FISHER-INFORMED NEURAL OPERATORS FOR PDE-CONSTRAINED INVERSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific inverse problems often require repeated evaluations of high-dimensional PDE forward models, motivating neural surrogates as efficient alternatives to accurate numerical solvers. However, even when a surrogate closely approximates a numerical solver in L^2, using it as the forward model can yield highly inaccurate parameter estimates. Here, we study gradient-based minimization of observation error for parameter inversion and propose training neural surrogates using the leading eigenvectors of an amortized Fisher information matrix. The resulting Fisher-informed Neural Operator (FINO) is a derivative-informed neural operator that learns the forward model Jacobian only along directions relevant to maximum-likelihood optimization. For computational efficiency, FINO uses randomized SVD to obtain an amortized set of Fisher-information eigenvectors. Our theoretical analysis relates Fisher directions to maximum-likelihood estimate recovery and characterizes how the effective rank of the inverse-problem Jacobian enters through spectral truncation and conditioning. We evaluate FINO on three forward problems with widely varying sensitivity structure: Darcy flow, two-dimensional Navier–Stokes flow, and seismic imaging with the acoustic wave equation. On Navier–Stokes, FINO reduces mean parameter RMSE by 42.8% and 25.1% relative to FNO trained only with L^2 and DINO, respectively. Controlled spectral-decay experiments show how FINO's benefit depends on the Fisher spectrum. Its amortizable offline cost also makes FINO practical for scientific inverse problems.

1 INTRODUCTION

PDE-constrained inverse problems infer physical parameters from noisy indirect measurements. Neural operators can amortize the repeated forward PDE solves that inversion requires, but accurate prediction of the solution field (forward) does not necessarily imply accurate parameter recovery: a gradient-based inverse method also depends on the surrogate's sensitivity to its input. This distinction matters especially when many parameter perturbations are weakly observed. The question is therefore not only how accurately a surrogate predicts the field, but also which parameter derivatives it must learn when derivative labels are costly.

Derivative-informed surrogates improve sensitivity accuracy by augmenting value fitting with selected Jacobian information (Czarnecki et al., 2017; O'Leary-Roseberry et al., 2023; Qiu et al., 2024; Yao et al., 2026). Existing methods use several dimensionality reduction and sensitivity geometries. FINO derives its supervision geometry from the downstream observation likelihood. Full-field sensitivity measures how strongly a parameter perturbation changes the complete solution, whereas likelihood-informed sensitivity measures its effect after observation and noise weighting (Cui et al., 2014; Cui & Zahm, 2021; Cao et al., 2025). Physics-informed operators offer a complementary approach: they constrain predicted solutions using the governing PDE (Raissi et al., 2019; Li et al., 2024). However, neither low field error nor a small PDE residual identifies which input derivatives control the accuracy of the resulting maximum-likelihood estimate. Figure 1 illustrates this: comparable forward error can still yield substantially different inverse solutions.

We propose Fisher-informed neural operators (FINO), which use an amortized Fisher basis to allocate full-field directional derivative labels. Fisher information plays two distinct roles. First, its leading eigenvectors determine *which* parameter directions receive the limited derivative-supervision

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.

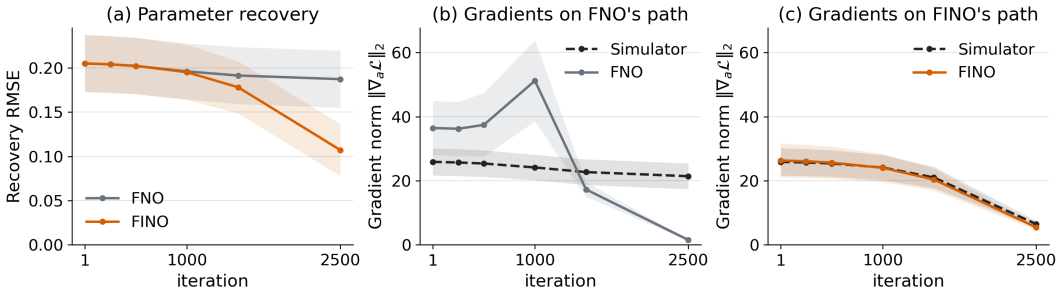


Figure 1: **Similar forward accuracy, different inverse behavior.** FNO and FINO have nearly identical Navier–Stokes test relative L^2 error (Table 1), yet FINO attains lower recovery RMSE (a). Panels (b) and (c) follow each surrogate's own inverse trajectory and evaluate, at the same saved iterates, the likelihood-gradient norm under the surrogate (solid) and under the simulator (dashed). FNO's gradient norm vanishes while the simulator's remains large, so FNO stalls at a point that is not stationary for the true likelihood (b); FINO's gradient norm tracks the simulator's throughout (c). Lines show means over 100 targets; bands denote one standard deviation.

budget. Second, the decay of its eigenvalues indicates *how broadly* observation-relevant sensitivity is distributed and, consequently, how many directions may benefit from supervision. The Fisher subspace serves only to allocate labels: FINO retains the full parameter-to-field map. This separation preserves a reusable full-field representation while adapting derivative supervision to the observation model. Thus, FINO repurposes likelihood-informed geometry from dimension reduction to observation-aware derivative supervision.

Contributions.

- We introduce Fisher-Informed Neural Operators: a type of derivative-informed neural operator that is trained with directional derivatives of the Jacobian along a matrix-free amortized Fisher basis.
- We derive recovery bounds for maximum-likelihood inversion in terms of FINO's forward and derivative errors, establishing conditions for recovering the identifiable component of a simulator-based MLE and extending the analysis to full-dimensional parameter recovery.
- We separate the directional and spectral roles of Fisher information: its eigenvectors select where derivatives are supervised, while its eigenvalues weight how errors in those directions affect likelihood-gradient accuracy. A controlled spectral-decay experiment supports this distinction.
- We demonstrate that FINO can yield substantially different, problem-dependent parameter recovery from FNO despite held-out forward relative L_2 errors within 0.2%, and demonstrate sparse directional supervision for higher-dimensional seismic imaging experiment.

2 Preliminaries

2.1 Observation model and maximum likelihood

Let $\boldsymbol{a} \in \mathbb{R}^{d_{\text{in}}}$ denote the parameter vector optimized during inversion, expressed in the chosen computational coordinates, and let μ denote its reference distribution used for training and basis construction. Let $\boldsymbol{m}_0$ be a fixed reference physical-parameter field and let $\boldsymbol{L}$ be a fixed linear map from these coordinates to physical-parameter perturbations. The corresponding physical parameter field is $\boldsymbol{m}(\boldsymbol{a}) := \boldsymbol{m}_0 + \boldsymbol{L}\boldsymbol{a}$. If $\mathcal{F}_{\text{phys}}$ denotes the PDE solution operator, we define the parameter-to-field map $\mathcal{F}(\boldsymbol{a}) := \mathcal{F}_{\text{phys}}(\boldsymbol{m}(\boldsymbol{a}))$, whose output is the full discretized solution field. For a native physical parameterization, $\boldsymbol{L} = \boldsymbol{I}$. For a whitened Gaussian parameterization, $\boldsymbol{a} \sim \mu = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$; consequently, the induced physical parameter satisfies $\boldsymbol{m} \sim \mathcal{N}\left(\boldsymbol{m}_0, \boldsymbol{L}\boldsymbol{L}^\top\right)$. Let $\boldsymbol{a}^\star$ denote the unknown true parameter coordinates, $\mathcal{R}$ the observation operator, and $\boldsymbol{\Sigma}_{\text{obs}} \succ 0$ the observation-noise covariance. With $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\text{obs}})$, the observed data are $\boldsymbol{d}_{\text{obs}} = \mathcal{R}\mathcal{F}(\boldsymbol{a}^\star) + \boldsymbol{\epsilon}$. Up to an

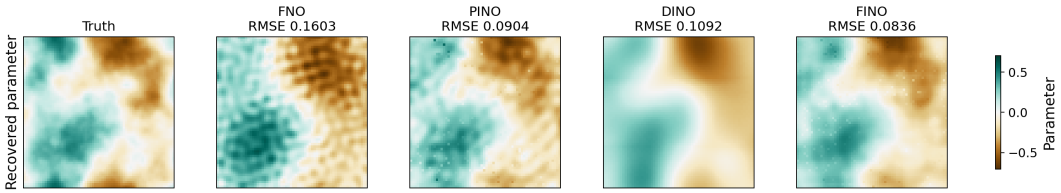


Figure 2: **Navier–Stokes parameter recovery.** FNO and FINO are evaluated under the same inversion setting and differ by less than 0.2% in validation L^2 error. FINO reduces mean RMSE of parameter error by 42.8% relative to FNO (Table 1). Titles report parameter RMSE for this representative target; Table 1 reports aggregate results over 100 paired targets.

additive constant, the negative log-likelihood is

$$\mathcal{L}(\boldsymbol{a}; \boldsymbol{d}_{\text{obs}}) = \frac{1}{2} \left\| \boldsymbol{\Sigma}_{\text{obs}}^{-1/2} \left(\mathcal{R}\mathcal{F}(\boldsymbol{a}) - \boldsymbol{d}_{\text{obs}} \right) \right\|_2^2 . \tag{1}$$

2.2 Sensitivity metrics and Fisher information

The likelihood (1) weights observed changes by their noise precision, so directions that strongly perturb the full solution field need not be most informative for inversion. Define $\boldsymbol{J}_{\mathcal{F}}(\boldsymbol{a}) := D_{\boldsymbol{a}}\mathcal{F}(\boldsymbol{a})$, $\widetilde{\boldsymbol{J}}(\boldsymbol{a}) := \boldsymbol{\Sigma}_{\text{obs}}^{-1/2}\mathcal{R}\boldsymbol{J}_{\mathcal{F}}(\boldsymbol{a})$, where $D_{\boldsymbol{a}}$ denotes differentiation with respect to the parameter coordinates. For a direction $\delta\boldsymbol{a} \in \mathbb{R}^{d_{\text{in}}}$, $\boldsymbol{L}\delta\boldsymbol{a}$ is the corresponding physical-parameter perturbation, $\boldsymbol{J}_{\mathcal{F}}(\boldsymbol{a})\delta\boldsymbol{a}$ is its first-order full-field effect, and $\widetilde{\boldsymbol{J}}(\boldsymbol{a})\delta\boldsymbol{a}$ is its noise-whitened observed effect. Writing $\boldsymbol{r}(\boldsymbol{a}) := \boldsymbol{\Sigma}_{\text{obs}}^{-1/2}\big(\mathcal{R}\mathcal{F}(\boldsymbol{a}) - \boldsymbol{d}_{\text{obs}}\big)$, the likelihood gradient is $\nabla_{\boldsymbol{a}}\mathcal{L} = \widetilde{\boldsymbol{J}}(\boldsymbol{a})^\top \boldsymbol{r}(\boldsymbol{a})$. Under the additive Gaussian observation model above, let $\boldsymbol{y} \mid \boldsymbol{a} \sim \mathcal{N}(\mathcal{R}\mathcal{F}(\boldsymbol{a}), \boldsymbol{\Sigma}_{\text{obs}})$, where $\boldsymbol{d}_{\text{obs}}$ is one realization. The pointwise Fisher information is then

$$\boldsymbol{H}_{\text{F}}(\boldsymbol{a}) := \mathbb{E}_{\boldsymbol{y}|\boldsymbol{a}}\big[\nabla_{\boldsymbol{a}} \log p(\boldsymbol{y} \mid \boldsymbol{a}) \nabla_{\boldsymbol{a}} \log p(\boldsymbol{y} \mid \boldsymbol{a})^\top\big] = \widetilde{\boldsymbol{J}}(\boldsymbol{a})^\top \widetilde{\boldsymbol{J}}(\boldsymbol{a}).$$

For an output mass or quadrature matrix $\boldsymbol{M}_{\text{out}} \succ 0$, define

$$\boldsymbol{H}_{\text{field}} := \mathbb{E}_{\boldsymbol{a}\sim\mu}\big[\boldsymbol{J}_{\mathcal{F}}(\boldsymbol{a})^\top \boldsymbol{M}_{\text{out}} \boldsymbol{J}_{\mathcal{F}}(\boldsymbol{a})\big] , \ \bar{\boldsymbol{H}}_{\text{F}} := \mathbb{E}_{\boldsymbol{a}\sim\mu}[\boldsymbol{H}_{\text{F}}(\boldsymbol{a})] = \mathbb{E}_{\boldsymbol{a}\sim\mu}\Big[\widetilde{\boldsymbol{J}}(\boldsymbol{a})^\top \widetilde{\boldsymbol{J}}(\boldsymbol{a})\Big] . \tag{2}$$

The subscripts field and F distinguish full-field and Fisher sensitivity. Unlike $\boldsymbol{H}_{\text{field}}$, $\bar{\boldsymbol{H}}_{\text{F}}$ assigns zero weight to perturbations invisible to the sensors and downweights noisy measurements. Thus, $\bar{\boldsymbol{H}}_{\text{F}}$ is the Fisher information averaged over the reference parameter distribution μ. Only the outer expectation over $\boldsymbol{a} \sim \mu$ is approximated by sampling. Because $\bar{\boldsymbol{H}}_{\text{F}}$ depends on the forward model, acquisition, noise, and μ, but not on a realized dataset, it can be constructed offline and reused.

Both quantities in (2) are Jacobian second-moment matrices; $\bar{\boldsymbol{H}}_{\text{F}}$ applies this construction to the noise-whitened observation map. When related metrics are used as dimension-reduction basis (Qiu et al., 2024; Cui et al., 2014), FINO uses this geometry to allocate full-output directional derivative labels and analyzes their effect on parameter recovery. If $\boldsymbol{\Sigma}_{\text{obs}} \succ 0$, whitening preserves rank:

$$\operatorname{rank} \boldsymbol{H}_{\text{F}}(\boldsymbol{a}) = \operatorname{rank} \widetilde{\boldsymbol{J}}(\boldsymbol{a}) = \operatorname{rank}(\mathcal{R}\boldsymbol{J}_{\mathcal{F}}(\boldsymbol{a})) \le \min\{\operatorname{rank} \boldsymbol{J}_{\mathcal{F}}(\boldsymbol{a}), \operatorname{rank} \mathcal{R}\} .$$

Thus, Fisher rank counts observation-visible parameter effects, whereas its eigenvalues measure their noise-weighted strength. Figure 17 and Appendix Figure 7 visualize the two metrics. Appendix E gives an illustrative, rather than universal, two-mode example in which field-sensitive supervision lowers forward error while Fisher-directed supervision lowers inverse error. Appendices A, B.3, D, and F provide notation, discretization details, connections to related subspaces, and rank bounds; Section 4.1 connects the spectrum to inversion uncertainty.

3 Fisher-informed derivative supervision

FINO uses the amortized Fisher basis only to select derivative labels, not to reduce the operator or inverse search space. Equations (2)–(5) assume that the operator-input and inversion coordinates coincide; Appendix B.2 describes the Darcy setting, where supervision uses ambient field-space directions and inversion uses a separate parameterization.

3.1 MATRIX-FREE BASIS AND DERIVATIVE TARGETS

To select directions for derivative supervision without forming the full Jacobian or Fisher information matrix (FIM), we estimate the leading Fisher subspace using random observation-space probes and vector–Jacobian products (VJP). Choose a supervision rank m, draw independent inputs for basis constructions $\boldsymbol{a}_i \overset{\text{iid}}{\sim} \mu$, $i = 1, \ldots, M$, and use ℓ independent observation-space probes per input, with $M\ell \geq m$. For each i, draw $\boldsymbol{Z}_i \in \mathbb{R}^{d_{\text{obs}} \times \ell}$ with independent $\mathcal{N}(0,1)$ entries and form:

$$\boldsymbol{Q}_i := \widetilde{\boldsymbol{J}}(\boldsymbol{a}_i)^\top \boldsymbol{Z}_i = \boldsymbol{J}_{\mathcal{F}}(\boldsymbol{a}_i)^\top \mathcal{R}^\top \boldsymbol{\Sigma}_{\text{obs}}^{-1/2} \boldsymbol{Z}_i, \qquad \boldsymbol{Q} := \frac{1}{\sqrt{M\ell}} [\boldsymbol{Q}_1 \ \cdots \ \boldsymbol{Q}_M]. \tag{3}$$

Since $\mathbb{E}[\boldsymbol{Z}_i \boldsymbol{Z}_i^\top / \ell] = \boldsymbol{I}$, $\widehat{\boldsymbol{H}}_{\text{F}} := \boldsymbol{Q}\boldsymbol{Q}^\top$ is an unbiased Monte Carlo estimator of $\bar{\boldsymbol{H}}_{\text{F}}$. Its leading m eigenvectors are the leading left singular vectors of $\boldsymbol{Q}$, denoted $\hat{\boldsymbol{V}}_m \in \mathbb{R}^{d_{\text{in}} \times m}$, with $\hat{\boldsymbol{V}}_m^\top \hat{\boldsymbol{V}}_m = \boldsymbol{I}_m$.

Each column of $\boldsymbol{Q}_i$ is computed by one VJP with cotangent $\mathcal{R}^\top \boldsymbol{\Sigma}_{\text{obs}}^{-1/2} \boldsymbol{z}$. Thus, basis acquisition forms neither a full Jacobian nor the $d_{\text{in}} \times d_{\text{in}}$ Fisher matrix. The construction requires samples from μ, not its density, and the Gaussian probes are algorithmic rather than an assumption on μ. Appendix B.5 derives the estimator and gives its computation and storage costs.

For a training input $\boldsymbol{a}_n \overset{\text{iid}}{\sim} \mu$, $n = 1, \ldots N$, let $\boldsymbol{m}_n = \boldsymbol{m}_0 + \boldsymbol{L}\boldsymbol{a}_n$. Its directional derivative target is

$$\boldsymbol{T}_n := \boldsymbol{J}_{\mathcal{F}}(\boldsymbol{a}_n) \hat{\boldsymbol{V}}_m = D_{\boldsymbol{m}} \mathcal{F}_{\text{phys}}(\boldsymbol{m}_n) \big(\boldsymbol{L}\hat{\boldsymbol{V}}_m\big) \in \mathbb{R}^{d_{\text{out}} \times m}. \tag{4}$$

The columns of $\boldsymbol{T}_n$ are full-dimensional output responses along the selected physical directions $\boldsymbol{L}\hat{\boldsymbol{V}}_m$. These directions are formed once and reused across training inputs. Here $M\ell$ is the number of VJP probe directions used to construct the basis, whereas m is the number of JVP derivative targets generated at each labeled training input.

3.2 FINO TRAINING OBJECTIVE

Having computed the Fisher-selected derivative targets in Section 3.1, we train FINO to match both the simulator's forward outputs and its directional derivatives. Let $\boldsymbol{J}_{\mathcal{F},\text{nn}} := D_{\boldsymbol{a}} \mathcal{F}_{\text{nn}}$. For training pairs $(\boldsymbol{a}_n, \boldsymbol{u}_n)$, with $\boldsymbol{u}_n = \mathcal{F}(\boldsymbol{a}_n)$, let $\mathcal{I}_D \subseteq \{1, \ldots, N\}$ contain the N_L inputs with derivative targets. We minimize

$$\mathcal{L}_{\text{train}} = \frac{1}{N} \sum_{n=1}^{N} \frac{\|\mathcal{F}_{\text{nn}}(\boldsymbol{a}_n) - \boldsymbol{u}_n\|_2^2}{\|\boldsymbol{u}_n\|_2^2} + \frac{\lambda_{\text{der}}}{N_L} \sum_{n \in \mathcal{I}_D} \Big\| \boldsymbol{J}_{\mathcal{F},\text{nn}}(\boldsymbol{a}_n) \hat{\boldsymbol{V}}_m - \boldsymbol{T}_n \Big\|_{\text{F}}^2, \tag{5}$$

where $\lambda_{\text{der}} > 0$ controls the derivative-loss weight. The first term measures mean squared relative forward-prediction error. The second averages, over derivative-labeled inputs, the summed squared errors along the m selected directions. When prior work like DINO applies value and full Jacobian losses to a reduced input–output map, FINO retains full-output JVP targets and selects only the input directions receiving derivative supervision. Appendix B.4 specifies the normalization used to set λ_{der}. We assume $\|\boldsymbol{u}_n\|_2 > 0$ for every training input.

4 THEORY: RECOVERY IN MEASUREMENT-SUPPORTED DIRECTIONS

The analysis addresses three questions. First, which r-dimensional parameter subspace is best conditioned by the observations? Second, how do surrogate forward and derivative errors within that subspace affect parameter recovery? Third, how do the Fisher eigenvectors, eigenvalues, and amortized-to-local subspace alignment determine the value of derivative supervision? Sections 4.1 and 4.2 address the first two questions, while Section 4.3 separates the directional, spectral, and transfer roles of Fisher information.

4.1 INVERSE-SPACE STABILITY

We begin with least-squares inversion using the numerical simulator in the affine setting, restricted to an r-dimensional Fisher subspace. Let the observed map have constant whitened Jacobian $\widetilde{\boldsymbol{J}}$,

with Fisher eigenvalues $\lambda_1 \geq \cdots \geq \lambda_{d_{\text{in}}} \geq 0$. Let $\boldsymbol{V} \in \mathbb{R}^{d_{\text{in}} \times r}$ contain the leading orthonormal eigenvectors, where $\lambda_r > 0$. For a reference $\boldsymbol{a}_{\text{ref}}$, define the restricted MLE by

$$\widehat{\boldsymbol{a}}_{\text{MLE}} = \boldsymbol{a}_{\text{ref}} + \boldsymbol{V}\widehat{\boldsymbol{z}}_{\text{MLE}}, \qquad \widehat{\boldsymbol{z}}_{\text{MLE}} := \arg\min_{\boldsymbol{z}\in\mathbb{R}^r} \mathcal{L}(\boldsymbol{a}_{\text{ref}} + \boldsymbol{V}\boldsymbol{z}).$$

Among all parameter subspaces of dimension r, the following proposition identifies the one that minimizes the largest variance due to measurement noise in the restricted estimate.

