SPECTRAL STATIONARITY PREDICTS NEURAL-OPERATOR TRANSFER ACROSS RESOLUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural operators may accept a new grid without preserving the physical response learned at the training resolution. We show that spectral stationarity predicts this transfer before target-resolution labels are available. At each physical frequency, it separates the stability of the learned response from its representability on the target grid. This yields an estimator using synthetic reference solutions from band-limited probes evaluated on the training grid and one additional grid. In controlled one-dimensional studies, the estimator reduces false acceptance of resolution-anchored models from 68.6% to 2.8%. In a fixed-architecture two-dimensional study, eight probes achieve 0.962 concordance in paired model selection, compared with 0.464 for global training error at identical 0.259 coverage. The analysis also distinguishes failures caused by target-grid sampling, discretization-anchored representations, and uninformative probes. Spectral stationarity therefore identifies unsafe grids and usable frequency bands before labels are generated at the deployment resolution.

1 INTRODUCTION

Neural operators are attractive surrogate solvers because they aim to learn maps between functions rather than maps between fixed-dimensional vectors (Li et al., 2021; Lu et al., 2021; Kovachki et al., 2023). This functional viewpoint suggests that a model trained on one discretization can be evaluated on another. Such transfer is operationally important: a coarse grid can reduce inference cost, a fine grid can support downstream analysis, and a new input regime can carry spectral content absent from training.

The functional description, however, does not ensure that a finite trained model preserves the same physical response after regridding. Recent studies show failures in both resolution directions and locate much of the error in frequency bands that are unlearned, aliased, or poorly represented (Sakarvadia et al., 2026; Shikhman, 2026). These evaluations answer whether transfer failed after a sweep over target grids with reference solutions. Deployment poses a different question: can we reject an unsafe grid, or identify the usable bands, before producing labels at that grid?

We study a fixed trained operator at training and target grid sizes N_{train} and N_{eval}. A small synthetic probe estimates fidelity in each physical band b, $g_{N_{\mathrm{train}}}(b) \in [0, 1]$. Target-grid sampling determines a separate representability factor $\mathrm{repr}(b; N_{\mathrm{eval}})$. Their product distinguishes two reasons for failure: the grid may not carry the band, or the model may not have learned to transmit it. The resulting danger score,

$$s(b; N_{\mathrm{eval}}) = 1 - \mathrm{repr}(b; N_{\mathrm{eval}}) g_{N_{\mathrm{train}}}(b),$$

contains no fitted coefficient. It replaces a labeled sweep over target grids with reference evaluations on a small, controlled probe set.

The transfer interpretation requires the learned response to remain attached to physical frequency rather than to the training discretization. We call this requirement *spectral stationarity*: within every commonly representable band, $g_N(b)$ should remain approximately invariant when the same trained operator is evaluated at a different grid N. A second synthetic probe grid turns the assumption into a test. Stationary responses support exporting the training-grid score over the tested grid change;

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.

an anchored response rejects that export. This distinction explains both success and failure in our experiments. Spectral and global-basis operators have stable band responses, whereas U-Net and anti-aliased convolutional neural operator (CNO) stress cases can be accurate at the training grid and fail in both off-grid directions.

The empirical design holds physical frequency fixed, compares variants within the same architecture and partial differential equation (PDE), and separates reverse from super-resolution. This avoids rewarding a model merely because low frequencies are easier. The experiments ask whether a second grid rejects anchored responses, whether the probe ranks variants on new random seeds within Fourier neural operator (FNO), deep operator network (DeepONet), and Galerkin families, and how accuracy changes with probe count. Two 2D PDEs and a fixed-architecture spectral-tradeoff experiment test whether the same idea survives a change of dimension.

Contributions. First, we identify spectral stationarity as the missing condition between function-space architecture and cross-grid behavior, and test it with matched probes at a second grid. Second, we separate learned band fidelity from target-grid representability in a coefficient-free decision statistic. Third, we evaluate the resulting decision on anchored failures, independently trained within-architecture choices, matched numbers of probes, and a separate 2D spectral tradeoff. Together, these results replace the question "is this architecture resolution-invariant?" with a measurable question about one trained model, one physical band, and one grid change.

2 Related Work

Multi-resolution operator failure. FNO introduced the modern discretization-flexible neural-operator setting (Li et al., 2021), alongside DeepONet and the broader neural-operator theory of function-space maps (Lu et al., 2021; Li et al., 2020; Kovachki et al., 2023). Message-passing and mesh-based simulators offer a complementary discretization-aware route (Brandstetter et al., 2022; Pfaff et al., 2021). Subsequent evaluations show that this flexibility is not enough for zero-shot resolution change. Sakarvadia et al. (2026) evaluate sub- and super-resolution across several operator families and connect failures to frequency extrapolation and aliasing. Shikhman (2026) build a broader failure atlas across PDE families and report that high in-distribution accuracy need not imply robustness under resolution shift. We use their observation as the starting point and ask a different question: can a small pre-deployment probe predict the band where failure will occur?

Discretization mismatch, aliasing, and CNOs. Theory and architecture work explains why discretization changes can break an operator implemented on a grid. Discretization mismatch errors accumulate when inference resolution differs from training resolution (Gao et al., 2025). Representation-equivalent neural operators formalize operator aliasing and propose alias-free constructions (Bartolucci et al., 2023). Convolutional neural operators were designed to preserve a continuous-operator interpretation while using convolutional structure (Raonić et al., 2023). Our experiments are consistent with the mechanism: the spectral and global operators in our suite have stationary band responses, whereas our U-Net and anti-aliased CNO constructions remain anchored to the training grid.

Spectral bias and operator benchmarks. Spectral bias in neural networks and Fourier-feature remedies provide the learning-theoretic backdrop for why resolution changes expose frequency-dependent error (Rahaman et al., 2019; Tancik et al., 2020). Spectral analyses of FNOs show that learned operators often transmit low frequencies more reliably than high frequencies (Qin et al., 2024). PDEBench provides broad benchmark tasks and metrics for scientific machine learning (Takamoto et al., 2022), and newer PDE benchmarks broaden the aggregate-evaluation setting (Gupta & Brandstetter, 2023; Koehler et al., 2024; Ohana et al., 2024). Data-driven discretization work also shows that numerical representations can be learned rather than fixed (Bar-Sinai et al., 2019). Reliable extrapolation work studies when neural operators remain accurate outside the training input distribution (Zhu et al., 2023). Existing resolution studies primarily explain what breaks or modify training to prevent it. We instead study a decision made after training and before a full target-grid sweep. The test measures a model-specific response in fixed physical bands, controls for the low-frequency prior, and checks whether that response persists on a second discretization. These elements turn spectral failure analysis into a prospective comparison between trained variants while retaining an explicit failure case when response remains tied to the training grid.

3 SPECTRAL RESOLUTION-TRANSFER TEST

3.1 BANDWISE SAFETY AND TRAINING-GRID PROBE

Let $G : a \mapsto u$ map an input field a to its periodic PDE solution u. We write $\hat{G}_{\theta,N}$ for weights θ evaluated on an N-point grid and u_N for the reference solution sampled on that grid. The physical Fourier axis is partitioned into fixed octave bands $b = [\ell_b, h_b)$ shared across resolutions. A represented band meets the tested criterion when its relative spectral error is below $\tau = 0.25$. For probe i among n_p inputs, Fourier mode k, Parseval weights w_k, and Fourier coefficient $\mathcal{F}_k[\cdot]$, the band error and fidelity are

$$e^{(i)}_{N,b} = \left[\frac{\sum_{k\in b} w_k |\mathcal{F}_k[\hat{G}_{\theta,N}(a^{(i)}) - u^{(i)}_N]|^2}{\sum_{k\in b} w_k |\mathcal{F}_k[u^{(i)}_N]|^2 + 10^{-12}} \right]^{1/2}, \qquad g_N(b) = \mathrm{clip}\Big(1 - \mathrm{median}^{n_p}_{i=1}\, e^{(i)}_{N,b}, 0, 1\Big). \tag{1}$$

Here $\mathrm{clip}(x, 0, 1)$ truncates x to $[0, 1]$. Without a grid subscript, g denotes $g_{N_{\mathrm{train}}}$. The primary paired selector uses $n_p = 8$; the direct stationarity check uses 24 matched probes at each grid. Inputs follow the deployment regime's spectral envelope, so fidelity is regime-specific. Reference outputs are generated only for probes at the train and check grids; no deployment case from the target-grid sweep is used. The effective bandwidth B_{eff} is the largest contiguous mode through which $g_{N_{\mathrm{train}}}(b) \geq 0.9$. This descriptive cutoff is stricter than the decision cutoff $1 - \tau = 0.75$; the predictor retains the full vector g.

