_g/n_g$. The portfolio score is

$$s(S) = \sqrt{\lambda_{\max}(M_S^{-1/2}GM_S^{-1/2})} + \frac{0.25}{\sqrt{d}}\sqrt{\operatorname{tr}(M_S^{-1}G)} + 0.10h(S), \tag{6}$$

Here I_d is the identity and $h(S)$ is the maximum distance from a candidate location to its nearest selected point, after scaling coordinates to $[0, 1]$. The selector evaluates sets proposed by pivoted QR, A/D/E-optimal, spectral log-determinant, and predictive A/E designs, and returns the set with the smallest score before any values are queried. The first two terms measure spectral and average predictive amplification; the last discourages poor spatial coverage. This score is an empirical design rule, while Theorems 2–3 describe its identifiability requirements.

For a nonlinear public residual F, let $F(\theta^\star) = 0$ and $J = DF(\theta^\star)$, with N an orthonormal basis for $\ker J$. Under a Lipschitz Jacobian and full-rank HN, Theorem 7 proves, within its stated neighborhood, $\|\theta - \theta^\star\| \le 2C_J\|F(\theta) - F(\theta^\star)\| + 2C_H\|H(\theta - \theta^\star)\|$, where $C_H = \sigma_{\min}(HN)^{-1}$ and $C_J = (1 + \|H\|_2C_H)/\sigma^+_{\min}(J)$; $\sigma^+_{\min}$ is the smallest positive singular value. In contrast,

$k < d = \dim \ker J$ cannot identify all first-order invisible directions. This predicts a phase transition at $k = d$ and a noise penalty controlled by $\sigma_{\min}(HN)$, without claiming global symbolic-search identifiability.

A separate guarded-exchange design uses a rank estimator that stacks three normalized public pilot Jacobians and selects the largest admissible trailing spectral gap; it returns nullity zero when no admissible gap exists. Starting from a uniform design, deterministic exchanges are accepted when the normalized fill radius is at most 1.10 times that of uniform and $\sigma_{\min}$ improves by at least 25%. Otherwise this design keeps the uniform placement. Calibration covers $d \in \{0, 1, 2, 4\}$.

Proposition 4 (Stable-gap recovery). *Let $J \in \mathbb{R}^{m\times p}$ have nullity $d \in \{0, \dots, p-1\}$ and smallest nonzero singular value $g = \sigma_{p-d}(J) > 0$. Order all p right singular values nonincreasingly, including zeros. If $\widehat{J} = J + E$ with $\|E\|_2 < g/2$, any threshold $\tau \in [\|E\|_2, g - \|E\|_2)$ identifies exactly d values at or below τ. For $d > 0$, their right-singular subspace obeys $\| \sin \Theta(N, \widehat{N})\|_2 \le \|E\|_2/(g - \|E\|_2)$.*

The proposition states when the pilot spectrum separates the ambiguity space from visible directions. It describes local rank recovery, not global uniqueness for a nonlinear objective.

4 Experimental Protocol

The evaluation has four parts. Matched omission–repair experiments use seven symbolic PDE families. Controlled reconstruction experiments pair eight designs on six ambiguity families with $d \in \{1, 2, 4\}$, with $k = d$ and 1% noise as the primary setting. The public-portfolio benchmark contains 16 PDE libraries, 13 designs, four budgets, three representation-mismatch levels, and five noise laws. Independent random-feature, finite-difference, PINN, and DeepXDE experiments test whether the mechanism survives a change of solver. Matched arms share seeds within each experiment; dense references are accessed only after prediction.

Fairness and inference. Matched specification arms share targets, initializations, optimizer steps, model capacity, and collocation budgets. Equal-cost placement arms also share an eight-observation count and loss weight. Inference is performed at the seed, case, or task level rather than treating repeated trials as independent. We use paired cluster bootstrap intervals, exact sign-flip tests when the number of clusters is small, and Holm correction for simultaneous design comparisons. Means, medians, 90th percentiles (P90), and catastrophic-error rates summarize central performance and tails.

Measures. Over the candidate pool we report Spearman rank correlation ρ between S and E. For each case and run, the returned error is the hidden error of the final generation's best-scoring accepted survivor. On a fixed pool $\mathcal{U}$, the selection regret $E(\arg\min_{u\in\mathcal{U}} S(u)) - \min_{u\in\mathcal{U}} E(u)$ measures the cost of selecting by the proxy rather than hidden error.

Selector variants. A selector $\sigma(u)$ replaces the outer-loop survivor objective while the inner L-BFGS fit always optimizes equation 2, so any effect is attributable to selection alone. Controls cover held-out, ensemble, and multislice residuals, higher-order and dual-weighted consistency, smoothness, parsimony, Borda aggregation, and Pareto survival. These controls change how exposed information is used; sparse observations change the information itself.

Comparison methods and reference-based controls. No-initial versus public-anchor arms test information sufficiency. Random and uniform placement are sanity baselines. Pivoted QR and A-, D-, E-, and G-optimal designs are strong equal-count comparators. For the information matrix, A-optimal minimizes the trace of its inverse, D-optimal maximizes its determinant, E-optimal maximizes its minimum eigenvalue, and G-optimal minimizes worst-case prediction variance. The lowest hidden error in a fixed candidate pool is an offline control. For the public portfolio, the primary tests are superiority to uniform and non-inferiority to A-optimal with margin $.01$. Figure 2 illustrates two multi-mode libraries.

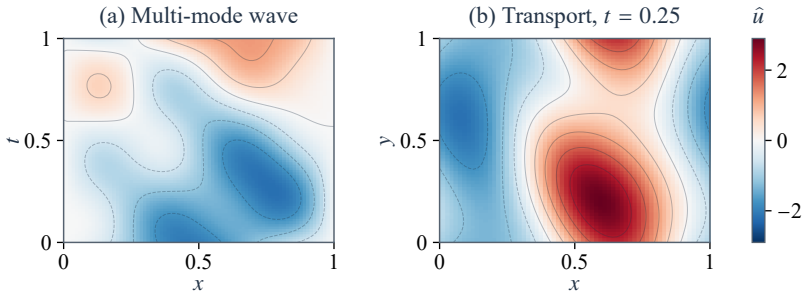

Figure 2: **Multi-mode reconstructed solution fields.** (a) Space–time evolution of a four-frequency wave. (b) Two-dimensional periodic advection–diffusion at $t = 0.25$. Both use ten public observations with 1% noise and zero representation mismatch in an eight-coefficient library. Appendix D supplies the reference fields and error maps.

Table 1: **Omission and repair.** Candidate-pool Spearman correlation ρ and returned relative L_2 over eight default and 16 omission/repair starts per case. The default exposes dense initial data. Omission admits a zero-proxy solution with error one. Nine public anchors recover useful solutions without exposing the hidden trajectory.

Case	ρ default	ρ anchors	L_2 default	L_2 no-init.	L_2 anchors	med. anchors
Allen–Cahn	0.543	0.800	0.073	1.000	0.108	0.064
Burgers	0.764	0.811	0.044	1.000	0.060	0.029
gCLM	0.267	0.235	0.107	1.000	0.085	0.086
Graveleau	0.356	0.327	0.096	1.000	0.082	0.075
Heat	0.990	0.948	0.048	1.000	0.078	0.002
Keller–Segel	0.800	0.569	0.106	1.000	0.114	0.106
NLS	0.823	0.817	0.044	1.000	0.089	0.023
Mean	0.649	0.644	0.074	1.000	0.088	0.055

5 Results

Results are reported for the prediction selected by the public objective; hidden error is computed only afterward. Correlations below are Spearman correlations between the exposed proxy and hidden relative L_2 error.

5.1 Complete specifications align proxy and solution error

With the complete public specification, every case has positive proxy–error correlation, from 0.267 on gCLM to 0.990 on heat, and mean returned error ranges from 0.044 to 0.107 (Table 1). Thus the baseline objective is imperfect but does not collapse to the zero solution.

5.2 Missing initial data makes the zero solution optimal

Removing only the dense initial condition leaves the hidden target unchanged but makes the zero function satisfy every exposed check:

$$\mathcal{R}[0] = 0, \qquad \mathcal{B}[0] = 0, \qquad \mathcal{R}[0]\big|_{t=0} = 0. \tag{7}$$

Every nonnegative proxy term is therefore zero. This is a global public optimum, not a ranking accident, whereas each nonzero hidden reference gives relative error one. All seven families reach this optimum in the matched stress test (Table 1). Adding nine public anchors penalizes the zero function before selection and restores mean error near the complete-specification regime.

