# MECHANISMS OFF THE DATA MANIFOLD: DISTRIBUTION-RELATIVE IDENTIFIABILITY IN NEURAL OPERATORS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The standard Navier–Stokes benchmark for operator learning contains a blind spot introduced by its own data pipeline: the generator enforces zero spatial-mean vorticity across all examples, whereas the true solution operator acts as the identity along this constant mode. Operators trained to benchmark accuracy recover that response at roughly one-eighth of its true strength, incurring a $28\%$ error under a mean offset that the true dynamics handle trivially. This failure is invisible to standard in-distribution evaluations, is predictable from the training covariance alone, and admits an exact, zero-cost structural repair once diagnosed. This reflects a broader phenomenon: function-value empirical training identifies local mechanisms only along directions excited by data leaving the remaining modes governed entirely by architectural inductive bias. For translation-equivariant model families, this completion is predictable from a continuity dichotomy between localized spatial kernels and per-mode spectral weights. Controlled experiments show that failure follows excitation rather than frequency, that common canonicalization removes the evidence required to resolve a symmetry's mechanism, and this vulnerability amplifies as physical dissipation weakens. The diagnosis is actionable: training covariance flags which data to acquire, while a known symmetry enables exact structural repairs that degrades gracefully under misspecification.

## 1 INTRODUCTION

Interpreting a neural network often involves intervening on its representations and measuring the resulting behavioral change. While causal interventions are stronger than correlational probes, they do not guarantee that the observed behavior was identified by the training objective. An intervention can easily push inputs off the training distribution. Two models can match on every training example yet respond entirely differently to the same intervention.This problem is particularly acute in neural operators (Li et al., 2021; Kovachki et al., 2023). While an operator is trained to map between discretized functions, mechanistic questions typically concern its local derivatives: which input perturbations are amplified, advected, or coupled? Because numerical solvers and governing PDEs provide ground-truth derivatives, neural operators offer an ideal testbed for a broader interpretability question: when is an intervention response identified by function-value training data, and when is it merely an artifact of an unconstrained off-support completion? The consequences are direct and practical. The standard Navier–Stokes benchmark manufactures such a blind spot during routine data generation, along a direction where the exact mechanism is known in closed form. Section 2 begins with an audit of this benchmark. The issue is not adversarial, it arises directly from standard preprocessing, and illustrates a general, predictable failure mode of function-value training.

**Is this simply out-of-distribution generalization?** No, for two decisive senses (see Appendix A for a full discussion). First, the failure is invisible in principle to standard i.i.d. evaluation: held-out test sets share the training support, so no volume of in-distribution test data can detect an incorrect support-normal derivative. Second, the failure is predictable before test time. Directional training excitation reveals which mechanisms remain unidentified, guides targeted data collection, and, when a direction corresponds to a known symmetry, yields an exact structural repair. Prior work on underspecification and predictive multiplicity (D'Amour et al., 2022; Marx et al., 2020; Semenova



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.

et al., 2022) shows that models matching on training data can diverge under distribution shift. Here, I provide a localized, direction-resolved account that isolates exactly which responses are unconstrained and provides a framework to repair them. I distinguish three objects: the target mechanism $DG_x$ (the true derivative of the physical operator), the data-identified mechanism (its restriction to directions excited by training data), and the completed mechanism (how a model interpolates or extrapolates into weak or absent directions). Standard loss functions constrain only the second object, yet mechanistic claims often implicitly assume the third.

The main contributions are:

1. **A benchmark audit with an exact reference key:** The standard Navier–Stokes benchmark removes a spatial mode whose exact response is the identity. Standard trained operators learn this response with a gain of only 0.13, incurring a 28% error under constant offsets. This error can be repaired at zero cost using an exact structural completion. Furthermore, this vulnerability widens as physical dissipation weakens, expanding from neutral modes at $\nu = 10^{-3}$ to weakly excited frequency bands at $\nu = 10^{-4}$.
2. **Excitation, not frequency, governs mechanistic fidelity:** By constructing 1D and 2D covariances that decouple excitation from spatial frequency, I show that generalization tracks excitation energy rather than smoothness. A standard ridge-shrinkage explanation fails quantitatively, pointing instead to inductive biases in the architecture.
3. **Canonicalization removes mechanisms and symmetry restores them:** Common preprocessing steps, such as mean removal and unit normalization, eliminate the evidence needed to learn a mechanism. Small perturbations along the group orbit re-identify it, while exact structural completions, requiring knowledge only of the symmetry, not the solution, repair the model for free and degrade gracefully under misspecification.
4. **Architecture as a predictable completion prior:** A continuity dichotomy between localized spatial kernels and per mode spectral weights predicts the completion chosen by different model families. I verify this across random seeds and show with a non-PDE toy task that the phenomenon is general.
5. **Diagnostic tools and formal framework:** I provide a normal-completion theorem characterizing off-support ambiguity, a signed phase-aware Fourier-JVP evaluation protocol, and a covariance-guided data acquisition strategy.

Finally, I document ten preregistered negative results in Appendix H, including the failure of literal ridge shrinkage and two informative null results that delineate when off-support completions matter.

## 2 Motivating Audit: The standard Navier–Stokes benchmark

I generate data following the standard 2D Navier–Stokes benchmark specification (Li et al., 2021) (vorticity formulation, $\nu = 10^{-3}$, sinusoidal forcing, unit-time propagation at $64^2$; Appendix E.9) and train standard Fourier Neural Operators (12 modes, width 20, 4 layers, relative-$L_2$ loss). Across eight random seeds, models achieve $0.2$–$0.6\%$ one-step test error.

This benchmark provides an ideal audit because its exact physical response is known in closed form along an unexcited mode. The data generator enforces zero-mean vorticity ($k = 0$), and because the forcing also has zero mean, empirical excitation along the constant mode is negligible ($\sim 10^{-12}$). On the periodic torus, a constant shift in vorticity induces zero velocity, yielding the exact identity map $G(w + c) = G(w) + c$ (verified numerically to machine precision, $2 \times 10^{-16}$); thus, the row sums of the Jacobian are all equal to 1. The benchmark pipeline therefore systematically eliminates a direction whose true mechanism is known analytically. (While this constant mode represents a discretization degree of freedom in periodic vorticity formulations rather than a distinct physical flow state, the Burgers blind spots analyzed in Section 5 are fully physical; see Appendix E.9 for details.)

As predicted by the empirical covariance, trained operators fail to learn this completion (Figure 1; Table 9). The mean directional gain along the constant mode is only $0.130 \pm 0.019$, far below the true value of 1. Adding a modest mean offset of 0.25 (one-third of the field's standard deviation) increases prediction error to $28\%$, representing a 50- to 90-fold degradation over in-distribution performance. The resulting pointwise error field forms a nearly uniform shift of $\approx -0.22$, matching the unpropagated fraction of the offset.

Across the evaluated spectrum, tangent gain error strongly tracks training excitation (Spearman $\rho = -0.69 \pm 0.04$). Gain error correlates positively with spatial wavenumber under the natural power spectrum (generalized in Section 4). An exact structural completion, $C(w) = f(w - m(w)) + m(w)$ where $m(w)$ denotes spatial mean, restores the constant-mode tangent gain identically to 1 and reduces the offset error from $28\%$ to $0.3\%$ at zero in-distribution cost (Section 7). Auditing the officially distributed benchmark dataset directly, at native resolution without regeneration, with reference licensed by the measured mean conservation, recovers the same blind spot and structural repair (Appendix I.6).

**When do incorrect completions matter?** Two preregistered tests of finite-amplitude perturbations along weakly excited directions yielded informative null results (Appendix I.1). First, evolved test trajectories forget their initial prior because the training support acts as an attractor. Second, weak-band perturbations incur no higher error than strong-band ones despite a $10^7$-fold excitation gap and measurably inaccurate tangent gains. At $\nu = 10^{-3}$, these modes contract by $\sim e^{-4}$ per unit time, so both the true and learned responses decay to near zero. Consequently, unconstrained completions surface where dynamics do not contract: along conserved quantities and symmetry generators—the precise neutral directions eliminated by canonicalization.

**Low viscosity broadens vulnerability.** This masking effect stems from strong physical dissipation rather than a limitation of the theory. At $\nu = 10^{-4}$, where viscous decay at $|k| \sim 10$ is only $e^{-0.4}$ per unit time, the same weakly excited band incurs $2.1\times$, $3.1\times$, and $3.9\times$ higher error than the matched strongly excited band at perturbation amplitudes of $0.05$, $0.1$, and $0.2$ (three seeds; the constant-mode blind spot and structural repair replicate; Appendix I.1). Model vulnerability is therefore regime-dependent: neutral directions remain permanently exposed, while ordinary weakly excited modes become vulnerable as physical dissipation weakens.

One might object that addressing these blind spots requires knowing the governing symmetry in advance, in which case one would simply build an equivariant architecture. Three points address this concern. First, benchmark practice shows that known symmetries are routinely overlooked (the benchmark data generator stripped a known symmetry direction). Second, the training covariance audit identifies unexcited directions without prior knowledge of the symmetry, which is needed only to evaluate ground-truth error in closed form. Third, post-hoc structural wrappers repair existing, pre-trained models where retraining or architectural modification is impractical.

## 3 Distribution-relative identifiability

I now formalize what function-value training can and cannot identify. Let $F : \mathcal{X} \to \mathcal{Y}$ be a target operator and let the training distribution $\mu$ be supported on a closed $C^2$ embedded submanifold $\mathcal{M} \subset \mathcal{X}$. I define the risk equivalence class such that two models $F$ and $G$ if and only if they are identical almost surely. The following result characterizes what this risk equivalence leaves undetermined.