3.2 TARGET-GRID DANGER SCORE

In 1D, the target grid contributes the mode-count representability term

$$\mathrm{repr}(b; N_{\mathrm{eval}}) = \frac{\max(0, \min(h_b, N_{\mathrm{eval}}/2 + 1) - \ell_b)}{h_b - \ell_b}. \tag{2}$$

It is one for fully represented bands and zero above target Nyquist; 2D uses radial shells with the same represented-mode interpretation. The predicted danger score is

$$s(b; N_{\mathrm{eval}}) = 1 - \mathrm{repr}(b; N_{\mathrm{eval}})\, g_{N_{\mathrm{train}}}(b). \tag{3}$$

The statistic has no fitted coefficient: repr records what the target grid can carry, while $g_{N_{\mathrm{train}}}$ records what this trained model transmits on the probe distribution. For a fully represented band with median probe error below one, $1 - g$ is exactly that median error. The product is therefore an operational decision statistic, not a calibrated estimate of target-grid error. An input-support factor tested during development collapsed to the probe-only score under an alias-free spectrum and is excluded (Appendix D).

3.3 TESTING SPECTRAL STATIONARITY

Equation (3) assumes that the training-grid response remains attached to the same physical band after regridding. Let $g_N(b)$ denote probe fidelity at grid N. The transfer condition is

$$g_{N_{\mathrm{eval}}}(b) \approx g_{N_{\mathrm{train}}}(b) \quad \text{for all representable bands } b. \tag{4}$$

To test it without deployment labels, we repeat matched synthetic probes at a check grid and compare only bands represented by both grids. With $\delta = 0.10$, a band passes when

$$q(b) = \mathbf{1}\{\min(g_{N_{\mathrm{train}}}(b), g_{N_{\mathrm{check}}}(b)) \geq 1 - \tau,\ |g_{N_{\mathrm{train}}}(b) - g_{N_{\mathrm{check}}}(b)| \leq \delta\}. \tag{5}$$

The primary check is $N_{\mathrm{train}}/2$; we also test $N_{\mathrm{train}}/4$ and $2N_{\mathrm{train}}$ to measure direction sensitivity. A failed check prevents exporting the training-grid verdict. Passing one check measures agreement at the sampled grids; interval-wide guarantees require additional regularity and probe-coverage assumptions (Appendix A).

3.4 WHAT THE TEST RETURNS

The output is the tuple $(g_{N_{\text{train}}}(b), q(b), \text{repr}(b; N_{\text{eval}}))$, not one architecture rank. It distinguishes a narrow but stationary operator, a model that appears wide only at its training grid, and a wide stationary operator. We compare $1 - g_{N_{\text{train}}}$ with a band prior, a Nyquist rule, global training error, architecture class, and B_{eff} rank. No per-PDE regressor or architecture-specific coefficient is fitted. The interpretation depends on three observable conditions: the probes must cover the deployed spectral envelope, the learned response must remain stationary across the tested grid change, and relative spectral error must match the downstream quantity of interest. Regime-matched probes, a second-grid check, and a task-specific decision functional address these conditions in turn.

4 EXPERIMENTAL SETUP

Systems and models. The 1D suite contains periodic advection–diffusion, viscous Burgers, and Kuramoto–Sivashinsky operators. A balanced matrix spanning FNO, DeepONet, Galerkin, U-Net, and CNO separates resolution-stationary operators from cases whose response remains tied to the training grid. Within-architecture selection then varies only FNO mode count or DeepONet/Galerkin capacity, with nine new PDE–seed clusters per family. The 2D study covers Navier–Stokes vorticity and advection–diffusion. Its central test holds the FNO-2D architecture fixed and varies only a radial Fourier-loss tilt in $\{-2, -1, 0, 1, 2\}$, yielding ten PDE–seed clusters, six generated after the analysis choices were fixed. Reference solutions are self-converged pseudo-spectral simulations; the appendix gives model counts, training-set sizes, update counts, and the complete matrix.

Resolutions and metrics. The 1D models are trained at $N_{\text{train}} = 128$ and evaluated at grids from 8 to 1024 ($r = N_{\text{eval}}/N_{\text{train}} \in [0.0625, 8]$). We report reverse ($r < 1$) and super-resolution ($r > 1$) separately. The reverse direction is the genuine forecast because target representability changes; super-resolution mainly amortizes a training-grid measurement. The primary label is $e_b \geq \tau$ with $\tau = 0.25$. Within each fixed band, we ask whether $1 - g(b)$ separates safe from unsafe trained operators, making a frequency prior constant with area under the receiver operating characteristic curve (AUC) 0.5. Macro AUC uses only bands containing both labels. Uncertainty is aggregated across independent PDE–seed clusters rather than correlated grid cells, and the 1D and 2D conclusions remain separate. The targeted 1D reruns use disjoint training and evaluation fields across three spectral regimes and 13 target grids. Probe-count curves use nested prefixes of one shared 48-probe sequence, so comparisons across probe counts see identical inputs. All thresholds and check grids are fixed before the target sweep. Exact sample counts, threshold sensitivity, and matched wall-clock accounting appear in the appendix.

For within-architecture model selection, a cell fixes architecture, PDE, seed, input regime, target grid, and physical band. We compare every matched pair whose log-error gap is at least 0.02 for retrospective scoring, and let the probe abstain when $|g_i(b) - g_j(b)| < 0.02$. The primary estimand is concordance among non-abstained pairs; coverage is their fraction of eligible pairs. Target-error eligibility is used only for scoring; deployment uses the probe gap. In addition to evaluating global training relative-L^2 error E_i^{train} on the probe-selected cells, we give it its own confidence $|\log_{10} E_i^{\text{train}} - \log_{10} E_j^{\text{train}}|$ and retain exactly the probe's coverage without using target labels; tied cutoff cells receive their uniform-random-tie-breaking weight. Confidence intervals resample PDE–seed clusters and one probe repeat within a sampled cluster. The direct stationarity study has nine independent PDE–training-seed clusters, with architectures, grids, regimes, and bands kept inside each cluster. Margins, independent training runs, matched probe prefixes, and decision thresholds were fixed before evaluation. Equal-weight macros, leave-one-cluster-out analyses, sign-flip tests, and the complete estimand specification are reported in the appendix.

5 RESULTS

5.1 THE PROBE SELECTS WITHIN ARCHITECTURE ON NEW RANDOM SEEDS

Across the original 45 resolution-stationary 1D settings, within-band AUC of $1 - g$ is 0.845 (per-seed mean 0.842 ± 0.018 standard deviation over five seeds), compared with 0.73 for architecture class, 0.69 for global training error, 0.693 for B_{eff} rank, and 0.5 for the constant band prior. Treating

Table 1: **Four questions determine whether the spectral stationarity test is useful.** Training-error comparisons use either probe-selected cells or an independent ranking at matched coverage; FA/FR denote false acceptance/rejection.

Question	Setting and independent unit	Probe result	Comparison
Can the probe rank variants within one architecture?	Three 1D PDEs; nine PDE–seed clusters per family	**0.990** FNO, **0.871** DeepONet, **0.909** Galerkin	Training error: 0.794, 0.662, 0.392
Does the ordering survive a fixed-architecture 2D tradeoff?	Two PDEs; five FNO-2D loss tilts; ten PDE–seed clusters	**0.978** concordance at 0.265 coverage	Training error 0.528; model-pair macro 0.967
Does a second grid expose an anchored response?	Nine 1D PDE–seed clusters; stationary and anchored families	Anchored FA 68.6%→**2.8%**	Stationary FR 1.3%→3.7%
How many probes are needed?	Two 2D PDEs; ten clusters and five matched sequences	Eight probes: **0.962** concordance at 0.259 coverage	Training error 0.464; 24 probes 0.974; cost ratio 0.33×

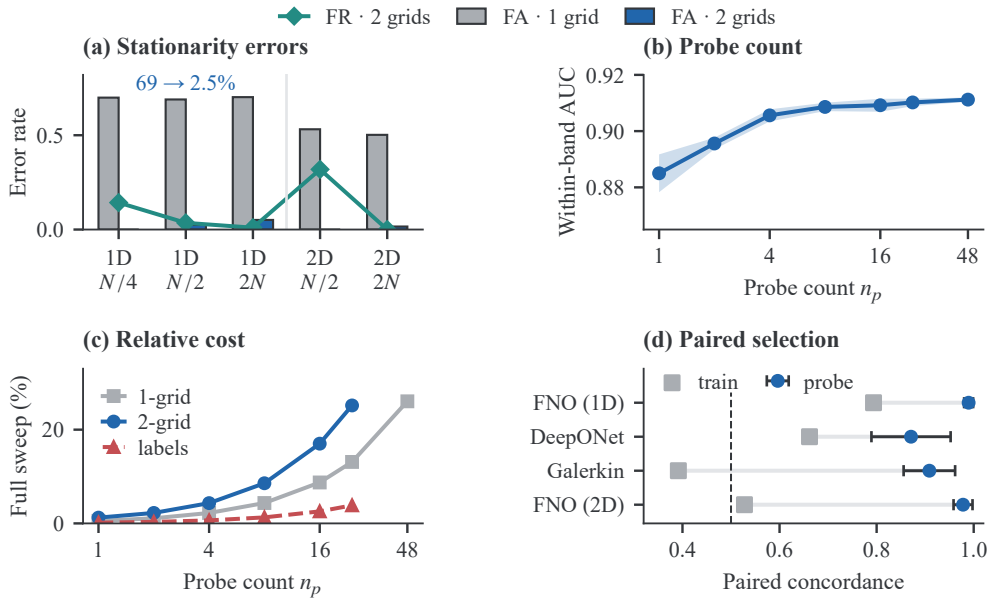


Figure 1: Spectral stationarity across error, sample efficiency, cost, and selection. (A) One- versus two-grid FA/FR. (B) AUC versus probe count. (C) Cost relative to a labeled sweep. (D) Paired concordance against training-error selection. Intervals use independent PDE–seed clusters; the dashed line marks chance.

safety as the positive class at the fixed operating point, precision and specificity are 1.00, recall is 0.655, and balanced accuracy is 0.827. This asymmetry favors precise declarations of safety while leaving some safe bands unselected.