Table 2: **Observation-budget, noise, and location sensitivity.** Each row pools seven cases and 16 starts. A single anchor leaves high-error outliers. Three to nine anchors, noisy values, and off-slice observations preserve the repair.

Public measurement	mean L_2	med. L_2	max L_2
1 anchor	0.182	0.106	1.000
3 anchors	0.086	0.080	0.568
5 anchors	0.085	0.081	0.457
9 anchors	0.089	0.083	0.487
9 anchors, 5% noise	0.098	0.096	0.476
3 off-slice observations	0.087	0.072	0.482
5 off-slice observations	0.083	0.041	0.542
9 off-slice observations	0.077	0.043	0.577

5.3 A FEW OBSERVATIONS REPAIR THE AMBIGUITY

Residual-based selector variants produce only small changes on complete problems and cannot recover an omitted condition. Adding information changes the outcome: Table 2 shows that one anchor leaves large outliers. Three to nine anchors reduce mean error substantially, including under noise and away from the initial slice, but some individual runs remain inaccurate.

The same pattern appears when the hidden reference comes from a numerical solver. In a reaction–diffusion case, a finite-difference trajectory is used only for evaluation. With the full public specification, the baseline proxy has $\rho = 0.982$ and mean returned $L_2 = 0.105$. Removing the dense public initial condition again returns the zero solution, with mean $L_2 = 1.000$ and $\rho = 0.035$. Sparse public anchors repair that delayed-reference case, reducing mean returned L_2 to 0.053 (median 0.016).

The two-dimensional heat example gives the same result: no-initial gives $L_2 = 1.000$ and $\rho = 0.043$, while 25 off-slice observations restore high rank alignment ($\rho = 0.977$) and lower mean returned L_2 to 0.292 (0.203 with held-out selection).

5.4 A NONZERO GAUGE AMBIGUITY PRODUCES A SECOND FAILURE FAMILY

The zero solution is not the only specification error that can defeat a public score. A one-dimensional Poisson–Neumann case has hidden target $u^\star(x) = \cos(2\pi x)$. The public operator is $-u_{xx} = (2\pi)^2 \cos(2\pi x)$ with zero normal derivative at both boundaries. These conditions identify $u^\star + c$ for any constant c, so an omitted additive gauge leaves an entire continuum of low-residual, nonzero wrong solutions.

For this case the ambiguity is exact, not a sampled-collocation artifact. For every constant c, $-(u^\star + c)_{xx} = (2\pi)^2 \cos(2\pi x)$ and $(u^\star + c)_x|_{x\in\{0,1\}} = 0$. Using the continuous $L_2([0,1])$ norm, the hidden relative error is exactly

$$\frac{\|(u^\star + c) - u^\star\|_2}{\|u^\star\|_2} = \sqrt{2}\,|c|, \tag{8}$$

although every member has the same exposed residual and boundary score. The identity in equation 8 shows that the gauge selects an otherwise indistinguishable member. It is not merely a numerical regularizer.

Across 16 starts, the public zero-mean gauge returns mean selected error 7.80×10^{-4}. Removing only that gauge raises the returned error to 0.373, while the best candidate encountered in the same runs has mean error 2.77×10^{-3}. Thus the public residual remains nearly equally rank-correlated with error ($\rho = 0.421$ without the gauge versus 0.404 with it), yet its final selected error is about $478\times$ larger. One fixed public value $u(1/4) = 0$ restores mean selected error to 1.00×10^{-3}.

Replications at two target frequencies preserve the ordering (Appendix B): omitting the gauge gives mean error 0.285–0.361, versus 5.24×10^{-4}–2.95×10^{-3} after either public repair, despite positive proxy–truth correlations as high as 0.466.

Table 3: Public-portfolio benchmark at budget 10, 5% representation mismatch, and 1% Gaussian noise. Results cover 16 PDE libraries and 32 seeds per task. Wins count task means below Uniform(10); catastrophic error is relative $L_2 > 1$. Holm-adjusted p values compare each design with the portfolio.

Design	Mean	Median	P90	Cat. (%)	Wins/16	Holm p
Predictive A	.05668	.05317	.06748	0	16	1.0000
A-optimal	.05676	.05317	.06801	0	16	1.0000
D-optimal	.05703	.05305	.06840	0	15	1.0000
Public portfolio	.05874	.05735	.06981	0	15	–
E-optimal	.05883	.05724	.06723	0	15	1.0000
Predictive E	.05902	.05735	.06991	0	15	1.0000
Pivoted QR	.06017	.06059	.06825	0	15	1.0000
Spectral logdet	.07009	.06332	.08881	0	16	.0295
Coverage guard	.08721	.08043	.12921	0	15	.00165
Maximin	7.27645	.27150	6.68442	18.75	2	.000366
Uniform	7.28051	.22816	6.68442	18.75	0	.000610
Random	17.91152	.58564	5.33312	22.66	4	.000366
Leverage	22701.87241	.77970	1216.25554	40.63	9	.00171

Table 4: Exact-null phase transition at 1% observation noise. Each row pools 1,800 paired trials (six families, 12 operator seeds, and 25 trials). With $k < d$, zero of 72 independent family–operator units is identifiable. Means are omitted because random, uniform, and E-optimal have extreme tails. Rank recovery alone does not control reconstruction error.

Design at $k = d$	Rank (%)	Median	P90	$L_2 > 1$ (%)
Random	94.89	.01043	.93353	9.00
Uniform	83.33	.01845	317.80492	16.56
Pivoted QR	100.00	.00809	.01549	0
A-optimal	100.00	.00820	.01579	0
D-optimal	100.00	.00804	.01533	0
E-optimal	100.00	.01143	83.58109	31.61
D–E hybrid	100.00	.00788	.01538	0

5.5 Rank and conditioning explain when repair transfers

Table 3 compares the public portfolio with all 12 equal-budget alternatives on 16 PDE libraries. The portfolio improves on uniform in 15/16 tasks, has no observed catastrophic failures, and meets the .01 non-inferiority margin to A-optimal. Its mean error, .05874, lies in the tight cluster formed by Predictive A, A/D/E-optimal, Predictive E, and pivoted QR (.05668–.06017); their pairwise differences from the portfolio are not significant under the reported test. In contrast, unconditioned designs have large tails.

Mechanism tests. Table 4 confirms the predicted transition. With $k < d$, none of 72 family–operator units is identifiable. At $k = d$, QR/A/D/D–E recover rank in every trial with P90 .0153–.0158, whereas random, uniform, and E-optimal retain substantial failure rates. A same-budget causal ablation gives mean error .00961 for spanning QR and .61173 for wrong-subspace QR. Increasing pilot rows from d to $16d$ reduces subspace error and reconstruction error from .05490 to .01359; full results are in Appendix C.

5.6 Independent and third-party solver replication

Residual-adaptive PINN. Across 2,016 runs, a separate hyperbolic tangent (tanh) PINN with fixed or R3 sampling at widths 16/32 reduces mean error from .99705–1.00187 to .06642–.06663 with eight public values. All four $84/84$ paired comparisons favor repair, consistent with Corollary 1. Budget and three-value checks appear in Appendix C.

Standard PDEs in DeepXDE. An unmodified DeepXDE 1.15.0 library (Lu et al., 2021) with our adapter adds 72 runs on standard heat, Burgers, and Allen–Cahn equations (Table 5). Matched arms

Table 5: **DeepXDE replication.** Relative L_2 mean $\pm$ sample standard deviation over eight matched seeds per PDE. Wins compare Initial(8) with no initial information. All eight Allen–Cahn repair runs still exceed error 0.5.

Standard PDE	Full (160)	No initial	Initial(8)	Wins/8
Heat	$.0175 \pm .0153$	$1.0002 \pm .0022$	$.0235 \pm .0111$	8
Burgers	$.2293 \pm .1016$	$.9996 \pm .0022$	$.2624 \pm .0597$	8
Allen–Cahn	$.8440 \pm .0484$	$1.4502 \pm .0016$	$.7741 \pm .0751$	8

share initialization, collocation, and 8,000 Adam steps. References are opened after all three models are trained. Initial(8) improves all 24 seed pairs, but Allen–Cahn remains inaccurate even with full initial data. The experiment therefore separates successful information repair from the remaining optimization error (Appendix C).