Let's raise **theorem 1**. Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}^p$. For any $C^1$ map $F : \mathcal{X} \to \mathcal{Y}$ and any $C^1$ bundle map $A_x : N_x\mathcal{M} \to \mathcal{Y}$, there exists a $C^1$ extension $\widetilde{F} : \mathcal{X} \to \mathcal{Y}$ that matches $F$ and its tangential derivatives on $\mathcal{M}$, but perturbs normal derivatives arbitrarily:

$$\widetilde{F}|_{\mathcal{M}} = F|_{\mathcal{M}}, \quad D\widetilde{F}_x|_{T_x\mathcal{M}} = DF_x|_{T_x\mathcal{M}}, \quad \text{and} \quad D\widetilde{F}_x|_{N_x\mathcal{M}} = DF_x|_{N_x\mathcal{M}} + A_x. \tag{1}$$

Thus, function values on $\mathcal{M}$ impose no constraints on normal directional derivatives.

Within a tubular neighborhood of $\mathcal{M}$, let $\pi$ denote the nearest-point projection and $n(z) = z - \pi(z)$ the normal coordinate. Choosing a smooth cutoff $\chi$ supported inside this neighborhood, the perturbation $H(z) = \chi(z)A_{\pi(z)}n(z)$ vanishes on $\mathcal{M}$ along with its tangential derivatives, while evaluating to $A_x$ on normal vectors. Setting $\widetilde{F} = F + H$ yields the desired completion (see Appendix A for the complete proof).

This is an identifiability limit: an architectural inductive bias can consistently select a specific normal completion—which may happen to be correct—even though the training distribution $\mu$ provides no empirical signal for it. For weakly excited directions, ridge regression provides intuition for why under-excited modes decay toward zero (Appendix B); Section 4 confirms this rank-order trend while showing that the literal ridge model fails quantitatively.

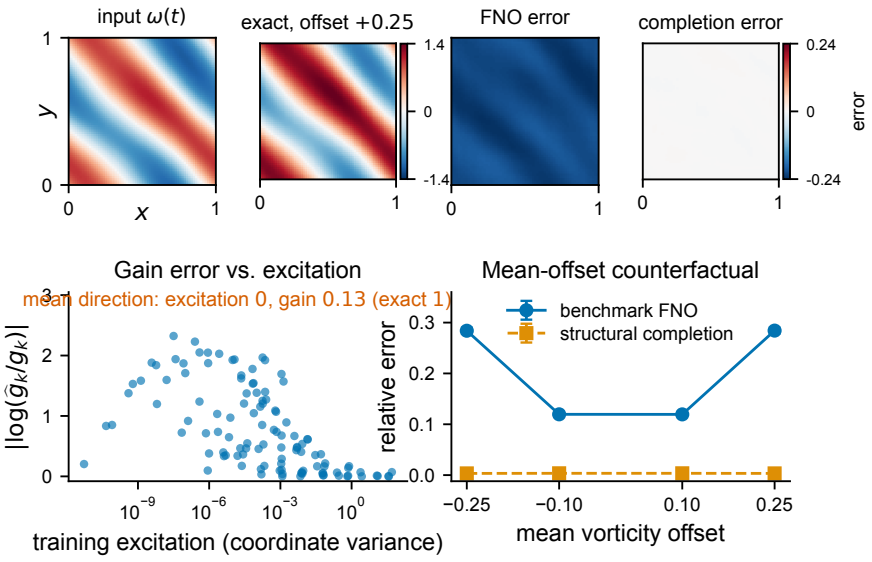


Figure 1: **Standard FNOs miss an analytically known mechanism, and structural completion restores it.** Top: representative test sample (offset error 0.280 at the seed's test mean; shared color scale). Bottom: tangent gain error versus training excitation, alongside offset counterfactuals with and without the completion (error bars indicate standard deviation over eight seeds).

**Canonicalization.** When an operator $F$ is equivariant under a continuous one-parameter symmetry group $g_t$, differentiating the equivariance condition $F(g_t^{\mathcal{X}} x) = g_t^{\mathcal{Y}} F(x)$ at $t = 0$ yields the infinitesimal relation:

$$DF_x\, \xi_{\mathcal{X}}(x) = \xi_{\mathcal{Y}}(F(x)), \tag{2}$$

where $\xi$ denotes the infinitesimal generator. When data preprocessing canonicalizes inputs onto a slice transverse to the group orbit, the generator direction $\xi_{\mathcal{X}}(x)$ is eliminated from the tangent space $T_x\mathcal{M}$. Theorem 1 shows that training data restricted to this canonical slice cannot identify the directional derivative in (2). Incorporating symmetry or equivariance therefore provides genuinely new physical constraints that cannot be learned from canonicalized data alone.

## 4 Does failure follow excitation or frequency?

All measurements use a signed, phase-aware tangent audit. Projecting onto the orthonormal real Fourier basis via synthesis matrices $Q_{\mathcal{X}}$ and $Q_{\mathcal{Y}}$, I evaluate the Jacobian $J_F(x) = Q_{\mathcal{Y}}\, DF_x\, Q_{\mathcal{X}}^{\top}$ column-by-column using exact forward-mode Jacobian-vector products (JVPs). I report signed relative errors, per-mode directional gains, and finite counterfactual errors. Preserving sign and phase is essential: magnitude-only diagnostics conflate within-frequency phase shifts with cross-frequency energy leakage. The diagnostic passes six analytic validation tests, identity, diffusion, advection, squaring, finite differences, 2D Hermitian (Table 1).

While theorem 1 governs directions with zero excitation, standard regularized regression models (Appendix B) predict that weakly excited directions degrade continuously with diminishing variance, regardless of their semantic meaning (e.g., spatial frequency, scale, or coordinate location). In natural datasets, frequency and variance are strongly confounded because power spectra decay with wavenumber. To separate these variables, I decouple them by construction. I fix a mode-dependent spectral target operator and train the same grid-free FNO across four controlled input covariance spectra: natural (variance decays with wavenumber), inverted (variance grows with wavenumber), permuted (the same variance values assigned non-monotonically across modes), and flat (equal variance across all modes). Target operator gains, total input energy, dataset size, observation noise, and model architecture remain strictly fixed.

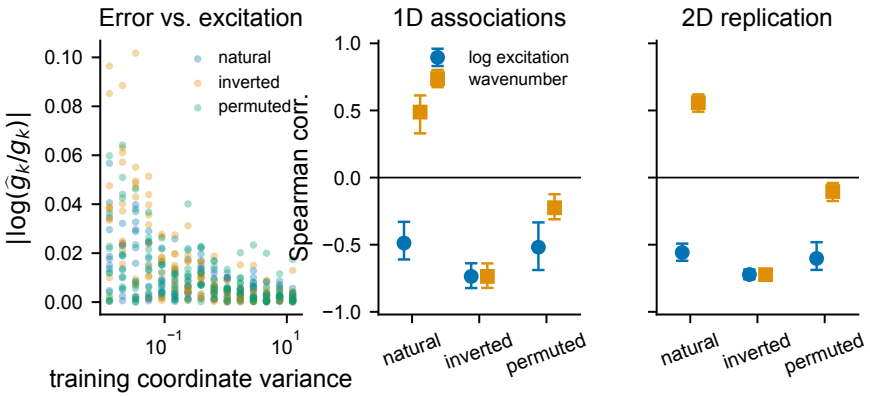


Figure 2: **Training excitation, not spatial frequency, predicts derivative error.** Tangent gain error tracks training variance across all conditions (left). The correlation with wavenumber reverses under spectral inversion and collapses under permutation, whereas the correlation with excitation persists in both 1D (center) and 2D (right). Error bars denote model-cluster bootstrap intervals across ten random seeds.

Figure 2 summarizes the results across ten paired random seeds in 1D and a 2D replication ($32^2$ grid, radial multiplier target, 40 models total; wavenumber denotes $|k|$). Within each model, tangent gain error correlates negatively with training excitation across all informative conditions in both dimensions. In contrast, the association with wavenumber reverses sign when the input spectrum is inverted (paired contrasts of $-1.224$ in 1D and $-1.281$ in 2D, with cluster-robust confidence intervals excluding zero; Table 6 and Appendix G).

Because spectral inversion also inverts the relationship between excitation and frequency, the permuted condition provides the decisive test. With the two variables decoupled, the negative correlation between error and excitation persists (Spearman $\rho = -0.52$ in 1D, $-0.60$ in 2D), whereas the correlation with wavenumber collapses towards zero. Furthermore, dropping log-excitation from pooled regression models reduces explained variance ($R^2$) by an order of magnitude more than dropping wavenumber or target gain.

**The ridge model is an intuition, not a physical mechanism.** Although ridge regression correctly predicts the rank-order dependence on variance, fitting the literal scalar shrinkage curve $s/(s+\lambda)$ fails quantitatively: it explains virtually none of the gain-ratio variance across training, and half or more of the weakly excited modes are amplified rather than attenuated (Appendix I.3). Thus, while training excitation governs which directions remain unconstrained, the specific profile of the off-support completion is determined by architectural inductive bias (Section 6).

## 5 Canonicalization Creates Symmetry Blind Spots

In practice, directional excitation is most commonly eliminated by data preprocessing: canonicalizing continuous symmetries restricts training data to a cross-sectional slice, which, by (2), removes the exact variations required to identify the infinitesimal generator's mechanism. I demonstrate, quantify, and repair two such blind spots using exact symmetry baselines.