The harder question is whether g selects the better variant *within* one architecture. Across nine new PDE–seed clusters per family, equal-weight seed/PDE macro concordance is 0.990 for FNO (coverage 0.370; small-cluster 95% confidence interval (CI) $[0.980, 0.999]$), 0.871 for DeepONet (0.416; $[0.789, 0.952]$), and 0.909 for Galerkin (0.535; $[0.856, 0.962]$). Global training error reaches only 0.794, 0.662, and 0.392 on the same selected cells. Model-pair macro concordance is 0.988, 0.854, and 0.884, while the corresponding worst leave-one-cluster-out estimates are 0.987, 0.854, and 0.898. The comparison therefore isolates within-architecture selection at the level of independent PDE–seed clusters.

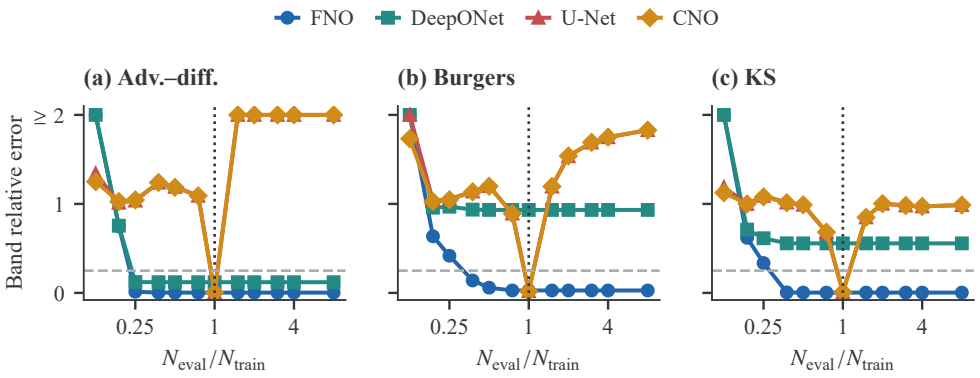


Figure 2: Physical-band relative errors across resolution. U-Net and anti-aliased CNO show training-grid minima; FNO and DeepONet are flatter away from coarse-grid limits.

5.2 A SECOND GRID EXPOSES RESOLUTION-ANCHORED RESPONSES

U-Net and anti-aliased CNO stress models are accurate at the training grid but fail away from it. In band $[8, 16)$, their training-grid errors are below 0.02 on most 1D PDEs, while off-grid means approach or exceed one. This produces the V-shaped error curve in Figure 2: a one-grid probe appears safe precisely where the architecture is anchored.

The direct two-grid experiment quantifies this failure. At the primary $N/2$ check, PDE–seed macro false acceptance falls from 68.6% to 2.8%, an absolute reduction of 65.8 percentage points with cluster-bootstrap 95% CI $[57.2, 74.2]$ percentage points. The reduction is positive for every PDE, and its worst leave-one-cluster-out value is 63.7 percentage points. On stationary operators, macro false rejection rises from 1.3% to 3.7%. Across the nine independent PDE–seed clusters, the reduction is consistent, although the Student-t upper endpoints (6.0% for FA and 6.3% for FR) remain wider than the point estimates. The experiment therefore establishes a cluster-robust reduction in false acceptance; more PDE–seed clusters would be needed to estimate a population-level 5% bound. Per-PDE counts and intervals appear in the appendix. Check direction is not interchangeable. The more aggressive $N/4$ check eliminates anchored false accepts but rejects 14.3% of stationary safe cases (cell-pooled, Table 11); a $2N$ check leaves a 5.1% anchored false-accept rate and cannot expose all reverse-resolution failures. The check therefore identifies a supported grid interval and direction rather than assigning a global invariance label.

5.3 EIGHT PROBES NEARLY MATCH THE 24-PROBE SELECTOR

Across the 27 stationary models in the probe-count study, within-band AUC rises from 0.885 ± 0.007 with one probe to 0.906 ± 0.002 with four and 0.911 ± 0.001 with 48 (five probe-prefix draws). Four probes therefore recover 79% of the improvement from one to 48; this curve measures discrimination of threshold-defined safety labels, not paired model selection.

The new matched curve measures paired model selection itself. Across ten PDE–seed clusters and five probe-sequence repeats, eight probes reach macro concordance 0.962 (hierarchical 95% CI $[0.945, 0.973]$) at coverage 0.259. Twenty-four probes reach 0.974 at coverage 0.263, only 0.012 higher. Median marginal deployment time (one shared reference plus inference and scoring of five candidates in three regimes) is 11.36 s versus 34.61 s, an eight-to-24-probe cost ratio of 0.33. Reference cost is PDE-dependent: at eight probes the median is 0.318 s for linear advection–diffusion and 23.190 s for Navier–Stokes. At eight probes, the selector reaches the targeted accuracy and coverage using one-third of the measured 24-probe cost. The 0.978 primary estimate below uses the original 24-probe sequence; the 0.974 curve point averages five new sequences on the same trained models. Their agreement is an independent probe-draw replication, not reuse of the primary probes.

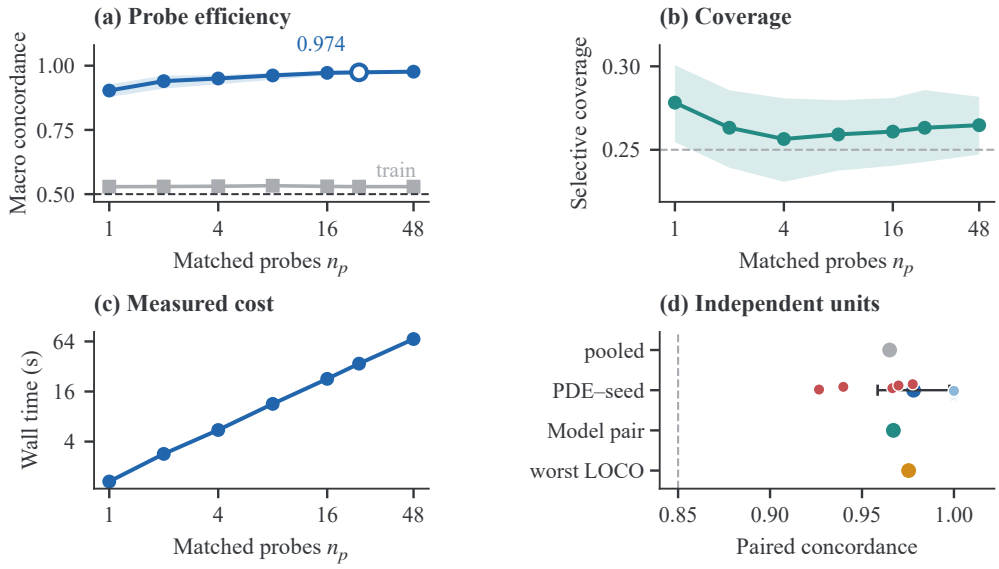


Figure 3: Probe-count sensitivity and aggregation. (A,B) Paired concordance and coverage for matched prefixes. (C) Marginal cost with shared references. (D) Pooled, macro, pairwise, and worst leave-one-cluster-out (LOCO) estimates. Bands are hierarchical 95% intervals; dots are independent PDE–seed clusters.

5.4 The decision is stable across margins, aggregators, and components

The operating point lies in a broad positive neighborhood (Table 1; Figure 5). Across margins 0.005–0.04, eight-probe concordance increases from 0.944 to 0.974 while coverage traces the expected risk–coverage frontier. At the selected 0.02 margin, concordance is 0.962 at 0.259 coverage. Mean, trimmed mean, winsorized mean, and median aggregation reach 0.962–0.973 concordance on identical probe prefixes, showing that the ordering persists across estimator choice.

The component ablations also recover the intended decomposition. Adding the check-grid level cuts anchored false acceptance from 0.690 to 0.049; the drift condition reaches 0.025, and the full check combines that rate with 0.037 stationary false rejection. On the same labels, the product score attains AUC 0.937 in 1D reverse resolution and 0.894 in 2D, improving on representability alone (0.847 and 0.720). The paired ordering further persists for all finite represented pairs (0.937), adjacent candidates (0.953), and the worst leave-one-tilt-out analysis (0.932). Appendix tables and Figures 4–5 report the complete grids, intervals, and probe-spectrum controls.

5.5 Direction and two-dimensional replication

In 1D super-resolution ($r > 1$), the probe-only score and full score reach AUC 0.998 and 0.997: the new grid mostly repeats a measured response. Reverse resolution is the forecast regime because representability changes; there the probe reaches 0.810 against 0.766 for the band prior.