6 Related Work

Proxy optimization and scientific residuals. Goodhart work establishes degradation under proxy pressure (Amodei et al., 2016; Skalse et al., 2022; Gao et al., 2023; Laidlaw et al., 2025). PINNs can converge to trivial solutions, and sampled residual can decouple from global error (Leiteritz & Pflüger, 2021; Krishnapriyan et al., 2021; Rahaman et al., 2019; Wang et al., 2024). Remedies strengthen training, verification, or sampling (Liu et al., 2023; Eiras et al., 2024; Daw et al., 2023). These methods address how well an exposed condition is optimized or checked. Our matched omission–repair test instead asks whether the exposed conditions determine the target at all.

Error control and experimental design. Our specification test complements discretization-specific estimators (Becker & Rannacher, 2001; Hillebrecht & Unger, 2022) and established observability, optimal design, and sparse sensing (Cohn & Dee, 1988; Allen-Zhu et al., 2017; Alexanderian et al., 2014; 2016; Manohar et al., 2018). We target the public operator's ambiguity space and connect its rank and conditioning directly to failure in proxy-driven PDE search. The portfolio is a practical way to choose among established design criteria without using target values or validation error.

7 Discussion, Limitations, and Conclusion

Practical interpretation. First check whether the exposed equations and conditions distinguish the target, then whether observations span the remaining ambiguity with adequate conditioning. Persistent error can then be examined through representation, fitting, sampling, and noise. In the Poisson–Neumann example, more residual queries leave the additive constant invisible, whereas one informative value fixes it. DeepXDE shows the complementary limit: public values improve predictions without ensuring accurate optimization. Reports should therefore pair observation count with estimated ambiguity dimension, conditioning, and returned error, especially under noise or library mismatch.

Limitations. The external study covers three standard PDEs and no laboratory data. Allen–Cahn remains inaccurate even with full initial data. The nonlinear theorem is local: full rank in a finite library does not establish uniqueness in a richer class, and an inaccurate pilot can miss ambiguity directions. Portfolio non-inferiority is specific to the tested tasks, noise, observation counts, and absolute margin; it does not imply universal superiority over its constituent designs. Application-dependent acquisition costs may exceed the number of queried values.

Conclusion. Sparse public observations repair the tested zero and gauge ambiguities. Rank and conditioning explain when that repair is possible and stable. The public portfolio turns this analysis into a target-independent observation policy that matches strong conditioned designs while avoiding the large tails of simple placement rules. The broader lesson is straightforward: before improving an optimizer or residual sampler, check whether the public problem identifies the solution being sought.

Reproducibility Statement

The accompanying implementation provides case definitions, selection rules, experiment settings, and result aggregation. Search and reference-based evaluation are separated; complete outcomes are retained for each experimental design. The PINN/R3 and DeepXDE summaries are computed from the recorded training runs. Proofs are provided in Appendix A. Experiment-specific settings, computational costs, and inferential units appear in the corresponding appendices, including the external solver protocol in Appendix C.

References

Alen Alexanderian, Noemi Petra, Georg Stadler, and Omar Ghattas. A-optimal design of experiments for infinite-dimensional bayesian linear inverse problems with regularized ℓ_0-sparsification. *SIAM Journal on Scientific Computing*, 36(5):A2122–A2148, 2014. doi: 10.1137/130933381.

Alen Alexanderian, Noemi Petra, Georg Stadler, and Omar Ghattas. A fast and scalable method for A-optimal design of experiments for infinite-dimensional bayesian nonlinear inverse problems. *SIAM Journal on Scientific Computing*, 38(1):A243–A272, 2016. doi: 10.1137/140992564.

Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal design of experiments via regret minimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 126–135. PMLR, 2017. URL https://proceedings.mlr.press/v70/allen-zhu17e.html.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. URL https://arxiv.org/abs/1606.06565.

Roland Becker and Rolf Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 10:1–102, 2001. doi: 10.1017/S0962492901000010.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113.

Stephen E. Cohn and Dick P. Dee. Observability of discretized partial differential equations. *SIAM Journal on Numerical Analysis*, 25(3):586–617, 1988. doi: 10.1137/0725037.

Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv preprint arXiv:2305.01582*, 2023. URL https://arxiv.org/abs/2305.01582.

Arka Daw, Jie Bu, Sifan Wang, Paris Perdikaris, and Anuj Karpatne. Mitigating propagation failures in physics-informed neural networks using retain-resample-release (R3) sampling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7264–7302. PMLR, 2023. URL https://proceedings.mlr.press/v202/daw23a.html.

Francisco Eiras, Adel Bibi, Rudy R. Bunel, Krishnamurthy Dj Dvijotham, Philip Torr, and M. Pawan Kumar. Efficient error certification for physics-informed neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12318–12347. PMLR, 2024. URL https://proceedings.mlr.press/v235/eiras24a.html.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Birgit Hillebrecht and Benjamin Unger. Certified machine learning: A posteriori error estimation for physics-informed neural networks. *arXiv preprint arXiv:2203.17055*, 2022. URL https://arxiv.org/abs/2203.17055.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of AI ingenuity. Google DeepMind Blog, 2020. URL https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.

Aditi S. Krishnapriyan, Amir Gholami, Shandian Zhe, Robert M. Kirby, and Michael W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 26548–26560, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/df438e5206f31600e6ae4af72f2725f1-Abstract.html.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=msEr27EejF.

Raphael Leiteritz and Dirk Pflüger. How to avoid trivial solutions in physics-informed neural networks. *arXiv preprint arXiv:2112.05620*, 2021. URL https://arxiv.org/abs/2112.05620.

Shuheng Liu, Xiyue Huang, and Pavlos Protopapas. Residual-based error bound for physics-informed neural networks. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 1284–1293. PMLR, 2023. URL https://proceedings.mlr.press/v216/liu23b.html.

Lu Lu, Xuhui Meng, Zhiping Mao, and George Em Karniadakis. DeepXDE: A deep learning library for solving differential equations. *SIAM Review*, 63(1):208–228, 2021. doi: 10.1137/19M1274067.

David Manheim and Scott Garrabrant. Categorizing variants of Goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018. URL https://arxiv.org/abs/1803.04585.

Krithika Manohar, Bingni W. Brunton, J. Nathan Kutz, and Steven L. Brunton. Data-driven sparse sensor placement for reconstruction: Demonstrating the benefits of exploiting known patterns. *IEEE Control Systems Magazine*, 38(3):63–86, 2018. doi: 10.1109/MCS.2018.2810460.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025. URL https://arxiv.org/abs/2506.13131.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=JYtwGwIL7ye.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 2019. URL https://proceedings.mlr.press/v97/rahaman19a.html.

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. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024. doi: 10.1038/s41586-023-06924-6.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K. Reddy. LLM-SR: Scientific equation discovery via programming with large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=m2nmp8P5in.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9460–9471, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d719fee332caa23d5038b8a90e81796-Abstract-Conference.html.

Silviu-Marian Udrescu and Max Tegmark. AI feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020. doi: 10.1126/sciadv.aay2631.

Tao Wang, Bo Zhao, Sicun Gao, and Rose Yu. Understanding the difficulty of solving Cauchy problems with PINNs. In *Proceedings of the 6th Annual Learning for Dynamics and Control Conference*, volume 242 of *Proceedings of Machine Learning Research*, pp. 453–465. PMLR, 2024. URL https://proceedings.mlr.press/v242/wang24b.html.

Zongmin Yu and Liu Yang. Agentic symbolic search: Characterizing PDEs beyond hand-crafted expressions, meshes, and neural networks. *arXiv preprint arXiv:2606.20467*, 2026. URL https://arxiv.org/abs/2606.20467.

Artificial Intelligence (AI) Use Statement

OpenAI Codex was used for coding assistance and language editing. The authors reviewed all AI-assisted content and take full responsibility for the manuscript, implementation, and reported results.

A Proofs and Scope of the Guarantees

Theorem 5 (Public-transcript indistinguishability)**.** *Let u_0, u_1 be admissible targets in a normed prediction space with the same complete public transcript $P(u_0) = P(u_1)$. Any deterministic algorithm whose output depends only on P returns a common $\widehat{u}$ and satisfies*

$$\max_{j\in\{0,1\}} \|\widehat{u} - u_j\| \geq \tfrac{1}{2}\|u_0 - u_1\|. \tag{9}$$

The same bound holds for the larger expected error of a randomized public-only algorithm.