Proposition 1 (Minimum worst case variance in restricted inversion)**.** *For* $1 \leq r \leq \operatorname{rank} \widetilde{\boldsymbol{J}}$,

$$\max_{\dim(\mathcal{S})=r} \min_{\substack{\boldsymbol{v}\in\mathcal{S} \\ \|\boldsymbol{v}\|_2=1}} \|\widetilde{\boldsymbol{J}}\boldsymbol{v}\|_2^2 = \lambda_r,$$

attained by $\mathcal{S} = \operatorname{range} \boldsymbol{V}$*. Equivalently, under Gaussian observation noise, the leading Fisher subspace minimizes the largest directional variance of the restricted estimate, which equals* λ_r^{-1}*.*

Thus, $\sqrt{\lambda_r}$ is the weakest retained observed sensitivity. Appendix G.1 proves the proposition, and Appendix G discusses variance and spectral energy based choices of r.

4.2 Recovery error

Let $\boldsymbol{Q} := \boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^\top$, $\boldsymbol{t}_{\text{trunc}} := \boldsymbol{Q}(\boldsymbol{a}^\star - \boldsymbol{a}_{\text{ref}})$, and $\boldsymbol{\xi} := \boldsymbol{\Sigma}_{\text{obs}}^{-1/2}\boldsymbol{\epsilon}$. Write $\widehat{\boldsymbol{a}} := \widehat{\boldsymbol{a}}_{\text{MLE}}$ and $\widehat{\boldsymbol{a}}_{\text{nn}} := \widehat{\boldsymbol{a}}_{\text{MLE,nn}}$ for simulator and surrogate MLEs on the same restricted space. Define

$$\begin{aligned}
\widetilde{\boldsymbol{J}}_{\text{nn}} &:= \boldsymbol{\Sigma}_{\text{obs}}^{-1/2}\mathcal{R}D_{\boldsymbol{a}}\mathcal{F}_{\text{nn}}, \\
\varepsilon_{\text{sim}} &:= \|\boldsymbol{r}(\widehat{\boldsymbol{a}})\|_2, \\
\varepsilon_{\text{nn,forward}} &:= \|\boldsymbol{\Sigma}_{\text{obs}}^{-1/2}\mathcal{R}(\mathcal{F}_{\text{nn}} - \mathcal{F})(\widehat{\boldsymbol{a}})\|_2, \\
\varepsilon_{\text{nn,Fisher}} &:= \|[\widetilde{\boldsymbol{J}}_{\text{nn}} - \widetilde{\boldsymbol{J}}](\widehat{\boldsymbol{a}})\boldsymbol{V}\|_{\text{op}}.
\end{aligned}$$

Theorem 1 (Recovery within and beyond a Fisher subspace)**.** *Assume that the simulator's observed map is affine and* $\lambda_r > 0$*.*

(i) Restricted simulator. *The restricted simulator MLE is unique and*

$$\|\widehat{\boldsymbol{a}} - \boldsymbol{a}^\star\|_2^2 = \underbrace{\|\boldsymbol{t}_{\text{trunc}}\|_2^2}_{\text{truncation}} + \underbrace{\|(\widetilde{\boldsymbol{J}}\boldsymbol{V})^\dagger\boldsymbol{\xi}\|_2^2}_{\text{noise}}. \quad (6)$$

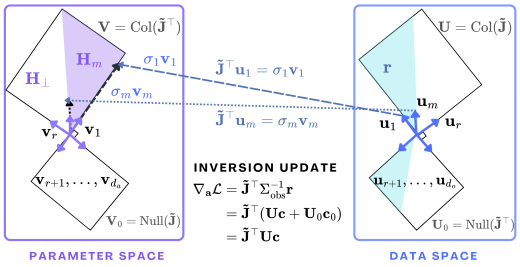


Figure 3: **Only measurement-supported directions drive a local inverse update.** For the noise-whitened Jacobian $\widetilde{\boldsymbol{J}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, each observable data direction $\boldsymbol{u}_i$ maps to $\boldsymbol{v}_i$ with strength σ_i. Residual components in $\operatorname{Null}(\widetilde{\boldsymbol{J}}^\top)$ do not enter $\nabla_{\boldsymbol{a}}\mathcal{L} = \widetilde{\boldsymbol{J}}^\top\boldsymbol{r}$; parameter directions in $\operatorname{Null}(\widetilde{\boldsymbol{J}})$ are locally unidentifiable. The shaded space contains the leading Fisher directions.

(ii) Restricted surrogate. *If the surrogate observed map is affine on the restricted space and* $\varepsilon_{\text{nn,Fisher}} < \sqrt{\lambda_r}$*, set* $\delta_r := \sqrt{\lambda_r} - \varepsilon_{\text{nn,Fisher}} > 0$*. Then its restricted MLE is unique and*

$$\|\widehat{\boldsymbol{a}}_{\text{nn}} - \widehat{\boldsymbol{a}}\|_2 \leq E_{\text{nn}} := \underbrace{\frac{\varepsilon_{\text{nn,forward}}}{\delta_r}}_{\text{forward error}} + \underbrace{\frac{\varepsilon_{\text{nn,Fisher}}\,\varepsilon_{\text{sim}}}{\delta_r^2}}_{\text{derivative error in Fisher subspace}}. \quad (7)$$

(iii) Full space. *Under the assumptions of part (ii), if both observed maps are affine on the full space, define, for any* $\boldsymbol{a}_t$*,*

$$\boldsymbol{q}_t := \boldsymbol{Q}(\boldsymbol{a}_t - \boldsymbol{a}_{\text{ref}}), \quad c_r := \|(\widetilde{\boldsymbol{J}}_{\text{nn}}\boldsymbol{V})^\dagger\widetilde{\boldsymbol{J}}_{\text{nn}}\boldsymbol{Q}\|_{\text{op}}, \quad g_{\text{F,t}} := \|\boldsymbol{V}^\top\nabla\mathcal{L}_{\text{nn}}(\boldsymbol{a}_t)\|_2.$$

Then

$$\|\boldsymbol{a}_t - \boldsymbol{a}^\star\|_2^2 \leq \underbrace{\|\boldsymbol{q}_t - \boldsymbol{t}_{\text{trunc}}\|_2^2}_{\text{off subspace error}} + \left(\|(\widetilde{\boldsymbol{J}}\boldsymbol{V})^\dagger\boldsymbol{\xi}\|_2 + E_{\text{nn}} + \underbrace{c_r\|\boldsymbol{q}_t\|_2}_{\text{cross-coupling}} + \underbrace{g_{\text{F,t}}/\delta_r^2}_{\text{optimization error}}\right)^2. \quad (8)$$

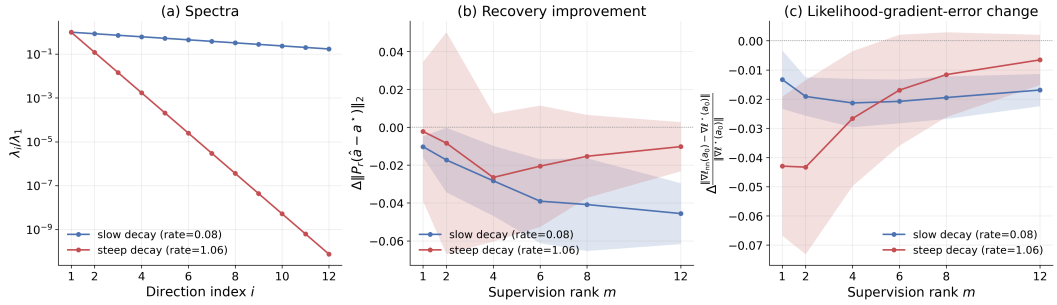


Figure 4: **Supervision rank under controlled observed-sensitivity decay.** Two 16-dimensional maps share a four-dimensional exact null space and differ in singular-value decay. (a) Normalized Fisher spectra. (b) Paired change in observable-subspace recovery error relative to value-only training ($m = 0$), with each model inverted on its own path. (c) Paired change in relative likelihood-gradient error at the same initial parameter for each case. Negative changes favor derivative supervision. Curves average 64 paired targets in each of three training seeds; bands are 95% intervals from the three seed-level means. Construction and uncertainty details are in Appendix B.1.

The bound identifies the two surrogate errors that directly affect recovery: forward error, amplified by δ_r^{-1} and directional-derivative error, amplified by δ_r^{-2}. Thus, as shown in Table 1, even forward-error-matched surrogates can produce substantially different parameter estimates when the inverse problem is weakly conditioned. FINO supplements value fitting by directly reducing the directional-derivative term. If both surrogate errors vanish while δ_r remains bounded away from zero, the surrogate estimate converges to the simulator estimate in the supervised Fisher coordinates. When $r = \mathrm{rank}(\widetilde{\boldsymbol{J}})$, these coordinates span all observation-sensitive directions, so the limit agrees with the identifiable component of a full-space simulator MLE (Appendix Corollary 2). Full-space convergence additionally requires the error term from unsupervised subspace, coupling, and optimization terms to vanish. Refer to Appendix H for the full proof.

4.3 Directional and Spectral Roles of Fisher Information

Derivative supervision controls an otherwise unconstrained error. The leading Fisher eigenvectors identify parameter perturbations that produce the largest noise-weighted changes in the observations. FINO uses these directions to allocate a limited derivative-label budget. This targets the retained observed-Jacobian error $\varepsilon_{\mathrm{nn,Fisher}}$ appearing in Theorem 1. If $\varepsilon_{\mathrm{nn,Fisher}} \le \sqrt{\lambda_r}/2$, then Theorem 1 gives: $\|\widehat{\boldsymbol{a}}_{\mathrm{MLE,nn}} - \widehat{\boldsymbol{a}}_{\mathrm{MLE}}\|_2 \le \frac{2\varepsilon_{\mathrm{nn,forward}}}{\sqrt{\lambda_r}} + \frac{4\varepsilon_{\mathrm{nn,Fisher}}\varepsilon_{\mathrm{sim}}}{\lambda_r}$.

Thus, as the weakest selected sensitivity λ_r decreases, both surrogate errors are amplified, with derivative error receiving the stronger factor, amplified by δ_r^{-2}. Value fitting alone does not control this term: surrogates can agree at all training inputs while having different directional derivatives. Appendix I formalizes this distinction.

Fisher spectral decay determines the effective dimension of inverse sensitivity. Direction selection and spectral decay answer different questions. The Fisher eigenvectors determine *which* parameter derivatives receive supervision, whereas the eigenvalues quantify the observation strength of those directions. A slowly decaying spectrum distributes observation-relevant sensitivity across more directions; a rapidly decaying spectrum concentrates it in fewer directions. To quantify the contribution of each retained direction, consider a point where a baseline and a derivative-trained surrogate both match the simulator's observed value. For $b \in \{\mathrm{base}, \mathrm{deriv}\}$, define the following error metrics: $e_{b,i} := \frac{\|(\widetilde{\boldsymbol{J}}_b - \widetilde{\boldsymbol{J}})\boldsymbol{v}_i\|_2}{\sqrt{\lambda_i}}$, $\mathcal{G}_b := \mathbb{E}\left\|\boldsymbol{V}\boldsymbol{V}^\top(\nabla\mathcal{L}_b - \nabla\mathcal{L})\right\|_2^2$. For $b \in \{\mathrm{base}, \mathrm{deriv}\}$, let $e_{b,i}$ measure the relative observation-weighted Jacobian error along the Fisher direction $\boldsymbol{v}_i$, and let G_b measure the corresponding projected likelihood-gradient error. Under standardized residuals,

$$\mathcal{G}_{\mathrm{base}} - \mathcal{G}_{\mathrm{deriv}} = \sum_{i=1}^{r} \lambda_i \left(e_{\mathrm{base,i}}^2 - e_{\mathrm{deriv,i}}^2\right). \tag{9}$$

Table 1: **Forward accuracy and noiseless inverse recovery.** Validation and test entries are mean relative field errors. Inverse entries mean over the same 100 unseen targets paired across methods, using noiseless observations ($\sigma = 0$). FNO and FINO have nearly identical validation error within each PDE, but FINO improves recovery only on Navier–Stokes. Sample standard deviations and results with observation noise are in Appendix Table 6. Bold indicates the best mean for each PDE and metric.

		Forward error			Inverse recovery, $\sigma = 0$		
PDE	Method	Validation $L^2 \downarrow$	Test $L^2 \downarrow$	Spatial $H^1 \downarrow$	RMSE $\downarrow$	SSIM $\uparrow$	Data residual $\downarrow$
Navier–Stokes	FNO	0.05947	0.0624	0.1822	0.1873	0.2909	0.1403
	DINO	0.27604	0.2652	0.4709	0.1430	0.4091	0.0952
	PINO	**0.03776**	**0.0384**	**0.1166**	0.1226	0.4868	0.0562
	FINO	0.05939	0.0615	0.1681	**0.1071**	**0.5734**	**0.0490**
Darcy	FNO	0.009236	0.0086	0.2336	**0.1997**	**0.5549**	0.00487
	DINO	0.008804	0.0086	0.0908	0.2724	0.3559	**0.00456**
	PINO	**0.008558**	**0.0078**	**0.0514**	0.2113	0.5314	0.00457
	FINO	0.009222	0.0088	0.2196	0.2094	0.5422	0.00530

Derivative improvements therefore reduce likelihood-gradient error in proportion to the corresponding Fisher eigenvalues. Under slow spectral decay, errors across many directions can contribute, creating more opportunity for FINO. When the FIM is effectively low-rank and value-only training already captures its dominant directions, FINO has limited room to help. Consistent with this interpretation, Darcy exhibits faster decay as shown in Figure 6 and shows no corresponding recovery gain. Slow decay creates an opportunity rather than a guarantee: recovery also depends on forward error, noise, movement outside the selected subspace, and optimization. Accordingly, Figure 4 shows that recovery need not improve monotonically with supervision rank. Proposition 2 and Corollary 1 provide the corresponding spectral and noise-amplification results.

Subspace alignment transfers supervision to inversion. FINO supervises derivatives along the amortized basis $\widehat{\boldsymbol{V}}_m$, whereas the recovery bound depends on the local Fisher basis $\boldsymbol{V}$ encountered during inversion. Derivative accuracy in the training basis therefore transfers to recovery only when the amortized and local subspaces remain sufficiently aligned. For $m = r$, define $\delta_V := \left\| \hat{\boldsymbol{V}}_m \hat{\boldsymbol{V}}_m^\top - \boldsymbol{V}\boldsymbol{V}^\top \right\|_{\mathrm{op}}$, $\Delta \boldsymbol{J} := \boldsymbol{J}_{\mathcal{F},\mathrm{nn}} - \boldsymbol{J}_{\mathcal{F}}$. If, at the recovery point, $\left\| \Delta \boldsymbol{J} \hat{\boldsymbol{V}}_m \right\|_{\mathrm{F}} \leq \varepsilon_D$, $\|\Delta \boldsymbol{J}\|_{\mathrm{op}} \leq \varepsilon_{\mathrm{all}}$, then

$$\varepsilon_{\mathrm{nn,Fisher}} \leq \left\| \boldsymbol{\Sigma}_{\mathrm{obs}}^{-1/2} \mathcal{R} \right\|_{\mathrm{op}} (\varepsilon_D + \varepsilon_{\mathrm{all}} \delta_V) \,. \tag{10}$$

The bound separates derivative error in the supervised directions, ε_D, from the cost of amortized-to-local subspace mismatch, $\varepsilon_{\mathrm{all}} \delta_V$. Because training controls an empirical average at labeled inputs rather than accuracy along an inverse trajectory, held-out directional error and local subspace alignment are complementary transfer diagnostics. Appendix K proves the result.

5 Numerical Experiments

The experiments examine three distinct questions. First, can two surrogates with comparable forward accuracy produce different parameter recovery? We test this through forward-error-matched FNO and FINO comparisons on Navier–Stokes and Darcy. Second, how does the Fisher spectrum affect the breadth of useful derivative supervision? We isolate this using controlled maps with different spectral-decay rates. Third, can sparse directional supervision extend to a substantially higher-dimensional wave-imaging problem? We examine this through an illustrative seismic case study.

5.1 Experiment setup

Benchmarks. We consider recovery of initial vorticity on a 64×64 Navier–Stokes (Li et al., 2024) and log-permeability on a 128×128 Darcy (Beskos et al., 2017). FINO supervises 400 of

4096 parameter directions for Navier–Stokes (9.77%). For Darcy, FNO and FINO take a 128×128 log-permeability field as input ($d_{\text{field}} = 16{,}384$). FINO supervises 200 linearly independent directions in this ambient field-input space, corresponding to 1.22% of its discretized dimension. The fields are generated from 64 Karhunen–Loève/cosine coefficients (Beskos et al., 2017), and inversion optimizes these same 64 coefficients rather than 16,384 independent pixels. Appendix Table 2 summarizes these and the corresponding inversion settings. Appendix B gives the PDE definitions (Section B.2) and experimental setup (Section B.4); Appendix B.1 defines the synthetic map used in Figure 4.

Comparison methods. FNO is trained only on relative L^2 error in the predicted solution (Li et al., 2021); FINO adds supervision of the simulator's derivatives along selected Fisher directions across the full output dimension, as in (5). For a broader comparison, we include two baseline methods from prior work. DINO (O'Leary-Roseberry et al., 2023) learns a map between reduced input and output coefficients and supervises its Jacobian; we follow the DINO architecture and training procedure of Qiu et al. (2024). Our PINO baseline (Li et al., 2024) uses the FNO architecture with an additional PDE residual loss.

Inverse evaluation. Checkpoints are selected solely by held-out validation relative L^2 error. Within each PDE, FNO and FINO use the same experiment setting; their selected validation errors differ by less than 0.2% relatively. This controls forward accuracy when assessing the effect of directional derivative supervision. DINO and PINO are evaluated as complete methods because their representations and training objectives differ. All methods use the same paired targets, observations, noise realizations, initializations, and inversion budgets. Table 1 reports noiseless recovery, while Appendix Table 6 reports both $\sigma = 0$ and $\sigma = 0.01$.

Evaluation metrics. Validation and test L^2 errors average the relative solution error over samples. Test H^1 error similarly uses a relative spatial norm, with derivatives computed by finite differences. For inverse target t, let $\boldsymbol{m}_t^\star$ and $\widehat{\boldsymbol{m}}_t$ be the true and recovered physical parameter fields, and let d_{phys} be the number of scalar grid values. We report parameter RMSE (Chai & Draxler, 2014) and the data residual (Peherstorfer et al., 2018):

$$\text{RMSE}_t = \frac{\|\widehat{\boldsymbol{m}}_t - \boldsymbol{m}_t^\star\|_2}{\sqrt{d_{\text{phys}}}}, \qquad D_t = \frac{\|\mathcal{R}\mathcal{F}(\widehat{\boldsymbol{m}}_t) - \boldsymbol{d}_{\text{obs},t}\|_2}{\sqrt{d_{\text{obs}}}}.$$

The residual uses the numerical PDE solver $\mathcal{F}$. We also report $\text{SSIM}(\widehat{\boldsymbol{m}}_t, \boldsymbol{m}_t^\star)$ using 7×7 windows and a parameter range fixed from the training data for each benchmark and shared across methods (Wang et al., 2004). Table 1 reports means over the same 100 unseen inverse targets; Appendix Table 6 gives sample standard deviations across targets and additional forward metrics.

5.2 Results

Forward accuracy does not determine inversion quality. Table 1 shows that similar forward accuracy need not yield similar inverse recovery. On Navier–Stokes, FNO and FINO have nearly identical validation relative L^2 errors (0.05947 and 0.05939), yet FINO reduces the mean noiseless recovery RMSE from 0.1873 to 0.1071 (42.8%) and the simulator data residual from 0.1403 to 0.0490 (65.1%), while increasing SSIM from 0.2909 to 0.5734.

Appendix Figures 8 and 9 illustrate this contrast on representative targets: similar forward predictions can produce different recovered parameters and gradients. Aggregate diagnostics over 100 targets show that FINO more closely tracks the simulator gradient norm (Appendix Figure 12) and has lower relative gradient error than all three baselines along their respective trajectories (Appendix Figure 13).

The benefit does not extend uniformly to Darcy. FNO and FINO again have similar forward test errors (0.0086 and 0.0088), but FINO does not improve mean recovery RMSE (0.2094 for FINO versus 0.1997 for FNO). Appendix Fig-

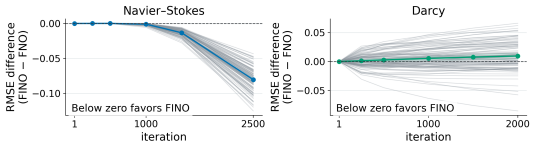


Figure 5: **FINO–FNO recovery differences along the inversion trajectories.** For each inversion experiment, the RMSE difference with respect to $\boldsymbol{m}^\star$ is computed before averaging. Negative values favor FINO. Thin lines show independent inversion experiments; bands show unadjusted pointwise 95% Student-t intervals over the 100 paired differences.

ure 10 shows a representative reconstruction. FINO's benefit therefore depends on the problem, but these two PDE benchmarks alone do not show why. We next examine their observed sensitivity and inversion gradients, then use a controlled synthetic experiment to test the role of spectral decay.

Fisher spectral decay and derivative supervision. Figure 4 shows that the effect of supervision rank depends on Fisher spectral decay and need not be monotone. Appendix C.2 provides spectral and trajectory diagnostics, while Appendix Table 8 further contrasts direction choices: Fisher-selected JVPs improve Navier–Stokes recovery despite worse forward error, whereas Darcy favors field selection.