Across 2D advection–diffusion and Navier–Stokes, FNO-2D modes $\{8, 16, 24\}$, U-Net-2D, and CNO-2D provide a separate radial-band test. The representable-band score reaches AUC 0.843 against 0.772 for the band prior; the per-PDE values are 0.791 versus 0.769 for advection–diffusion and 0.976 versus 0.934 for Navier–Stokes.

The within-architecture 2D test is stronger. Holding FNO-2D modes, width, depth, data, and parameter count fixed, five spectral-loss tilts induce lower–higher frequency tradeoffs. The six clusters generated after fixing the analysis alone reach PDE–seed macro concordance 0.981 (small-cluster 95% CI $[0.956, 1.000]$). Across all ten clusters, macro concordance is 0.978 at coverage 0.265 versus 0.528 for global training error. The interval is $[0.959, 0.998]$, model-pair macro is 0.967, and worst LOCO is 0.975. Concordance is 1.000 for advection–diffusion and 0.956 for Navier–Stokes; exact cluster sign-flip tests give $p = 0.0020$ against chance and $p = 0.0313$ against training error.

Table 2: Fixed-band response and error across FNO mode counts. Each entry is g / relative spectral error in band $[16, 32)$.

	8 modes	16 modes	32 modes
Advection–diffusion	$0.43/0.93$	$0.91/0.11$	$0.98/0.04$
Kuramoto–Sivashinsky	$0.23/1.02$	$0.96/0.05$	$0.96/0.05$

The result therefore survives a change of dimension and isolates band-specific selection from global model quality without treating correlated cells as independent evidence.

The baseline conclusion is stronger at matched abstention. At exactly 0.259 coverage in every PDE–seed–repeat deployment, the eight-probe selector reaches 0.962 versus 0.464 for the independently ranked global-training-error selector. The paired advantage is 0.498 (hierarchical 95% CI $[0.454, 0.547]$), remains positive on both PDEs, and has worst leave-one-cluster-out value 0.478. The methods select cells independently but retain the same weighted fraction in each deployment, so their coverage is identical.

5.6 Robustness and deployment consequences

An FNO trained with the independent `neuraloperator` library reproduces monotone g curves and finite learned cutoffs on all three 1D PDEs. Under input spectra extended to $2\times$ and $4\times$ the training bandwidth, the predicted safe endpoint contracts as unlearned high-frequency content appears. These checks support the same separation: representability describes the grid, while g describes the trained model.

The output is a usable band endpoint rather than one resolution label. For an FNO trained on Burgers at $N_{\text{train}} = 128$, $N = 64$ remains usable below the coarse Nyquist, while $N = 16$ is limited to mode 7. A DeepONet trained on Kuramoto–Sivashinsky remains high-band unsafe at $N = 512$ because a finer grid does not create learned bandwidth. The full atlas and extreme-spectrum maps are shown in the appendix.

The fixed-band comparison in Table 2 separates learned spectral capacity from grid representability. All entries use the same physical band $[16, 32)$, so the target grid carries the same frequencies in every column. Moving from eight to 16 retained modes sharply increases the measured response and reduces spectral error on both equations; moving from 16 to 32 modes produces little further change in these settings. The comparison shows why mode count, evaluation resolution, and learned bandwidth cannot be treated as interchangeable quantities.

6 Scope and Outlook

The evidence spans three periodic 1D PDEs, two periodic 2D PDEs, spectral, global-basis, convolutional, and independently implemented operator families. The same decomposition suggests a path to non-periodic domains, irregular meshes, and anisotropic outputs by replacing Fourier bands with a physical basis or a task-aligned partition. Testing that extension on three-dimensional operators and broader continuous-kernel constructions is an important next step.

The probes still require trusted reference solves. Their practical advantage is therefore not label-free evaluation, but replacing a sweep over deployment cases and target grids with a small controlled set at two grids. The present study compares this cost with global training error and with larger probe sets. A same-cost baseline built from labeled deployment cases would answer a different, complementary question and remains to be tested when such cases are available.

The selector intentionally reports both concordance and coverage: the margin identifies comparisons supported by a measurable spectral difference. Reference cost depends on the PDE: eight probes require 0.318 s for the linear 2D system and 23.190 s for Navier–Stokes. The median total selector cost is $0.33\times$ that of the matched 24-probe design. Additional check grids expand the tested interval, and downstream quantities of interest can combine their own spectral sensitivities with the reported band response. These extensions retain the same central decision object: learned response within a grid-representable physical band.

The experiments also separate three questions that are often collapsed in a resolution study. Representability asks whether the target grid contains the physical band of interest; stationarity asks whether the learned response over that band persists when the grid changes; selection asks whether the remaining difference is large enough to order two trained models. A model can therefore run on a finer grid without having learned the newly represented frequencies, or it can preserve its response while the comparison remains too close to call. Keeping these cases distinct explains why the method returns an interval and an abstention decision rather than a single architecture-level verdict.

This distinction also changes how the result should be used. The probe is most informative before a costly labeled sweep, when the deployment band is known but target solutions are scarce. It does not certify accuracy outside the measured band, replace stability analysis, or compare downstream quantities whose spectral sensitivities differ from the probe. Instead, it supplies a compact measurement of what one fitted operator preserves across grids. The matched probe study shows that this measurement can remain selective with eight inputs, while the two-dimensional experiment shows that it can discriminate models whose architecture, width, depth, and training data are held fixed.

The statistical unit is equally important. Pairwise cells that share a PDE, training seed, or probe sequence are not independent replications, so the main intervals aggregate within PDE–seed clusters and the leave-one-cluster-out analysis asks whether one cluster drives the ordering. The matched-coverage comparison controls the retained fraction while allowing each method to choose cells using its own confidence ranking. These choices make the reported concordance a statement about the tested collection of trained models rather than about a much larger collection of correlated pairs. More broadly, they show how a selective scientific test should be reported: together with its coverage, its independent units, and the measurement cost required to reach the decision.

7 Conclusion

Resolution transfer depends on what a trained operator has learned at each physical frequency, not simply on the resolution at which its architecture can run. Our synthetic probes separate that learned band response from what a target grid can represent. A second grid then reveals models whose apparent bandwidth is anchored to the training discretization.

Across controlled 1D operators, the $N/2$ check sharply reduces false acceptance of anchored models. For paired model choice, eight probes reach 0.962 concordance versus 0.464 for global training error at the same 0.259 coverage, while using one third of the measured 24-probe marginal cost. The same physical-band view also selects among fixed-architecture FNO-2D variants on two PDEs.

The output is a usable frequency interval together with the comparisons on which the probe abstains. This turns resolution transfer from an architecture label into a property that can be measured for a particular trained model, spectral regime, and pair of grids before running a labeled resolution sweep.

The two-grid measurement distinguishes anchored responses that fail the stationarity check, stationary but indistinguishable models that trigger abstention, and separated responses that yield a comparison over the shared physical band. It supports both forward and reverse resolution changes without assuming that a finer target grid is automatically safer. The result applies to the measured band and declared probe family; a new domain, boundary condition, or downstream functional requires a corresponding physical basis and a new measurement rather than an extrapolated architecture label.

The factorization also distinguishes what should change when transfer fails. If representability is limiting, the target grid cannot carry the requested band, so changing the fitted weights cannot recover it. If the response is grid-anchored, the model rather than the target discretization is responsible, which motivates testing anti-aliasing, multi-grid training, or a different operator parameterization. If both responses are stationary but too close to separate, the appropriate next step is to enrich the physical probes rather than declare either success or failure. These cases have similar downstream symptoms but different remedies. Separating them is the practical value of measuring the learned response and target-grid representability as distinct quantities. It also prevents two models with the same target-grid error from receiving the same transfer label when one fails through sampling and the other through its learned representation. Reporting the band, direction, and failure class therefore preserves information that a single error value discards. That mechanism-based distinction guides the next experiment: grid refinement tests representability, retraining tests anchoring, and additional probes test whether an apparent tie reflects limited measurement resolution.

AI USE STATEMENT

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

REPRODUCIBILITY STATEMENT

Sections 3 and 4 define the score, stationarity check, systems, resolution regimes, and evaluation metrics. The integrated appendix gives the complete deployment procedure, matched probe-cost design, cluster-level inference, and complete result tables. The anonymous artifact provides the implementation, numerical checks, experimental settings, and probe sequences needed to reproduce the reported comparisons.

REFERENCES

Yohai Bar-Sinai, Stephan Hoyer, Jason Hickey, and Michael P. Brenner. Learning data-driven discretizations for partial differential equations. *Proceedings of the National Academy of Sciences*, 116(31):15344–15349, 2019. doi: 10.1073/pnas.1814058116.

Francesca Bartolucci, Emmanuel de Bézenac, Bogdan Raonić, Roberto Molinaro, Siddhartha Mishra, and Rima Alaifari. Representation equivalent neural operators: a framework for alias-free operator learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 69661–69672, 2023. URL https://openreview.net/forum?id=7LSEkvEGCM.

Johannes Brandstetter, Daniel Worrall, and Max Welling. Message passing neural PDE solvers. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.03376.