**Mean removal eliminates the Galilean generator.** The viscous Burgers equation satisfies Galilean equivariance, $G_T(v+c)(x) = G_T(v)(x-cT)+c$, with the infinitesimal response:

$$DG_T(v)\mathbf{1} = \mathbf{1} - T\partial_x G_T(v), \tag{3}$$

whereas standard benchmark data generators enforce zero spatial mean. Introducing mean excitation $\sigma_c \in \{0, 0.01, 0.05, 0.2\}$ along the group orbit recovers this mechanism across a sharp transition between $\sigma_c = 0.01$ and $0.05$, increasing directional gain from $0.028$ to $0.963$ and reducing offset error from $1.008$ to $0.106$ (eight paired seeds; Table 4). However, broad data augmentation incurs an on-slice accuracy penalty, meaning empirical orbit sampling is not cost-free. This behavior replicates in

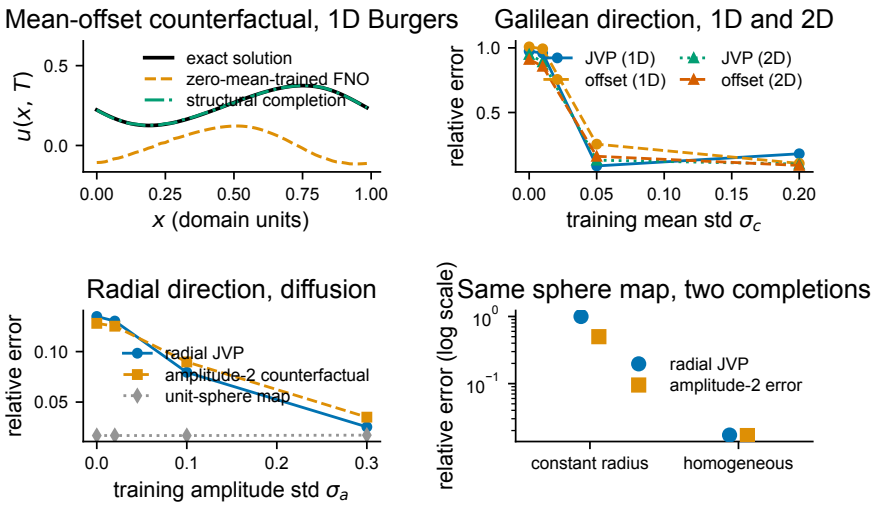


Figure 3: **Data canonicalization creates symmetry blind spots; orbit excitation restores them.** (a) Solution profiles under a constant mean offset: the zero-mean-trained FNO fails, whereas the completion recovers the true dynamics. (b) Sharp Galilean repair threshold in 1D and 2D. (c) Radial repair under amplitude excitation. (d) Two completions of the same learned sphere representation yield radial errors of 1.000 versus 0.017.

2D nonlinear Burgers, whose $(1, 1)$ Galilean identity is verified numerically to $1.1 \times 10^{-6}$: directional gain recovers from 0.086 to 0.943 and offset error drops from 0.911 to 0.161 at $\sigma_c = 0.05$ across eight seeds (Figure 3b, Table 7).

**Unit normalization eliminates the radial generator.** For linear periodic diffusion, homogeneity implies $D(au) = aD(u)$, but unit-RMS normalization removes the radial generator. Adding an amplitude excitation of $\sigma_a = 0.3$ reduces radial JVP error from 0.135 to 0.025 and resolves the $2\times$ amplitude counterfactual without degrading performance on the unit sphere (eight paired seeds; Table 5). Most decisively, the constant extension $C_{\text{const}}(u) = f(u/\|u\|)$ and the homogeneous extension $C_{\text{hom}}(u) = \|u\| f(u/\|u\|)$ derived from the same trained sphere model match identically on the training manifold, yet yield radial JVP errors of 1.000 and 0.017, respectively. The baseline FNO falls between these extremes (0.135 error): architectural inductive bias implicitly selects a partially correct completion, demonstrating that off-support behavior is governed by prior assumptions rather than empirical identifiability.

## 6 Architecture as a Completion Prior

To isolate off-support behavior from in-distribution accuracy, I attach linear mean extensions $C_\alpha(u) = f(u^\circ) + \alpha m(u)$ (with $\alpha \in \{0, 1, 2\}$, plus an exact Galilean completion) to eight zero-mean-trained Burgers models ($u^\circ = u - m(u)$). All four extensions match observed data to within $1.2 \times 10^{-7}$, yet their directional derivative errors along the mean span from 1.000 to 0.018 (Table 11). Standard validation loss cannot distinguish among them.

**Distinct architectures select different completions.** To test learned completions, I train an accuracy-matched panel of different architectures on identical zero-mean data: standard FNOs, grid-free FNOs, and a localized dilated CNN, matched to common validation error (test map errors of 0.059–0.061, Appendix F). Along the unexcited mean direction, the FNOs select near-zero completions (gains of 0.025 and 0.028), whereas the CNN partially transfers the offset ($0.171 \pm 0.014$), exceeding seed variation by a factor of seven. Derivative disagreement between models rises steeply as excitation falls with training excitation (Spearman $\rho = -0.95$; Figure 4), with cross-architecture variance dominating seed-level variance. Where data excite the input space, all architectures con-

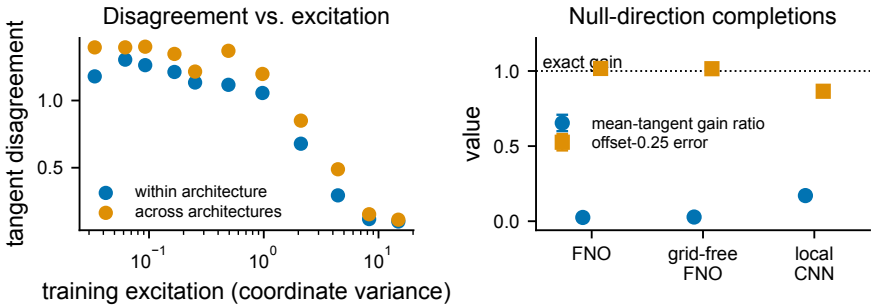


Figure 4: **Disagreement concentrates in weakly excited directions, and completions diverge across model families.** Left: per-mode tangent disagreement versus training excitation. Right: zero-mode completion responses at matched test error (error bars indicate standard deviation across seeds).

verge on the true derivative; where data are absent, intervention responses reflect only architectural parameterization (Table 10).

**Completions are predictable via a continuity dichotomy.** This divergence stems from layer parameterization in translation-equivariant models. A localized spatial convolution has an analytic transfer function (a trigonometric polynomial in wavenumber $k$), tethering its $k = 0$ gain to neighboring, well-excited wavenumbers through Lipschitz continuity. Conversely, spectral operators assign an independent complex weight to each Fourier mode, uncoupling the $k = 0$ multiplier from excited frequencies. Nonlinear models confirm this mechanism across eight seeds (Table 16): the CNN shifts its $k = 0$ gain substantially from initialization lands near its trained $k = 1$ response (relative continuity gap of $0.50 \pm 0.11$), whereas FNO responses remain an order of magnitude lower (gaps of $0.80$–$0.82$), with $k = 0$ weights updating $4\times$–$5\times$ less than excited modes. The FNO constant mode is not strictly static, but its update is decoupled from non-zero wavenumbers. Architectural inductive bias is therefore not merely a generic prior; for these model families, its completion is predictable from first principles.

**The phenomenon extends beyond operators.** This ambiguity appears in simple tabular regression. Training 2 MLPs and a self-attention model on mean-centered feature vectors with an underlying linear mean dependence (target slope $+2$) yields low test error on in-distribution features ($2.6$–$4.9\%$), yet mean-direction intervention slopes range from $-1.5$ to $+3.7$ across architectures and random seeds (Table 17, Appendix I.5). Without a continuity prior, unconstrained completions are unstable even across initializations of the same model family.

**Retraining stability provides only one-way evidence.** While seed-to-seed variance increases as excitation decreases, unexcited directions can remain stably incorrect. Across seeds, the unexcited mean mode yields an FNO gain of $0.027 \pm 0.006$ against the true value of $1.0$. High retraining variance indicates that a direction is unconstrained, but low variance does not imply empirical identification; shared inductive bias can yield stable, incorrect predictions.

## 7 Prospective Acquisition and Targeted Repair

### 7.1 Selecting Queries Without Target Derivatives

Can the covariance audit predict which function values to acquire to resolve missing mechanisms? In eight independent trials on a non-monotone spectral problem, I select four modes before observing any target JVPs using four criteria under an identical 64-query budget: ensemble JVP disagreement, lowest empirical covariance, highest wavenumber, and uniform random sampling. Covariance-based selection reduces weak-mode error to $0.010$, outperforming predictive disagreement ($0.024$) and random sampling ($0.083$), and approaching the oracle baseline ($0.005$; paired confidence intervals

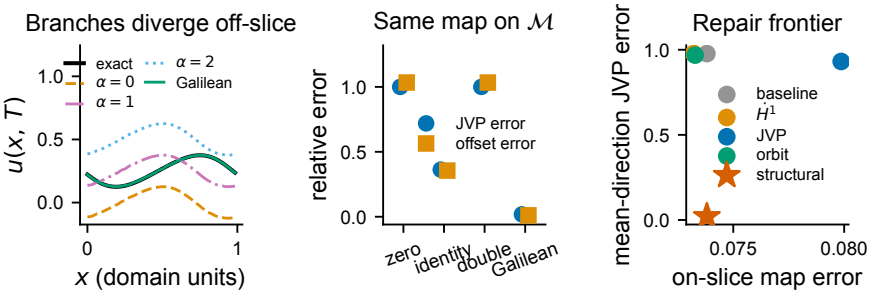


Figure 5: **Observationally equivalent completions diverge off-slice, while structural completion restores the mechanism at zero cost.** (a) Solution profiles under a constant mean offset for four completions of the same baseline model; only the Galilean branch matches the exact solution. (b) Summary of completion branches. (c) Repair Pareto frontier, with the structural completion denoted by a star.

exclude zero; Table 12). Empirical covariance, rather than spatial frequency or ensemble disagreement, provides the decisive acquisition signal for targeting unidentified modes in controlled settings.

### 7.2 Can the missing Burgers mechanism be repaired for free?

I evaluate four intervention strategies to recover the missing Burgers mechanism: additional zero-mean samples, a homogeneous $\dot{H}^1$ Sobolev loss, Galilean orbit augmentation, and direct mean-direction JVP supervision. Interventions are evaluated under a Pareto criterion (validation map error within 20%; three tuning seeds, five untouched test seeds). Direct JVP supervision yields modest tangent improvements at a small validation penalty, weak orbit augmentation provides smaller gains with no validation loss, and the $\dot{H}^1$ penalty is indistinguishable from the baseline because its weight at $k = 0$ is identically zero (Table 13). Models that accurately recover the tangent derivative consistently fail the validation constraint, even with increased network width.