To prove equation 9, note that $P(u_0) = P(u_1)$, so a deterministic public-only algorithm returns the same $\widehat{u}$ under both targets. The triangle inequality gives

$$\|u_0 - u_1\| \leq \|u_0 - \widehat{u}\| + \|\widehat{u} - u_1\|, \tag{10}$$

so equation 10 implies equation 9. For a randomized algorithm, couple the two tasks with the same internal random seed. The output distribution is identical because the transcript is identical. Take expectations in the triangle inequality and apply the same argument.

Theorem 6 (Adaptive public-query lower bound)**.** *Let V be a d-dimensional linear ambiguity subspace of a normed prediction space. Consider any deterministic policy that, after observing previous answers, adaptively chooses at most $k < d$ continuous linear public queries and then returns an estimate. For every $R > 0$, there exist two targets $u_+, u_- \in V$ with $\|u_+ - u_-\| = 2R$ that generate the same adaptive transcript, and the policy has error at least R on one of them.*

Run the policy on the all-zero transcript, and denote its resulting queries by $\ell_1, \ldots, \ell_k$. Rank–nullity gives a nonzero $v \in V \cap \bigcap_{i=1}^{k} \ker \ell_i$. Rescale it to $\|v\| = R$ and set $u_+ = v$, $u_- = -v$. Inductively, both targets answer zero to every adaptively chosen query, so the policy returns the same $\widehat{u}$ for both. The triangle inequality gives $2R = \|u_+ - u_-\| \leq \|u_+ - \widehat{u}\| + \|\widehat{u} - u_-\|$, proving the claim. The theorem concerns deterministic query policies. Randomized policies require an explicit minimax distribution and are not claimed here.

Proof of Corollary 1. Couple the two executions with the same initialization, optimizer randomness, and resampling randomness. At round zero their public transcripts agree by assumption. If the transcripts agree through round r, the coupled algorithm chooses the same next collocation distribution and the same parameter update. The residual response at every sampled coordinate is again identical, so the transcripts agree through round $r + 1$. Induction gives identical output distributions. For

each coupled output, the triangle inequality yields $\|u_0 - u_1\| \le \|u_0 - \widehat{u}\| + \|\widehat{u} - u_1\|$. Taking expectations proves the bound. When $H(u_0 - u_1) \neq 0$, the two public transcripts differ as soon as the declared measurement is queried, so the coupling premise fails. This does not guarantee successful optimization. It states precisely what new information removes the impossibility argument.

Proof of Theorem 2. Decompose $e = Nc + e_\perp$, where $e_\perp \perp \ker A$. The Moore–Penrose inverse gives $\|e_\perp\|_2 \le \|A^\dagger\|_2 \|Ae\|_2$. Moreover, $HNc = He - He_\perp$. If HN has full column rank,

$$\|c\|_2 \le \sigma_{\min}(HN)^{-1}\{\|He\|_2 + \|H\|_2\|e_\perp\|_2\}, \tag{11}$$

and orthonormality of N turns equation 11 into the stated bound. Conversely, if $\operatorname{rank}(HN) < d$, a nonzero c exists with $HNc = 0$. Then Nc is invisible to both A and H, so identification is impossible.

Proof of Theorem 3. By the orthogonal Procrustes characterization of principal angles, an orthogonal matrix Q exists such that $\|N - \widehat{N}Q\|_2 = 2\sin(\theta_{\max}/2) = \eta(\delta)$, where $\sin\theta_{\max} = \delta$. Right multiplication by Q preserves singular values. Hence

$$\|HN - H\widehat{N}Q\|_2 \le \|H\|_2 \eta(\delta). \tag{12}$$

Combining equation 12 with Weyl's perturbation inequality gives $\sigma_{\min}(HN) \ge \sigma_{\min}(H\widehat{N}Q) - \|H\|_2\eta(\delta) = \gamma - \|H\|_2\eta(\delta) > 0$. Substitution into Theorem 2 proves the claimed stability bound.

Theorem 7 (Local nonlinear specification stability)**.** *Suppose $F : \mathbb{R}^p \to \mathbb{R}^m$ is continuously differentiable, $F(\theta^\star) = 0$, and DF is L-Lipschitz on the radius-r ball $B_r(\theta^\star)$. Let H be a linear observation operator. Let $J = DF(\theta^\star)$ have nullity $d > 0$, and let N be an orthonormal basis for $\ker J$, set $\alpha = \sigma^+_{\min}(J) > 0$ and $\gamma = \sigma_{\min}(HN) > 0$, and define $C_J = \alpha^{-1}(1 + \|H\|_2/\gamma)$ and $C_H = \gamma^{-1}$. For $e = \theta - \theta^\star$ with $\|e\|_2 \le \min\{r, (C_J L)^{-1}\}$ (the second radius is infinite if $L = 0$),*

$$\|e\|_2 \le 2C_J\|F(\theta) - F(\theta^\star)\|_2 + 2C_H\|He\|_2. \tag{13}$$

Conversely, an observation operator with $k < d = \dim\ker J$ rows cannot provide first-order identification, since its product with N lacks full rank.

Write $e = Nc + e_\perp$ with $e_\perp \perp \ker J$. Because the smallest positive singular value of J is α, $\|e_\perp\|_2 \le \alpha^{-1}\|Je\|_2$. Moreover, $HNc = He - He_\perp$ and $\sigma_{\min}(HN) = \gamma$, so

$$\|e\|_2 \le C_J\|Je\|_2 + C_H\|He\|_2. \tag{14}$$

The Lipschitz Jacobian assumption and the integral form of Taylor's theorem give $\|F(\theta) - F(\theta^\star) - Je\|_2 \le (L/2)\|e\|_2^2$. Hence

$$\|e\|_2 \le C_J\|F(\theta) - F(\theta^\star)\|_2 + C_H\|He\|_2 + (C_J L/2)\|e\|_2^2. \tag{15}$$

Under the stated radius condition, the last term is at most $\|e\|_2/2$ and can be absorbed into the left side, yielding equation 13. If $k < d$, rank–nullity gives $\operatorname{rank}(HN) \le k < d$, so a nonzero first-order invisible direction remains.

Proof of Proposition 4. If $d = 0$, Weyl's inequality gives $\sigma_p(\widehat{J}) \ge g - \|E\|_2 > \tau$, so no direction is declared weak. If $d > 0$, Weyl gives $\sigma_{p-d}(\widehat{J}) \ge g - \|E\|_2 > \tau$ and $\sigma_{p-d+1}(\widehat{J}) \le \|E\|_2 \le \tau$. Hence exactly d singular values lie at or below the threshold. For $d > 0$, Wedin's sin–theta theorem applied across this gap yields $\|\sin\Theta(N, \widehat{N})\|_2 \le \|E\|_2/(g - \|E\|_2)$.

B Observation Design, Calibration, and Robustness

Anchor-design benchmark. The design universe contains six fixed near-null families, sensor counts d and $d + 2$, noise fractions 0, 1%, and 5%, three near-null strengths, ten seed clusters, and 100 trials per cell. Coefficients and standard-normal noise are paired across all methods. The primary comparison is full-field relative error at $k = d$ and 1% noise. Table 6 shows that full rank alone is insufficient: E-optimal recovers rank but retains large tail errors, while QR, A-, and D-optimal remain stable.

Runs use one central processing unit (CPU) thread. Each design fixes its k locations before requesting values; dense trajectories are accessed only after prediction. The ten seed clusters, rather than the 100 repeated trials within each cell, provide the independent replication units for this benchmark.

Table 6: Controlled primary endpoint aggregated across 18,000 paired trials per design. Heavy-tailed failures remain visible in P90 even when the median is small.