Computational cost and application to realistic inverse problem. Appendix Table 9 gives cost scaling, while Tables 10–11 give measured costs in separate cohorts. In the Navier–Stokes study, FINO costs 20.143 hours offline, versus 0.319 hours for FNO. Its mean inverse runtime is 87.79 seconds, versus 3556.20 seconds for direct simulation, yielding a 21-inversion runtime break-even.

Seismic velocity recovery is a consequential inverse problem in exploration: wave-equation solves are expensive, and the acquisition geometry illuminates only part of the subsurface. We extended the Fisher-informed supervision principle to learn reverse-time migration operator (Park et al., 2026). On 256×256 velocity grids, supervising only 8 or 64 directions improved reconstruction and the reported data fit over a forward-only FNO over 20 test samples (Appendix Table 7). Figure 16 summarizes the dataset and visually compare reconstructions. Details are available in Appendix C.3.

6 Related Work

Sobolev training augments value fitting with Jacobian information (Czarnecki et al., 2017). O'Leary-Roseberry et al. (2023) compresses Jacobian information and emphasizes reduced-basis operators; Qiu et al. (2024) selects input directions using full-field sensitivity; and Yao et al. (2026) applies directional-derivative training to a full-field FNO. Li et al. (2024) instead imposes PDE-residual constraints (Raissi et al., 2019). Likelihood-informed methods use Fisher or Gauss–Newton geometry to reduce parameter or posterior representations during inference (Cui et al., 2014; Cui & Zahm, 2021), while DINO-based analyses bound MAP error for reduced input–output surrogates (Cao et al., 2025; Go & Chen, 2025). FINO repurposes this geometry for derivative-label allocation, not dimension-reduction: an amortized, noise-weighted Fisher basis determines where a limited budget of full-field JVP labels is spent. Our analysis expands recovery error in the same basis. For affine observed maps, the positive-eigenvalue Fisher subspace contains all likelihood-sensitive directions, and the discrepancy between surrogate and simulator maximum-likelihood solutions is controlled by forward error and directional-derivative error. As both errors vanish, the surrogate solution converges to the simulator solution in the retained coordinates; if they span the complete positive Fisher eigenspace, the limit recovers the identifiable component of a full-space simulator MLE. A full-space extension quantifies error from outside of the eigenspace and optimization error.

7 Discussion and conclusion

FINO uses Fisher geometry to allocate full-field derivative labels without reducing the learned operator or the inverse search space. For affine observed maps, the positive Fisher eigenspace contains all likelihood-sensitive parameter directions. Our bounds show that two types of neural surrogate-induced error, forward and directional-derivative errors, control the discrepancy between surrogate- and simulator-based maximum-likelihood solutions; under identifiability, this discrepancy vanishes with both errors. Value-only training directly constrains only the former, whereas FINO targets both.

The empirical benefit is problem dependent. At comparable forward accuracy, FINO reduces Navier–Stokes parameter RMSE by approximately 43% and data residual by 65% relative to FNO, but provides no comparable gain on Darcy. Controlled experiments support spectral decay as one mechanism behind this problem dependence. A preliminary seismic study further shows that sparse directional supervision is feasible at high dimension. **Limitations.** The exact analysis is affine; nonlinear applications require local Jacobian stability and amortized-to-local subspace alignment. Derivative labels incur offline cost and depend on the observation model. Future work should examine adaptive bases, trajectory-level gradient capture, and matched-cost comparisons with reduced-basis derivative-informed operators.

AI USE STATEMENT

During preparation of the work, the authors used generative AI (Claude, Anthropic; Codex, OpenAI) tools to refine sentence structure, improve readability and get support with the implementation code related tasks. The main theoretical derivations and results are not based on Generative AI tools. The authors have reviewed, verified and edited all AI-assisted text and implementation changes and take full responsibility for the manuscript.

ETHICS STATEMENT

We do not identify any ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

The main paper and Appendices F, G, H, I, J, and K provide the definitions, assumptions, derivations, and complete proofs of the theoretical results, and Appendix D relates the Fisher construction to existing sensitivity subspaces. Appendices B and C describe the benchmark PDEs, the Fisher-basis construction, the training and inversion protocol, all hyperparameters, and the full experimental results, including the seismic case study and recorded computational costs. Specifically, Appendix B.1 defines the controlled spectral-decay experiment, and Appendix E gives a closed-form two-mode example that illustrates the mechanism analyzed in the theory. Together, these materials provide the information needed to reproduce both the theoretical and empirical results reported in this work.

An anonymous Git repository containing the implementation of FINO and of all comparison methods, together with the data-generation, training, and inversion scripts, is available at `https://anonymous.4open.science/r/FINO_MWE-7EB4`.

All models were trained and evaluated on a single NVIDIA DGX node equipped with eight NVIDIA A100 GPUs (40 GB each) and dual-socket AMD EPYC processors (64 cores per socket, 512 GB of RAM per socket).

REFERENCES

Alexandros Beskos, Mark Girolami, Shiwei Lan, Patrick E. Farrell, and Andrew M. Stuart. Geometric mcmc for infinite-dimensional inverse problems. *Journal of Computational Physics*, 335:327–351, April 2017. ISSN 0021-9991. doi: 10.1016/j.jcp.2016.12.041. URL http://dx.doi.org/10.1016/j.jcp.2016.12.041.

Lianghao Cao, Thomas O'Leary-Roseberry, and Omar Ghattas. Derivative-informed neural operator acceleration of geometric MCMC for infinite-dimensional Bayesian inverse problems. *Journal of Machine Learning Research*, 26(78):1–68, 2025. URL https://jmlr.org/papers/v26/24-0745.html.

T. Chai and R. R. Draxler. Root mean square error (rmse) or mean absolute error (mae)? – arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7:1247–1250, 2014. doi: 10.5194/gmd-7-1247-2014.

Paul G. Constantine, Eric Dow, and Qiqi Wang. Active subspace methods in theory and practice: Applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–A1524, 2014. doi: 10.1137/130916138. URL https://arxiv.org/abs/1304.2070.

Paul G. Constantine, Carson Kent, and Tan Bui-Thanh. Accelerating Markov chain Monte Carlo with active subspaces. *SIAM Journal on Scientific Computing*, 38(5):A2779–A2805, 2016. doi: 10.1137/15M1042127. URL https://arxiv.org/abs/1510.00024.

Tiangang Cui and Olivier Zahm. Data-free likelihood-informed dimension reduction of Bayesian inverse problems. *Inverse Problems*, 37(4):045009, 2021. doi: 10.1088/1361-6420/abeafb. URL https://arxiv.org/abs/2102.13245.

Tiangang Cui, James Martin, Youssef M. Marzouk, Antti Solonen, and Alessio Spantini. Likelihood-informed dimension reduction for nonlinear inverse problems, 2014. URL https://arxiv.org/abs/1403.4680v2. arXiv:1403.4680, version 2.

Wojciech Marian Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Świrszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/file/758a06618c69880a6cee5314ee42d52f-Paper.pdf.

Jinwoo Go and Peng Chen. Accurate, scalable, and efficient bayesian optimal experimental design with derivative-informed neural operators. *Computer Methods in Applied Mechanics and Engineering*, 438:117845, 2025. doi: 10.1016/j.cma.2025.117845.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4 edition, 2013. URL https://www.press.jhu.edu/books/title/10678/matrix-computations.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011. doi: 10.1137/090771806. URL https://arxiv.org/abs/0909.4061.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991. doi: 10.1017/CBO9780511840371.

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

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM/IMS Journal of Data Science*, 1(3), 2024. doi: 10.1145/3648506.

M. Louboutin, M. Lange, F. Luporini, N. Kukreja, P. A. Witte, F. J. Herrmann, P. Velesko, and G. J. Gorman. Devito (v3.1.0): an embedded domain-specific language for finite differences and geophysical exploration. *Geoscientific Model Development*, 12(3):1165–1187, 2019. doi: 10.5194/gmd-12-1165-2019. URL https://www.geosci-model-dev.net/12/1165/2019/.

Mathias Louboutin, Philipp Witte, Ziyi Yin, Henryk Modzelewski, Kerim, Carlos da Costa, and Peterson Nogueira. slimgroup/judi.jl: v3.2.3, March 2023. URL https://doi.org/10.5281/zenodo.7785440.

Fabio Luporini, Mathias Louboutin, Michael Lange, Navjot Kukreja, Philipp Witte, Jan Hückelheim, Charles Yount, Paul H. J. Kelly, Felix J. Herrmann, and Gerard J. Gorman. Architecture and performance of devito, a system for automated stencil computation. *ACM Trans. Math. Softw.*, 46(1), apr 2020. ISSN 0098-3500. doi: 10.1145/3374916. URL https://doi.org/10.1145/3374916.

Per-Gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. A randomized algorithm for the decomposition of matrices. *Applied and Computational Harmonic Analysis*, 30(1):47–68, 2011. doi: 10.1016/j.acha.2010.02.003. URL https://www.sciencedirect.com/science/article/pii/S1063520310000242.

Thomas O'Leary-Roseberry, Peng Chen, Umberto Villa, and Omar Ghattas. Derivative-informed neural operator: An efficient framework for high-dimensional parametric derivative learning, 2023. URL https://arxiv.org/abs/2206.10745v4. arXiv:2206.10745, version 4.

Jeongjin Park, Huseyin Tuna Erdinc, Grant Bruer, Richard Rex, N. Chandramoorthy, and Felix J. Herrmann. Neural operators with accurate jacobian for high-fidelity image-domain seismic inversion. In *International Meeting for Applied Geoscience and Energy*. SEG/AAPG, 2026.

Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3):550–591, 2018. doi: 10.1137/16M1082469.

Yuan Qiu, Nolan Bridges, and Peng Chen. Derivative-enhanced deep operator network. In *Advances in Neural Information Processing Systems*, 2024. URL https://neurips.cc/virtual/2024/poster/94851.

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

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

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, 2026. URL https://arxiv.org/abs/2512.14086v2. arXiv:2512.14086, version 2.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. doi: 10.1093/biomet/asv008. URL https://arxiv.org/abs/1405.0680.

Olivier Zahm, Paul G. Constantine, Clémentine Prieur, and Youssef M. Marzouk. Gradient-based dimension reduction of multivariate vector-valued functions. *SIAM Journal on Scientific Computing*, 42(1):A534–A558, 2020. doi: 10.1137/18M1221837. URL https://arxiv.org/abs/1801.07922.

APPENDIX CONTENTS

A Notation **14**

B Benchmark PDEs and experiment setup **14**

B.1 Controlled spectral-decay toy in Figure 4 14

B.2 PDE definitions 16

B.3 Field and observation metrics 17

B.4 Data, training, and inversion protocol 17

B.5 Score-probe Fisher construction and label cost 19

B.6 Hyperparameter setting 20

C Results **21**

C.1 Inversion Results from Table 1 21

C.2 Spectral and inverse-trajectory diagnostics 21

C.3 Realistic Inverse Problem: Seismic Imaging 26

C.4 Comparison study: field-sensitivity vs Fisher-sensitivity 29

C.5 Computational cost and amortization 31

D General sensitivity constructions **33**

E A two-mode advection–diffusion example **33**

F Field–measurement geometry **34**

G Inverse rank and the bias–variance tradeoff **35**

G.1 Proof of Proposition 1 35

H Proof of the recovery theorem **35**

I Why value fitting is insufficient **36**

J Additional spectral consequences **37**

K From amortized training to local accuracy **37**

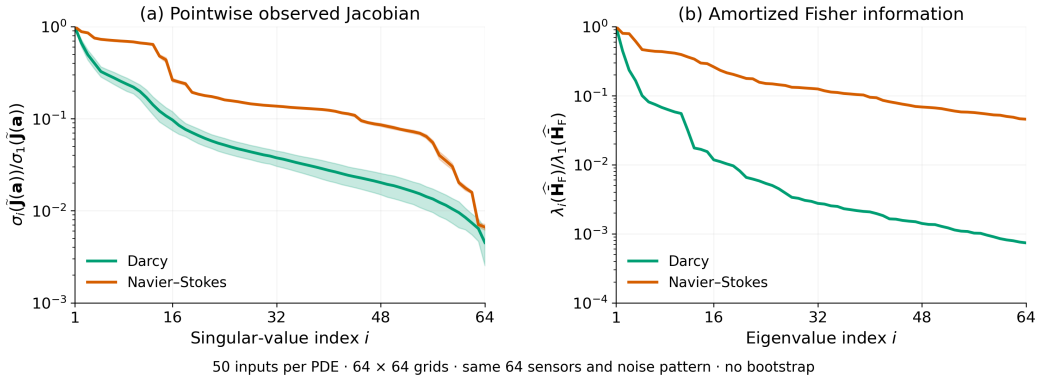


Figure 6: Pointwise and amortized observation sensitivities for Darcy and Navier–Stokes on 64×64 grids, using 50 inputs per PDE and the same 64 sensor locations and noise model. **(a)** Singular values of the noise-whitened observed Jacobian, normalized by the leading singular value at each input. Lines and shaded bands show the mean and one standard deviation across inputs. **(b)** Eigenvalues of the Fisher matrix averaged over all 50 inputs, normalized by its leading eigenvalue. Both panels display the first 64 modes. Navier–Stokes exhibits slower spectral decay than Darcy under this observation model, indicating that observation sensitivity is distributed across more parameter directions.

A Notation

Symbol	Meaning
$\boldsymbol{m}, \boldsymbol{a}, \boldsymbol{L}$	Physical input, sampling coordinates, and covariance root; $\boldsymbol{m} = \boldsymbol{m}_0 + \boldsymbol{L}\boldsymbol{a}$, $\boldsymbol{C} = \boldsymbol{L}\boldsymbol{L}^\top$.
$\mathcal{F}, \boldsymbol{J}_\mathcal{F}$	Full-field map and its Jacobian $\boldsymbol{J}_\mathcal{F} = (D_{\boldsymbol{m}}\mathcal{F}_{\text{phys}})\boldsymbol{L}$.
$\mathcal{R}, \boldsymbol{\Sigma}_{\text{obs}}$	Observation map and noise covariance.
$\widetilde{\boldsymbol{J}}, \boldsymbol{r}, \mathcal{L}$	$\boldsymbol{\Sigma}_{\text{obs}}^{-1/2}\mathcal{R}\boldsymbol{J}_\mathcal{F}$, $\boldsymbol{\Sigma}_{\text{obs}}^{-1/2}(\mathcal{R}\mathcal{F} - \boldsymbol{d}_{\text{obs}})$, and likelihood $\|\boldsymbol{r}\|^2/2$.
$\mu, \boldsymbol{M}_{\text{out}}$	input distribution and full-field metric.
$\boldsymbol{H}_{\text{field}}, \bar{\boldsymbol{H}}_{\text{F}}$	full-field sensitivity and amortized Fisher matrices in (2); hats denote empirical estimates.
$\boldsymbol{H}_{\text{F}}(\boldsymbol{a}), \lambda_i$	Pointwise Fisher matrix $\widetilde{\boldsymbol{J}}(\boldsymbol{a})^\top\widetilde{\boldsymbol{J}}(\boldsymbol{a})$ and its ordered eigenvalues at the simulator estimate.
$r, \boldsymbol{V}, \boldsymbol{Q}$	Inverse rank, local Fisher basis, and complementary projector $\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^\top$.
$\varepsilon_{\text{nn,forward}}, \varepsilon_{\text{sim}}, \varepsilon_{\text{nn,Fisher}}$	Observed forward error, simulator residual, and retained observed-Jacobian error.
$d_{\text{in}}, d_{\text{out}}, d_{\text{obs}}, N, M, N_L$	Input/output/observation dimensions, training count, construction count, and number of fields receiving derivative labels.
$m, \ell, \hat{\boldsymbol{V}}_m$	Supervision rank, number of observation-space probes per construction input, and acquired amortized Fisher basis.
$\lambda_D, E_{\text{train}}$	Derivative-loss weight and centered training-field energy.
$\|\cdot\|, \|\cdot\|_F, \|\cdot\|_{\text{op}}$	Euclidean, Frobenius, and induced operator norms.

B Benchmark PDEs and experiment setup

This section shows the physical benchmark definitions and experiment setup used in the main paper. All reported parameter errors are in the physical domain, $\boldsymbol{m}$, not in coordinates, $\boldsymbol{a}$.

B.1 Controlled spectral-decay toy in Figure 4

Fix orthogonal $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{16\times 16}$, shared across every seed and both regimes below, and rank $r = 12$ (a 4-dimensional exact null space). For decay rate $\rho \in \{\rho_{\text{slow}}, \rho_{\text{steep}}\} = \{0.08, 1.06\}$, let $\sigma_i(\rho) =$

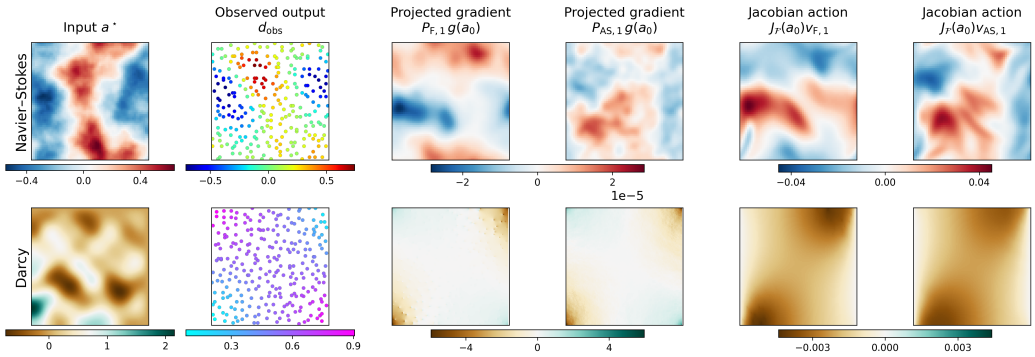


Figure 7: **Field and Fisher geometries select different sensitivity directions.** Representative Navier–Stokes and Darcy inputs are shown by row. Columns display the input, noisy observations, the initial likelihood gradient projected onto the leading Fisher and full-field sensitivity subspaces, and the full-field Jacobian actions along the leading Fisher and full-field sensitivity directions. Color limits are shared within each corresponding pair. The figure visualizes how the output metric changes the selected parameter directions and their field responses; it does not by itself imply that either basis yields superior inversion performance.

$e^{-\rho(i-1)}$ for $i = 1, \dots, r$, and define the mildly nonlinear map

$$\mathcal{F}_\rho(\boldsymbol{a}) = \boldsymbol{U}_{:,1:r}\,\mathrm{diag}(\boldsymbol{\sigma}(\rho))\,\mathcal{C}\left(\boldsymbol{V}_{:,1:r}^\top \boldsymbol{a}\right), \qquad \mathcal{C}(\boldsymbol{z}) = \boldsymbol{z} + \gamma \boldsymbol{z}^{\circ 2}, \quad \gamma = 0.1,$$

with $\boldsymbol{z}^{\circ 2}$ the elementwise square and $\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{16})$. Observations are $\boldsymbol{d}_{\text{obs}} = \mathcal{F}_\rho(\boldsymbol{a}^\star) + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma_{\text{obs}}^2 \boldsymbol{I}_{16})$, $\sigma_{\text{obs}} = 0.005$. Because the noise is isotropic, the sensitivity matrix $\boldsymbol{H}[\mathcal{F}_\rho, \sigma_{\text{obs}}^{-2}\boldsymbol{I}; \mu]$ of Section 2.2 has $\boldsymbol{V}_{:,1:r}$ as its exact top-r eigenvectors, with eigenvalues $\sigma_i(\rho)^2/\sigma_{\text{obs}}^2$; the top-$m$ supervision directions are therefore the first m columns of $\boldsymbol{V}$, known analytically rather than estimated. Write $P_r = \boldsymbol{V}_{:,1:r}\boldsymbol{V}_{:,1:r}^\top$ and $P_{\text{null}} = \boldsymbol{I} - P_r$.

A surrogate $\mathcal{F}_{\text{nn}}$ (two hidden tanh layers, width 64) is trained by minimizing

$$\mathcal{L}_{\text{train}}(m) = \mathbb{E}\|\mathcal{F}_{\text{nn}}(\boldsymbol{a}) - \mathcal{F}_\rho(\boldsymbol{a})\|^2 + \lambda \cdot \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}\,\|D\mathcal{F}_{\text{nn}}(\boldsymbol{a})\boldsymbol{v}_j - D\mathcal{F}_\rho(\boldsymbol{a})\boldsymbol{v}_j\|^2, \qquad \boldsymbol{v}_j = \boldsymbol{V}_{:,j},$$

with derivative weight $\lambda = 0.8$; $m = 0$ (empty sum) is the value-only baseline. For each $m \in \{0, 1, 2, 4, 6, 8, 12\}$ and 3 independent training seeds, otherwise-identical networks are trained on 400 samples (200 held out for validation).