Wenhan Gao, Ruichen Xu, Yuefan Deng, and Yi Liu. Discretization-invariance? on the discretization mismatch errors in neural operators. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=J9FgrqOOni.

Jayesh K. Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized PDE modeling. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=dPSTDbGtBY.

Felix Koehler, Simon Niedermayr, Rüdiger Westermann, and Nils Thuerey. APEBench: A benchmark for autoregressive neural emulators of PDEs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d9875ebcf74bccdc5076acab0dbee62c-Abstract-Datasets_and_Benchmarks_Track.html.

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. URL https://jmlr.org/papers/v24/21-1524.html.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Andrew Stuart, Kaushik Bhattacharya, and Anima Anandkumar. Multipole graph neural operator for parametric partial differential equations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6755–6766, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4b21cf96d4cf612f239a6c322b10c8fe-Abstract.html.

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. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021. doi: 10.1038/s42256-021-00302-5.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond B. Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich R. Kerswell, Suryanarayana Maddu, Jonah Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam H. Parker, Miles Cranmer, and Shirley Ho. The well: a large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1430. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4f9a5acd91ac76569f2fe291b1f4772b-Abstract-Datasets_and_Benchmarks_Track.html.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W. Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.03409.

Shaoxiang Qin, Fuyuan Lyu, Wenhui Peng, Dingyang Geng, Ju Wang, Xing Tang, Sylvie Leroyer, Naiping Gao, Xue Liu, and Liangzhu Leon Wang. Toward a better understanding of fourier neural operators from a spectral perspective. *arXiv preprint arXiv:2404.07200*, 2024.

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

Bogdan Raonić, Roberto Molinaro, Tim De Ryck, Tobias Rohner, Francesca Bartolucci, Rima Alaifari, Siddhartha Mishra, and Emmanuel de Bézenac. Convolutional neural operators for robust and accurate learning of PDEs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77187–77200, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f3c1951b34f7f55ffaecada7fde6bd5a-Abstract-Conference.html.

Mansi Sakarvadia, Kareem Hegazy, Amin Totounferoush, Kyle Chard, Yaoqing Yang, Ian Foster, and Michael W. Mahoney. The false promise of zero-shot super-resolution in machine-learned operators. In *The Fourteenth International Conference on Learning Representations*, 2026. doi: 10.48550/arXiv.2510.06646. URL https://openreview.net/forum?id=hkF7ZM7fEp.

Lennon Shikhman. Diagnosing failure modes of neural operators across diverse PDE families. *Transactions on Machine Learning Research*, 2026. doi: 10.48550/arXiv.2601.11428. URL https://openreview.net/forum?id=0S1LWZHQYn.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1596–1611, 2022.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/55053683268957697aa39fba6f231c68-Abstract.html.

Min Zhu, Handi Zhang, Anran Jiao, George Em Karniadakis, and Lu Lu. Reliable extrapolation of deep neural operators informed by physics or sparse observations. *Computer Methods in Applied Mechanics and Engineering*, 412:116064, 2023. doi: 10.1016/j.cma.2023.116064.

A Mathematical Foundations and Limits

This section gives sufficient conditions behind the transfer test, rather than asserting that a finite two-grid check proves discretization invariance. All statements concern a fixed trained operator and a specified probe law. The function-space interpretation of neural operators (Kovachki et al., 2023) motivates comparing physical frequencies; the finite-grid conclusions below require additional assumptions explicitly stated in each proposition. The arguments also identify counterexamples that the empirical test cannot rule out.

A.1 Deterministic stationarity and representability

Let P_b be the orthogonal Fourier projection onto a fully represented band b. Grid outputs are lifted into the same finite-dimensional trigonometric space with Parseval-normalized coefficients. Write $v_N(a)$ for the lifted prediction and $u(a) = P_b G(a)$ for the common band-limited reference. Assume $\|u(a)\|_2 > 0$ and define the unregularized relative error

$$e_N(a) = \frac{\|P_b v_N(a) - u(a)\|_2}{\|u(a)\|_2}. \tag{6}$$

The identical-reference condition matters: a discretization-dependent reference solver can otherwise be mistaken for a change in the trained model.

Proposition 1 (response perturbation). If $\|P_b(v_N(a) - v_M(a))\|_2 \le \eta_b \|u(a)\|_2$ for every matched probe, then $|e_N(a) - e_M(a)| \le \eta_b$. For any finite matched probe set, the empirical fidelities, defined by clipping one minus the median of these errors, satisfy

$$|\widehat{g}_N(b) - \widehat{g}_M(b)| \le \eta_b. \tag{7}$$

Proof. The reverse triangle inequality gives $|\|P_b v_N - u\|_2 - \|P_b v_M - u\|_2| \le \|P_b(v_N - v_M)\|_2$. Divide by $\|u\|_2$. If paired numbers obey $|x_i - y_i| \le \eta_b$, their jth order statistics obey $|x_{(j)} - y_{(j)}| \le \eta_b$: at least j of the y_i are at most $x_{(j)} + \eta_b$, and exchange the roles for the reverse inequality. This applies to either central order statistic and to their average when the sample size is even. Finally, $x \mapsto \mathrm{clip}(1 - x, 0, 1)$ is 1-Lipschitz. □

The proof remains valid with the same positive denominator regularizer at both grids. If approximate references u_N, u_M are used instead, assume $\|u_N\|, \|u_M\| \ge \rho > 0$, $\|u_N - u_M\| \le \zeta$, and $\|P_b v_M - u_M\| \le K$. Then the corresponding error difference is at most

$$|e_N - e_M| \le \frac{\|P_b(v_N - v_M)\| + \zeta}{\rho} + \frac{K\zeta}{\rho^2}. \tag{8}$$

Indeed, subtract the two norm ratios, apply the reverse triangle inequality to the numerators and denominators, and use their lower bounds. This explains why extremely low-energy bands need an explicit eligibility rule: an absolute reference error can become an arbitrarily large relative error as $\rho \to 0$.

The perturbation bound becomes exact for physical Fourier multipliers under the following structural assumptions.

Proposition 2 (a sufficient structural condition). Consider a linear reference multiplier $H(k)$ and a learned multiplier $\widehat{H}(k)$ whose parameters depend on physical mode k, not grid size. For a continuous input with coefficients a_k, suppose both grids sample the input without aliasing and fully represent b. Then

$$e^2_{N,b}(a) = \frac{\sum_{k \in b} w_k |\widehat{H}(k) - H(k)|^2 |a_k|^2}{\sum_{k \in b} w_k |H(k)|^2 |a_k|^2} = e^2_{M,b}(a). \tag{9}$$

Proof. Exact trigonometric sampling identifies the same physical Fourier coefficients at both resolutions. Apply each diagonal multiplier, then Parseval's identity. Every term in the ratio is independent of N and M. Hence the errors, their medians, and clipped fidelities agree. □

The hypothesis concerns a linear diagonal action; it is not a theorem that an arbitrary nonlinear FNO is exactly stationary. Nonlinear activations mix frequencies and can alias even when the retained

spectral layer is diagonal. For an approximately diagonal operator, Proposition 1 bounds the residual cross-grid discrepancy whenever it can be controlled in the band norm. For comparison, a one-grid-point circular shift has multiplier $\exp(2\pi ik/N)$. Its phase is grid dependent at fixed physical k, so small training error for a physical shift does not imply transfer of the same discrete stencil.

Representability and energy geometry. Let Q_N retain the modes representable on the target grid and let $u_b = P_b u$. Any target prediction v_N supported in Q_N obeys

$$\frac{\|v_N - u_b\|^2}{\|u_b\|^2} = \frac{\|v_N - Q_N u_b\|^2}{\|u_b\|^2} + 1 - R_E(b,N) \;\geq\; 1 - R_E(b,N), \qquad R_E(b,N) = \frac{\|Q_N u_b\|^2}{\|u_b\|^2}. \tag{10}$$

Proof. The represented and omitted Fourier subspaces are orthogonal. Decompose $v_N - u_b = (v_N - Q_N u_b) - (I - Q_N)u_b$ and use Pythagoras. □ Here the error is against the *full-band* reference. It differs from the paper's scored error on the represented portion of a partially available band.

The mode-count quantity $R = \mathrm{repr}(b; N)$ equals R_E only under a uniform-energy convention, with consistent conjugate-mode counting. More generally, if the energy per counted mode lies between $e_{\min} > 0$ and $e_{\max}$, with $\kappa = e_{\max}/e_{\min}$, then

$$\frac{R}{R + \kappa(1-R)} \leq R_E \leq \frac{\kappa R}{\kappa R + 1 - R}. \tag{11}$$

To obtain the lower bound, assign minimum energy to represented modes and maximum energy to omitted modes; reverse these assignments for the upper bound. If the omitted modes may carry nearly all energy, mode counts alone give no nontrivial universal error bound. Thus $1 - Rg$ is a monotone score combining two failure sources, not the energy lower bound in (10) and not a calibrated target-error estimate.

A.2 Finite-probe guarantees and their limits

Let $E_1, \ldots, E_n$ be independent errors from the declared probe distribution, with distribution function F, unique median m, and empirical median $\widehat{m}$. Suppose for $0 < t \leq t_0$ there is a constant $c > 0$ such that $F(m+t) \geq 1/2 + ct$ and $F(m-t) \leq 1/2 - ct$. This local mass condition excludes an arbitrarily flat or nonidentifiable median.