Design	Median L_2	P90	Median $\sigma_{\min}$	Identifiable (%)	Nullity (%)
Random	.0110	.9116	.0243	95.9	100.0
Uniform	.0176	319.8	1.60e−5	83.3	100.0
Leverage	.0109	533.8	.0902	83.3	100.0
Pivoted QR	.0076	.0156	.1107	100.0	100.0
A-optimal	.0076	.0157	.1124	100.0	100.0
D-optimal	.0076	.0155	.1107	100.0	100.0
E-optimal	.0112	89.4596	.0752	100.0	100.0
D–E hybrid	.0077	.0155	.1114	100.0	100.0

Stable-gap rank estimation. The rank-estimation benchmark covers seven cases, two solvers, six eight-observation designs, and 24 seeds, for 2,016 returned candidates. The final estimator stacks three normalized public pilot Jacobians and selects the largest admissible trailing spectral gap. It returns a one-dimensional weak space throughout this benchmark, consistent with the constructed ambiguity. Guarded refinement improves 13 of 14 case–solver means; on Graveleau with symbolic search, mean error increases by .0140, while the paired random-feature result improves by .00750. This exception separates a gain in most case–solver means from a uniform improvement.

Component, budget, and noise ablations. The ablation matrix crosses seven cases, two solvers, 27 configurations, and seeds 40–63, producing 9,072 returned candidates. Transfer experiments add finite-difference reaction–diffusion and two-dimensional heat cases. Each registered configuration contributes one returned prediction per case, solver, and seed. Tables 8 and 9 show that the clean gain concentrates at smaller budgets and does not persist uniformly under noise or transfer. Table 7 isolates the fill constraint: removing it worsens mean error, while removing only the gain gate improves the mean but worsens the tail. Here and in the following symbolic-search tables, catastrophic error (Cat.) means $L_2 > 0.2$. The paired Uniform/full-guard rates are denoted U/F. Differences Δ subtract the uniform mean; CI denotes confidence interval. "Accept" reports how often the modified design is accepted instead of uniform placement.

Table 7: Component ablation (seeds 40–63). Each row contains 336 returns with eight public observations. Removing the fill guard reverses the gain. Removing only the gain gate lowers the mean but increases the catastrophic rate.

Arm	Mean	Median	P90	$L_2>0.2$ (%)	Δ [95% CI]	Accept (%)
Uniform(8)	.0542	.0467	.1058	1.49	0	0.0
Full guard	**.0490**	.0377	.0943	1.49	$-.00523\ [-.00852, -.00252]$	41.7
No fill guard	.0812	.0622	.1671	5.06	$+.02695\ [+.02077, +.03380]$	91.7
No gain gate	.0479	.0308	.0896	2.08	$-.00628\ [-.00935, -.00322]$	86.3
No guards	.0812	.0622	.1671	5.06	$+.02695\ [+.02085, +.03379]$	92.3
Conditioning only	.2182	.1272	.4443	41.37	$+.16396\ [+.13278, +.19979]$	0.0

Table 8: Budget ablation on the seven-case main tier. Improvement is largest below the eight-observation operating point and saturates at twelve observations.

Budget	Uniform	Full guard	Relative gain	Δ 95% CI	Cat. U/F (%)
$k = 6$	.33410	.21071	36.93%	$[-.14371, -.10485]$	50.60/40.77
$k = 8$	.05422	.04899	9.64%	$[-.00852, -.00252]$	1.49/1.49
$k = 12$	.04874	.04907	−0.68%	$[-.00485, +.00473]$	1.79/1.19

Search pressure, equal-cost comparisons, and external tasks. Seeds 64–87 are used for search-pressure and equal-cost comparisons; external seeds 100–123 cover eight additional PDE solution libraries. Table 10 tests the gauge intervention at two target frequencies. Table 11 shows that extra generations cannot repair omission, while Table 12 compares placements at equal observation count and loss weight. The external comparisons in Tables 13 and 14 show stable conditioned reconstructions across noise models. Catastrophic error is $L_2 > 0.2$ in these comparisons.

Table 9: Noise and transfer ablations. Only the clean controlled comparison has a confidence interval strictly below zero; the remaining conditions are statistically unresolved.

Condition	Uniform	Full guard	Δ 95% CI	Cat. U/F (%)
Clean main	.05422	.04899	$[-.00852, -.00252]$	1.49/1.49
1% Gaussian	.05632	.05608	$[-.00365, +.00504]$	1.79/2.38
5% Gaussian	.07991	.07683	$[-.00700, +.00084]$	2.08/1.79
10% contamination	.08293	.08216	$[-.00523, +.00499]$	3.87/4.46
Numerical/two-dimensional clean	.23430	.23417	$[-.00040, .00000]$	70.83/70.83

Table 10: Gauge replication. Each cell is candidate-pool Spearman ρ followed by mean selected L_2 over 16 starts. Positive rank correlation does not identify the additive offset.

Replication and target	Zero-mean gauge	No additive gauge	One public point
Original, $\cos(2\pi x)$	$0.404,\ 7.80\text{e}{-4}$	$0.421,\ 0.373$	$0.390,\ 1.00\text{e}{-3}$
Repeat, $\cos(2\pi x)$	$0.412,\ 5.24\text{e}{-4}$	$0.404,\ 0.285$	$0.411,\ 6.49\text{e}{-4}$
Frequency shift, $\cos(3\pi x)$	$0.322,\ 2.23\text{e}{-3}$	$0.466,\ 0.361$	$0.323,\ 2.95\text{e}{-3}$

Table 11: Search-pressure comparison over 2,688 runs. Increasing generations improves the specified and repaired arms but never dislodges the perfect wrong optimum admitted by the missing specification.

Specification	Gen.	Mean	Median	P90	$L_2>0.2$ (%)
Default	3	.15612	.10180	.29614	22.62
Default	6	.14673	.08910	.34551	13.69
Default	12	.12758	.08493	.21389	11.31
Default	24	.09338	.08665	.18246	5.36
No initial condition	3	1.00000	1.00000	1.00000	100.00
No initial condition	6	1.00000	1.00000	1.00000	100.00
No initial condition	12	1.00000	1.00000	1.00000	100.00
No initial condition	24	1.00000	1.00000	1.00000	100.00
Eight anchors, weight 2	3	.14379	.12007	.23846	14.88
Eight anchors, weight 2	6	.13404	.11418	.18639	8.33
Eight anchors, weight 2	12	.12118	.11355	.18213	7.74
Eight anchors, weight 2	24	.08881	.09821	.14819	2.38
Eight anchors, weight 20	3	.11372	.09212	.21555	11.90
Eight anchors, weight 20	6	.08969	.08662	.13349	2.98
Eight anchors, weight 20	12	.08811	.08713	.13773	4.76
Eight anchors, weight 20	24	.07295	.08223	.13559	.60

Table 12: Equal-budget, equal-weight comparison over 2,352 runs. Differences and seed-cluster intervals use Uniform(8) as the comparator.

Arm	Mean	Median	P90	Cat. (%)	Δ	95% CI	Wins/14
Fixed initial(8)	.10179	.10863	.15534	3.87	$-.01191$	$[-.03157, +.00504]$	8
Random(8)	.11858	.09340	.25302	16.96	$+.00488$	$[-.01380, +.02374]$	5
Uniform(8)	.11370	.08686	.22122	14.29	0	$[0, 0]$	0
Maximin(8)	.12757	.09052	.30237	20.24	$+.01387$	$[-.00424, +.03299]$	3
Condition only(8)	.27117	.20708	.55237	50.60	$+.15747$	$[+.12295, +.19337]$	2
Coverage D(8)	.13343	.11442	.29762	19.05	$+.01973$	$[+.00261, +.03779]$	3
Guarded hybrid(8)	.11626	.09364	.21887	13.10	$+.00256$	$[-.00088, +.00581]$	9

Public-portfolio evaluation. Selection uses only coordinates, the declared eight-column public library, and a public reconstruction grid. Coefficients, omitted components, noise, targets, and errors remain hidden. The score and candidate designs are defined in equation 6. Seed errors are averaged within task before a 50,000-replicate paired task bootstrap and an exact task-level sign-flip test. At the primary point (budget 10, 5% mismatch, 1% Gaussian noise), superiority to uniform requires a negative upper confidence endpoint, 12/16 task wins, and no higher catastrophic-error rate. Non-inferiority to A-optimal uses an absolute relative-L_2 margin of .01. The primary comparison appears in Table 3; Table 15 varies budget and noise. Across 60 registered observation-count, mismatch, and

Table 13: External-PDE comparison under 1% Gaussian noise. Wins are task-family mean improvements over Uniform(10). The leverage mean is shown without clipping.