For 64 inverse targets per seed (with the same truth, noise, and initialization reused across every m), each full-space surrogate inverse estimate $\widehat{\boldsymbol{a}}_m$ minimizes $\|\mathcal{F}_{\text{nn},m}(\boldsymbol{a}) - \boldsymbol{d}_{\text{obs}}\|^2$ by 500 Adam steps from $\boldsymbol{a}_0 \sim 0.1 \cdot \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{16})$. We report the observable- and null-space recovery errors $\|\boldsymbol{P}_r(\widehat{\boldsymbol{a}}_m - \boldsymbol{a}^\star)\|_2$, $\|\boldsymbol{P}_{\text{null}}(\widehat{\boldsymbol{a}}_m - \boldsymbol{a}^\star)\|_2$, and the surrogate/simulator likelihood-gradient relative error at $\boldsymbol{a}_0$,

$$E_{\text{grad}}(m) = \frac{\|\nabla \ell_{\text{nn,m}}(\boldsymbol{a}_0) - \nabla \ell^\star(\boldsymbol{a}_0)\|}{\|\nabla \ell^\star(\boldsymbol{a}_0)\|}, \qquad \ell_{\text{nn,m}}(\boldsymbol{a}) = \|\mathcal{F}_{\text{nn},m}(\boldsymbol{a}) - \boldsymbol{d}_{\text{obs}}\|^2, \quad \ell^\star(\boldsymbol{a}) = \|\mathcal{F}_\rho(\boldsymbol{a}) - \boldsymbol{d}_{\text{obs}}\|^2.$$

Panel (b) of Figure 4 reports paired differences in $E_{\text{rec}}(m) := \|\boldsymbol{P}_r(\widehat{\boldsymbol{a}}_m - \boldsymbol{a}^\star)\|_2$; panel (c) reports paired differences in $E_{\text{grad}}(m)$. Each comparison reuses $\boldsymbol{a}^\star$, $\boldsymbol{\epsilon}$, and $\boldsymbol{a}_0$ across ranks; differences are averaged first within seeds and then across the three seeds.

Shaded bands are 95% confidence intervals built in two steps: within each of the 3 training seeds, we first average that seed's 64 paired differences down to a single seed-level mean, then compute the interval from the mean and standard error of those 3 seed-level means, using a Student-t critical value appropriate for only 2 degrees of freedom (t=4.303) rather than a large-sample normal approximation.

Under slow decay, recovery improves monotonically over the tested ranks, while gradient-error improvement is non-monotone. Steep decay peaks in benefit near rank 4 (close to its noise-floor crossover, σ_i = noise at $i = 6$) and then erodes as additional, near-noise-floor directions are

supervised—consistent with those directions degrading the surrogate's overall value fit, not only its derivative accuracy. With only 3 seeds (df=2), these intervals are wide; treat the point estimates' qualitative trend (monotonic for slow, peak-then-erode for steep) as the primary evidence, not the interval width alone.

B.2 PDE DEFINITIONS

Transport geometry diagnostic. The transport diagnostic underlying Figure 17 is posed on the periodic domain $[0, 2\pi)^2$ and represents the spatially distributed input in physical grid coordinates. The observation operator samples the solution at the fixed sensor locations shown in the figure. The full-field and Fisher sensitivity constructions use identical inputs, sensors, and rank; only the output metric differs. The isotropic case uses a common sensor variance, whereas the anisotropic case uses sensor standard deviations 0.01 and 0.05, indicated by the two marker sizes. This controlled diagnostic isolates the effect of observation noise on the selected parameter subspace.

Initial-vorticity Navier–Stokes. On the periodic domain $[0, 2\pi)^2$, the initial vorticity m determines the terminal vorticity through

$$\begin{aligned} \partial_t \omega + \boldsymbol{v} \cdot \nabla \omega = 100^{-1} \Delta \omega + f, \qquad & \boldsymbol{v} = (\partial_{x_2} \psi, -\partial_{x_1} \psi), \\ -\Delta \psi = \omega, \qquad & \omega(\boldsymbol{x}, 0) = m(\boldsymbol{x}), \end{aligned} \tag{11}$$

with $f(\boldsymbol{x}) = 0.1[\sin(x_1 + x_2) + \cos(x_1 + x_2)]$. The forward map returns $\omega(\cdot, 2)$ on a 64^2 Fourier grid, using Crank–Nicolson/Heun time stepping with $\Delta t = 0.01$. Initial conditions are drawn from the zero-mean Gaussian random field with parameters $\alpha = 2.5$ and $\tau = 7$ used by the simulator; the associated covariance root, including its null constant Fourier mode, defines $\boldsymbol{L}$. The restored recovery configuration observes 256 fixed terminal-vorticity values with independent noise standard deviation 0.01.

Darcy parameterization. The numerical forward model is discretized on a 128×128 grid. Each log-permeability field supplied to FNO and FINO is generated from 64 KL/cosine coefficients:

$$\boldsymbol{a} = \boldsymbol{B}\boldsymbol{z}, \qquad \boldsymbol{B} \in \mathbb{R}^{16{,}384 \times 64}, \qquad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{64}).$$

FNO and FINO therefore retain a full-grid neural-operator representation, but the inverse problem optimizes $\boldsymbol{z} \in \mathbb{R}^{64}$ rather than 16,384 independent field values.

FINO's derivative labels are constructed in the operator-input coordinates. Specifically,

$$\boldsymbol{V}_{\text{field}} \in \mathbb{R}^{16{,}384 \times 200}, \qquad \boldsymbol{T} = D_{\boldsymbol{a}} \mathcal{F}_{\text{phys}}(\boldsymbol{a}) \boldsymbol{V}_{\text{field}}.$$

Thus, the 200 supervised directions are linearly independent field-space directions, not KL-coordinate directions or randomized score probes. During inversion, the optimized objective is

$$\min_{\boldsymbol{z} \in \mathbb{R}^{64}} \mathcal{L}_\theta(\boldsymbol{B}\boldsymbol{z}),$$

and its gradient satisfies

$$\nabla_{\boldsymbol{z}} \mathcal{L}_\theta = \boldsymbol{B}^\top \nabla_{\boldsymbol{a}} \mathcal{L}_\theta.$$

DINO's rank-16 input representation introduces a further reduction within this 64-dimensional inverse parameterization.

Pullback to the inverse coordinates. Let

$$\widetilde{\boldsymbol{J}}_{\boldsymbol{a}}(\boldsymbol{a}) := \boldsymbol{\Sigma}_{\text{obs}}^{-1/2} \boldsymbol{R} D_{\boldsymbol{a}} \mathcal{F}(\boldsymbol{a})$$

denote the noise-whitened Jacobian with respect to the ambient field input. Because $\boldsymbol{a}(\boldsymbol{z}) = \boldsymbol{B}\boldsymbol{z}$, the chain rule gives the inverse-coordinate Jacobian

$$\widetilde{\boldsymbol{J}}_{\boldsymbol{z}}(\boldsymbol{z}) = \widetilde{\boldsymbol{J}}_{\boldsymbol{a}}(\boldsymbol{B}\boldsymbol{z}) \boldsymbol{B}.$$

Consequently, the corresponding pointwise Fisher matrices satisfy

$$\begin{aligned} \boldsymbol{H}_{\boldsymbol{z}}(\boldsymbol{z}) &:= \widetilde{\boldsymbol{J}}_{\boldsymbol{z}}(\boldsymbol{z})^\top \widetilde{\boldsymbol{J}}_{\boldsymbol{z}}(\boldsymbol{z}) \\ &= \boldsymbol{B}^\top \boldsymbol{H}_{\boldsymbol{a}}(\boldsymbol{B}\boldsymbol{z}) \boldsymbol{B}, \qquad \boldsymbol{H}_{\boldsymbol{a}}(\boldsymbol{a}) := \widetilde{\boldsymbol{J}}_{\boldsymbol{a}}(\boldsymbol{a})^\top \widetilde{\boldsymbol{J}}_{\boldsymbol{a}}(\boldsymbol{a}). \end{aligned}$$

Since $\boldsymbol{B}$ is fixed, the amortized Fisher matrices similarly satisfy

$$\overline{\boldsymbol{H}}_{\boldsymbol{z}} = \boldsymbol{B}^\top \overline{\boldsymbol{H}}_{\boldsymbol{a}} \boldsymbol{B}.$$

Thus, the field-space Fisher matrix governs sensitivity with respect to perturbations of the full operator input, whereas $\overline{\boldsymbol{H}}_{\boldsymbol{z}}$ governs sensitivity in the 64 coordinates optimized during Darcy inversion. The 200 supervised directions and the ratio $200/16{,}384$ refer to the former.

B.3 Field and observation metrics

The full-field metric includes spatial quadrature weights: $\boldsymbol{M}_{\mathrm{out}} = (2\pi/64)^2 \boldsymbol{I}$ for scalar Navier–Stokes output and $\boldsymbol{M}_{\mathrm{out}} = 64^{-2}\boldsymbol{I}$ for the Darcy grid. The Fisher metric uses the observations and noise of the corresponding inverse problem, $\mathcal{R}^\top \boldsymbol{\Sigma}_{\mathrm{obs}}^{-1} \mathcal{R}$.

For construction inputs $\boldsymbol{a}_i$, define

$$\widehat{\boldsymbol{H}}_{\mathrm{field}} = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{J}_i^\top \boldsymbol{M}_{\mathrm{out}} \boldsymbol{J}_i, \qquad \widehat{\boldsymbol{H}}_{\mathrm{F}} = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{J}_i^\top \mathcal{R}^\top \boldsymbol{\Sigma}_{\mathrm{obs}}^{-1} \mathcal{R} \boldsymbol{J}_i,$$

where $\boldsymbol{J}_i = D_{\boldsymbol{a}}\mathcal{F}(\boldsymbol{a}_i) = (D_{\boldsymbol{m}}\mathcal{F}_{\mathrm{phys}})\boldsymbol{L}$. Both matrices are estimated using the same construction inputs and coordinates, so Figures 7 and 17 isolate how the output metric changes the selected parameter subspace.

Pointwise and amortized spectra are not interchangeable. At any fixed input, $\operatorname{rank} \boldsymbol{H}_{\mathrm{F}}(\boldsymbol{a}) \le d_{\mathrm{obs}}$, but the rank of $\widehat{\boldsymbol{H}}_{\mathrm{F}}$ can exceed d_{obs} because observable subspaces may rotate across construction inputs. Accordingly, plots of sample-averaged pointwise singular values are labeled as local Jacobian spectra; plots used to diagnose the acquired FINO basis report eigenvalues of the empirical amortized matrices above.

B.4 Data, training, and inversion protocol

FNO and FINO share the full-field backbone, value data, optimizer, and training budget.DINO uses the reduced-basis formulation described in Section 3. No inverse target or noisy observation is used for basis construction, training, or checkpoint selection.

Checkpoints are selected solely by held-out forward error; FNO and FINO validation errors differ by less than 0.2% relatively, whereas DINO and PINO are not matched to that range. After selection, weights are frozen and methods are evaluated on paired targets, observations, noise, and optimization budgets. FNO and FINO use the same inverse variables within each benchmark. Navier–Stokes uses direct field coordinates, whereas Darcy optimizes the 64 KL/cosine coefficients defining its full-grid operator input (Beskos et al., 2017). DINO instead performs inversion in its native learned reduced input coordinates. Ground truth parameter error is used only for evaluation, never for optimization or stopping.

Forward metrics are relative L^2 and spatial H^1 errors. Inverse metrics are physical-parameter RMSE and SSIM and the simulator-evaluated data residual defined in Section 5. Recovery and data fit are reported separately because a low observation residual need not identify the physical parameter. Main-table uncertainty is across inverse targets; the synthetic rank study reports uncertainty across training seeds.

Table 2 summarizes the parameter dimensions, derivative-supervision ranks, and inversion budgets. Although both inputs are high-dimensional fields, the supervised Fisher directions constitute less than 10% of the Navier–Stokes parameter space and less than 1.3% of the ambient operator-input dimension for Darcy.

Derivative-loss normalization. In the reported experiments, the effective weight in (5) is set to $\lambda_{\mathrm{der}} = \lambda_D^{\mathrm{raw}}/E_{\mathrm{train}}$, where

$$\bar{\boldsymbol{u}}_{\mathrm{train}} := \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{u}_n, \qquad E_{\mathrm{train}} := \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{u}_n - \bar{\boldsymbol{u}}_{\mathrm{train}}\|_2^2.$$

Table 2: **Operator-input and inverse-coordinate dimensions.** The supervision ratio $m_{\text{field}}/d_{\text{field}}$ refers to directions in the discretized field supplied to the neural operator; it is not the fraction of inverse variables supervised.

Setting	Navier–Stokes	Darcy
Operator-input resolution	64×64	128×128
Field-input dimension, d_{field}	4096	16,384
Field-space supervision rank, m_{field}	400	200
Ambient dimension ratio, $m_{\text{field}}/d_{\text{field}}$	9.77%	1.22%
Inverse-coordinate dimension, d_{inv}	4096	64
Inversion iterations	2500	2000

Table 3: **Notation and training configurations.** Panel (a) distinguishes the number of value samples, derivative-labeled inputs, Fisher-basis construction samples, score probes, and retained supervision directions. Panel (b) reports the configurations used for the two PDE benchmarks.

(a) Definitions

Quantity	Meaning
N	Number of value-training input–output pairs.
N_L	Number of training inputs carrying derivative labels; this is not the number of supervised directions.
M	Number of parameter samples used to construct the amortized Fisher basis.
m	Number of retained Fisher directions supervised during training.
ℓ	Number of randomized score probes per basis-construction sample; this is distinct from m, M, and the number of sensors.
λ_D^{raw}	Coefficient multiplying the derivative term before normalization. Its value is meaningful only together with the normalization in the training loss.

(b) Benchmark configurations

Setting	Navier–Stokes	Darcy
Discretization	64×64	128×128
Value-training pairs, N	400	400
Derivative-labeled inputs, N_L	400	400
Supervision rank, m	400	200
Directional derivative labels, $N_L m$	160,000	80,000
Fisher-basis parameter samples, M	200	200
Score probes per basis sample, ℓ	405	205
Raw derivative coefficient, λ_D^{raw} (`reg_param`)	10^{-4}	10^{-3}
Batch size per GPU	100	50
Number of GPUs	2	1
Effective global batch size	200	50
Optimizer	AdamW	AdamW
Learning rate	8×10^{-4}	8×10^{-4}
Weight decay	10^{-5}	10^{-5}
Configured training duration	500 epochs	200 epochs
Selected checkpoint epoch (zero-based)	210	198
Optimizer updates at selected checkpoint	422	1,592

Each derivative-labeled input contributes targets in m directions, so $N_L m$ is the total number of directional derivative labels. Basis construction uses $M\ell$ score-probe VJPs under the notation of Section 3.

Here λ_D^{raw} is the unnormalized coefficient and $E_{\text{train}} > 0$ measures variation in the simulated training outputs. Both statistics are computed once from the training set and remain fixed during optimization. This is algebraically equivalent to using $\lambda_D^{\text{raw}}/(N_L E_{\text{train}})$ in front of the derivative-error sum.

Observations and inversion (Darcy). The observation operator extracts pressure at the same 256 saved sensor locations for every target and method. The clean comparison uses the saved noise-free

observations. Each inversion starts from $\boldsymbol{\theta}_0 = \mathbf{0}$ and runs 2,000 ordinary fixed-step gradient-descent updates with $\eta = 0.1$, retaining the final iterate. The numerical objective is

$$\mathcal{L}_{\mathrm{num}}(\boldsymbol{\theta}) = \frac{1}{2n_{\mathrm{obs}}s_{\mathcal{L}}^2}\|\mathcal{R}\mathcal{F}_{\mathrm{phys}}(\boldsymbol{K}\boldsymbol{\theta}) - \boldsymbol{d}_{\mathrm{obs}}\|_2^2, \qquad n_{\mathrm{obs}} = 256, \quad s_{\mathcal{L}} = 0.01,$$

with the surrogate replacing $\mathcal{F}_{\mathrm{phys}}$ during surrogate inversion.

Coordinate-dependent diagnostics (Darcy). For $\boldsymbol{m} = \boldsymbol{K}\boldsymbol{\theta}$, gradients satisfy $\nabla_{\boldsymbol{\theta}}\mathcal{L} = \boldsymbol{K}^\top \nabla_{\boldsymbol{m}}\mathcal{L}$. A nonzero physical-field gradient can therefore coexist with a small gradient in the actual optimization coordinates. Physical-field and KL-coordinate diagnostics are reported separately. The common-sensor spectral comparison is a separate 64×64, 64-sensor diagnostic, not the discretization or observation mask of this 128×128 inversion benchmark.

B.5 Score-probe Fisher construction and label cost

Let $\boldsymbol{h}(\boldsymbol{a}) := \mathcal{R}\mathcal{F}(\boldsymbol{a})$ and

$$\widetilde{\boldsymbol{J}}(\boldsymbol{a}) := \boldsymbol{\Sigma}_{\mathrm{obs}}^{-1/2} D_{\boldsymbol{a}}\boldsymbol{h}(\boldsymbol{a}).$$

Under the Gaussian observation model,

$$\boldsymbol{z} := \boldsymbol{\Sigma}_{\mathrm{obs}}^{-1/2}\big(\boldsymbol{y} - \boldsymbol{h}(\boldsymbol{a})\big) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d_{\mathrm{obs}}}).$$

The conditional score is

$$\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{y}) := \nabla_{\boldsymbol{a}} \log p(\boldsymbol{y} \mid \boldsymbol{a}) = \widetilde{\boldsymbol{J}}(\boldsymbol{a})^\top \boldsymbol{z}, \tag{12}$$

and therefore

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{a}}\left[\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{y})\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{y})^\top\right] = \widetilde{\boldsymbol{J}}(\boldsymbol{a})^\top \widetilde{\boldsymbol{J}}(\boldsymbol{a}) = \boldsymbol{H}_{\mathrm{F}}(\boldsymbol{a}). \tag{13}$$

Thus, Gaussian score probes provide unbiased rank-one samples of the pointwise Fisher matrix.

For construction inputs $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M \overset{\mathrm{iid}}{\sim} \mu$, draw independent matrices

$$\boldsymbol{Z}_i = [\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{i\ell}] \in \mathbb{R}^{d_{\mathrm{obs}} \times \ell}, \qquad \boldsymbol{z}_{ik} \overset{\mathrm{iid}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d_{\mathrm{obs}}}),$$

and define

$$\boldsymbol{q}_{ik} := \widetilde{\boldsymbol{J}}(\boldsymbol{a}_i)^\top \boldsymbol{z}_{ik}, \qquad \boldsymbol{Q} := \frac{1}{\sqrt{M\ell}}[\boldsymbol{q}_{11}, \ldots, \boldsymbol{q}_{M\ell}].$$

The resulting estimator is

$$\widehat{\boldsymbol{H}}_{\mathrm{F}} := \boldsymbol{Q}\boldsymbol{Q}^\top = \frac{1}{M\ell}\sum_{i=1}^{M}\sum_{k=1}^{\ell} \boldsymbol{q}_{ik}\boldsymbol{q}_{ik}^\top. \tag{14}$$

Conditioned on the construction inputs,

$$\mathbb{E}_{\boldsymbol{Z}}\left[\widehat{\boldsymbol{H}}_{\mathrm{F}} \mid \boldsymbol{a}_1, \ldots, \boldsymbol{a}_M\right] = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{H}_{\mathrm{F}}(\boldsymbol{a}_i), \tag{15}$$

and consequently

$$\mathbb{E}_{\boldsymbol{a},\boldsymbol{Z}}\left[\widehat{\boldsymbol{H}}_{\mathrm{F}}\right] = \bar{\boldsymbol{H}}_{\mathrm{F}}. \tag{16}$$

If $\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{W}^\top$, then

$$\widehat{\boldsymbol{H}}_{\mathrm{F}} = \boldsymbol{U}\boldsymbol{S}^2\boldsymbol{U}^\top.$$

Hence the acquired Fisher basis is

$$\hat{\boldsymbol{V}}_m := \boldsymbol{U}_{:,1:m}.$$

A thin or randomized SVD may be used to compute these vectors (Halko et al., 2011; Martinsson et al., 2011). The observation-probe count ℓ controls Monte Carlo accuracy; it is not the oversampling parameter of a randomized SVD.

Each $\boldsymbol{q}_{ik}$ requires one VJP with observation-space cotangent

$$\mathcal{R}^\top \boldsymbol{\Sigma}_{\mathrm{obs}}^{-1/2}\boldsymbol{z}_{ik}.$$

Table 4: Neural-network architectures for Navier–Stokes (64×64 discretization). FNO and FINO share the same architecture but differ in training loss. Parameter counts include trainable real scalars and exclude fixed coordinate features, input bases, and POD bases.

Setting	FNO	FINO	DINO	PINO
Backbone	PhysicsNeMo FNO	PhysicsNeMo FNO	Reduced residual MLP	Coordinate FNO
Input field	Initial vorticity	Initial vorticity	Initial vorticity	Initial vorticity
Output field	Final vorticity	Final vorticity	Final vorticity	Final vorticity
Physical input/output channels	$1/1$	$1/1$	$1/1$	$1/1$
Input/output reduction ranks	None	None	$16/64$	None
Fourier blocks	5	5	–	3
Retained Fourier modes per block	32×32	32×32	–	12×12
Hidden width	64	64	256	16
Additional coordinate channels	2	2	0	4
Input lifting / first affine map	$3 \to 32 \to 64$	$3 \to 32 \to 64$	$16 \to 256$	$5 \to 16$
Output decoder / last affine map	$64 \to 128 \to 1$	$64 \to 128 \to 1$	$256 \to 64$ + fixed POD	$16 \to 1$
Hidden activation	GELU	GELU	ELU	$\tanh$
Decoder hidden activation	SiLU	SiLU	–	–
Spatial padding	0	0	–	0
Dropout / batch normalization	None	None	None	None
Derivative-supervision directions	0	400	16	0
Trainable parameters used in inversion	83,917,569	83,917,569	152,384	443,297
Total trainable parameters during training	83,917,569	83,917,569	152,384	1,773,188

Table 5: Neural-network architectures for Darcy (128×128 discretization). The parameter field is log-permeability, and the predicted state is the hydraulic-head field. Parameter counts follow Table 4.