**Structural completion provides zero-cost repair.** Knowing the unidentified direction enables an exact, zero-cost structural repair. Wrapping each held-out model $f$ with the analytical Galilean completion,

$$C(u) = \text{shift}(f(u - m(u)),\, m(u)T) + m(u), \tag{4}$$

requires no additional training data or loss terms and matches $f$ on zero-mean inputs to within $1.5 \times 10^{-7}$. Across the five test seeds, this post-hoc wrapper reduces tangent error from $0.977$ to $0.021$ and offset error from $1.016$ to $0.013$ with zero degradation in validation accuracy (Table 8; Table 14). As shown in Figure 5, spectral weighting provides no improvement, learned supervision sacrifices in-distribution accuracy, and structural completion restores the mechanism for free. The primary challenge is identifying which direction is unconstrained, which the covariance audit resolves directly.

**Circularity and robustness under symmetry misspecification.** This structural repair is not circular: the wrapper requires only the symmetry group action, without evaluating solution values, derivatives, or solver queries. It post-hoc imposes the same inductive bias that an equivariant architecture encodes a priori. Furthermore, the repair degrades gracefully under physical misspecification. When trained on damped Burgers ($u_t + uu_x = \nu u_{xx} - \gamma u$), which breaks Galilean invariance by a tunable parameter $\gamma$, applying the exact Galilean wrapper yields offset errors that match the physical symmetry violation to three decimal places across all $\gamma$ (Table 15). The wrapped model remains superior to the raw baseline until the symmetry violation reaches order one.

## 8 Related work and novelty boundary

**Neural-operator sensitivities.** Spectral auditing demonstrates that accurate operator predictions can coexist with distorted gains and spurious mode coupling (Gao et al., 2026). While derivative-

informed, sensitivity-constrained, and Sobolev training incorporate derivative supervision (Cao et al., 2025; Behroozi et al., 2025; Czarnecki et al., 2017; Cho et al., 2024), neither the diagnostic probe nor derivative training constitutes the primary contribution (Li et al., 2021; Kovachki et al., 2023). Instead, I address a prior question: which directional sensitivities are identified by function values alone, and can unidentified directions be predicted before querying expensive derivatives?

**Symmetry and canonicalization.** Lie point symmetry data augmentation (Brandstetter et al., 2022) corresponds to the orbit excitation repair, deployed for identification rather than boost baseline accuracy. While equivariant and canonicalized neural operators (Shumaylov et al., 2025; Yang et al., 2024; Xu et al., 2026) enforce known symmetries directly, I show that canonicalization acts as an information bottleneck that systematically eliminates the empirical evidence needed to identify the symmetry generator's derivative. I note that the "non-identifiability" of boundary-indexed operator families (Shikhman, 2026) concerns model-class capacity—a notion distinct from and complementary to the analysis of statistical identifiability for local operator derivatives.

**Underspecification and predictive multiplicity.** The divergence of risk-equivalent models under distribution shift is formalized by underspecification (D'Amour et al., 2022) and Rashomon sets of near-optimal candidates (Semenova et al., 2022; Marx et al., 2020). I provide a localized, direction-resolved refinement of this phenomenon: while underspecification establishes that equivalent models diverge off-support, the analysis specifies exactly where they diverge (along support-normal directions), by how much (ordered by excitation), and how to resolve the ambiguity (via targeted data acquisition or exact structural completion).

**Off-distribution interventions in interpretability.** Prior interpretability studies have noted that subspace patching can activate dormant circuits (Makelov et al., 2024), corruptions push representations off-distribution (Zhang & Nanda, 2024), simplified proxies fail out-of-distribution (Friedman et al., 2024), and causal claims require explicit identification assumptions (Lin & Liu, 2026). I complement these critiques from a data-support perspective within physical systems that admit closed-form ground truth. Non-PDE experiment (Appendix I.5) demonstrates that this same off-support ambiguity arises in the MLP and attention architectures studied in mechanistic interpretability. While recent benchmarks test interpretations against synthetic ground-truth circuits (Gupta et al., 2024; Palumbo et al., 2025), I delineate when an intervention response is empirically identified in the first place.

## 9 Conclusion

**Limitations.** The formal completion theorem is framed in finite dimensions; empirical validations focus on spectral operators in 1D and 2D, Burgers equations, and linear diffusion; and the non-PDE example is intentionally stylized. Experiments are conducted at moderate spatial resolutions (grids up to $128^2$), which is necessary to evaluate exact numerical reference derivatives, leaving production-scale 3D foundation models for future audits. In addition, the dissipation analysis evaluates two viscosity regimes, targeted acquisition is tested on discrete modes, and the attention architecture is evaluated separately from the matched comparison panel (though an unconstrained sensitivity pass reproduces the panel's conclusions; Appendix F).

An intervention response does not automatically constitute a learned mechanism: it is identified only along directions constrained by the training distribution or enforced by structural priors. Neural operators render this distinction empirically verifiable through exact mathematical and physical references. The framework is predictive across each stage: training covariance reveals which directional mechanisms are absent, the governing physics dictates which missing directions matter, the architectural inductive bias determines the off-support completion, and known symmetries enable zero-cost structural repairs that degrade gracefully under misspecification. Interpretability evaluations should therefore report training data support, assess causal interventions across observationally equivalent model initializations, and separate data-identified responses from architectural completions. Stable behavior can reflect an inductive bias, accurate responses can be accidental, and neither proves empirical identification without data excitation.

## REPRODUCIBILITY STATEMENT

All experimental hypotheses and falsification criteria were preregistered in experiment ledgers before the decisive runs, and ten negative results are cataloged in Appendix H. All audits are preceded by solver-level verification, including symmetry checks down to $10^{-16}$, grid-resolution consistency tests, and calibration against six analytic operations.

## AI USE STATEMENT

Following ICLR disclosure policies, generative AI was used to refine experimental configurations from author-specified research questions, assist in scripting experimental pipelines, and polish prose and citation formatting. Generative AI was not used to synthesize data (all datasets originate from numerical PDE solvers or the official benchmark distribution) or produce any reported quantitative values.

## REFERENCES

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

Johannes Brandstetter, Max Welling, and Daniel E. Worrall. Lie point symmetry data augmentation for neural PDE solvers. In *International Conference on Machine Learning*, 2022.

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.

Namkyeong Cho, Junseung Ryu, and Hyung Ju Hwang. Sobolev training for operator learning. *arXiv preprint arXiv:2402.09084*, 2024.

Wojciech M. Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Świrszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Dan Friedman, Andrew Lampinen, Lucas Dixon, Danqi Chen, and Asma Ghandeharioun. Interpretability illusions in the generalization of simplified models. In *International Conference on Machine Learning*, 2024.

Zhiwei Gao, Liu Yang, and George Em Karniadakis. Spectral audit of in-context operator networks. *arXiv preprint arXiv:2606.02427*, 2026.

Rohan Gupta, Iván Arcuschin, Thomas Kwa, and Adrià Garriga-Alonso. InterpBench: Semi-synthetic transformers for evaluating mechanistic interpretability techniques. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

John M. Lee. *Introduction to Smooth Manifolds*. Springer, second edition, 2013.

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

Zezheng Lin and Fengming Liu. Position: Mechanistic interpretability must disclose identification assumptions for causal claims. *arXiv preprint arXiv:2605.08012*, 2026.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International Conference on Machine Learning*, 2020.

Nils Palumbo, Ravi Mangal, Zifan Wang, Saranya Vijayakumar, Corina S. Pasareanu, and Somesh Jha. Validating mechanistic interpretations: An axiomatic approach. In *International Conference on Machine Learning*, 2025.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *ACM Conference on Fairness, Accountability, and Transparency*, 2022.

Lennon J. Shikhman. One operator to rule them all? On boundary-indexed operator families in neural PDE solvers. *arXiv preprint arXiv:2603.01406*, 2026.

Zakhar Shumaylov, Peter Zaika, James Rowbottom, Ferdia Sherry, Melanie Weber, and Carola-Bibiane Schönlieb. Lie algebra canonicalization: Equivariant neural operators under arbitrary Lie groups. In *International Conference on Learning Representations*, 2025.

Jiaxiao Xu, Changhong Mou, Yeyu Zhang, and Fengxiang He. Physics-aligned canonical equivariant Fourier neural operator under symmetry-induced shifts. *arXiv preprint arXiv:2605.18606*, 2026.

Jianke Yang, Wang Rao, Nima Dehmamy, Robin Walters, and Rose Yu. Symmetry-informed governing equation discovery. In *Advances in Neural Information Processing Systems*, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024.

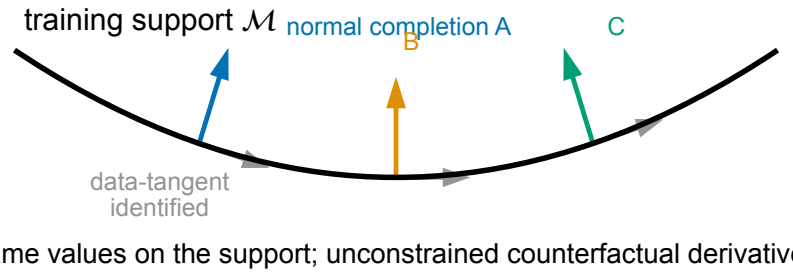


Figure 6: **Distribution-relative mechanism identification.** Function values on a support manifold constrain data-tangent derivatives (gray) but not transverse responses (colored); an architecture or structural assumption chooses the completion.

## A Full Proof of Normal Non-Identifiability

**The full distinction from OOD generalization.** Beyond the two senses in the introduction: (i) the object of study is infinitesimal and mechanistic — signed directional derivatives against exact references — not aggregate error under a macroscopic shift; a model can have every tangent wrong in a direction and still score perfectly on any OOD benchmark that does not move along it; (ii) the failure is invisible in principle to i.i.d. evaluation, since test sets inherit the training support; (iii) the organizing variable is measured per direction (training excitation, disentangled from frequency by the adversarial designs of Section 4) rather than a qualitative notion of shift; (iv) the account is predictive and constructive, specifying in advance which directions are unidentified, which data to acquire, and how to repair a known-symmetry direction exactly.