External design, $k = 10$	Mean	Median	P90	Cat. (%)	Wins/8
Random	2.22377	.16502	1.35484	43.75	4
Uniform	2.44128	.27250	4.62058	58.33	0
Maximin	2.40900	.23173	4.62058	53.65	1
Leverage	37817.97335	.56908	120655.16831	62.50	3
Pivoted QR	.06674	.06721	.08258	0	7
A-optimal	.06159	.06203	.06940	0	8
D-optimal	.06171	.05988	.07344	0	7
E-optimal	.06691	.06683	.08414	0	7
Coverage D	.11100	.08975	.18543	8.85	5
Full guarded	.07214	.06827	.11017	0	6
No fill guard	.05921	.05801	.06930	0	7

Table 14: External noise sensitivity for uniform placement, the full guard, and its fill-guard ablation. Conditioned designs remain stable across noise models; removing the fill guard gives the lowest mean within each noise condition.

Noise	Design	Mean	P90	Cat. (%)
None	Uniform	.20544	.64153	37.50
None	Full guarded	.07158	.11155	0
None	No fill guard	.05893	.06978	0
1% Gaussian	Uniform	2.44128	4.62058	58.33
1% Gaussian	Full guarded	.07214	.11017	0
1% Gaussian	No fill guard	.05921	.06930	0
5% Gaussian	Uniform	11.57060	23.47767	70.31
5% Gaussian	Full guarded	.08175	.10599	0
5% Gaussian	No fill guard	.06675	.08241	0
Heteroscedastic	Uniform	5.86825	10.71213	59.38
Heteroscedastic	Full guarded	.07300	.10946	0
Heteroscedastic	No fill guard	.05966	.06928	0
5% contamination	Uniform	6.72993	9.03017	61.98
5% contamination	Full guarded	.07652	.11256	0
5% contamination	No fill guard	.06339	.07593	0

noise settings, 58 bootstrap intervals for portfolio-minus-uniform mean error lie below zero. The other two settings give noiseless exact-recovery ties to numerical precision. In the table, P/U denotes portfolio/uniform catastrophic rates.

Table 15: Budget and noise ablations at 5% mismatch. The $k = 10$, 1% Gaussian row is the primary condition. Across all 60 registered settings, 58 task-bootstrap intervals for the portfolio-minus-uniform mean error are strictly below zero. The two remaining cells are noiseless, zero-mismatch exact-recovery ties at approximately 10^{-15}.

Condition	Portfolio	A-opt.	Uniform	Δ vs. U. [95% CI]	Wins	Cat. P/U (%)
$k = 8$, 1% Gaussian	.06161	.06063	.52491	$[-.68487, -.27302]$	16	0/18.75
$k = 10$, 1% Gaussian	.05874	.05676	7.28051	$[-20.57593, -.21433]$	15	0/18.75
$k = 12$, 1% Gaussian	.05713	.05627	.14592	$[-.15048, -.04480]$	16	0/0
$k = 16$, 1% Gaussian	.05666	.05571	.11453	$[-.08193, -.03567]$	15	0/0
$k = 10$, no noise	.05828	.05640	7.49603	$[-21.21704, -.20983]$	15	0/18.75
$k = 10$, 5% Gaussian	.06783	.06568	9.20938	$[-25.80796, -.36637]$	16	0/25.00
$k = 10$, heteroscedastic	.05933	.05726	7.36897	$[-20.76671, -.24145]$	16	0/19.14
$k = 10$, contamination	.06186	.05980	8.50479	$[-23.99359, -.29011]$	16	0/21.68

Mechanism, budget, and external benchmarks. The combined benchmark contains 530,472 runs. The phase-transition experiment crosses six ambiguity families, eight designs, seeds 200–211, 25 paired trials, three near-null strengths, three noise levels, and budgets $\max(0, d-1), d, d+2, 2d$ (518,400 rows). The mechanism tier crosses seven cases, seeds 160–171, no/3/5/8/full initial information and an eight-point wrong-slice control under three search-pressure regimes (1,512

rows). The external tier crosses eight PDE libraries, seeds 124–147, five noise models, and eleven equal-budget designs (10,560 rows).

The inferential units are 72 family–operator seeds, 21 case–regime endpoints, and 40 task–noise endpoints. Contrasts use cluster bootstrap intervals and two-sided sign-flip tests; all eight stochastic mechanism contrasts remain significant after Holm correction (maximum adjusted $p = 7.99992 \times 10^{-5}$). The rank transition is summarized by exact counts and pooled tail rates in Table 4. Table 16 tests information quantity and observation time, showing that the repair is not specific to eight initial-slice values. Table 17 compares all external designs across the five noise laws without discarding unstable baselines.

Table 16: Mechanism and budget ablation. Entries are mean returned L_2 over seven cases and 12 held-out seeds. The last column is the largest regime-specific P90. Initial(8) beats no initial information in 7/7 cases in each regime ($\Delta = -.93115$, 95% CI $[-.95657, -.90627]$, $p = 1.99998 \times 10^{-5}$). Initial(3) is already sufficient on average, and observations away from the initial slice often match Initial(8), limiting slice-specific claims. Feature-1 and Feature-3 denote the two random-feature regimes; Symbolic denotes evolutionary program search.

Public specification	Symbolic	Feature-1	Feature-3	Worst P90
No initial information	1.00000	1.03041	1.04221	1.31190
Initial 3	.09471	.08145	.08108	.15546
Initial 5	.10013	.08627	.08611	.14750
Initial 8	.10398	.08760	.08758	.14793
Wrong-slice 8	.08999	.08021	.07997	.17960
Full initial condition	.10559	.08146	.08110	.18088

Table 17: External equal-budget comparison pooled across all eight PDE tasks, five noise laws, and 24 seeds. Wins count task–noise means below Uniform(10). A-optimal versus Uniform has $\Delta = -6.59922$ (95% CI $[-13.92035, -1.36296]$). G-optimal has -6.59854 ($[-13.91922, -1.36269]$). The no-fill guard has the lowest mean.

Equal-budget design	Mean	Median	P90	$L_2 > 1$ (%)	Wins/40
Random	14.05646	.19148	2.47282	18.12	13
Uniform	6.66268	.25279	9.37917	18.85	0
Maximin	6.67884	.25769	9.37917	20.42	5
Leverage	93734.4	.66711	2742.17	49.48	21
Pivoted QR	.06885	.06830	.08382	0	34
A-optimal	.06345	.06375	.07207	0	39
D-optimal	.06363	.06071	.07428	0	34
E-optimal	.06896	.06754	.08474	0	37
Coverage D	.13679	.09017	.21409	.73	31
No-fill guard	.06141	.05927	.07121	0	35
G-optimal	.06413	.06161	.07334	0	38

C Mechanism Tests and Independent Solvers

Spanning, subspace estimation, and solver transfer. This benchmark uses seeds 300–323 and contains 373,824 runs: 151,200 same-budget spanning tests, 172,800 finite-sample subspace tests, 2,304 random-feature solver runs, and 47,520 finite-difference runs. Table 18 tests spanning against equally sized but uninformative measurements. Table 19 varies pilot information and reports both subspace angle and reconstruction error. Tables 20 and 21 test transfer to independent solvers. Together they separate causal support for spanning from the the scope of solver transfer supported by the available task clusters.

The spanning and random-feature contrasts remain significant after correction over all 13 comparisons. The finite-difference designs improve all four tasks, but four task clusters give an exact two-sided sign-flip value of .125. These results support transfer across independent discretizations without establishing prevalence over a larger population of PDEs.

PINN and residual resampling. Every row below is produced by a boundary-conforming tanh PINN with analytic derivatives. R3 changes only public residual collocation points. Dense references

Table 18: Causal spanning ablation at 1% noise, pooled over six ambiguity dimensions, 24 held-out seeds, and 50 paired trials per cell (7,200 rows per arm). Against the wrong-subspace control, spanning QR has cluster-mean $\Delta = -.60211$, 95% CI $[-.64681, -.55507]$, 133/144 wins, and sign-flip $p < 10^{-5}$.

Same-budget arm	Mean L_2	Median	P90	$L_2 > 1$ (%)
Spanning QR	.00961	.00913	.01560	0
Anti-span	.11645	.02973	.24417	1.71
Uniform	.12310	.05627	.32309	.54
Random	.28187	.04805	.35163	1.43
Under-budget QR ($d-1$)	.41170	.27148	1.00000	.26
Rank deficient	.61173	.71859	.98522	.69
Wrong-subspace QR	.61173	.71859	.98522	.69

Table 19: Finite-sample estimated-subspace experiment at pilot noise .05 for the estimated-QR arm (2,880 trials per ratio). "Mean sine" is the mean sine of the largest principal angle between the true and estimated subspaces. Moving from d to $16d$ pilot rows gives $\Delta = -.04131$, 95% CI $[-.04584, -.03735]$, with 142/144 dimension–seed wins and sign-flip $p < 10^{-5}$.