Proposition 3 (median concentration). Under these conditions,

$$\Pr(|\widehat{m} - m| > t) \leq 2\exp(-2nc^2t^2), \qquad 0 < t \leq t_0. \tag{12}$$

Proof. If the median exceeds $m + t$, at least half the observations exceed $m + t$, although their probability is at most $1/2 - ct$. For a Bernoulli variable Z of mean p, the centered log moment-generating function has second derivative at most $1/4$, so $\mathbb{E}\exp(\lambda(Z-p)) \leq \exp(\lambda^2/8)$. Independence, exponential Markov inequality and minimization at $\lambda = 4ct$ give $\Pr(n^{-1}\sum Z_i - p \geq ct) \leq \exp(-2nc^2t^2)$. Apply the same argument to the lower median tail and add the bounds. For an even-sample averaged median, a strict tail event still requires at least half the observations in the corresponding tail. □

For M fixed model–band–grid combinations indexed by j and failure probability $\alpha \in (0,1)$, dependence *between* combinations is permitted: a union bound yields simultaneous median accuracy

$$|\widehat{m}_j - m_j| \leq \epsilon_n, \qquad \epsilon_n = \frac{1}{c}\sqrt{\frac{\log(2M/\alpha)}{2n}}, \tag{13}$$

with probability at least $1 - \alpha$, provided the common lower mass bound holds and $\epsilon_n \leq t_0$. Independent probes within each combination are essential; repeated evaluation of the same field does not increase n. Clipping is nonexpansive, so this also bounds fidelity error. If two measured fidelities differ by more than $2\epsilon_n$, their population clipped-fidelity ordering is correct on this event. With an additional cross-grid fidelity bound η for each candidate, a margin greater than $2(\epsilon_n + \eta)$ preserves that ordering on the target grid. Strict ordering of clipped fidelities corresponds to ordering of median errors; ties at zero contain no usable ordering information.

A conditional safety certificate. If the target median satisfies $|m_{\text{target}} - m_{\text{train}}| \le \eta$ and $\widehat{g}_{\text{train}} > 0$, then on the concentration event

$$m_{\text{target}} \le 1 - \widehat{g}_{\text{train}} + \epsilon_n + \eta. \tag{14}$$

Consequently $\widehat{g}_{\text{train}} > 1 - \tau + \epsilon_n + \eta$ is sufficient for target median error below τ. This is a median guarantee, not a guarantee for every input or an upper bound on the unsafe-input probability. The empirical 0.02 abstention margin and eight-probe operating point do not establish a distribution-free certificate: c and η are not known a priori. Their reported reliability comes from held-out experiments. Equation (13) explains both the benefit of more probes and why small probe counts cannot provide strong worst-case guarantees without distributional information.

The preceding guarantees require quantitative assumptions. Two counterexamples show why two grids and a restricted probe family alone are insufficient.

Counterexample 1 (an untested-grid failure). Fix train and check grids N_0, N_1. Let G be the identity and define $\widehat{G}_N = (1 + h(N))G$, where $h(N) = C(N - N_0)(N - N_1)$ for a constant C. Train and check errors are exactly zero, so their level and drift tests pass. At any third grid N_2 distinct from both, the relative error is $|C(N_2 - N_0)(N_2 - N_1)|$ and can be made arbitrarily large. Even smooth dependence on N therefore does not make two measurements sufficient without a quantitative regularity bound.

Counterexample 2 (an unexcited subspace). In a finite Fourier space, let all probes lie in a proper subspace S and define $\widehat{G} = G + CP_{S^\perp}$ with $G = I$. Every probe is predicted exactly on every grid. A unit target input in $S^\perp$ has relative error $|C|$. More probes from the same restricted support cannot expose this failure. This establishes a coverage requirement on probe spectra independently of the number of probes.

For a quantified extension, assume the population fidelity is L-Lipschitz in $x = \log N$ throughout an interval. If every target x lies within distance h of a measured check point, then its fidelity is at least the smallest measured fidelity minus $\epsilon_n + Lh$ on the simultaneous event. This follows directly from the Lipschitz inequality and provides a principled reason to add check grids. Two observed points alone do not estimate a valid global L. These limits support the paper's use of the check as falsification evidence within a declared physical regime, rather than an unconditional invariance theorem.

B Experimental Protocol and Reproducibility

The appendix groups mathematical foundations, experimental protocol, selection results, stationarity and transfer analyses, and expanded replications into five themes. The experimental stages, seed lists, margins, and decision criteria are included in the anonymous artifact. These choices were made before the later seeds and probe sequences were examined.

The distinction between descriptive cells and inferential units is central. A *cell* fixes a PDE, training seed, deployment regime, target grid, physical band, and model pair. Resolution–band cells from the same trained models are correlated. The independent unit used for uncertainty is therefore the PDE–training-seed cluster. Cell-pooled statistics are retained only to connect to the original cell-pooled analysis.

B.1 Probe construction and paired selection

Resolution-independent fields. In 1D, a probe input is the continuous periodic field

$$a(x) = \sum_{k=1}^{k_{\max}} \big(\alpha_k \cos(2\pi kx/L) + \beta_k \sin(2\pi kx/L)\big), \qquad \alpha_k, \beta_k \overset{\text{iid}}{\sim} \mathcal{N}(0, A^2 k^{-\gamma}). \tag{15}$$

Here L is the spatial period, A the amplitude, γ the spectral decay exponent, and $k_{\max}$ the frequency cutoff. The field is rescaled to root-mean-square (RMS) amplitude A. The 2D construction draws one coefficient from each conjugate pair in the integer disk $0 < \|\mathbf{k}\|_2 \le k_{\max}$ with independent

real and imaginary parts proportional to $\|\mathbf{k}\|_2^{-\gamma/2}$, adds its Hermitian conjugate, and applies the same RMS normalization. The same continuous draw is sampled on every probe grid; changing N therefore changes only the discretization, not the input realization.

The three deployment regimes use $k_{\max} \in \{k_{\text{train}}, 2k_{\text{train}}, 4k_{\text{train}}\}$. At probe grid N, $k_{\max}$ is clipped to $N/2 - 1$. Each candidate receives identical probe inputs and reference outputs. The selector uses $n_p = 8$ and the direct stationarity test uses 24 matched probes at each grid. Trusted reference solves are needed only for these synthetic probes at N_{train} and N_{check}; target-grid reference solutions are withheld until retrospective scoring.

Band error and fidelity. For a 1D band or 2D radial shell b, the Parseval-consistent relative error is

$$e_b = \left[\frac{\sum_{\mathbf{k}\in b} w_{\mathbf{k}} |\widehat{\hat{G}_\theta(a) - u}_{\mathbf{k}}|^2}{\sum_{\mathbf{k}\in b} w_{\mathbf{k}} |\hat{u}_{\mathbf{k}}|^2 + 10^{-12}} \right]^{1/2}, \qquad g(b) = \text{clip}\big(1 - \text{median}_{i\le n_p} e_b^{(i)}, 0, 1\big). \tag{16}$$

The weights $w_{\mathbf{k}}$ count conjugate frequencies correctly in the real FFT layout. A target cell is eligible only when more than half of its band is represented. A band absent from a probe grid is recorded as unrepresented and receives no transferable verdict; it is never imputed from another resolution.

Controlled 2D spectral tradeoff. The decisive paired experiment holds the FNO-2D architecture fixed at 16 modes, width 32, and depth 4. Only the training loss changes:

$$\mathcal{L}_t = 0.25\mathcal{L}_{\text{rel}} + 0.75 \left[\frac{\sum_{\mathbf{k}} (1 + \|\mathbf{k}\|_2)^t |\widehat{\hat{u} - u}_{\mathbf{k}}|^2}{\sum_{\mathbf{k}} (1 + \|\mathbf{k}\|_2)^t |\hat{u}_{\mathbf{k}}|^2 + 10^{-8}} \right]^{1/2}, \qquad t \in \{-2, -1, 0, 1, 2\}. \tag{17}$$

Here $\mathcal{L}_{\text{rel}}$ is relative-L^2 prediction loss and t is the spectral-loss exponent. This creates low–high-frequency crossovers without changing modes, width, depth, data, or parameter count. The anchored stress set uses circular convolutions and Fourier truncation/zero padding for U-Net down/up-sampling; the CNO variant additionally applies upsample–GELU–low-pass–downsample anti-aliased activations. Both enter the same balanced two-grid comparison.

Paired selection and target quantities. For a trained operator j, physical band β, and an ordered probe sequence, the n_p-probe fidelity is

$$g_j^{(n_p)}(\beta) = \text{clip}\Big(1 - \text{median}_{i\le n_p} e_{j,\beta}^{(i)}, 0, 1\Big). \tag{18}$$

For two variants j and k of the same architecture, trained on the same PDE and data seed, the selector chooses k in band β when $g_k^{(n_p)}(\beta) > g_j^{(n_p)}(\beta)$. An eligible target cell must be off the training grid, more than half represented, finite for both variants, and separated by at least 0.02 in log-error. The selector abstains when $|g_k^{(n_p)}(\beta) - g_j^{(n_p)}(\beta)| < 0.02$.