**Theorem 1** (Arbitrary normal completion)**.** *Let $\mathcal{M}$ be a closed $C^2$ embedded submanifold of $\mathbb{R}^d$, let $F : \mathbb{R}^d \to \mathbb{R}^p$ be $C^1$, and let $A : N\mathcal{M} \to \mathbb{R}^p$ be a $C^1$ vector-bundle morphism, i.e., $A_x : N_x\mathcal{M} \to \mathbb{R}^p$ is linear for each $x$. Then there is a $C^1$ map $\widetilde{F} : \mathbb{R}^d \to \mathbb{R}^p$ satisfying, for every $x \in \mathcal{M}$,*

$$\widetilde{F}(x) = F(x), \tag{5}$$

$$D\widetilde{F}_x|_{T_x\mathcal{M}} = DF_x|_{T_x\mathcal{M}}, \tag{6}$$

$$D\widetilde{F}_x|_{N_x\mathcal{M}} = DF_x|_{N_x\mathcal{M}} + A_x. \tag{7}$$

*Proof.* By the tubular-neighborhood theorem (Lee, 2013), there is an open neighborhood $U$ of $\mathcal{M}$ and a $C^1$ projection $\pi : U \to \mathcal{M}$ such that every $z \in U$ has a unique decomposition

$$z = \pi(z) + n(z), \qquad n(z) \in N_{\pi(z)}\mathcal{M}. \tag{8}$$

Choose a $C^1$ bump function $\chi : \mathbb{R}^d \to [0, 1]$ supported in $U$ and equal to one on a smaller neighborhood of $\mathcal{M}$. Define

$$H(z) = \begin{cases} \chi(z)A_{\pi(z)}n(z), & z \in U, \\ 0, & z \notin U. \end{cases} \tag{9}$$

Because $\chi$ vanishes before the boundary of $U$, this piecewise definition is $C^1$. For $x \in \mathcal{M}$, $n(x) = 0$, so $H(x) = 0$.

The derivative of $n(z) = z - \pi(z)$ at $x \in \mathcal{M}$ is the orthogonal projection $P_{N_x\mathcal{M}}$. Differentiating $H$ at $x$, all terms involving derivatives of $\chi$, $A$, or $\pi$ multiply $n(x) = 0$. Hence

$$DH_x w = A_x P_{N_x\mathcal{M}} w. \tag{10}$$

It follows that $DH_x v = 0$ for $v \in T_x\mathcal{M}$ and $DH_x n = A_x n$ for $n \in N_x\mathcal{M}$. Set $\widetilde{F} = F + H$. □

**Risk consequence.** If $\mu(\mathcal{M}) = 1$ and a loss depends only on function values, then $F$ and $\widetilde{F}$ have exactly the same population risk. Therefore, no amount of function-value data sampled only from $\mathcal{M}$ can select $A$.

**Finite versus infinite dimensions.** The theorem applies directly to every finite grid used in this paper. In a Hilbert or Banach function space, the same construction requires a split $C^1$ tubular neighborhood and smooth bump functions. These are additional geometric assumptions and are not automatic. I therefore state the main result for discretized operator learning rather than claiming unrestricted Banach-space generality.

**What the theorem does not say.** It does not imply that optimization is unstable, that every architecture can realize every $A$, or that learned normal derivatives must be inaccurate. Architecture and optimization can select a deterministic completion. The theorem says only that this selection is not identified by values on $\mathcal{M}$.

# B Weak Excitation under Ridge Regression

**Proposition 1.** *Let $x \in \mathbb{R}^d$ have covariance $\Sigma \succeq 0$ and let $y = Ax$. The minimizer of*

$$\min_B \ \mathbb{E}\|Bx - Ax\|_2^2 + \lambda \|B\|_F^2, \qquad \lambda > 0, \tag{11}$$

*is*

$$B_\lambda = A\Sigma(\Sigma + \lambda I)^{-1}. \tag{12}$$

*If $\Sigma v_i = s_i v_i$, then*

$$B_\lambda v_i = \frac{s_i}{s_i + \lambda} A v_i, \qquad \|(B_\lambda - A)v_i\| = \frac{\lambda}{s_i + \lambda}\|Av_i\|. \tag{13}$$

*Proof.* The first-order condition is $(B - A)\Sigma + \lambda B = 0$. Rearranging and right-multiplying by $(\Sigma + \lambda I)^{-1}$ gives the solution. Applying it to an eigenvector gives the directional formula. □

This exact result is used as an explanatory model. It is not presented as a theory of nonlinear FNO training. The adversarial covariance experiment tests its qualitative directional prediction in a nonlinear parameterization.

# C Symmetry Identities

## C.1 Burgers Galilean Identity

Consider periodic viscous Burgers

$$u_t + uu_x = \nu u_{xx}. \tag{14}$$

If $u(x,t)$ is a solution, then

$$u_c(x,t) = u(x - ct, t) + c \tag{15}$$

is the solution initialized at $u(x,0) + c$. Thus, for the horizon-$T$ solution operator,

$$G_T(v + c)(x) = G_T(v)(x - cT) + c. \tag{16}$$

Differentiating with respect to $c$ at zero yields

$$DG_T(v)\mathbf{1} = \mathbf{1} - T\partial_x G_T(v). \tag{17}$$

For the normalized DC Fourier coordinate, both sides are divided by $\sqrt{N}$. I evaluate the right side spectrally, without differentiating through the numerical time integrator.

## C.2 Positive Homogeneity Identity

The periodic diffusion map at horizon $T$ is

$$D_T u = \mathcal{F}^{-1}\left[\exp\{-\nu(2\pi k)^2 T\}\mathcal{F}u\right]. \tag{18}$$

It is linear and therefore positively homogeneous: $D_T(au) = aD_T u$. Differentiating at $a = 1$ gives

$$DD_T(u)[u] = D_T u. \tag{19}$$

When the radial probe direction is $u/\|u\|_2$, the reference response is $D_T u/\|u\|_2$.

# D SIGNED REAL FOURIER TANGENT REPRESENTATION

## D.1 BASIS CONSTRUCTION

For a one-dimensional real field of length $N$, each non-self-conjugate frequency $k$ contributes

$$q_{k,c}(x_j) = \sqrt{\frac{2}{N}} \cos(2\pi kj/N), \tag{20}$$

$$q_{k,s}(x_j) = \sqrt{\frac{2}{N}} \sin(2\pi kj/N). \tag{21}$$

DC and, for even $N$, Nyquist frequencies have one real coordinate. The multidimensional basis keeps one representative from each Hermitian conjugate pair. The resulting synthesis matrix has orthonormal rows.

For each basis direction $q_j$, I evaluate

$$DF_x q_j = \frac{\mathrm{d}}{\mathrm{d}\epsilon} F(x + \epsilon q_j)\Big|_{\epsilon=0} \tag{22}$$

with forward-mode automatic differentiation and project the response onto the output basis. This produces a signed real matrix, not a matrix of magnitudes.

## D.2 VALIDATION CONTROLS

| control | expected property | outcome |
|---|---|---|
| identity | signed identity tangent | numerical precision |
| diffusion | analytic diagonal gains | numerical precision |
| translation | rotation within each Fourier block | block fraction 1.000 |
| translation | coordinate diagonality need not be one | 0.568 |
| square map | $D(u^2)[v] = 2uv$ | numerical precision |
| finite difference | agrees with autograd JVP | passed |
| 2D basis | orthonormal, no conjugate duplication | passed |

Table 1: Calibration of the phase-aware tangent representation.

The translation control is important. A spatial shift multiplies a complex Fourier coefficient by a phase; in real coordinates this is a $2 \times 2$ rotation within the cosine/sine block. Magnitude-only or coordinate-diagonal scores can incorrectly interpret this physical phase response as cross-frequency coupling.

## D.3 METRICS

For estimated and reference tangents $\widehat{J}, J$:

$$e_{\tan}(x) = \frac{\|\widehat{J} - J\|_F}{\|J\|_F}, \qquad \cos(\widehat{J}, J) = \frac{\langle \widehat{J}, J \rangle_F}{\|\widehat{J}\|_F \|J\|_F}. \tag{23}$$

For a real Fourier mode $k$ with one or two coordinate columns $I_k$, its gain is $\|J_{:,I_k}\|_F$. All main claims use signed JVP error or gain ratios. Magnitude-only matrices are not used as evidence.

# E EXPERIMENTAL PROTOCOLS

## E.1 COMMON FNO

The controlled experiments use a one-dimensional FNO with lifting width 24, three spectral layers, 12–16 retained modes, pointwise $1 \times 1$ paths, GELU, and a two-layer projection head. Models are optimized with AdamW, weight decay $10^{-4}$, cosine learning-rate decay, and batch size 32. Conditions share data draws and initialization seeds wherever possible.

| experiment | data units | model units | training | primary held-out quantities |
|---|---|---|---|---|
| stored checkpoint | 8 probe inputs | 1 checkpoint | existing 120-epoch model | signed tangent, nonzero-mode tangent, mean offsets |
| mean excitation | 256 train, 64 test | 4 stds $\times$ 8 seeds | 50 epochs | mean JVP, zero-mean map error, offsets |
| variance–frequency | 256 train, 64 test | 4 assignments $\times$ 10 seeds | 60 epochs | per-model excitation/error and wavenumber/error associations |
| variance–frequency 2D | 256 train, 64 test | 4 assignments $\times$ 10 seeds | 60 epochs | same associations on a $32^2$ grid with $\lvert k\rvert$ as wavenumber |
| mean excitation 2D | 256 train, 64 test | 4 stds $\times$ 8 seeds | 60 epochs | Galilean JVP, zero-mean map error, offsets (nonlinear 2D Burgers) |
| architecture panel | 256 train, 64 val, 64 test | 4 architectures $\times$ 8 seeds | checkpoint-matched | null-direction completions, per-mode tangent disagreement |
| structural completion | 256 train, 64 test | 5 paired seeds | 60 epochs | map/JVP/offset of the exact Galilean wrapper |
| amplitude | 256 train, 64 test | 4 stds $\times$ 8 seeds | 60 epochs | radial JVP, unit sphere, amplitudes 0.5–2 |
| completion priors | 256 train, 64 test | 4 branches $\times$ 8 cores | 60 epochs/core | exact on-slice agreement, mean JVP, offset |
| prospective acquisition | 128 train, 32 selection, 64 test | 5 methods $\times$ 8 trials | 50 epochs | weak-mode JVP and flat-spectrum counterfactual |
| repair selection | 256 train, 64 validation, 64 untouched test | 3 tuning + 5 test seeds | 60 epochs | map/JVP/offset Pareto performance |

Table 2: Experimental units and data splits. Modes within one trained model are not treated as independent inferential samples.