Pilot rows	Mean sine	Mean error	P90 error	Valid (%)	Coverage (%)
d	.09325	.05490	.08402	100	100
$2d$	.04933	.03365	.05408	100	100
$4d$	.02997	.02229	.03168	100	100
$8d$	.01998	.01664	.02279	100	100
$16d$	.01376	.01359	.01818	100	100

Table 20: Independent random-feature collocation solver, pooled over eight PDE cases, three feature widths, and 24 held-out seeds (576 rows per specification). Initial(8) beats no initial information in all 24 case–width clusters: $\Delta = -.91345$, 95% CI $[-.93809, -.88424]$, sign-flip $p < 10^{-5}$. The full condition remains descriptively stronger, so sparse anchors are a repair rather than an information-free optimum.

Public specification	Mean L_2	Median	P90	$L_2 > 1$ (%)	Opt. success (%)
No initial information	1.00000	1.00000	1.00000	0	100
Initial(3)	.23033	.07680	.75751	.17	100
Initial(8)	.08655	.07030	.23969	0	100
Full initial condition	.06047	.05326	.11001	0	100

Table 21: Independent finite-difference target benchmark at budget 10, mismatch .05, and Gaussian noise .01 (96 rows per design). Wins count PDE-task means below Uniform. All seven conditioned designs improve all four tasks. Their task-cluster bootstrap intervals against Uniform exclude zero. With only four independent task clusters, however, the exact two-sided sign-flip value is .125 for every comparison. These are consistent directional effects, not a significance or unique-dominance claim.

Equal-budget design	Mean L_2	Median	P90	$L_2 > 1$ (%)	Wins/4
Random	9.84163	.14573	2.34749	20.83	0
Uniform	.29851	.25651	.59896	0	–
Maximin	.29851	.25651	.59896	0	0
Leverage	272.198	27.8191	1122.09	98.96	0
Pivoted QR	.03723	.02931	.07238	0	4
A-optimal	.03505	.02674	.07247	0	4
D-optimal	.03133	.02104	.07115	0	4
E-optimal	.03923	.03539	.06730	0	4
D–E hybrid	.03666	.03018	.06699	0	4
No-fill guard	.03029	.02029	.06734	0	4
G-optimal	.03352	.02380	.06997	0	4

are opened after training. The main and ablation matrices use disjoint seeds 300–311 and 312–317. Each job is single-threaded CPU-only, uses at most 2 gigabytes (GB), and runs 160 optimizer

iterations. Fixed and R3 arms share width, public specification, and collocation budget. Table 22 shows that residual resampling leaves the information gap intact at both widths. Catastrophic error here means $L_2 > 1$. Collocation budgets 96/192/384 and an Initial(3) ablation preserve the same

Table 22: Independent PINN comparison over seven cases and 12 held-out seeds. Entries are mean returned relative L_2 errors, except the catastrophic rate. R3 is a residual-only adaptive sampler, not an information repair.

Sampler	Width	Full initial	No initial	Initial(8)	Cat. Initial(8) (%)
Fixed	16	0.06623	0.99705	0.06642	0.00
Fixed	32	0.06694	1.00187	0.06649	0.00
R3	16	0.06722	0.99750	0.06663	0.00
R3	32	0.06743	1.00071	0.06656	0.00

omission–repair gap, showing that the result is not tied to one collocation budget or exactly eight values.

DeepXDE equations and implementation. The unmodified DeepXDE 1.15.0 library (Lu et al., 2021) supplies differentiation and training. Our adapter implements the standard equations below, not the earlier surrogates. It differs from official demonstrations by using soft condition losses, so omitted initial information cannot leak through an output transform.

DeepXDE equations and references. All spatial domains are $x \in [-1, 1]$. Heat uses $u_t - u_{xx} = 0$, zero Dirichlet boundary conditions, $u(x, 0) = \sin(\pi x)$, and $t \in [0, 1]$. Its analytic trajectory $e^{-\pi^2 t} \sin(\pi x)$ is evaluated on a 101×101 grid. Burgers uses $u_t + uu_x - (.01/\pi)u_{xx} = 0$, zero Dirichlet boundaries, $u(x, 0) = -\sin(\pi x)$, and $t \in [0, .99]$. Allen–Cahn uses $u_t - .001u_{xx} - 5(u - u^3) = 0$, $u(\pm 1, t) = -1$, $u(x, 0) = x^2 \cos(\pi x)$, and $t \in [0, 1]$. Burgers and Allen–Cahn are evaluated on the original DeepXDE v1.15.0 reference grids, 100×256 and 101×201, respectively, including $t = 0$. The source is the tagged DeepXDE example-data directory. Reference orientation and the initial slice were verified locally; both numerical initial slices agree with the declared functions to within 10^{-6}.

DeepXDE information intervention and fairness. Seeds 500–507 define eight matched runs per PDE. Each of the 24 task–seed cells trains full-initial, no-initial, and Initial(8) models before loading any dense reference. Full-initial exposes values at 160 sampled initial locations. Initial(8) exposes only $x_j = -1 + 2j/9$, $j = 1, \ldots, 8$, at $t = 0$. No-initial exposes none. Values are computed from the declared public initial condition, not from a dense-reference file. Every arm uses the same 1,024 interior, 100 spatial-boundary, and 160 initial-slice residual locations, sampled pseudorandomly and held fixed. The latter enforce only the PDE residual in the no-initial arm. Loss terms are mean-square residual, boundary, and, when present, initial data error, each with unit weight. Hence full and sparse data terms have equal aggregate weight, not equal numbers of observations.

The network is $[2, 32, 32, 32, 1]$ with tanh activations, Glorot-normal initialization, and 2,241 parameters. Training uses float32, Adam at 10^{-3} for exactly 8,000 steps, with no L-BFGS continuation, validation-based stopping, or hidden-error model selection. Initialization and residual locations match across all three arms in every cell. The final checkpoint is evaluated on each complete reference grid. Python 3.11.11 and PyTorch 2.4.0 were used with graphics processing unit (GPU) visibility disabled.

DeepXDE results. Table 23 reports per-task tails, public losses, and training costs. All three task means and all 24 paired seeds favor Initial(8) over omission. Pooled mean relative L_2 decreases from 1.14999 to .35333 ($\Delta = -.79666$). Per-task differences are $-.97670$, $-.73718$, and $-.67609$. All Allen–Cahn Initial(8) and full-initial runs exceed the error threshold .5: the added information removes the omission failure, but the optimizer still does not recover an accurate Allen–Cahn solution. Initial(8) is therefore a repair for missing information, not a general substitute for the full condition. The inferential unit is the PDE, of which there are only three. Seed standard deviations describe optimization variation and do not increase the number of independent tasks.

DeepXDE computational cost. Each training task used one CPU, with at most two tasks running concurrently and no GPU. Peak resident memory was approximately 528 mebibytes (MiB). Summed model-training time was 8,820 seconds.

Table 23: Per-task tails and costs over eight seeds per cell. Train (s) and public loss are means. Times exclude import, compilation, and hidden-reference evaluation.

PDE	Specification	Median L_2	P90	$L_2 > .5$ (%)	Train (s)	Public loss
Heat	Full (160)	.01251	.02768	0	206.1	4.00e−4
	No initial	.99992	1.00161	100	188.0	4.67e−6
	Initial(8)	.02203	.03795	0	219.4	3.10e−4
Burgers	Full (160)	.22429	.36581	0	39.6	5.25e−3
	No initial	.99997	1.00178	100	35.0	3.94e−8
	Initial(8)	.26925	.33981	0	36.2	4.30e−3
Allen–Cahn	Full (160)	.82837	.89127	100	131.2	2.41e−3
	No initial	1.44989	1.45229	100	120.6	1.69e−4
	Initial(8)	.74749	.85087	100	126.5	1.79e−3

D Multi-Mode Solution Visualizations

Figure 3 provides reference fields and errors for the two local reconstructions in Figure 2. These examples use the public-portfolio design on eight-coefficient libraries, ten public observations, and 1% Gaussian observation noise. Seed 125 is the smallest held-out seed and was chosen by index, not error. Zero library mismatch isolates the represented solution structure. It is not the primary 5% mismatch condition in Table 3.