Setting	FNO	FINO	DINO	PINO
Backbone	PhysicsNeMo FNO	PhysicsNeMo FNO	Reduced residual MLP	Coordinate FNO
Input field	Log-permeability	Log-permeability	Log-permeability	Log-permeability
Output field	Hydraulic head	Hydraulic head	Hydraulic head	Hydraulic head
Physical input/output channels	$1/1$	$1/1$	$1/1$	$1/1$
Input/output reduction ranks	None	None	$16/64$	None
Fourier blocks	4	4	–	4
Retained Fourier modes per block	20×20	20×20	–	20×20
Hidden width	64	64	256	64
Additional coordinate channels	2	2	0	4
Input lifting / first affine map	$3 \to 32 \to 64$	$3 \to 32 \to 64$	$16 \to 256$	$5 \to 64$
Output decoder / last affine map	$64 \to 256 \to 1$	$64 \to 256 \to 1$	$256 \to 64$ + fixed POD	$64 \to 1$
Hidden activation	GELU	GELU	ELU	$\tanh$
Decoder hidden activation	SiLU	SiLU	–	–
Spatial padding	0	0	–	0
Dropout / batch normalization	None	None	None	None
Derivative-supervision directions	0	200	16	0
Trainable parameters	26,250,177	26,250,177	152,384	26,231,489

Thus, basis construction uses $M\ell$ VJP directions. If the physical parameterization is $\boldsymbol{m} = \boldsymbol{m}_0 + \boldsymbol{L}\boldsymbol{a}$, then

$$\boldsymbol{q}_{ik} = \boldsymbol{L}^\top D_{\boldsymbol{m}} \mathcal{F}_{\mathrm{phys}}(\boldsymbol{m}_i)^\top \mathcal{R}^\top \boldsymbol{\Sigma}_{\mathrm{obs}}^{-1/2} \boldsymbol{z}_{ik}. \tag{17}$$

After fixing $\hat{\boldsymbol{V}}_m$, generating the full-output directional targets in (4) requires $N_L m$ JVP directions and stores $N_L d_{\mathrm{out}} m$ derivative entries. These operation counts distinguish score-probe VJPs for basis construction from full-output JVPs for label generation.

B.6 Hyperparameter setting

FNO and FINO. Both methods use the same PhysicsNeMo FNO architecture within each benchmark. Two spatial coordinates in $[0, 1]^2$ are concatenated with the scalar input field. The lifting network comprises two pointwise affine maps with an intervening GELU activation. Each Fourier block adds a spectral convolution to a learned pointwise linear transformation. GELU follows every Fourier block except the last. The pointwise output decoder has one SiLU hidden layer and a linear scalar output. The archived FNO and FINO models use unnormalized physical input and output fields.

Table 6: **Inverse recovery with and without observation noise.** Values are mean ± sample standard deviation over 100 unseen inverse targets paired across methods. The noisy observations use $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ with $\sigma = 0.01$, corresponding to 1% of the RMS training-observation scale; $\sigma = 0$ denotes noiseless observations. Checkpoints, targets, initializations, and inversion settings are unchanged between noise levels. Validation L^2 errors are included to document checkpoint selection. Bold indicates the best inverse-recovery mean for each PDE and noise level.

			Noiseless, $\sigma = 0$			Noisy, $\sigma = 0.01$		
PDE	Method	Validation $L^2 \downarrow$	RMSE ↓	SSIM ↑	Residual ↓	RMSE ↓	SSIM ↑	Residual ↓
Navier–Stokes	FNO	0.05947	0.1873 ± 0.0323	0.2909 ± 0.0801	0.1403 ± 0.0244	0.1873 ± 0.0323	0.2909 ± 0.0801	0.1403 ± 0.0244
	DINO	0.27604	0.1430 ± 0.0272	0.4091 ± 0.1008	0.0952 ± 0.0127	0.1430 ± 0.0272	0.4091 ± 0.1008	0.0952 ± 0.0127
	PINO	0.03776	0.1226 ± 0.0323	0.4868 ± 0.1111	0.0562 ± 0.0117	0.1226 ± 0.0323	0.4866 ± 0.1111	0.0563 ± 0.0117
	FINO	0.05939	$\mathbf{0.1071 \pm 0.0294}$	$\mathbf{0.5734 \pm 0.1183}$	$\mathbf{0.0490 \pm 0.0067}$	$\mathbf{0.1071 \pm 0.0294}$	$\mathbf{0.5733 \pm 0.1182}$	$\mathbf{0.0491 \pm 0.0067}$
Darcy	FNO	0.009236	$\mathbf{0.1997 \pm 0.0344}$	$\mathbf{0.5549 \pm 0.0887}$	0.00487 ± 0.00110	$\mathbf{0.2146 \pm 0.0352}$	$\mathbf{0.5176 \pm 0.0876}$	$\mathbf{0.00736 \pm 0.00077}$
	DINO	0.008804	0.2724 ± 0.0359	0.3559 ± 0.0890	$\mathbf{0.00456 \pm 0.00081}$	0.2739 ± 0.0372	0.3536 ± 0.0901	0.00737 ± 0.00062
	PINO	0.008558	0.2113 ± 0.0387	0.5314 ± 0.0846	0.00457 ± 0.00255	0.2266 ± 0.0394	0.4910 ± 0.0871	$\mathbf{0.00736 \pm 0.00217}$
	FINO	0.009222	0.2094 ± 0.0394	0.5422 ± 0.0792	0.00530 ± 0.00116	0.2259 ± 0.0419	0.4960 ± 0.0837	0.00767 ± 0.00090

DINO. DINO (O'Leary-Roseberry et al., 2023; Qiu et al., 2024) uses a rank-16 input basis and a rank-64 POD output basis. Its reduced network has three hidden layers of width 256, ELU activations, and one residual connection:

$$\boldsymbol{h}_1 = \mathrm{ELU}(\boldsymbol{W}_1 \boldsymbol{z} + \boldsymbol{b}_1), \qquad \boldsymbol{h}_2 = \mathrm{ELU}(\boldsymbol{W}_2 \boldsymbol{h}_1 + \boldsymbol{b}_2),$$
$$\boldsymbol{h}_3 = \mathrm{ELU}(\boldsymbol{h}_1 + \boldsymbol{W}_3 \boldsymbol{h}_2 + \boldsymbol{b}_3), \qquad \widehat{\boldsymbol{c}} = \boldsymbol{W}_4 \boldsymbol{h}_3 + \boldsymbol{b}_4.$$

The full-field prediction is

$$\widehat{\boldsymbol{u}} = \bar{\boldsymbol{u}} + s_u \boldsymbol{U}_{\mathrm{POD}} \widehat{\boldsymbol{c}}, \qquad s_u = \max\left\{ \left(\frac{1}{N d_{\mathrm{out}}} \sum_{i=1}^{N} \|\boldsymbol{u}_i - \bar{\boldsymbol{u}}\|_2^2 \right)^{1/2}, 10^{-8} \right\},$$

where the mean, scalar scale, and POD basis are computed from training outputs only. For Navier–Stokes, the physical input is whitened by the prior-root pseudoinverse and projected onto the rank-16 input basis. For Darcy, the encoder is $\boldsymbol{z} = \boldsymbol{V}_{\mathrm{AS}}^{\top} \boldsymbol{K}^{\dagger} \boldsymbol{m}$, where $\boldsymbol{K} \in \mathbb{R}^{128^2 \times 64}$ maps KL coefficients to the log-permeability field and $\boldsymbol{V}_{\mathrm{AS}} \in \mathbb{R}^{64 \times 16}$ is the input basis. This encoding is linear; it does not apply a tanh transformation.

PINO. PINO (Li et al., 2024) uses a coordinate-augmented Fourier network with a pointwise linear lifting layer, spectral-plus-pointwise hidden blocks with tanh activations, and a single linear output projection. For Navier–Stokes, the four coordinate features are $(\sin x_1, \cos x_1, \sin x_2, \cos x_2)$. Four independent networks predict the states at $t \in \{0.5, 1, 1.5, 2\}$, while the state at $t = 0$ is the given input. Each network uses width 16, three Fourier blocks, and 12 retained modes per spatial axis; inversion uses the terminal-time network. The Darcy network uses width 64, four Fourier blocks, and 20 modes per axis. Its coordinate features are $(\xi_1, \xi_2, \xi_1^2, \xi_2^2)$, with $\xi_j = 2(j + \frac{1}{2})/128 - 1$ for grid indices $j = 0, \ldots, 127$; input and output normalization constants are computed from training data. Thus, FNO and FINO form the architecture-matched comparison; DINO and PINO are architectural references rather than additional architecture-matched ablations.

C Results

C.1 Inversion Results from Table 1

Figures 8–11 compare the parameters recovered by four surrogate methods and the likelihood gradients evaluated at the initial and final inversion iterates. The fourth and fifth rows evaluate the surrogate and numerical-simulator gradients at the same method-specific final iterate, thereby isolating gradient disagreement from differences in the evaluation point.

C.2 Spectral and inverse-trajectory diagnostics

Fisher spectral decay. Figure 6 compares Navier–Stokes and Darcy under a common 64-sensor diagnostic. Under this shared observation and noise model, the Navier–Stokes sensitivity spectrum

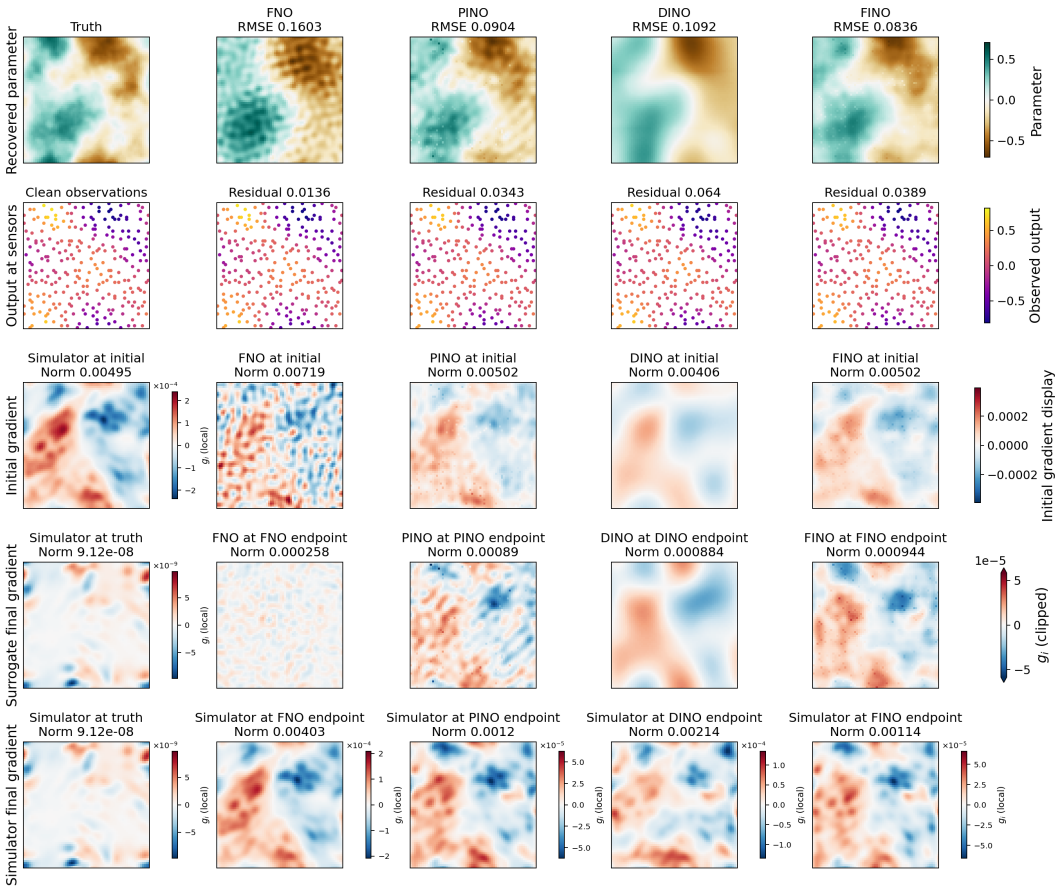


Figure 8: **Representative Navier–Stokes inversion.** Columns: truth or simulator, then FNO, PINO, DINO, and FINO. *Row 1:* recovered initial vorticity with parameter RMSE; FINO is lowest (0.0836), FNO highest (0.1603). *Row 2:* predicted outputs at the 256 sensors with data residual; FNO fits the data best (0.0136) while recovering the parameter worst. *Row 3:* likelihood gradient at the common initial iterate under the simulator (left) and each surrogate; FNO's gradient is high-frequency noise with norm 0.00719 against the simulator's 0.00495, whereas FINO's matches in structure and norm (0.00502). *Rows 4 and 5:* gradient at each method's own endpoint, under the surrogate (row 4, shared clipped scale) and under the simulator at the same point (row 5); the left column shows the simulator at the truth. FNO's own gradient nearly vanishes (0.000258) while the simulator gradient at its endpoint is the largest of all methods (0.00403), so FNO stops where the true likelihood is not stationary. At FINO's endpoint the two gradients agree (0.000944 and 0.00114) and the simulator gradient is the smallest. Norms are Euclidean over the grid. Values are for this target; Table 1 reports means over 100 targets.

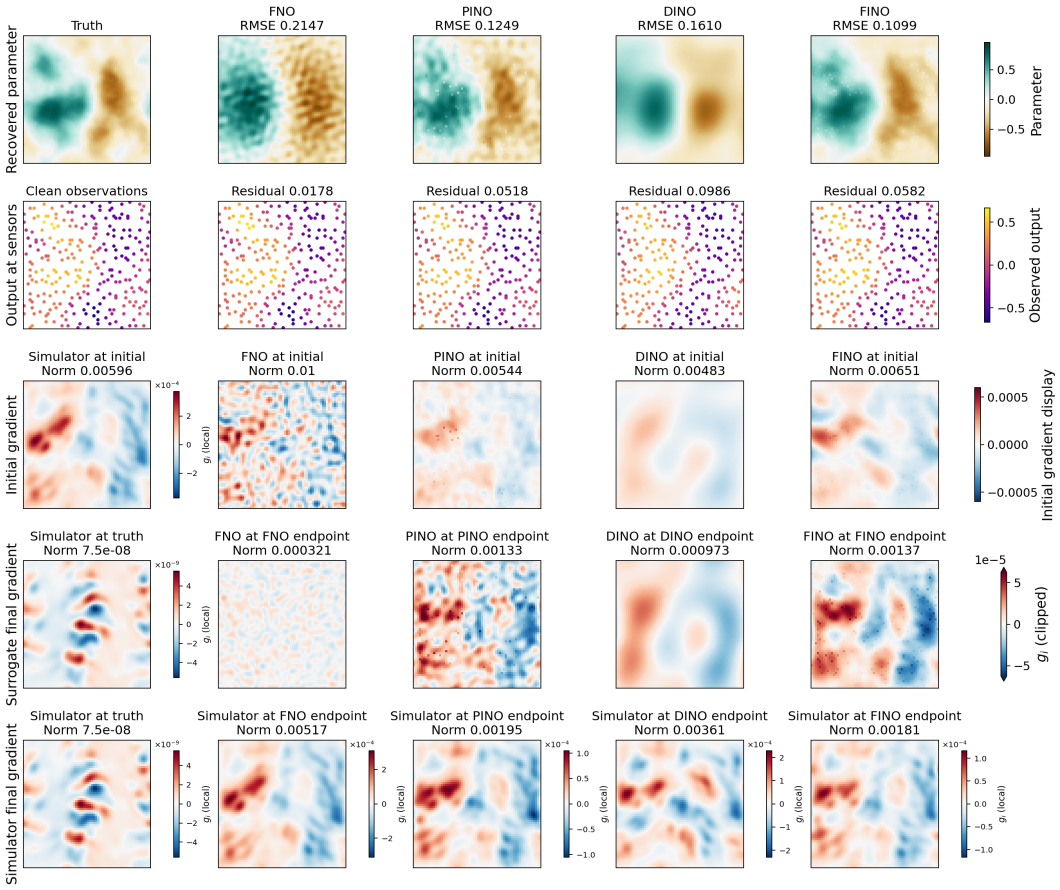


Figure 9: **Second Navier–Stokes inversion example.** Layout as in Figure 8. The same pattern holds on a different target: FINO recovers the parameter best (RMSE 0.1099 versus 0.2147 for FNO) while FNO again has the smallest data residual (0.0178). At the initial iterate, FNO's gradient norm (0.01) is nearly double the simulator's (0.00596) and is spatially noisy; FINO's (0.00651) tracks the simulator. At the endpoints, FNO's own gradient is 0.000321 but the simulator gradient there is 0.00517, the largest of all methods; at FINO's endpoint the surrogate and simulator gradients are 0.00137 and 0.00181, the closest pair and the smallest simulator gradient.

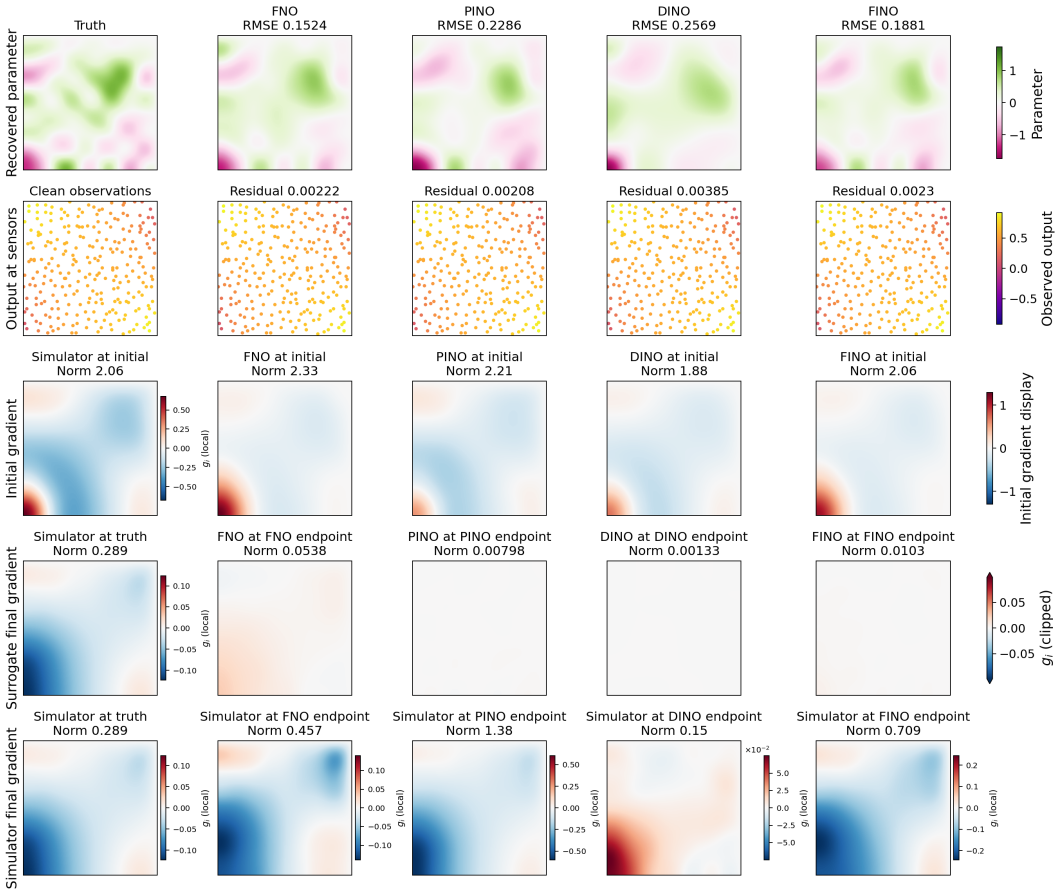


Figure 10: **Representative Darcy inversion.** Layout as in Figure 8, with recovered log-permeability in row 1 and pressure at the 256 sensors in row 2; inversion optimizes the 64 KL coefficients. Unlike Navier–Stokes, FNO recovers the parameter best (RMSE 0.1524), and FNO, PINO, and FINO reach nearly the same data residual (0.0021 to 0.0023). All methods underestimate the amplitude of the truth. *Row 3:* at the initial iterate every surrogate already reproduces the simulator gradient in spatial structure, and FINO's norm equals the simulator's (2.06): the leading observation-sensitive directions are captured by value-only training, so Fisher-directed supervision has little to add. *Rows 4 and 5:* at each method's own endpoint the surrogate gradient has collapsed (0.0013 to 0.054) while the simulator gradient at the same point remains large (0.15 to 1.38) and exceeds its value at the truth. All four surrogates, FINO included, stop at points that are not stationary for the true likelihood, and FINO's endpoint gradient under the simulator (0.709) is not smaller than FNO's (0.457). Surrogate stationarity therefore does not certify recovery on this problem, and the remaining differences between methods are not explained by initial-gradient accuracy. Table 1 reports means over 100 targets.