### E.2 Burgers Data

The base data solve $u_t + uu_x = \nu u_{xx}$ on the periodic unit interval using a pseudospectral integrating-factor RK4 solver with $2/3$ dealiasing. Initial conditions are Gaussian random fields with zero DC mode. Stored 256-point fields are downsampled by four in the controlled follow-ups; viscosity is $0.05$ and horizon is $T = 1$. The same base trajectories are used for all mean-orbit conditions.

### E.3 Variance–Frequency Target

The target is the fixed spectral multiplier

$$G(u) = \mathcal{F}^{-1}[g(k)\mathcal{F}u], \qquad g(k) = 0.25 + 0.65\min(k/15, 1)^2. \tag{24}$$

Nonzero real Fourier-coordinate variances span three orders of magnitude. Natural, inverted, flat, and fixed-permutation conditions have equal expected field energy. Training targets include Gaussian observation noise with standard deviation $0.05$. The output target and architecture never change across covariance assignments.

### E.4 Amplitude Target

Unit-RMS random fields have a decaying input spectrum and zero mean. The target is periodic diffusion with $\nu = 0.01$, $T = 0.2$. Training amplitudes are $1 + \sigma z$, clipped below at $0.2$, for $\sigma \in \{0, 0.02, 0.1, 0.3\}$. Unit-sphere inputs are held fixed across conditions.

### E.5 2D Variance–Frequency Replication

The 2D target is the fixed radial spectral multiplier

$$G(u) = \mathcal{F}^{-1}[g(|k|)\mathcal{F}u], \qquad g(|k|) = 0.25 + 0.65\min(|k|/10, 1)^2 \tag{25}$$

on a $32 \times 32$ periodic grid. The probe covers the unique real Fourier modes with $|k_x|, |k_y| < 8$ (113 modes, 225 coordinates), and the same variance multiset is assigned to modes sorted by $|k|^2$ in natural, inverted, flat, and fixed-permutation orders with equal expected field energy and observation noise $0.05$. The learner is a grid-free FNO2d (three spectral layers, eight retained modes per dimension, width 24). All forty models reach $0.8$–$2.2\%$ test error. Wavenumber is the Euclidean norm $|k|$.

### E.6 2D Burgers Galilean Excitation

Scalar 2D Burgers, $u_t + u(u_x + u_y) = \nu \Delta u$ with $\nu = 0.05$ and horizon $T = 0.5$, is solved pseudospectrally (integrating-factor RK4, $2/3$ dealiasing, $dt = 5 \times 10^{-4}$) at $64^2$ and downsampled to $32^2$ for training. Initial conditions are zero-mean Gaussian random fields normalized to unit standard deviation. The Galilean identity $G_T(v + c)(x, y) = G_T(v)(x - cT, y - cT) + c$ is verified on the solver before training: constructing $G_T(v + c)$ directly versus by shift-and-add of $G_T(v)$ agrees to $1.1 \times 10^{-6}$ relative error at offsets $0.25$ and $-0.4$. The trained model is an FNO2d with coordinate channels (three layers, eight modes, width 24); the mean-direction probe uses the exact tangent $DG_T(v)\mathbf{1} = \mathbf{1} - T(\partial_x + \partial_y)G_T(v)$ evaluated spectrally.

### E.7 Accuracy-Matched Architecture Panel

Four architectures are trained on identical zero-mean Burgers data (grid 64, 256 train, 64 validation, 64 test): a standard FNO with coordinate channel, a grid-free FNO (both: three spectral layers, 12 modes, width 24), a local CNN (five circular-padded convolutions, kernel 5, dilations $1, 2, 4, 8, 1$, width 48, receptive field covering the grid), and a pre-norm softmax-attention operator (three blocks, four heads, model width 48, explicit attention so forward-mode differentiation is exact). Accuracy is matched by checkpoint selection: the target validation error $E^*$ is the across-seed median best validation error of the least accurate architecture, and every run is deterministically retrained to its earliest epoch with validation error at or below $E^*$; runs that never reach $E^*$ are excluded from matched comparisons and reported. Signed tangents are evaluated at four shared probe inputs over the real Fourier modes with $|k| < 12$; per-mode disagreement between two models is the relative Frobenius difference of the signed input-mode column blocks, and excitation is the empirical per-mode coordinate variance of the training inputs.

### E.8 Structural Galilean Completion

The structural arm wraps each held-out baseline core $f$ (trained on the zero-mean slice with the protocol of the repair experiment, same seeds) with

$$C(u) = \mathrm{shift}(f(u - m(u)),\, m(u)T) + m(u), \tag{26}$$

where $m(u)$ is the spatial mean and the shift is implemented exactly by a Fourier phase. $C$ consumes no additional data or loss terms and agrees with $f$ on the zero-mean slice to numerical precision.

### E.9 Navier–Stokes Benchmark Audit

Data follow the published benchmark specification: vorticity form on the periodic unit torus with $\nu = 10^{-3}$, forcing $0.1(\sin 2\pi(x+y) + \cos 2\pi(x+y))$, and initial vorticity from the Gaussian random field $N(0,\, 7^{3/2}(-\Delta + 49I)^{-2.5})$ whose $k = 0$ coefficient the official generator sets to zero. Trajectories are solved pseudospectrally (integrating-factor RK4, $2/3$ dealiasing, $dt = 10^{-3}$) at $128^2$ and downsampled to $64^2$; unit-time input/target pairs are taken at $t \in [5, 20)$ from 60 training and 20 test trajectories (900/300 pairs). The operator is a benchmark-style FNO2d (12 retained modes per dimension, width 20, four layers, coordinate channels) trained with the relative-$L_2$ loss, eight seeds. Before any training, two solver controls run: the mean identity $G(w + c) = G(w) + c$ (exact on the torus because the $k = 0$ vorticity mode produces no velocity and the zero-mean forcing conserves it) and coarse/fine resolution consistency. The probe covers real Fourier modes with $|k_x|, |k_y| < 8$; the solver reference tangent is a batched float64 central difference ($\epsilon = 10^{-4}$) through the $64^2$ solver, and excitation is the empirical per-mode coordinate variance of the 900 training inputs.

**Physical status of the DC direction.** On the torus a nonzero mean vorticity corresponds to no periodic velocity field, so the DC direction is a discretization degree of freedom rather than a realizable

flow state — the exactly-conserved endpoint of the spectrum of weakly excited directions. The solver treats it consistently (the $k = 0$ mode is excluded from the stream-function solve and conserved by the zero-mean forcing), which is what makes $G(w + c) = G(w) + c$ exact. Nothing in the account depends on this special status: the 1D and 2D Burgers blind spots lie along the Galilean boost, a fully physical direction, and the low-viscosity band experiment exposes ordinary weakly excited physical directions.

### E.10 Prospective Acquisition

The target is an irregular, fixed spectral multiplier:

$$g(k) = 0.2 + 0.7\{0.5 + 0.5\sin(1.7k + 0.3)\}. \tag{27}$$

Input coordinate variances span four orders of magnitude and are assigned by a fixed permutation. Each trial trains a three-member ensemble. Selection scores are: (1) standard deviation of model-JVP mode gain across the ensemble; (2) empirical training-coordinate variance; (3) wavenumber; (4) a seeded random ordering; and (5) true pre-acquisition JVP error (oracle only). Each method selects four nonzero modes. Sixty-four new inputs are formed by adding signed basis perturbations of amplitude three to sampled base inputs; only function values of the target are added. Target JVPs are computed after selection for evaluation.

### E.11 Repair Selection

The tuning grid contains widths 24 and 48; JVP weights $0.01, 0.03, 0.1, 0.3$ (a subset at width 48); and orbit amplitudes $0.025, 0.05, 0.1, 0.25$. The JVP objective uses a centered finite difference along normalized DC. The $\dot{H}^1$ baseline weights output errors by squared wavenumber and therefore gives DC exactly zero weight. For each method, selection minimizes mean validation tangent error among configurations satisfying $e^{\mathrm{val}}_{\mathrm{map}} \le 1.2\, e^{\mathrm{val}}_{\mathrm{map,paired\ baseline}}$.

## F Complete Numerical Results

### F.1 Stored Checkpoint Audit

Table 3 reports the corrected signed audit of the stored Burgers checkpoint.

| quantity | value |
|---|---|
| in-distribution relative $L_2$ | 0.0628 |
| full signed tangent error | 0.9852 |
| full signed tangent cosine | 0.1725 |
| nonzero-input tangent error | 0.1663 |
| nonzero-input tangent cosine | 0.9873 |
| nonzero block-pattern correlation | 0.9965 |
| solver mean gain | 1.075 |
| FNO mean gain | 0.0107 |
| offset $-0.25$ error | $\approx 1.02$ |
| offset $+0.25$ error | $\approx 1.02$ |

Table 3: Corrected signed audit of the stored Burgers checkpoint.

### F.2 Mean/Galilean Excitation

Table 4 reports the 1D create-and-repair sweep for the mean direction.