The primary performance estimand is concordance among non-abstained cells; coverage is the fraction of eligible cells retained. Global training error is evaluated on the exact same selected cells. The confirmatory point estimate first averages probe-sequence repeats within a PDE–seed cluster, then gives equal weight to seeds within a PDE and equal weight to PDEs. A separate model-pair macro average checks that one easy tilt pair does not dominate.

Nested probes and deployment cost. For each 2D $(\text{PDE}, \text{seed}, \text{regime}, \text{repeat})$ tuple, one ordered sequence of 48 reference-labelled probes is generated and shared by all five trained loss-tilt models. Five independently seeded sequences are used per PDE–seed cluster. The points

$$n_p \in \{1, 2, 4, 8, 16, 24, 48\}$$

are exact prefixes, so a curve point differs only by adding matched probes. The hierarchical 95% interval resamples PDE–seed clusters and then samples one of the five probe sequences within each drawn cluster.

Wall time is measured after GPU warm-up. One deployment compares five candidate models across three spectral regimes and is defined as

$$\text{shared input/reference generation} + \text{five-model inference and band scoring.}$$

Loading model weights, imports, model training, and the target-grid evaluation sweep are excluded from this marginal selector cost. CUDA synchronization brackets timed inference. Reference, candidate-evaluation, and total costs are retained separately for every PDE–seed–repeat deployment.

B.2 Reproducibility and artifact contents

The original anonymous artifact contains source code, tests, per-model records, the 10 matched probe sequences, cluster and model-pair summaries, figure source data, and this LaTeX source. Trained weights and cached training arrays are omitted for size, but all resumable training drivers are present. A checksum manifest supports verification of the released files.

The reference solvers are tested against analytic linear solutions and self-convergence; spectral measurements are tested for Parseval consistency and known Fourier modes; operators are tested across resolutions; and the inference code includes a constructed pseudo-replication counterexample in which a cell-pooled concordance of 0.9 correctly becomes a cluster macro of 0.5. Another regression test verifies that shared reference generation is counted once rather than once per candidate model.

The additional replication is distributed separately with its protocol, raw records, checkpoints, and analysis scripts; it is not covered by the original artifact manifest.

C Selection Performance and Robustness

C.1 Probe efficiency and matched-coverage comparisons

Figure 3 gives the complete result. Eight probes reach paired concordance 0.962 (95% CI $[0.945, 0.973]$) at coverage 0.259, compared with 0.974 at 24 probes. Their median total selector times are 11.36 and 34.61 seconds, a ratio of 0.33. Thus the paired selector used for the strong 2D result is measured on the same cost curve as its selection accuracy; the earlier threshold-AUC curve is not reused for this claim. The expanded confirmatory point estimate 0.978 uses the 24-probe sequence chosen before confirmation. The 24-probe curve value 0.974 averages five newly drawn sequences on the same trained models, so the two numbers are an independent probe-draw replication. At 24 probes, coverage is 0.130 for advection–diffusion and 0.397 for Navier–Stokes; the lower linear-PDE coverage is visible rather than absorbed by cell pooling.

Table 3: Matched paired-selector probe-count curve. Time is the median total deployment time across PDE–seed–repeat instances.

n_p	concordance [95% CI]	coverage [95% CI]	time (s)
1	0.903 $[0.878, 0.927]$	0.278 $[0.255, 0.301]$	1.32
2	0.940 $[0.911, 0.961]$	0.263 $[0.239, 0.286]$	2.84
4	0.950 $[0.928, 0.963]$	0.256 $[0.231, 0.281]$	5.51
8	0.962 $[0.945, 0.973]$	0.259 $[0.238, 0.280]$	11.36
16	0.972 $[0.962, 0.980]$	0.261 $[0.240, 0.281]$	22.71
24	0.974 $[0.967, 0.981]$	0.263 $[0.243, 0.286]$	34.61
48	0.977 $[0.970, 0.983]$	0.265 $[0.247, 0.282]$	68.38

Matched-coverage baseline. The primary comparison uses the chosen $n_p = 8$, $\Delta_g = 0.02$, and target log-error separation $\delta_e = 0.02$. For each PDE–training-seed–probe-repeat deployment, let K_{keep} be the mass retained by the probe selector. The independent global-training-error selector predicts the candidate with lower training relative-L^2 error and ranks eligible cells by

$$c_{\text{train}}(j,k) = \left|\log_{10} E_j^{\text{train}} - \log_{10} E_k^{\text{train}}\right|. \tag{19}$$

It retains the highest-confidence cells with total weight K_{keep}. If the cutoff splits a confidence tie, every boundary cell receives the same fractional weight needed to make the retained mass exactly K_{keep}. This is the expectation under uniform random tie-breaking and prevents target labels or row order from resolving the tie. Coverage therefore matches to numerical precision in every deployment, not just after pooling.

Table 4 shows a decisive separation: at exactly 0.259 coverage, the probe reaches 0.962 concordance versus 0.464 for training error. The paired advantage is 0.498 with hierarchical 95% CI $[0.454, 0.547]$ and small-cluster Student-t interval $[0.217, 0.779]$. It is positive for both PDEs, and its worst leave-one-PDE–seed-cluster-out value is 0.478. Thus neither unequal refusal rates nor correlated target cells explain the probe advantage. Model-pair aggregation over pairs selected by both rules is retained as a descriptive common-support summary; its coverage is not, by construction, the matched deployment coverage.

Table 4: Eight-probe selector versus global training error at exactly matched coverage. The final row is a hierarchical interval that resamples PDE–seed clusters and one of five probe repeats within cluster.

stratum	probe	matched train error	difference	coverage
advection–diffusion	1.000	0.868	0.498	0.129
Navier–Stokes	0.924	0.060	0.864	0.389
PDE–seed macro	0.962	0.464	0.498	0.259
hierarchical 95% CI	[0.945, 0.973]	[0.419, 0.505]	[0.454, 0.547]	–

C.2 Independent-cluster and model-pair inference

The original new-only 2D replication contains 6 previously untouched PDE–seed clusters. Its cluster macro concordance is 0.981 with Student-t interval $[0.956, 1.000]$; the model-pair macro is 0.963 and the worst leave-one-cluster-out estimate is 0.977. Combining these with the four earlier confirmatory clusters gives 10 clusters. The expanded PDE–seed macro is 0.978 versus 0.965 cell-pooled; its small-cluster interval is $[0.959, 0.998]$, model-pair macro is 0.967, and worst leave-one-cluster-out concordance is 0.975. Per-PDE estimates are 1.000 for advection–diffusion and 0.956 for Navier–Stokes. The exact two-sided cluster sign-flip test against chance gives $p = 0.0020$.

Table 5: Robust 2D aggregation. Model-pair aggregation is descriptive; the uncertainty interval remains at the PDE–seed level.

set	clusters	PDE–seed macro	model-pair macro	worst LOCO
new untouched	6	0.981	0.963	0.977
expanded confirmatory set	10	0.978	0.967	0.975

Table 6: All expanded fixed-architecture 2D PDE–seed clusters. Cells are descriptive comparison cells, not independent replicates.

PDE	seed cluster	cells	coverage	concordance
advection–diffusion	s10	554	0.162	1.000
advection–diffusion	s11	521	0.086	1.000
advection–diffusion	s7	538	0.132	1.000
advection–diffusion	s8	537	0.132	1.000
advection–diffusion	s9	540	0.143	1.000
Navier–Stokes	s5	625	0.414	0.927
Navier–Stokes	s6	621	0.385	0.967
Navier–Stokes	s7	594	0.448	0.940
Navier–Stokes	s8	611	0.380	0.970
Navier–Stokes	s9	610	0.366	0.978

Expanded 1D confirmation. The 1D untouched confirmation now has nine independent PDE–seed clusters per family. Equal-weight PDE–seed macro concordance is 0.990 for FNO, 0.871 for DeepONet, and 0.909 for Galerkin, at respective coverages 0.370, 0.416, and 0.535. Their small-cluster intervals are $[0.980, 0.999]$, $[0.789, 0.952]$, and $[0.856, 0.962]$. The corresponding worst leave-one-cluster-out estimates are 0.987, 0.854, and 0.898. Development-augmented estimates are archived separately and are not silently pooled into this confirmatory set.

C.3 Decision-component ablations and robustness

The ablations were specified before their analysis. They replay the same ten 2D PDE–seed clusters, five loss-tilt candidates, five matched ordered probe sequences, target grids, and target labels used in the cost curve. No model is retrained, and no candidate or cell is added after inspecting the results. The component comparison reuses the complete 1D two-grid records; the score factorization uses 81 stationary ICLR 1D records and the 12 stationary FNO-2D records from the independent radial-band matrix. Anchored models remain in the stationarity evaluation but are not mixed into the factorization. All outputs, including the low-coverage endpoint, are retained in Tables 7–10.