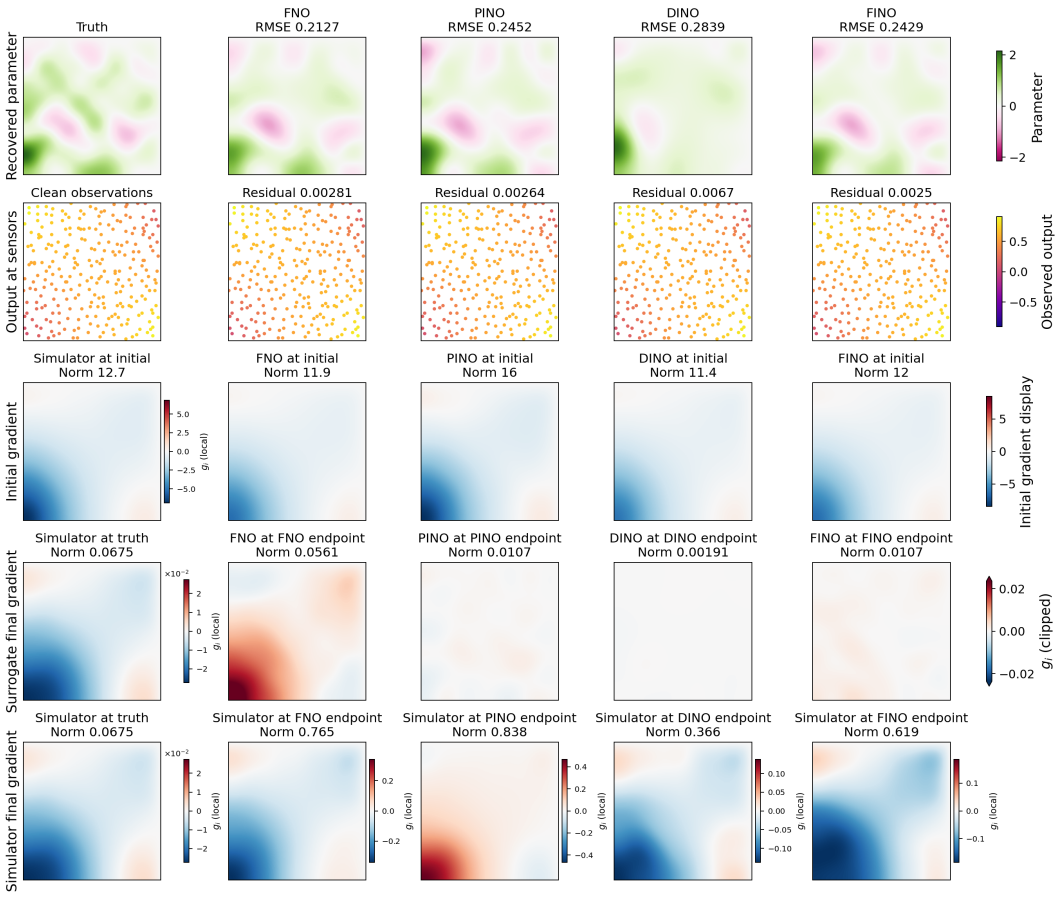


Figure 11: **Second Darcy inversion example.** Layout as in Figure 10. FNO again recovers the parameter best (RMSE 0.2127), with FINO attaining the smallest data residual (0.0025). The initial gradients of all surrogates match the simulator's (norms 11.4 to 16 against 12.7), and at every endpoint the surrogate gradient is small (0.0019 to 0.056) while the simulator gradient is large (0.37 to 0.84). The picture is the same as in Figure 10: derivative supervision changes neither the initial gradient agreement nor the endpoint mismatch.

decays more slowly, indicating that inverse-relevant sensitivity is distributed across more parameter directions. This diagnostic is separate from the benchmark inversion configurations and is intended to compare spectral structure rather than recovery performance directly.

Figure 4 isolates this mechanism using synthetic maps with the same dimension, singular vectors, and null space but different spectral decay. Relative to value-only training, recovery improves through larger supervision ranks under slow decay, whereas the steep-decay improvement is largest near $m = 4$ and then weakens. Likelihood-gradient error follows a different rank dependence. Equation (9) explains that improvements in directional derivative accuracy are weighted by the corresponding Fisher eigenvalues, while Theorem 1 connects directional error to likelihood-solution recovery. Neither result guarantees monotone improvement with m, because forward error, weakly informed directions, and optimization behavior may also change. Because the experiment uses only three training seeds, its intervals support a qualitative spectral-decay conclusion rather than a precise optimal-rank claim.

Inverse trajectories. Figures 13 and 15 evaluate each surrogate and the numerical simulator at states along that surrogate's own inverse trajectory. On Navier–Stokes, FINO's lower recovery error accompanies smaller likelihood-gradient disagreement than FNO's. On Darcy, the surrogate gradient can become small while the simulator gradient remains nonzero; FINO does not obtain a corresponding recovery gain. Figures 13 and 15 provide the corresponding gradient-norm comparisons.

These diagnostics illustrate why surrogate stationarity or a low data residual cannot alone certify parameter recovery. They probe the nonlinear mechanisms represented by the full-space terms in Theorem 1, but they are not direct measurements of its affine bound. Because each method follows its own trajectory, comparisons across methods are made at different parameter states except at their common initialization.

Diagnostic conventions. All trajectory diagnostics summarize the same 100 paired inverse targets using one fixed checkpoint per method. At each saved iterate, the surrogate and numerical-simulator quantities are evaluated at the same parameter state on that method's own trajectory. The initial state is shared across methods, but subsequent states generally differ; cross-method trajectory comparisons are therefore not fixed-state ablations. Unless stated otherwise, curves show means and bands show one sample standard deviation across targets. Paired-difference panels instead show mean within-target differences with pointwise 95% Student-t intervals; negative values favor FINO.

C.3 Realistic Inverse Problem: Seismic Imaging

We applied Fisher-guided derivative supervision to image-domain seismic inversion, an exploration workflow that repeatedly evaluates wave-equation imaging operators (Park et al., 2026). The surrogate maps a velocity model and a smooth background model to a reverse-time-migration (RTM) image. The averaged Born Gauss–Newton matrix is used to construct the derivative label, which admits a Fisher interpretation under Gaussian noise. Let $\mathcal{F}_{\mathrm{seis}}(\boldsymbol{m})$ map a discretized velocity model to seismic receiver data and let $\boldsymbol{B}(\boldsymbol{m}) = D_{\boldsymbol{m}}\mathcal{F}_{\mathrm{seis}}(\boldsymbol{m})$ be the Born operator. The reported direction matrix is

$$\bar{\boldsymbol{G}} := \mathbb{E}_{\boldsymbol{m}\sim\mu}\big[\boldsymbol{B}(\boldsymbol{m})^\top \boldsymbol{B}(\boldsymbol{m})\big].$$

For the Gaussian data model $\boldsymbol{d} = \mathcal{F}_{\mathrm{seis}}(\boldsymbol{m}) + \boldsymbol{\eta}$ with $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \sigma_d^2 \boldsymbol{I})$, the amortized seismic-data Fisher matrix is $\sigma_d^{-2}\bar{\boldsymbol{G}}$. Thus the reported Gauss–Newton and data-Fisher bases have identical eigenvectors under this *modeling interpretation*.

The Fisher matrix of an *RTM-image* observation model is a different construction. Write an RTM map with a model-dependent Born operator as $\mathcal{G}_{\mathrm{RTM}}(\boldsymbol{m}) = \boldsymbol{B}(\boldsymbol{m})^\top[\boldsymbol{d} - \mathcal{F}_{\mathrm{seis}}(\boldsymbol{m})]$. With residual $\boldsymbol{r} = \boldsymbol{d} - \mathcal{F}_{\mathrm{seis}}(\boldsymbol{m})$, differentiation gives

$$D_{\boldsymbol{m}}\mathcal{G}_{\mathrm{RTM}} = \sum_i r_i \nabla^2_{\boldsymbol{m}}(\mathcal{F}_{\mathrm{seis}})_i - \boldsymbol{B}^\top\boldsymbol{B} \approx -\boldsymbol{B}^\top\boldsymbol{B},$$

where the approximation drops the residual-dependent curvature term. For a Gaussian image-error model with precision $\boldsymbol{\Sigma}_{\mathrm{img}}^{-1}$, this leads to the approximate amortized image-Fisher matrix

$$\bar{\boldsymbol{H}}_{\mathrm{img}} \approx \mathbb{E}_\mu\Big[(\boldsymbol{B}^\top\boldsymbol{B})\boldsymbol{\Sigma}_{\mathrm{img}}^{-1}(\boldsymbol{B}^\top\boldsymbol{B})\Big].$$

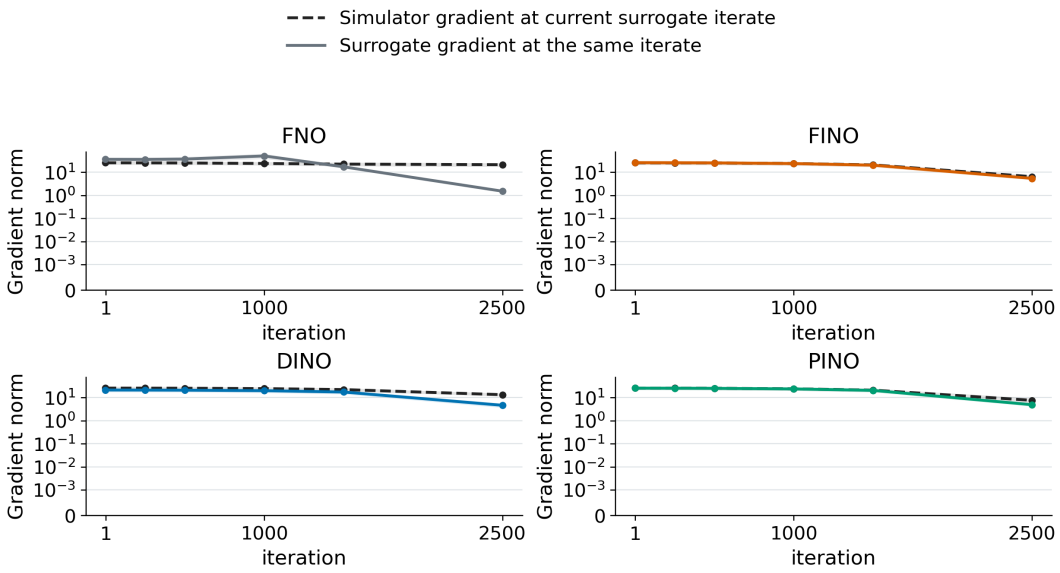


Figure 12: **Navier–Stokes simulator and surrogate gradient norms.** Dashed and solid curves respectively show the numerical-simulator and surrogate likelihood-gradient norms evaluated at the same saved states on each method's own inverse trajectory. FINO tracks the simulator gradient throughout the trajectory, whereas the FNO surrogate gradient falls well below the simulator gradient near termination. DINO and PINO show smaller late-trajectory discrepancies. Curves and bands denote mean $\pm$ sample standard deviation over 100 targets. Raw gradient norms are diagnostic quantities rather than constrained optimality residuals.

Table 7: **Seismic image-domain inversion over 20 different unseen parameter samples**

Method	Relative data residual ↓	SSIM ↑	Relative model error ↓	Model RMSE ↓
FNO	0.3047 ± 0.0262	0.3914 ± 0.0576	0.0944 ± 0.0061	0.2766 ± 0.0269
FINO ($m = 8$)	0.2746 ± 0.0281	0.3955 ± 0.0587	0.0924 ± 0.0073	0.2707 ± 0.0295
FINO ($m = 64$)	$\mathbf{0.2144 \pm 0.0272}$	$\mathbf{0.4329 \pm 0.0677}$	$\mathbf{0.0889 \pm 0.0075}$	$\mathbf{0.2608 \pm 0.0315}$

The accompanying derivation proposes probes $\boldsymbol{q} = (\boldsymbol{B}^\top \boldsymbol{B})\boldsymbol{W}_{\text{img}}\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$. Their expected outer product equals the expression above when $\boldsymbol{W}_{\text{img}}\boldsymbol{W}_{\text{img}}^\top = \boldsymbol{\Sigma}_{\text{img}}^{-1}$. If the image weight contains an exact mute, it instead defines a positive-semidefinite weighted sensitivity metric. In general, $\mathbb{E}[\boldsymbol{B}^\top \boldsymbol{B}]$ and $\mathbb{E}[(\boldsymbol{B}^\top \boldsymbol{B})\boldsymbol{\Sigma}_{\text{img}}^{-1}(\boldsymbol{B}^\top \boldsymbol{B})]$ need not share eigenvectors.

We used 256×256 discretized subsurface velocity models, having 202 training samples, and 22 test samples. Its two derivative-trained FNOs supervised 8 and 64 directions, approximately 0.012% and 0.098% of the 65,536 model coordinates. The finite-difference calculations and experiment set up was constructed using softwares such as Devito and JUDI Louboutin et al. (2023; 2019); Luporini et al. (2020).

During image-domain inversion, the background model was held fixed and the velocity model was updated for 1200 Adam steps from a smoothed version of the truth, subject to a total-variation constraint. For inversion, we randomly select 20 of the 22 held-out velocity models. Table 7 reports aggregate results over these 20 targets, while Figure 16 shows a representative reconstruction. FINO improves recovery relative to FNO across all reported metrics.

The study demonstrates the feasibility of sparse directional supervision in a 256×256 wave-equation imaging workflow relevant to exploration. The RTM operator, total-variation constraint, and inversion initialization also differ from the PDE experiments in this paper.

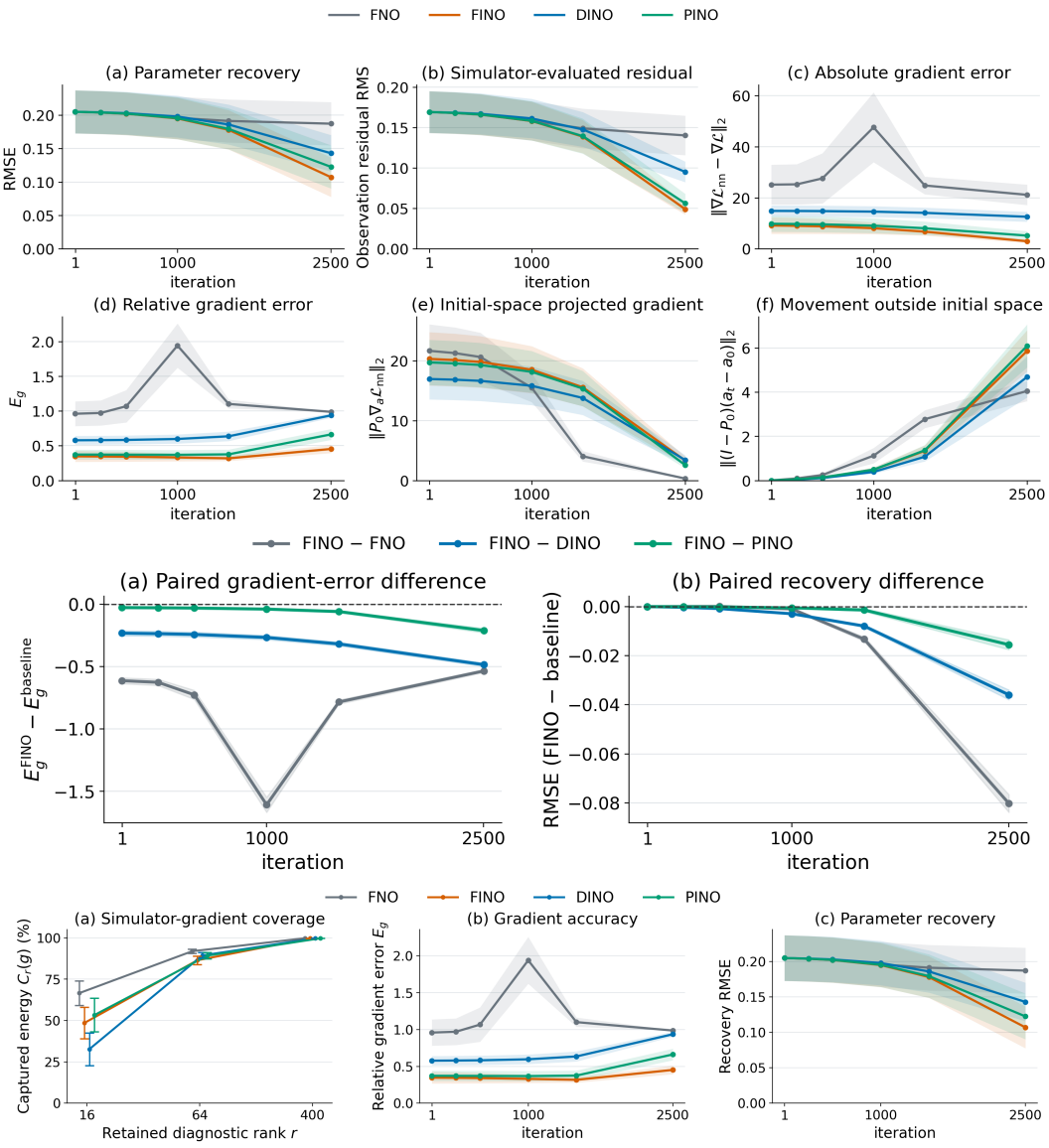


Figure 13: **Navier–Stokes inverse-trajectory diagnostics.** The top block reports parameter RMSE, simulator-evaluated observation residual, absolute and relative likelihood-gradient errors, gradient magnitude in the rank-16 local Fisher space at initialization, and movement outside that fixed space. FINO maintains the smallest likelihood-gradient disagreement and attains the lowest recovery error, whereas FNO develops substantially larger gradient error. The middle block shows paired FINO-minus-baseline differences in relative gradient error and recovery RMSE; negative values favor FINO. The bottom block reports simulator-gradient energy captured by nested ranks of the fixed FINO basis, together with the corresponding gradient-error and recovery trajectories. The diagnostic rank is varied only in post-processing and does not retrain any surrogate. These nonlinear diagnostics are consistent with, but are not direct measurements of, the affine recovery bound.

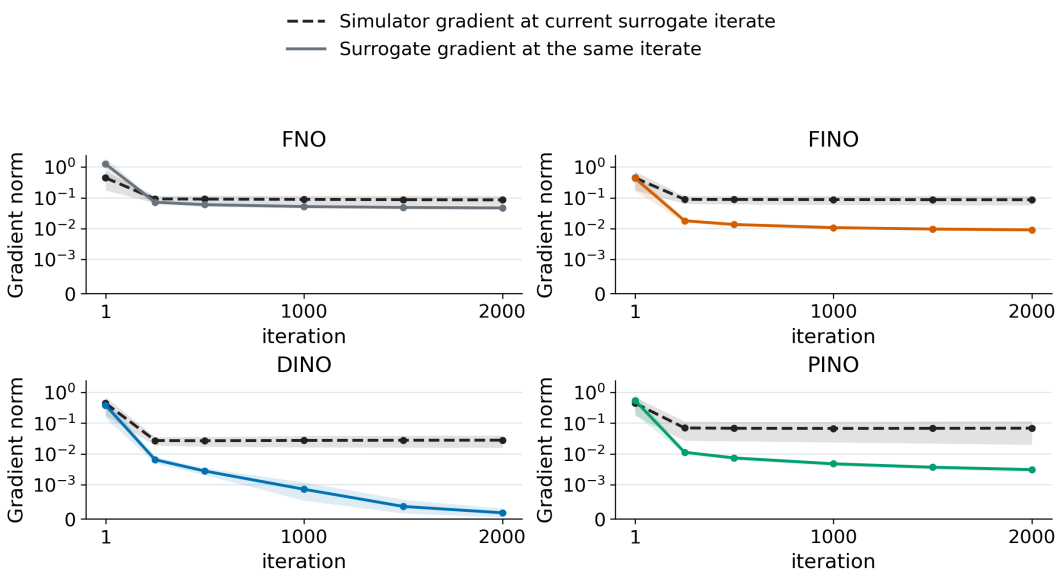


Figure 14: **Darcy physical-field gradient norms.** Dashed and solid curves respectively show simulator and surrogate likelihood-gradient norms with respect to the 128^2 physical log-permeability entries, evaluated at the same states on each method's own trajectory. For FINO, DINO, and PINO, the surrogate field-gradient norm becomes much smaller than the simulator gradient at the same state, demonstrating that surrogate stationarity need not imply simulator stationarity. These are field-space sensitivity diagnostics, not the gradients used by the 64-dimensional KL optimizer, which satisfy $\nabla_{\boldsymbol{z}}\mathcal{L} = \boldsymbol{B}^\top \nabla_{\boldsymbol{a}}\mathcal{L}$. Curves and bands denote mean $\pm$ sample standard deviation over 100 targets.

Table 8: **Fisher- versus field-sensitivity-based derivative supervision.** Each method uses its best saved validation checkpoint. Within each PDE and noise condition, the methods use identical inverse targets, observations, initializations, and optimization budgets. Recovery metrics are mean $\pm$ sample standard deviation over 100 targets for fixed trained networks; simulator residuals are evaluated using the numerical PDE solver. This is a separate matched direction-selection experiment and does not reuse the Table 1 checkpoints. Bold denotes the better mean within each PDE and condition.

			Clean observations			Noisy observations ($\sigma = 0.01$)		
PDE	Method	Test rel. L^2 error ↓	Parameter RMSE ↓	SSIM ↑	Simulator residual RMS ↓	Parameter RMSE ↓	SSIM ↑	Simulator residual RMS ↓
Navier–Stokes	FINO	0.03181	**0.15294 ± 0.03377**	**0.3819 ± 0.0962**	**0.09687 ± 0.02218**	**0.15292 ± 0.03377**	**0.3818 ± 0.0963**	**0.09688 ± 0.02217**
	Field-JVP	**0.01903**	0.17430 ± 0.03282	0.3149 ± 0.0835	0.12083 ± 0.02357	0.17431 ± 0.03281	0.3148 ± 0.0834	0.12086 ± 0.02353
Darcy	FINO	0.00726	0.20860 ± 0.03905	0.5446 ± 0.0796	0.00386 ± 0.00104	0.22513 ± 0.04116	0.5012 ± 0.0848	0.00676 ± 0.00070
	Field-JVP	**0.00336**	**0.18856 ± 0.03267**	**0.5915 ± 0.0768**	**0.00204 ± 0.00039**	**0.20578 ± 0.03501**	**0.5415 ± 0.0799**	**0.00584 ± 0.00029**

C.4 Comparison study: field-sensitivity vs Fisher-sensitivity

Interpretation. Table 8 illustrates that lower full-field prediction error does not necessarily imply better inverse recovery. On Navier–Stokes, Field-JVP attains a substantially lower test relative L^2 error than FINO (0.01903 versus 0.03181), yet FINO reduces parameter RMSE by 12.3% and simulator residual by 19.8%, while increasing SSIM by approximately 0.07 on the clean and noisy observations. This is consistent with the different roles of the two direction-selection metrics: Field-JVP prioritizes directions that strongly affect the complete solution field, whereas FINO prioritizes directions whose effects are visible through the observation operator and significant relative to the noise. The Darcy results show that this advantage is not universal: when the field-sensitive directions already capture the inverse-relevant structure, FINO need not improve recovery.