### F.3 Amplitude Excitation

Table 5 reports the eight-seed amplitude-normalization sweep.

| mean std | gain ratio | tangent error | $+0.25$ error |
|---|---|---|---|
| 0 | $0.028 \pm 0.005$ | $0.975 \pm 0.006$ | 1.008 |
| 0.01 | $0.043 \pm 0.011$ | $0.959 \pm 0.011$ | 0.993 |
| 0.05 | $0.963 \pm 0.012$ | $0.086 \pm 0.017$ | 0.255 |
| 0.2 | $0.968 \pm 0.009$ | $0.181 \pm 0.014$ | 0.106 |

Table 4: Create-and-repair sweep for the Burgers mean direction (eight paired seeds).

| amplitude std | sphere error | radial error | amplitude-two error |
|---|---|---|---|
| 0 | 0.0168 | 0.1348 | 0.1280 |
| 0.02 | 0.0169 | 0.1303 | 0.1252 |
| 0.1 | 0.0168 | 0.0795 | 0.0897 |
| 0.3 | 0.0172 | 0.0254 | 0.0351 |

Table 5: Eight-seed amplitude-normalization experiment.

### F.4 2D Variance–Frequency Associations

Figure 7 shows the raw per-mode data behind the 2D condition summaries. Table 6 summarizes the correlations.

| assignment | excitation assoc. | wavenumber assoc. |
|---|---|---|
| natural | $-0.558$ $[-0.620, -0.492]$ | $+0.558$ $[+0.491, +0.619]$ |
| inverted | $-0.722$ $[-0.757, -0.685]$ | $-0.723$ $[-0.758, -0.686]$ |
| permuted | $-0.602$ $[-0.688, -0.481]$ | $-0.104$ $[-0.174, -0.041]$ |
| flat | — | $-0.239$ $[-0.321, -0.160]$ |

Table 6: 2D replication, ten seeds per condition, cluster bootstrap over models. The paired inverted-minus-natural wavenumber contrast is $-1.281$ $[-1.348, -1.206]$. Pooled partial-$R^2$ drops: excitation $0.326$, wavenumber $0.024$, target gain $0.020$.

### F.5 2D Burgers Galilean Excitation

Table 7 reports the 2D Burgers sweep.

### F.6 Structural Galilean Completion

Table 8 reports the structural wrapper on held-out seeds.

### F.7 Navier–Stokes Benchmark Audit

Figure 8 illustrates on-slice accuracy; Table 9 provides eight-seed statistics.

### F.8 Accuracy-Matched Architecture Panel

Figure 9 and Table 10 document the architecture panel.

Table 11 reports the controlled completion branches over eight cores.

Table 12 reports the eight prospective acquisition trials.

Table 13 reports evaluation on five untouched test seeds after validation-only selection.

## G Statistical Analysis

**Sampling units and bootstrap procedures.** The independently trained condition–seed model is the unit for variance–frequency analysis. The independent trial is the unit for acquisition. Paired

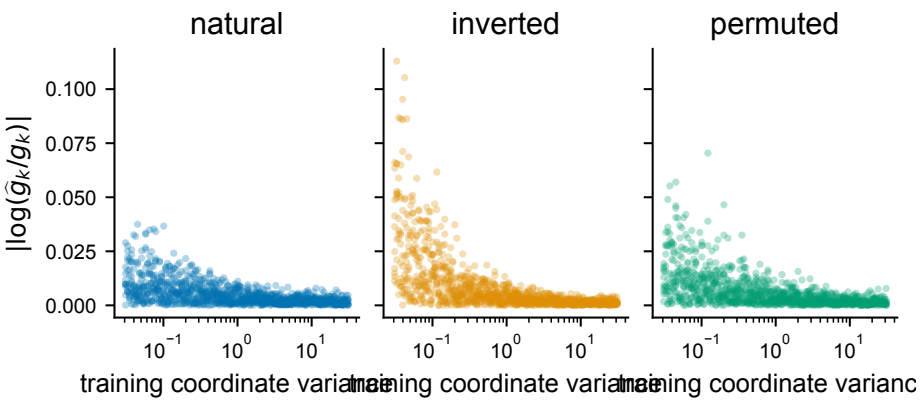


Figure 7: **2D replication, raw per-mode data.** Absolute log tangent-gain error against training-coordinate variance for each nonzero mode, all ten seeds per condition ($32^2$ grid, radial multiplier target).

| mean std | in-dist. error | gain ratio | tangent error | $+0.25$ error |
|---|---|---|---|---|
| 0 | 0.073 | $0.086 \pm 0.006$ | $0.949 \pm 0.006$ | 0.911 |
| 0.01 | 0.099 | $0.131 \pm 0.012$ | $0.892 \pm 0.011$ | 0.857 |
| 0.05 | 0.098 | $0.943 \pm 0.010$ | $0.131 \pm 0.011$ | 0.161 |
| 0.2 | 0.116 | $1.001 \pm 0.004$ | $0.101 \pm 0.009$ | 0.091 |

Table 7: Create-and-repair sweep for the 2D Burgers Galilean direction, eight paired seeds. The solver identity control passes to $1.1 \times 10^{-6}$ before training.

initialization seeds are the units for amplitude and repair. Fourier modes within one model are treated as repeated measurements. Cluster bootstrap resamples the 10 trained models within condition; paired bootstrap resamples paired trial-level differences (10,000 resamples for acquisition, 20,000 for paired statistics).

## H Negative Results and Claim Boundaries

1. **High-frequency attenuation is rejected.** The corrected stored checkpoint amplifies, rather than attenuates, several high-frequency tangent gains relative to the solver. Frequency is not the organizing variable.
2. **A local/identity-dominated covariance pilot is null.** Parameter sharing transfers pointwise sensitivity across weak frequencies. This rejects a universal covariance-shrinkage claim and motivates deliberately putting excitation and frequency in conflict.
3. **Low retraining disagreement is inconclusive.** The absent mean direction is stably wrong across seeds.
4. **Architectural diversity is famzeily-dependent.** Accuracy-matched standard and grid-free FNOs learn nearly identical (near-zero) mean completions. Diversity appears across families: the local CNN chooses a systematically different partial completion at matched accuracy, and attention cannot reach the accuracy band at this data budget.
5. **Generic spectral weighting does not identify DC.** A homogeneous $\dot{H}^1$ loss assigns zero weight to the missing mean mode and matches the baseline tangent.
6. **Strong repair is not free.** Near-complete orbit/JVP repair remains outside the predeclared map-error band, including at width 48.
7. **Disagreement is not the best selector.** It beats random prospectively, but empirical covariance is cheaper and more accurate in this controlled problem.
8. **The literal ridge form fails.** $s/(s+\lambda)$ does not fit trained gain ratios at any training stage or data size; weak directions are amplified as often as attenuated (Appendix I.3).

| condition | map error | tangent error | offset error |
|---|---|---|---|
| baseline | 0.0738 | 0.9774 | 1.0164 |
| structural wrapper | 0.0738 | 0.0211 | 0.0134 |

Table 8: Structural completion on the five held-out repair seeds. The maximum on-slice prediction difference between wrapper and core is $1.5 \times 10^{-7}$; map errors are identical by construction.

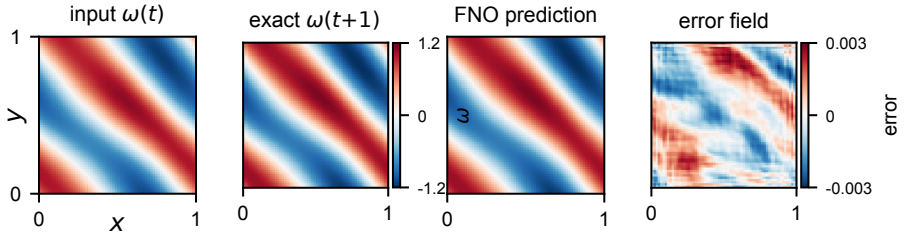


Figure 8: **The same benchmark model is excellent in distribution.** On-slice one-step prediction for the representative test sample of the main-text benchmark figure: the error field (right, own color scale $\pm 0.003$) is roughly $400\times$ smaller than the vorticity fields.

9. **Initial-prior shift does not move the input distribution.** Dissipation erases the prior; the training support is the attractor (Appendix I.1).
10. **Dissipated weak directions are masked at finite amplitude.** Band-limited physical perturbations in weakly excited directions cost no more than strongly excited ones at $\nu = 10^{-3}$, despite measurably wrong tangents; exposure is confined to neutral directions (Appendix I.1).

# I Additional Empirical Audits and Controls

## I.1 Instructive Nulls: When Weak Directions Are Masked

Two attempts to surface finite-amplitude failures in physically ordinary weak directions returned nulls at $\nu = 10^{-3}$:

1. *Prior shift*: Test trajectories evolved from rougher ($\alpha = 1.7$) and smoother ($\alpha = 3.5$) initial GRF priors yield one-step errors of $0.45\%$ and $0.58\%$ versus $0.55\%$ in-prior. Dissipation erases the prior, placing $\sim 10^{-5}$ of their energy in weak modes by the sampled time.
2. *Band perturbations*: Perturbing inputs by matched-norm fields in a weak band ($8 \le |k| \le 11$; per-coefficient training energy $1.6 \times 10^{-2}$) versus a strong band ($1 \le |k| \le 3$; $3.4 \times 10^{5}$) yields equal errors (ratios 0.9–1.2 at relative amplitudes 0.05–0.2). At $\nu = 10^{-3}$, these modes contract by $\sim e^{-4}$ per time unit, so true and learned responses are both near zero.

This masking lifts at lower viscosity: at $\nu = 10^{-4}$ (identity control $2.1 \times 10^{-16}$; resolution consistency $1.6 \times 10^{-3}$; energy gap $\sim 2000\times$), weak/strong error ratios rise to 2.10, 3.09, and 3.85 at relative amplitudes 0.05, 0.1, and 0.2 (three seeds; $2.8\%$ base error). The DC audit also replicates at $\nu = 10^{-4}$: gain 0.130, offset-0.25 error 0.208 reduced to 0.027 by the mean completion.

## I.2 Misspecified Structural Completion

Table 15 documents damped Burgers ($u_t + uu_x = \nu u_{xx} - \gamma u$). The wrapper's offset error equals the measured symmetry violation to three decimals across all $\gamma$, confirming graceful degradation.