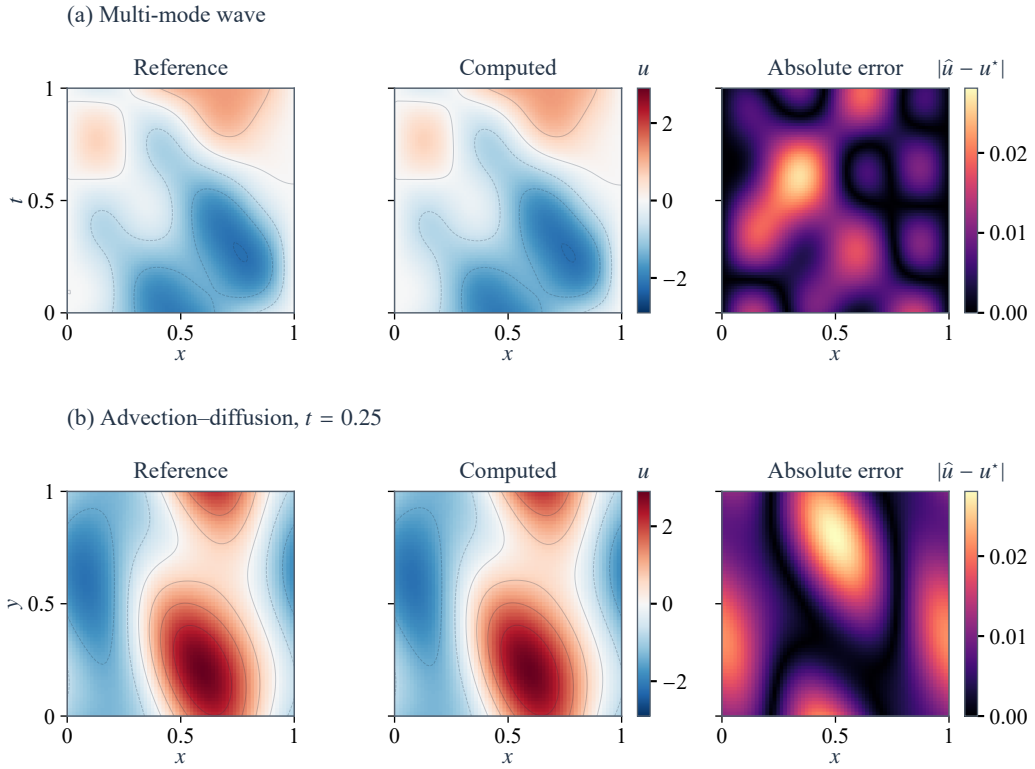

Figure 3: **Multi-mode reference fields, computed fields, and absolute errors.** Rows show the wave and two-dimensional advection–diffusion examples. Each row has two vertical color bars, aligned to the panel height: one beside the computed field, shared with the reference, and one beside the absolute error. Both scales are fixed across rows. The plotted grids are 81×81 and 65×65. These visualization examples use zero library mismatch.

Two distinct field structures. The one-dimensional wave satisfies $u_{tt} = 0.75^2 u_{xx}$ with homogeneous Dirichlet boundaries. Its library combines four spatial frequencies with sine and cosine temporal phases. The two-dimensional periodic field satisfies $u_t + 0.28u_x - 0.19u_y = 0.009\Delta u$ and uses spatial modes $(1, 0)$, $(0, 1)$, $(1, 1)$, and $(2, 1)$, each with two phases. The wave view

spans $(x, t) \in [0, 1]^2$. The transport view spans $(x, y) \in [0, 1]^2$ at $t = 0.25$. Multiple extrema and interference patterns replace the single-mode profiles, but the reconstruction problems remain finite-dimensional and linear.

Reconstruction error. Relative L_2 errors are 0.009729 and 0.009106 on the full-domain validation grids, including transport times beyond the displayed slice. These zero-mismatch examples are distinct from the primary aggregate benchmark.

Solution geometry and temporal evolution. Figures 4 and 5 show the same fitted fields with no additional reconstruction trials. References are used only for offline evaluation.

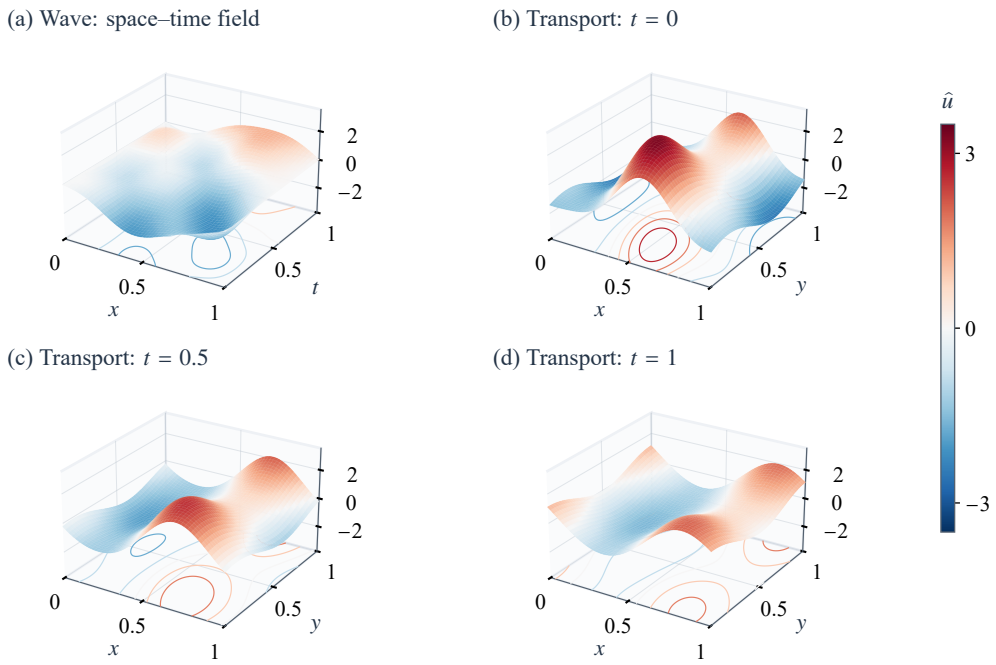

Figure 4: **Computed solution geometry and evolution.** The wave spans space and time; the periodic advection–diffusion field is shown at three times. All panels share an amplitude scale and evaluate the same fitted coefficients on 101×101 grids, with zero representation mismatch.

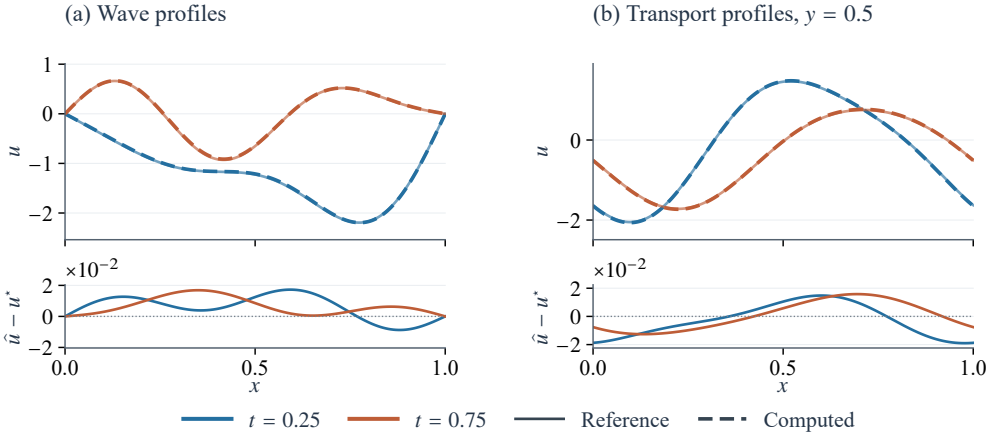

Figure 5: **Temporal profiles and resolved reconstruction errors.** Solid reference and dashed computed curves at $t = 0.25, 0.75$ use 601 points per profile; transport fixes $y = 0.5$. Lower strips resolve signed errors on separate vertical scales. All profiles use the same fitted coefficients.