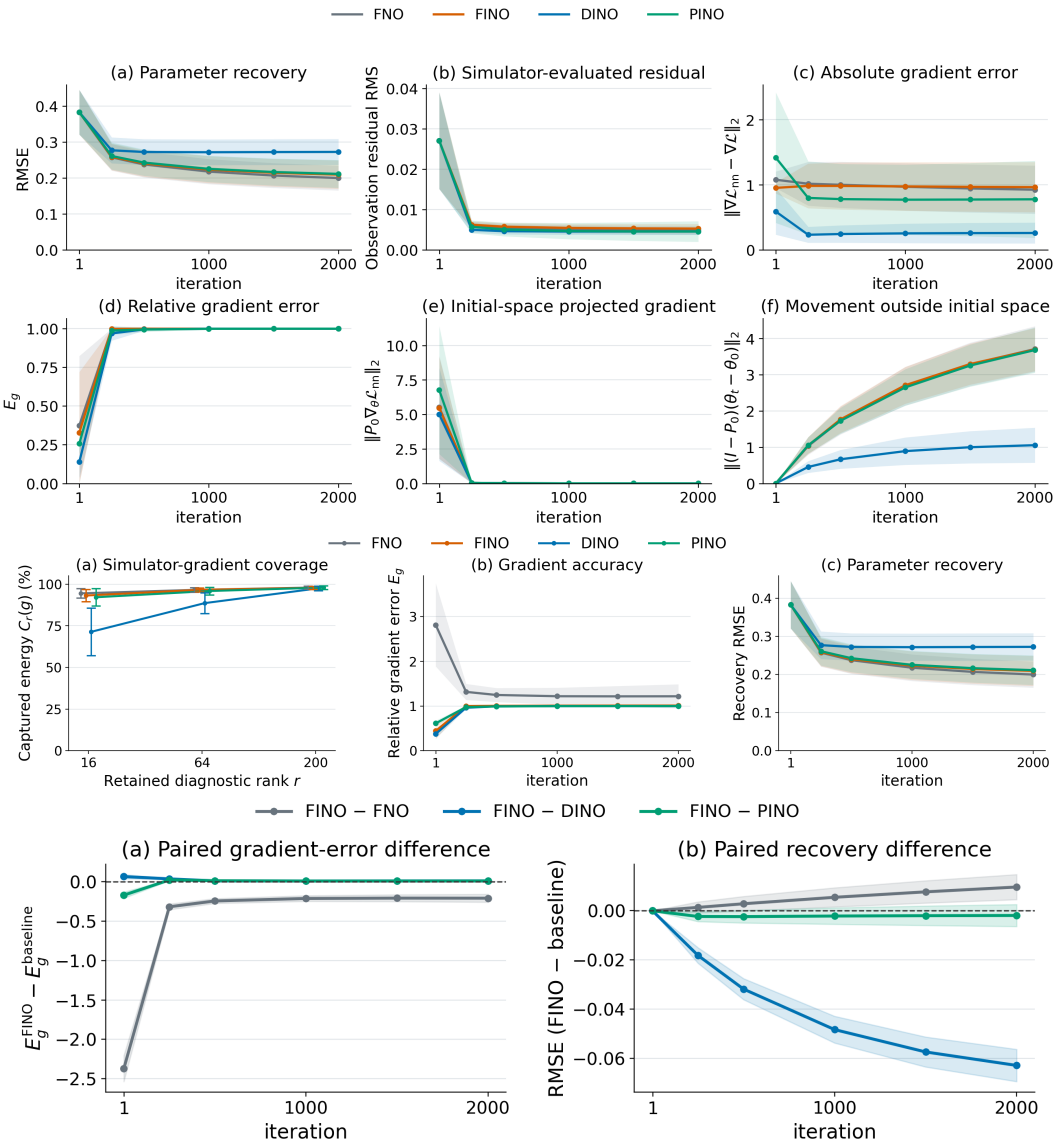
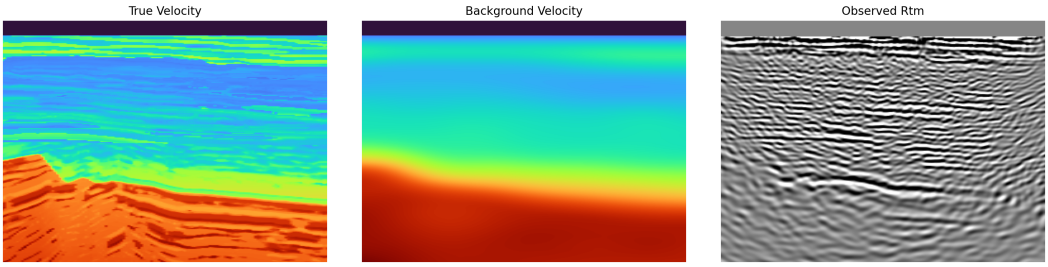


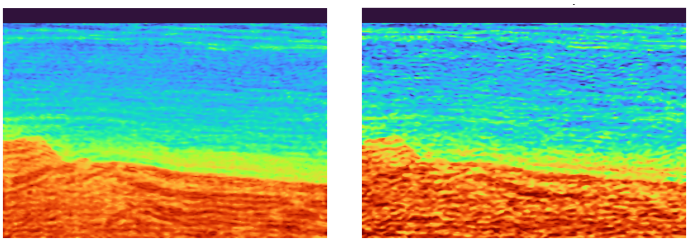
Figure 15: **Darcy inverse-trajectory diagnostics.** The top block reports trajectory quantities in the 64-dimensional KL coordinates used during inversion: parameter RMSE, simulator-evaluated observation residual, absolute and relative likelihood-gradient errors, projected gradient magnitude, and movement outside the fixed initial Fisher space. Although all methods rapidly reduce their surrogate objectives, the relative simulator–surrogate gradient error approaches one and the recovery ordering changes little. The middle block gives physical-field diagnostics: simulator-gradient energy captured by nested ranks of the fixed FINO field-space basis, relative physical-field gradient error, and recovery RMSE. These 128^2-dimensional gradients are not the KL-coordinate gradients used by the optimizer. The bottom block shows paired FINO-minus-baseline differences in physical-field gradient error and recovery RMSE; negative values favor FINO. High field-space subspace coverage alone therefore does not imply accurate optimizer gradients or parameter recovery.



(a) Ground truth subsurface velocity model

(b) Background velocity model which we use as a starting parameter

(c) Reverse Time Migration image serving as observation



(d) FINO reconstructed model

(e) FNO reconstructed model

Figure 16: **Representative seismic image-domain inversion.** Ground-truth and background velocity models, the observed RTM image, and the corresponding FNO and FINO reconstructions. Results show one out of twenty test cases; quantitative metrics are reported in Table 7. FINO trained with 64 different directional-derivatives, which accounts for 0.1% of parameter space, performed the best in inversion across all four metrics.

C.5 Computational cost and amortization

Table 9 describes per-input leading-order costs. Here n is the number of grid points, m and r_o are input and output basis sizes, w is the hidden width of the reduced-basis DINO network, and K is the number of saved states in PINO's PDE-residual loss. Input and output channel counts and the FNO width and depth are fixed, so their discretized dimensions scale as $O(n)$. Memory entries describe working arrays when all directions are processed together, not measured peak GPU usage. The full-Jacobian row is only a scaling reference. "Linearized solves" denotes tangent or adjoint sensitivity solves after linearization, not measured simulator calls. Basis construction is excluded from this table and included in the recorded costs below.

The measured study in Table 10 is separate from Table 1: its Navier–Stokes cohort uses 400 training inputs and $m = 400$, while the Helmholtz cohort uses 1024 inputs and $m = 24$. Data, basis, label, and fit times are disjoint and charged once per training seed. Fit time includes checks, checkpoint writes, and evaluation overhead; Navier–Stokes fit time is cumulative across continuations. PINO needs no Fisher basis, but its label entry counts the saved physics trajectories used by the PDE-residual loss.

For B inversions using one trained surrogate, let C_{offline} be its total offline time and t_{nn} and t_{sim} the mean times per surrogate and simulator inversion. The first integer at which surrogate use is no slower than direct simulation is

$$B^* = \begin{cases} \lceil C_{\text{offline}}/(t_{\text{sim}} - t_{\text{nn}}) \rceil, & t_{\text{nn}} < t_{\text{sim}}, \\ \infty, & t_{\text{nn}} \geq t_{\text{sim}}. \end{cases}$$

Table 11 applies this calculation to recorded runtimes. Each entry is mean $\pm$ sample SD across three training-seed means, each based on eight paired inverse targets. The Helmholtz and Navier–Stokes cohorts have different inversion setups; the latter uses zero initialization. The estimate assumes the same trained model and observation design can be reused. It compares elapsed time, not time to a common recovery error or convergence tolerance. Inversion timing excludes model loading and simulator diagnostics.

Table 9: **Per-input cost scaling by loss term.** FINO stores and trains on m full-output directional labels rather than the complete input–output Jacobian. For PINO, the final column is the PDE-residual loss, not derivative-label supervision. These leading-order bounds do not imply equal wall-clock costs across methods.

Method	Cost	Output loss	Derivative or PDE-residual loss
Full Jacobian (reference)	Data generation	1 PDE solve	$O(n)$ linearized solves
	Training memory	$O(n)$	$O(n^2)$
	Training computation	$O(n \log n)$	$O(n^2 \log n)$
FNO	Data generation	1 PDE solve	None
	Training memory	$O(n)$	—
	Training computation	$O(n \log n)$	—
DINO (O'Leary-Roseberry et al., 2023)	Data generation	1 PDE solve	$\min\{m, r_o\}$ linearized solves
	Training memory	$O(m + r_o + w)$	$O(mr_o + w(m + r_o))$
	Training computation	$O(w(m + r_o) + mr_o)$	$O(wmr_o)$
DIFNO (Yao et al., 2026)	Data generation	1 PDE solve	$\min\{m, r_o\}$ linearized solves
	Training memory	$O(n)$	$O(mn)$
	Training computation	$O(n \log n)$	$O(mn \log n + nmr_o)$
PINO (Li et al., 2024)	Data generation	1 PDE solve (K saved states)	None
	Training memory	$O(Kn)$	$O(Kn)$
	Training computation	$O(Kn \log n)$	$O(Kn \log n)$
FINO	Data generation	1 PDE solve	m linearized solves
	Training memory	$O(n)$	$O(mn)$
	Training computation	$O(n \log n)$	$O(mn \log n)$

Table 10: **Recorded offline costs.** Mean $\pm$ sample SD across three training seeds. NS denotes initial-vorticity Navier–Stokes. Total is the per-seed sum of data, basis, label, and fit costs, with seconds converted to hours. These are standalone costs, not matched-budget comparisons.

Cohort	Method	Data (s)	Basis (s)	Labels (s)	Fit (h)	Total (h)
NS	FNO	75.11 ± 1.04	0	0	0.298 ± 0.039	0.319 ± 0.039
NS	DINO (O'Leary-Roseberry et al., 2023)	75.11 ± 1.04	3287.21 ± 8.83	2352.22 ± 8.36	0.103 ± 0.002	1.691 ± 0.001
NS	DIFNO (Yao et al., 2026)	75.11 ± 1.04	3287.21 ± 8.83	2352.22 ± 8.36	12.103 ± 3.971	13.690 ± 3.971
NS	PINO (Li et al., 2024)	75.11 ± 1.04	0	97.68 ± 0.00	0.525 ± 0.014	0.573 ± 0.014
NS	FINO	75.11 ± 1.04	2716.95 ± 9.79	10890.08 ± 93.60	16.342 ± 6.336	20.143 ± 6.357

Table 11: **Recorded inversion times and runtime break-even.** B^* amortizes offline work against the measured mean simulator runtime; ∞ means no positive per-inversion time saving. Parameter RMSE is shown to expose the accuracy tradeoff. These are not equal-accuracy break-even guarantees or the main Table 1 checkpoints.

Method	Offline (h)	Inversion (s)	Parameter RMSE	B^*
Navier–Stokes, zero start				
FNO	0.319 ± 0.039	121.18 ± 49.96	0.1237 ± 0.0041	1
RB-DINO	1.691 ± 0.001	7.42 ± 2.32	0.1634 ± 0.0097	2
DIFNO	13.690 ± 3.971	79.29 ± 25.21	0.0790 ± 0.0146	15
FINO	20.143 ± 6.357	87.79 ± 26.31	0.0621 ± 0.0018	21
PINO	0.573 ± 0.014	119.19 ± 1.47	0.0812 ± 0.0093	1
Simulator	—	3556.20 ± 1050.49	0.0608 ± 0.0020	—

D General sensitivity constructions

Let $\mathcal{G} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^p$ be differentiable and let $\boldsymbol{W} \succeq 0$ satisfy $\mathbb{E}_\mu \|\boldsymbol{W}^{1/2} D\mathcal{G}\|_F^2 < \infty$. Define

$$\boldsymbol{H}[\mathcal{G}, \boldsymbol{W}; \mu] := \mathbb{E}_{\boldsymbol{a}\sim\mu} \left[D\mathcal{G}(\boldsymbol{a})^\top \boldsymbol{W} D\mathcal{G}(\boldsymbol{a}) \right].$$

For every direction $\boldsymbol{v}$,

$$\boldsymbol{v}^\top \boldsymbol{H}[\mathcal{G}, \boldsymbol{W}; \mu]\boldsymbol{v} = \mathbb{E}_\mu \|\boldsymbol{W}^{1/2} D\mathcal{G}(\boldsymbol{a})\boldsymbol{v}\|_2^2,$$

so the leading eigenvectors identify input directions with large average response in the chosen output metric (Zahm et al., 2020).

This construction includes scalar subspaces ($p = 1$, $\boldsymbol{W} = 1$) (Constantine et al., 2014); fixed-data misfit subspaces ($\mathcal{G} = \mathcal{L}$) (Constantine et al., 2016); unweighted observation subspaces ($\mathcal{G} = \mathcal{R}\mathcal{F}$, $\boldsymbol{W} = \boldsymbol{I}$); the full-field metric ($\mathcal{G} = \mathcal{F}$, $\boldsymbol{W} = \boldsymbol{M}_{\text{out}}$); and the noise-weighted metric ($\mathcal{G} = \mathcal{R}\mathcal{F}$, $\boldsymbol{W} = \boldsymbol{\Sigma}_{\text{obs}}^{-1}$). DE-DeepONet uses the full-field output geometry to select reduced input directions and then supervises the corresponding full-field directional derivatives (Qiu et al., 2024); DIFNO likewise targets derivatives of the PDE solution operator (Yao et al., 2026).

For completeness, under the Gaussian observation model,

$$\nabla_{\boldsymbol{a}} \log p(\boldsymbol{y} \mid \boldsymbol{a}) = \widetilde{\boldsymbol{J}}(\boldsymbol{a})^\top \boldsymbol{\Sigma}_{\text{obs}}^{-1/2} (\boldsymbol{y} - \mathcal{R}\mathcal{F}(\boldsymbol{a})).$$

Taking its conditional second moment gives

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{a}} \left[\nabla \log p \, \nabla \log p^\top \right] = \widetilde{\boldsymbol{J}}(\boldsymbol{a})^\top \widetilde{\boldsymbol{J}}(\boldsymbol{a}) = \boldsymbol{H}_{\text{F}}(\boldsymbol{a}),$$

and averaging over μ gives $\bar{\boldsymbol{H}}_{\text{F}}$. An unweighted observation subspace uses $\mathbb{E}_\mu[\boldsymbol{J}_{\mathcal{F}}^\top \mathcal{R}^\top \mathcal{R} \boldsymbol{J}_{\mathcal{F}}]$; it has the same eigenvectors when $\boldsymbol{\Sigma}_{\text{obs}} = \sigma^2 \boldsymbol{I}$, but may differ for heteroscedastic or correlated noise. Unlike a fixed-data misfit subspace, $\bar{\boldsymbol{H}}_{\text{F}}$ is independent of the realized residual.

E A two-mode advection–diffusion example

We isolate the choice of one derivative-supervision direction in a periodic initial-condition inverse problem. On $x \in [0, 2\pi)$, let

$$\partial_t u + b\partial_x u = \kappa \partial_{xx} u, \qquad u(0, x) = a_1\phi_1(x) + a_2\phi_4(x), \qquad \phi_j(x) := \frac{\cos(jx)}{\sqrt{\pi}}.$$

The modes are orthonormal in $L^2(0, 2\pi)$. Writing $s_j := e^{-\kappa j^2 T}$ and $\psi_j(x) := \phi_j(x - bT)$, the full solution field at time T is

$$\mathcal{F}(\boldsymbol{a})(x) = s_1 a_1 \psi_1(x) + s_4 a_2 \psi_4(x).$$

Choose $\kappa T = \log(10)/15$, so $s_1 = 10 s_4$. A sensor at $x_s = bT + \pi/2$ with $\mathcal{R}u := \sqrt{\pi}\, u(x_s)$ observes

$$y = \mathcal{R}\mathcal{F}(\boldsymbol{a}) + \eta = s_4 a_2 + \eta, \qquad \eta \sim \mathcal{N}(0, \tau^2).$$

The first mode dominates the full-field L^2 response but vanishes at the sensor. With the L^2 field metric, the constant-Jacobian sensitivity matrices are

$$\boldsymbol{H}_{\text{field}} = \text{diag}(s_1^2, s_4^2), \qquad \bar{\boldsymbol{H}}_{\text{F}} = \text{diag}(0, s_4^2/\tau^2).$$

Thus rank-one field sensitivity selects $\boldsymbol{e}_1$, whereas Fisher sensitivity selects $\boldsymbol{e}_2$. Their full-output JVP targets are $s_1\psi_1$ and $s_4\psi_4$, respectively.

To make the training effect explicit, use the same surrogate family $\mathcal{F}_{\text{nn}}(\boldsymbol{a}; x) = k a_1 \psi_1(x) + c a_2 \psi_4(x)$ for both selections. Let μ assign equal probability to independent $a_1 \in \{-1, 1\}$ and $a_2 \in \{-0.2, 0.2\}$, and minimize the same simplified quadratic objective, changing only the unit direction $\boldsymbol{v}$:

$$\mathcal{J}_{\boldsymbol{v}}(k, c) := \mathbb{E}_\mu \|\mathcal{F}_{\text{nn}} - \mathcal{F}\|_{L^2}^2 + \lambda_D \|(D\mathcal{F}_{\text{nn}} - D\mathcal{F})\boldsymbol{v}\|_{L^2}^2 + \rho(k^2 + c^2).$$

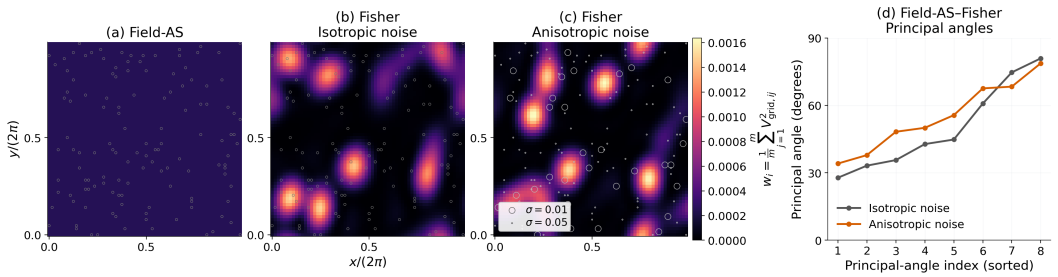


Figure 17: **Observation weighting changes the selected parameter directions.** For the advection–diffusion diagnostic, panels (a)–(c) plot the diagonal of the rank-8 input-subspace projector, $\ell_8(x) = \sum_{j=1}^{8} v_j(x)^2$, in spatial parameter coordinates. Brighter locations are more strongly represented in the selected subspace; circles mark observation sites. Panel (a) selects directions by solution-field sensitivity, whereas (b) and (c) use observed Fisher sensitivity with uniform and spatially varying noise. Panel (d) shows the principal angles between the field-sensitivity and Fisher subspaces. Even their smallest angles are about 28° and 34°, respectively: observation weighting changes directions, not merely their scale. This geometric diagnostic does not itself compare inversion accuracy.

The common value term is $(k - s_1)^2 + 0.04(c - s_4)^2$; the JVP term is $\lambda_D(k - s_1)^2$ for $\boldsymbol{v} = \boldsymbol{e}_1$ and $\lambda_D(c - s_4)^2$ for $\boldsymbol{v} = \boldsymbol{e}_2$. Hence

$$(k_{\text{field}}, c_{\text{field}}) = \left(\frac{(1 + \lambda_D)s_1}{1 + \lambda_D + \rho}, \frac{0.04 s_4}{0.04 + \rho} \right),$$
$$(k_{\text{F}}, c_{\text{F}}) = \left(\frac{s_1}{1 + \rho}, \frac{(0.04 + \lambda_D)s_4}{0.04 + \lambda_D + \rho} \right).$$

For $\lambda_D = 0.9$ and $\rho = 0.06$, these are approximately $(0.9694 s_1, 0.4 s_4)$ and $(0.9434 s_1, 0.94 s_4)$. Because $s_1 = 10 s_4$, the expected squared forward-field errors are $0.108 s_4^2$ and $0.321 s_4^2$, respectively: field-sensitivity supervision has the better forward-field fit.