## I.3 Failure of the Literal Ridge Law

Fitting $s/(s+\lambda)$ to per-mode gain ratios yields $R^2 \le 0.05$ in the natural, inverted, and permuted conditions at $N = 256$, at every training length in $\{5, 10, 20, 40, 80\}$ epochs, and at every training-set size in $\{64, \ldots, 1024\}$; 47–84% of weak-half modes exhibit gain ratios above one (amplification).

| quantity | mean ± std (8 seeds) |
|---|---|
| one-step test relative $L_2$ | $0.0036 \pm 0.0013$ |
| DC tangent gain (exact $= 1$) | $0.130 \pm 0.019$ |
| DC tangent relative error | $0.884 \pm 0.020$ |
| offset-0.25 error | $0.284 \pm 0.006$ |
| offset-0.1 error | $0.119 \pm 0.003$ |
| Spearman(log excitation, gain error) | $-0.691 \pm 0.041$ |
| Spearman($\lvert k \rvert$, gain error) | $+0.778 \pm 0.029$ |
| completion: DC gain | $1.000 \pm 0.000$ |
| completion: offset-0.25 error | $0.0034 \pm 0.0012$ |

Table 9: Benchmark audit over eight seeds. Solver controls: mean identity $1.9 \times 10^{-16}$, resolution consistency $2.8 \times 10^{-4}$; DC training excitation $1.3 \times 10^{-12}$; training-input standard deviation $0.754$; maximum on-slice difference between completion and core $4.8 \times 10^{-7}$.

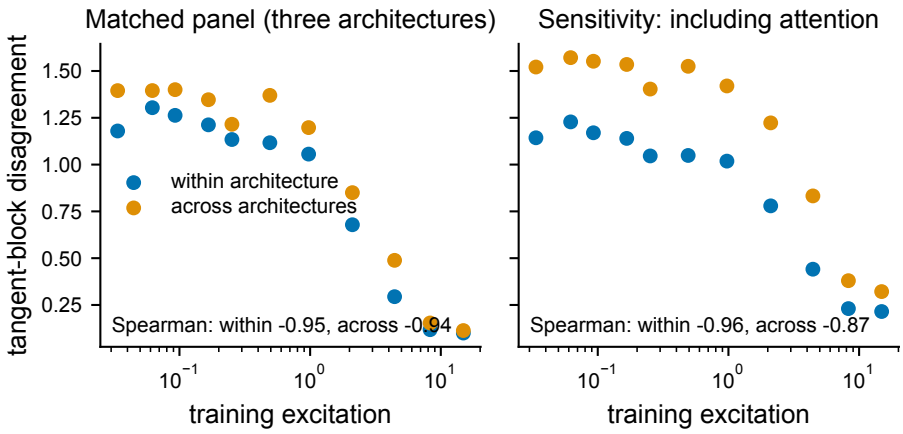


Figure 9: **Sensitivity of the disagreement–excitation relation.** Left: the accuracy-matched three-architecture panel. Right: including attention models at their best achievable checkpoints.

### I.4 Completion-Mechanism Measurements

Table 16 confirms the continuity dichotomy across eight seeds.

### I.5 Non-PDE Demonstration

Table 17 reports tabular regression on 32-dimensional zero-mean features with true off-slice mean slope $+2$ (five seeds each).

### I.6 Audit on Officially Distributed Data

Table 18 reports results using the official Zenodo (record 12825163) dataset at native $128^2$ resolution without modification.

**Compute resources.** Controlled experiments ran on standard Linux workstations using multi-core x86 CPUs and NVIDIA RTX A5000 GPU. Runtimes range from minutes to tens of minutes per experiment.

| architecture | matched $n$ | map error | mean gain | offset error |
|---|---|---|---|---|
| standard FNO | 4/8 | $0.061 \pm 0.008$ | $0.025 \pm 0.005$ | 1.017 |
| grid-free FNO | 8/8 | $0.059 \pm 0.005$ | $0.028 \pm 0.005$ | 1.015 |
| local CNN | 7/8 | $0.059 \pm 0.003$ | $0.171 \pm 0.014$ | 0.865 |
| attention (excluded) | 0/8 | $0.187 \pm 0.018$ | $0.085 \pm 0.037$ | 1.086 |

Table 10: Architecture panel at $E^* = 0.0478$. Matched $n$ counts runs whose validation curve reaches $E^*$. Disagreement associations with log excitation: within architecture $-0.945$ $[-0.991, -0.836]$, across architectures $-0.936$ $[-0.982, -0.827]$. Across-minus-within excess: $+0.152$ (weak half), $+0.111$ (strong half). Including unmatched attention models gives $-0.964$ (within) and $-0.873$ (across).

| completion | map error | tangent error | offset error |
|---|---|---|---|
| mean zero | 0.0503 | 1.000 | 1.034 |
| mean identity | 0.0503 | 0.363 | 0.355 |
| mean double | 0.0503 | 1.000 | 1.033 |
| Galilean | 0.0503 | 0.018 | 0.011 |

Table 11: Controlled branches over eight cores. Maximum prediction difference on the observed slice is $1.2 \times 10^{-7}$.

| selector | weak JVP error | flat error | precision |
|---|---|---|---|
| low covariance | 0.0101 | 0.0393 | 0.844 |
| disagreement | 0.0236 | 0.0478 | 0.656 |
| random | 0.0827 | 0.0891 | 0.313 |
| high wavenumber | 0.0913 | 0.0878 | 0.406 |
| target-JVP oracle | 0.0049 | 0.0437 | 1.000 |

Table 12: Eight prospective trials. Precision is overlap with the true pre-acquisition worst-error quartile, revealed after selection.

| method | map error | tangent error | offset error |
|---|---|---|---|
| baseline | 0.0738 | 0.9774 | 1.0164 |
| $\dot{H}^1$ | 0.0732 | 0.9780 | 1.0170 |
| selected JVP | 0.0799 | 0.9311 | 0.9698 |
| selected orbit | 0.0733 | 0.9685 | 1.0083 |

Table 13: Evaluation on five untouched seeds after validation-only selection.

| contrast | mean difference | 95% interval |
|---|---|---|
| amplitude std 0.3 minus 0, radial error | $-0.1094$ | $[-0.1246, -0.0961]$ |
| amplitude std 0.3 minus 0, amplitude-two error | $-0.0929$ | $[-0.1045, -0.0825]$ |
| amplitude std 0.3 minus 0, sphere error | $+0.0004$ | $[-0.0004, +0.0011]$ |
| covariance minus random, weak-mode error | $-0.0726$ | $[-0.1038, -0.0426]$ |
| disagreement minus random, weak-mode error | $-0.0590$ | $[-0.0826, -0.0340]$ |
| selected JVP minus baseline, tangent error | $-0.0463$ | $[-0.0625, -0.0302]$ |
| selected JVP minus baseline, map error | $+0.0061$ | $[+0.0036, +0.0085]$ |
| selected orbit minus baseline, tangent error | $-0.0089$ | $[-0.0107, -0.0070]$ |
| structural wrapper minus baseline, tangent error | $-0.9563$ | $[-0.9607, -0.9516]$ |
| structural wrapper minus baseline, offset error | $-1.0029$ | $[-1.0056, -1.0003]$ |
| structural wrapper minus baseline, map error | $+0.0000$ | $[-6 \times 10^{-9}, +6 \times 10^{-9}]$ |
| 2D inverted minus natural, wavenumber association | $-1.281$ | $[-1.348, -1.206]$ |

Table 14: Paired uncertainty for the main follow-up contrasts.

| $\gamma$ | violation | raw offset error | wrapper offset error | on-slice |
|---|---|---|---|---|
| 0 | 0.000 | 1.013 | 0.013 | 0.052 |
| 0.05 | 0.050 | 1.010 | 0.052 | 0.052 |
| 0.15 | 0.159 | 1.005 | 0.159 | 0.051 |
| 0.5 | 0.628 | 0.989 | 0.628 | 0.049 |

Table 15: Damped Burgers across three seeds (across-seed std $\leq 2 \times 10^{-4}$ on wrapper errors). Violation is the relative error of the symmetry prediction against direct numerical solves at offset 0.25.

| | standard FNO | grid-free FNO | local CNN |
|---|---|---|---|
| DC gain ratio at initialization | 0.004 | 0.006 | 0.002 |
| DC gain ratio after training | $0.022 \pm 0.004$ | $0.024 \pm 0.004$ | $0.173 \pm 0.014$ |
| continuity gap $\lvert g(0) - g(1)\rvert / \lvert g(1)\rvert$ | $0.82 \pm 0.04$ | $0.80 \pm 0.03$ | $0.50 \pm 0.11$ |
| $k$=0 weight movement / $k$>0 movement | 0.23 | 0.25 | — |

Table 16: Continuity dichotomy measurements across eight seeds per architecture.

| model | test RMSE | mean-intervention slope (true $+2$) | excited-direction slope (true 0.037) |
|---|---|---|---|
| MLP, depth 2 | $0.069 \pm 0.003$ | $+0.04 \pm 0.77$ | 0.028 |
| MLP, depth 4 | $0.067 \pm 0.003$ | $-0.55 \pm 0.83$ | 0.038 |
| attention | $0.037 \pm 0.004$ | $+1.65 \pm 1.45$ | 0.025 |

Table 17: Non-PDE demonstration (target standard deviation 1.41; relative test error 2.6–4.9%).

| quantity | mean $\pm$ std (8 seeds) |
|---|---|
| test relative $L_2$ | $0.114 \pm 0.004$ |
| DC tangent gain (exact $= 1$) | $0.479 \pm 0.148$ |
| offset-0.05 error | $0.129 \pm 0.005$ |
| offset-0.15 error | $0.212 \pm 0.016$ |
| completion: offset-0.05 error | $0.113 \pm 0.004$ |
| completion: offset-0.15 error | $0.111 \pm 0.004$ |

Table 18: Audit on officially distributed data ($128^2$, 1000 training pairs, input std 0.153).