Nevertheless, take $\boldsymbol{a}^\star = (0, 0.2)^\top$ and a zero-noise realization $y = 0.2 s_4$. Starting from $a_1 = 0$, surrogate maximum-likelihood inversion leaves the unobserved coordinate at zero and yields $\widehat{a}_2 = y/c$. Therefore

$$\widehat{a}_{2,\text{field}} = 0.5, \qquad \widehat{a}_{2,\text{F}} \approx 0.213, \qquad |\widehat{a}_{2,\text{field}} - a_2^\star| = 0.300, \quad |\widehat{a}_{2,\text{F}} - a_2^\star| \approx 0.0128.$$

Both surrogate observation residuals vanish, while their simulator-evaluated residuals are s_4 times the displayed recovery errors. At the shared initialization, relative likelihood-gradient errors in the a_2 direction are $|1 - c/s_4| = 0.60$ and 0.06. For nonzero observation noise, the error remains explicit:

$$\mathbb{E}_\eta(\widehat{a}_2 - a_2^\star)^2 = \left(\frac{s_4}{c} - 1 \right)^2 (a_2^\star)^2 + \frac{\tau^2}{c^2}.$$

For example, at $\tau = 0.01 s_4$, the two root-mean-squared recovery errors are approximately 0.301 and 0.0166.

This construction is illustrative, not the numerical FINO training protocol. Its diagonal two-mode map, exactly nodal sensor, and explicit weight decay make the mechanism transparent. Shifting the sensor by δ changes its two sensitivities to $-s_1 \sin\delta$ and $s_4 \cos(4\delta)$; thus the Fisher direction remains close to $\boldsymbol{e}_2$ when $10|\sin\delta| \ll |\cos(4\delta)|$, without an exact null. If $\rho = 0$, both surrogates fit this affine map exactly and their inversions coincide; the example does not assert universal FINO superiority.

F Field–Measurement Geometry

Lemma 1 (Field–measurement composition). *If $\boldsymbol{\Sigma}_{\text{obs}} \succ 0$, then*

$$\begin{aligned} \operatorname{rank} \boldsymbol{H}_{\text{F}} &= \operatorname{rank}(\mathcal{R}\boldsymbol{J}_{\mathcal{F}}) \\ &= \operatorname{rank} \boldsymbol{J}_{\mathcal{F}} - \dim(\operatorname{range} \boldsymbol{J}_{\mathcal{F}} \cap \ker \mathcal{R}) \le \min\{\operatorname{rank} \boldsymbol{J}_{\mathcal{F}}, \operatorname{rank} \mathcal{R}\}. \end{aligned}$$

Moreover, with $\mathcal{R}_{\mathrm{w}} = \boldsymbol{\Sigma}_{\mathrm{obs}}^{-1/2}\mathcal{R}$,

$$\lambda_{i+j-1}(\boldsymbol{H}_{\mathrm{F}}) \le \sigma_i(\mathcal{R}_{\mathrm{w}})^2\sigma_j(\boldsymbol{J}_{\mathcal{F}})^2, \qquad i,j \ge 1, \quad i+j-1 \le d_{\mathrm{in}}.$$

The singular values of the factors alone do not determine the Fisher spectrum; their singular-vector alignment also matters.

Proof. Let $\boldsymbol{B} = \mathcal{R}_{\mathrm{w}}\boldsymbol{J}_{\mathcal{F}}$. Because $\boldsymbol{H}_{\mathrm{F}} = \boldsymbol{B}^\top\boldsymbol{B}$, $\operatorname{rank}\boldsymbol{H}_{\mathrm{F}} = \operatorname{rank}\boldsymbol{B}$. Whitening is invertible, so $\ker\mathcal{R}_{\mathrm{w}} = \ker\mathcal{R}$. Rank–nullity applied to the restriction of $\mathcal{R}$ to $\operatorname{range}\boldsymbol{J}_{\mathcal{F}}$ proves the rank identity. The spectral inequality follows from $\lambda_k(\boldsymbol{H}_{\mathrm{F}}) = \sigma_k(\boldsymbol{B})^2$ and the multiplicative singular-value inequality (Horn & Johnson, 1991).

For alignment problem between $\boldsymbol{J}_{\mathcal{F}}$ and $\mathcal{R}_W$, take $\mathcal{R}_{\mathrm{w}} = [1\ 0]$ and

$$\boldsymbol{J}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \boldsymbol{J}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

The two Jacobians have identical singular values, but $\mathcal{R}_{\mathrm{w}}\boldsymbol{J}_1 = [1\ 0]$ whereas $\mathcal{R}_{\mathrm{w}}\boldsymbol{J}_2 = \mathbf{0}$, giving Fisher ranks one and zero. □

G Inverse rank and the bias–variance tradeoff

G.1 Proof of Proposition 1

Because $\|\widetilde{\boldsymbol{J}}\boldsymbol{v}\|_2^2 = \boldsymbol{v}^\top\boldsymbol{H}_{\mathrm{F}}\boldsymbol{v}$, the proposition is the Courant–Fischer characterization of λ_r (Golub & Van Loan, 2013). Repeated eigenvalues at the truncation boundary may make the maximizing subspace nonunique.

In the affine model, the noise-induced variance of the retained coefficient along $\boldsymbol{v}_i$ is λ_i^{-1}. A tolerated coordinate standard error u therefore suggests retaining $\lambda_i \ge u^{-2}$, with u interpreted in the chosen coordinate scaling. A percentage-energy rule instead chooses the smallest r with

$$\frac{\sum_{i=1}^{r}\lambda_i}{\sum_{i=1}^{d_{\mathrm{in}}}\lambda_i} \ge \eta.$$

The first criterion controls estimation uncertainty; the second only measures spectral concentration.

Corollary 1 (Retained rank and expected recovery)**.** *Let* $\widehat{\boldsymbol{a}}_{\mathrm{MLE},r}$ *be the affine simulator MLE restricted to the first* r *Fisher eigenvectors. For* $1 \le r < s \le \operatorname{rank}\widetilde{\boldsymbol{J}}$*,*

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}}\|\widehat{\boldsymbol{a}}_{\mathrm{MLE},s} - \boldsymbol{a}^\star\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}}\|\widehat{\boldsymbol{a}}_{\mathrm{MLE},r} - \boldsymbol{a}^\star\|_2^2 \\ = \sum_{i=r+1}^{s}\left[\lambda_i^{-1} - |\boldsymbol{v}_i^\top(\boldsymbol{a}^\star - \boldsymbol{a}_{\mathrm{ref}})|^2\right].\end{aligned} \tag{18}$$

Thus increasing the rank improves expected recovery if and only if the newly recovered truth energy exceeds $\sum_{i=r+1}^{s}\lambda_i^{-1}$.

Proof. Equation (6) and the Fisher eigenbasis give

$$\mathbb{E}_{\boldsymbol{\xi}}\|\widehat{\boldsymbol{a}}_{\mathrm{MLE},r} - \boldsymbol{a}^\star\|_2^2 = \sum_{i>r}|\boldsymbol{v}_i^\top(\boldsymbol{a}^\star - \boldsymbol{a}_{\mathrm{ref}})|^2 + \sum_{i=1}^{r}\lambda_i^{-1}.$$

Subtracting the expressions for s and r proves the result. □

If $r = \operatorname{rank}\widetilde{\boldsymbol{J}}$, the restricted solution is the simulator MLE closest to $\boldsymbol{a}_{\mathrm{ref}}$; for smaller r, it minimizes the likelihood only within the selected affine space.

H Proof of the recovery theorem

Let $\boldsymbol{A} := \widetilde{\boldsymbol{J}}\boldsymbol{V}$ and $\boldsymbol{B} := \widetilde{\boldsymbol{J}}_{\mathrm{nn}}\boldsymbol{V}$. Because $\lambda_r > 0$, $\boldsymbol{A}$ has full column rank and $\sigma_{\min}(\boldsymbol{A}) = \sqrt{\lambda_r}$.

Part (i). Write $\boldsymbol{z}^\star = \boldsymbol{V}^\top(\boldsymbol{a}^\star - \boldsymbol{a}_{\mathrm{ref}})$. Then $\boldsymbol{a}^\star = \boldsymbol{a}_{\mathrm{ref}} + \boldsymbol{V}\boldsymbol{z}^\star + \boldsymbol{t}_{\mathrm{trunc}}$. Along the restricted space, the whitened simulator residual is

$$\boldsymbol{r}(\boldsymbol{z}) = \boldsymbol{A}(\boldsymbol{z} - \boldsymbol{z}^\star) - \widetilde{\boldsymbol{J}}\boldsymbol{t}_{\mathrm{trunc}} - \boldsymbol{\xi}.$$

Since $\boldsymbol{V}$ is a Fisher eigenbasis and $\boldsymbol{t}_{\mathrm{trunc}} \in \operatorname{range}\boldsymbol{Q}$, $\boldsymbol{A}^\top\widetilde{\boldsymbol{J}}\boldsymbol{t}_{\mathrm{trunc}} = \boldsymbol{0}$. The normal equations therefore give $\widehat{\boldsymbol{z}}_{\mathrm{MLE}} - \boldsymbol{z}^\star = \boldsymbol{A}^\dagger\boldsymbol{\xi}$. Orthogonality of $\operatorname{range}\boldsymbol{V}$ and $\operatorname{range}\boldsymbol{Q}$ yields (6).

Part (ii). Let $\boldsymbol{E} := \boldsymbol{B} - \boldsymbol{A}$. By definition, $\|\boldsymbol{E}\|_{\mathrm{op}} \le \varepsilon_{\mathrm{nn,Fisher}}$, so Weyl's inequality gives $\sigma_{\min}(\boldsymbol{B}) \ge \delta_r > 0$. At the simulator MLE, write the surrogate residual as $\boldsymbol{r}_{\mathrm{nn}} = \boldsymbol{r} + \boldsymbol{e}$, where $\|\boldsymbol{e}\|_2 \le \varepsilon_{\mathrm{nn,forward}}$ and $\boldsymbol{A}^\top\boldsymbol{r} = \boldsymbol{0}$. The surrogate normal equations give

$$\widehat{\boldsymbol{z}}_{\mathrm{nn}} - \widehat{\boldsymbol{z}} = -(\boldsymbol{B}^\top\boldsymbol{B})^{-1}\boldsymbol{B}^\top(\boldsymbol{r} + \boldsymbol{e}).$$

The $\boldsymbol{e}$ term is bounded by $\varepsilon_{\mathrm{nn,forward}}/\delta_r$. Moreover, $\boldsymbol{B}^\top\boldsymbol{r} = (\boldsymbol{B} - \boldsymbol{A})^\top\boldsymbol{r}$, so the residual term is bounded by $\varepsilon_{\mathrm{nn,Fisher}}\varepsilon_{\mathrm{sim}}/\delta_r^2$. This proves (7).

Part (iii). Set $\boldsymbol{D} := \widetilde{\boldsymbol{J}}_{\mathrm{nn}}\boldsymbol{Q}$. For a fixed complementary displacement $\boldsymbol{q}$, the conditional surrogate minimizer in retained coordinates differs from the restricted surrogate minimizer by $-\boldsymbol{B}^\dagger\boldsymbol{D}\boldsymbol{q}$ and hence by at most $c_r\|\boldsymbol{q}\|_2$. At an arbitrary point $\boldsymbol{a}_t$, the distance in retained coordinates to this conditional minimizer is at most $g_{\mathrm{F,t}}/\sigma_{\min}(\boldsymbol{B})^2 \le g_{\mathrm{F,t}}/\delta_r^2$. Combining these two bounds with part (ii) and $\|\widehat{\boldsymbol{z}} - \boldsymbol{z}^\star\|_2 = \|\boldsymbol{A}^\dagger\boldsymbol{\xi}\|_2$ gives the retained-coordinate term in (8). The complementary error is $\boldsymbol{q}_t - \boldsymbol{t}_{\mathrm{trunc}}$ and is orthogonal to the retained error, proving the squared bound.

For fidelity to the restricted simulator MLE rather than to $\boldsymbol{a}^\star$, the same argument replaces $\|\boldsymbol{q}_t - \boldsymbol{t}_{\mathrm{trunc}}\|_2^2$ by $\|\boldsymbol{q}_t\|_2^2$ and omits the simulator-noise term.

Corollary 2 (Consistency with the simulator-based estimator)**.** *Suppose that the retained observed Jacobian is uniformly identifiable, with $\sqrt{\lambda_r} > \epsilon_{\mathrm{Fisher}}$, and let $\widehat{\boldsymbol{a}}$ and $\widehat{\boldsymbol{a}}_{\mathrm{nn}}$ denote the simulator- and surrogate-based maximum-likelihood solutions in the retained Fisher coordinates. If*

$$\epsilon_{\mathrm{forward}} \longrightarrow 0, \qquad \epsilon_{\mathrm{Fisher}} \longrightarrow 0,$$

then

$$\|\widehat{\boldsymbol{a}}_{\mathrm{nn}} - \widehat{\boldsymbol{a}}\|_2 \longrightarrow 0.$$

If the retained Fisher coordinates span the complete positive-eigenvalue Fisher subspace, this limit agrees with the observation-identifiable component of a full-space simulator-based maximum-likelihood solution.

I WHY VALUE FITTING IS INSUFFICIENT

Corollary 3 (Value matching does not control retained derivatives)**.** *Fix finitely many distinct inputs, a point $\boldsymbol{a}_c$, and an orthonormal $\boldsymbol{V} \in \mathbb{R}^{d_{\mathrm{in}} \times r}$ with $r \ge 1$. If $\boldsymbol{\Sigma}_{\mathrm{obs}} \succ 0$ and $\mathcal{R} \ne \boldsymbol{0}$, then for every $K > 0$ there exists a ReLU surrogate that matches $\mathcal{F}$ at those inputs and at $\boldsymbol{a}_c$, while*

$$\|[\widetilde{\boldsymbol{J}}_{\mathrm{nn}}(\boldsymbol{a}_c) - \widetilde{\boldsymbol{J}}(\boldsymbol{a}_c)]\boldsymbol{V}\|_{\mathrm{op}} = K.$$

Proof. Proposition 1 of Czarnecki et al. (2017) implies that, for a finite set of interpolation constraints, a sufficiently expressive ReLU network can match the prescribed values while independently prescribing a directional derivative at $\boldsymbol{a}_c$. Choose a nonzero direction in $\operatorname{range}\boldsymbol{V}$ and an output direction not annihilated by $\mathcal{R}$. Adding the resulting zero-value interpolant to a value-matching network leaves every prescribed value unchanged. Scaling this perturbation sets the whitened observed directional error to any $K > 0$. □

This is an existence statement, not a claim that every value-trained network has inaccurate derivatives. It shows only that value error alone cannot provide the derivative control required by Theorem 1.

J ADDITIONAL SPECTRAL CONSEQUENCES

The derivative–residual contribution in (7) exceeds the forward contribution exactly when

$$\varepsilon_{\mathrm{nn,Fisher}}\varepsilon_{\mathrm{sim}} > \varepsilon_{\mathrm{nn,forward}}\big(\sqrt{\lambda_r} - \varepsilon_{\mathrm{nn,Fisher}}\big).$$

This compares terms in an upper bound, not their realized contributions.

Proposition 2 (Spectral weighting of derivative improvements)**.** *At a point where the baseline and derivative-trained surrogates match the observed simulator value, hold their Jacobians fixed and average over standardized residuals $\mathbb{E}[\boldsymbol{\rho}\boldsymbol{\rho}^\top] = \boldsymbol{I}$. With $e_{b,i}$ and $\mathcal{G}_b$ defined in Section 4,*

$$\mathcal{G}_{\mathrm{base}} - \mathcal{G}_{\mathrm{deriv}} = \sum_{i=1}^{r} \lambda_i (e^2_{\mathrm{base,i}} - e^2_{\mathrm{deriv,i}}).$$

Proof. Observed value matching makes the two losses share the same standardized residual. Hence the retained gradient discrepancy for surrogate b is $\boldsymbol{V}\boldsymbol{V}^\top(\widetilde{\boldsymbol{J}}_b - \widetilde{\boldsymbol{J}})^\top\boldsymbol{\rho}$. Taking its squared norm and expectation gives

$$\mathcal{G}_b = \sum_{i=1}^{r} \|(\widetilde{\boldsymbol{J}}_b - \widetilde{\boldsymbol{J}})\boldsymbol{v}_i\|_2^2 = \sum_{i=1}^{r} \lambda_i e^2_{b,i}.$$

Subtracting the two expressions proves the claim. □

Thus, for common nonnegative improvements $\Delta_i = e^2_{\mathrm{base,i}} - e^2_{\mathrm{deriv,i}}$, a coordinatewise larger retained spectrum produces a larger reduction in projected gradient error. If the methods agree on the first r_0 directions and have errors $\alpha > \beta$ on the remaining directions, then

$$\mathcal{G}_{\mathrm{base}} - \mathcal{G}_{\mathrm{deriv}} = (\alpha^2 - \beta^2) \sum_{i=r_0+1}^{r} \lambda_i.$$

This isolates the Fisher spectral tail. For a common gradient step size η, the corresponding one-step squared update improvement is scaled by η^2.

Proposition 3 (Surrogate cross-coupling)**.** *Let $\varepsilon_{\mathrm{nn},\perp} := \|(\widetilde{\boldsymbol{J}}_{\mathrm{nn}} - \widetilde{\boldsymbol{J}})\boldsymbol{Q}\|_{\mathrm{op}}$ and $\delta_r = \sqrt{\lambda_r} - \varepsilon_{\mathrm{nn,Fisher}} > 0$. Then*

$$c_r \le \frac{\varepsilon_{\mathrm{nn},\perp}}{\delta_r} + \frac{\varepsilon_{\mathrm{nn,Fisher}}\sqrt{\lambda_{r+1}}}{\delta_r^2}, \qquad \lambda_{d_{\mathrm{in}}+1} := 0.$$

Proof. Set $\boldsymbol{A} = \widetilde{\boldsymbol{J}}\boldsymbol{V}$, $\boldsymbol{B} = \widetilde{\boldsymbol{J}}_{\mathrm{nn}}\boldsymbol{V}$, and $\boldsymbol{C} = \widetilde{\boldsymbol{J}}\boldsymbol{Q}$. The Fisher eigenbasis gives $\boldsymbol{A}^\top\boldsymbol{C} = 0$ and $\|\boldsymbol{C}\|_{\mathrm{op}} = \sqrt{\lambda_{r+1}}$. Therefore

$$\begin{aligned}\|\boldsymbol{B}^\dagger\widetilde{\boldsymbol{J}}_{\mathrm{nn}}\boldsymbol{Q}\|_{\mathrm{op}} &\le \|\boldsymbol{B}^\dagger(\widetilde{\boldsymbol{J}}_{\mathrm{nn}} - \widetilde{\boldsymbol{J}})\boldsymbol{Q}\|_{\mathrm{op}} + \|\boldsymbol{B}^\dagger\boldsymbol{C}\|_{\mathrm{op}} \\ &\le \frac{\varepsilon_{\mathrm{nn},\perp}}{\delta_r} + \frac{\varepsilon_{\mathrm{nn,Fisher}}\sqrt{\lambda_{r+1}}}{\delta_r^2},\end{aligned}$$

where the second term uses $\boldsymbol{B}^\top\boldsymbol{C} = (\boldsymbol{B} - \boldsymbol{A})^\top\boldsymbol{C}$. □

If the retained space contains every observable direction, $\lambda_{r+1} = 0$ and the second cross-coupling term vanishes. These spectral results describe gradient and stability mechanisms; parameter recovery still also depends on forward, truncation, noise, complementary, and optimization errors.

K FROM AMORTIZED TRAINING TO LOCAL ACCURACY

Let $\widehat{\boldsymbol{P}} = \hat{\boldsymbol{V}}_m\hat{\boldsymbol{V}}_m^\top$ and $\boldsymbol{P} = \boldsymbol{V}\boldsymbol{V}^\top$, with $m = r$, and set $\delta_V = \|\widehat{\boldsymbol{P}} - \boldsymbol{P}\|_{\mathrm{op}}$. At the recovery point, suppose

$$\|\Delta\boldsymbol{J}\hat{\boldsymbol{V}}_m\|_F \le \varepsilon_D, \qquad \|\Delta\boldsymbol{J}\|_{\mathrm{op}} \le \varepsilon_{\mathrm{all}}, \quad \Delta\boldsymbol{J} := \boldsymbol{J}_{\mathcal{F},\mathrm{nn}} - \boldsymbol{J}_{\mathcal{F}}.$$

Then (10) holds.

Proof. Using $\boldsymbol{P}\boldsymbol{V} = \boldsymbol{V}$,

$$\Delta \boldsymbol{J}\boldsymbol{V} = \Delta \boldsymbol{J}\widehat{\boldsymbol{P}}\boldsymbol{V} + \Delta \boldsymbol{J}(\boldsymbol{P} - \widehat{\boldsymbol{P}})\boldsymbol{V}.$$

The first term is bounded by $\|\Delta \boldsymbol{J}\hat{\boldsymbol{V}}_m\|_F$, and the second by $\varepsilon_{\mathrm{all}}\delta_V$. Left multiplication by $\boldsymbol{\Sigma}_{\mathrm{obs}}^{-1/2}\mathcal{R}$ proves the result. □

The mismatch admits the diagnostic decomposition

$$\|\widehat{\boldsymbol{P}} - \boldsymbol{P}(\boldsymbol{a})\|_{\mathrm{op}} \le \|\boldsymbol{P}(\boldsymbol{a}) - \bar{\boldsymbol{P}}\|_{\mathrm{op}} + \|\bar{\boldsymbol{P}} - \boldsymbol{P}_M\|_{\mathrm{op}} + \|\boldsymbol{P}_M - \widehat{\boldsymbol{P}}\|_{\mathrm{op}},$$

where the terms correspond to amortization across inputs, Monte Carlo estimation, and randomized eigensolver error. Spectral gaps convert matrix perturbation bounds into projector bounds (Yu et al., 2015). The FINO loss controls an empirical average of ε_D at labeled training inputs, whereas recovery requires local control along the inverse trajectory; this motivates the held-out derivative and alignment diagnostics in Section 5.