Abstention margin and probe aggregation. Table 7 evaluates the chosen margin grid at $n_p = 8$. The zero-margin rule retains all cells, including indistinguishable and exact-tie cases, but reaches only 0.576 concordance. The curve jumps above 0.94 at the first nonzero margin and reaches 0.962 at the fixed 0.02 point. Increasing the margin to 0.04 raises concordance to 0.974 while reducing coverage to 0.179; at 0.12, only 0.013 of cells remain and the interval becomes wide. Thus 0.02 is a risk–coverage compromise rather than the maximum observed point.

Table 7: Abstention-margin frontier at eight probes. Intervals are hierarchical 95% intervals over PDE–seed clusters and matched probe repeats.

Δ_g	concordance [95% CI]	coverage [95% CI]
0	0.576 [0.559, 0.591]	1.000 [1.000, 1.000]
0.005	0.944 [0.925, 0.958]	0.338 [0.318, 0.357]
0.01	0.949 [0.930, 0.963]	0.306 [0.287, 0.323]
0.02	0.962 [0.945, 0.973]	0.259 [0.237, 0.281]
0.04	0.974 [0.961, 0.987]	0.179 [0.161, 0.197]
0.08	0.977 [0.944, 1.000]	0.053 [0.035, 0.070]
0.12	0.906 [0.719, 1.000]	0.013 [0.001, 0.026]

We use the median to reduce the influence of occasional extreme probe errors. It is not uniquely favorable: ordinary mean, 20% trimmed mean, 20% winsorized mean, and median aggregation all exceed 0.96 concordance on identical prefixes, with nearly unchanged coverage (Table 8). Their observed spread is 0.011.

Table 8: Probe-error aggregation at $n_p = 8$ and $\Delta_g = 0.02$.

aggregator	concordance [95% CI]	coverage [95% CI]
mean	0.973 [0.965, 0.981]	0.261 [0.240, 0.282]
20% trimmed mean	0.970 [0.958, 0.980]	0.262 [0.243, 0.281]
20% winsorized mean	0.972 [0.962, 0.982]	0.260 [0.240, 0.280]
median	0.962 [0.945, 0.973]	0.259 [0.239, 0.280]

What each decision component contributes. The component comparison distinguishes a second-grid level check from the fidelity drift condition. On anchored models, checking both levels removes most false accepts (0.690 to 0.049); the drift condition closes the remaining gap (0.025). On stationary models, all variants retain cell-pooled false rejection below 0.04. Combining all components therefore passes both the 0.05 reference levels as a descriptive point estimate. Drift-only matches its anchored false-accept rate at this $N/2$ check and has lower false rejection (0.024); the full rule is retained because it was chosen in advance and additionally requires the check-grid fidelity itself to be safe. These data do not establish that the full conjunction is uniformly optimal across every check grid or PDE; the independent-unit intervals are reported below.

Table 9: Decision-component ablation at $N_{\text{check}} = N_{\text{train}}/2$.

decision	anchored FA	stationary FR
train-grid level	0.690	0.005
both levels	0.049	0.037
train level + drift	0.025	0.024
full: both levels + drift	0.025	0.037

The factor comparison evaluates $1 - \text{repr}$, $1 - g$, and $1 - \text{repr}\, g$ on identical labels (Table 10). Representability handles bands beyond the target grid, whereas the probe supplies model-specific ordering within represented bands. Their product is best in 1D reverse resolution and in the separate 2D matrix. In super-resolution, probe-only and full AUC are both approximately one because the finer grid mainly repeats a response already measured at training resolution.

Table 10: Coefficient-free score-factor ablation.

stratum	$1 - \text{repr}$	$1 - g$	$1 - \text{repr}\, g$
1D reverse, all bands	0.847	0.912	0.937
1D reverse, represented	0.589	0.818	0.830
1D super, represented	0.546	0.999	0.999
2D, all bands	0.720	0.885	0.894
2D, represented	0.597	0.842	0.848

Near ties, candidate hardness, and probe-spectrum conditioning. For candidates j and k with loss tilts $t_j, t_k \in \{-2, -1, 0, 1, 2\}$ and target relative errors E_j, E_k, let δ_e denote the minimum required target log-error separation $|\log E_j - \log E_k|$, and let $d_t = |t_j - t_k|$ measure candidate-pair hardness. Regime-matched probes use the same spectral envelope as the target regime; in-distribution-reused probes apply the in-distribution probe response to all three target regimes; regime-pooled probes aggregate the three probe spectra before ranking candidates. All three mappings use $n_p = 8$ and the same ten PDE–seed clusters, five candidates, and five matched probe-sequence repeats.

Figure 4 tests the paired ordering across these three independent perturbations. First, the target-error near-tie threshold is removed rather than increased. At $\delta_e = 0$, all finite represented pairs are eligible and macro concordance remains 0.937 (95% CI $[0.922, 0.951]$), compared with 0.962 at the chosen 0.02 threshold. The all-pair retained fraction, measured against the all-pair denominator rather than the filtered eligible set, is 0.227 at $\delta_e = 0$ and 0.183 at the chosen point. The high concordance therefore does not require deleting nearly tied target errors.

Second, adjacent loss tilts reach 0.953 ($[0.930, 0.974]$), essentially matching the 0.962 all-pair result. Even the worst individual pair reaches 0.855 ($t = -1$ versus $t = 0$; 95% CI $[0.736, 0.937]$; coverage 0.150), and the worst of the five leave-one-tilt-out analyses (omitting $t = 2$) reaches 0.932 ($[0.906, 0.955]$; coverage 0.229). Extreme tilt pairs therefore do not dominate the selector result. Finally, regime-matched, in-distribution-reused, and regime-pooled probes reach 0.962, 0.948, and 0.975, respectively. All exceed 0.90 and their spread is 0.026. The ordering is therefore stable across the three ways of matching probe spectra to the deployment regime.

D Stationarity and Resolution-Transfer Analysis

D.1 Two-grid tests, thresholds, and mechanisms

Let $g_N(b)$ denote the probe fidelity when the same model is evaluated on grid N. For unsafe-error threshold $\tau = 0.25$ and stationarity tolerance $\delta = 0.10$, a commonly represented band passes the two-grid test when

$$q(b) = \mathbf{1}\{\min(g_{N_{\text{train}}}(b), g_{N_{\text{check}}}(b)) \geq 1 - \tau, \quad |g_{N_{\text{train}}}(b) - g_{N_{\text{check}}}(b)| \leq \delta\}\,. \tag{20}$$

The primary check is $N_{\text{train}}/2$. When the test fails, the training-grid verdict is withheld; target cells are never relabelled retrospectively.

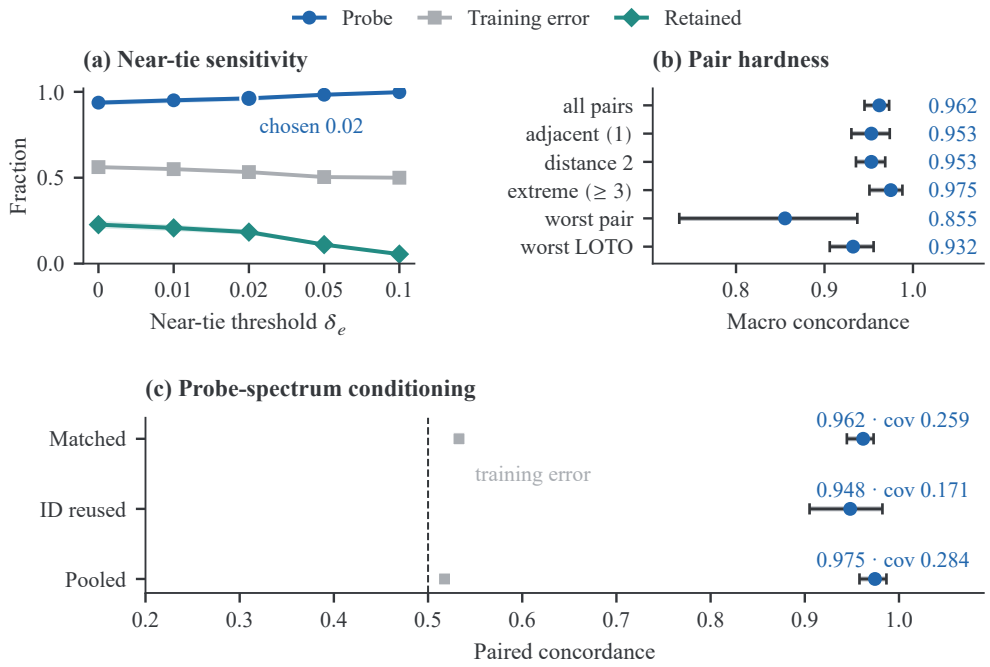


Figure 4: Across ten PDE–seed clusters and five matched probe repeats at $n_p = 8$, the paired conclusion is stable to target near ties, hard candidate pairs, and probe-spectrum changes. (a) Probe and training-error concordance as the minimum target log-error separation δ_e varies; the teal curve is the selected fraction relative to all finite represented pairs. Intervals are hierarchical 95% intervals over PDE–seed clusters and probe repeats. (b) Concordance by loss-tilt distance, followed by the worst individual pair and worst leave-one-tilt-out (LOTO) result. (c) Regime-matched probes compared with reusing in-distribution probes and pooling all three probe spectra; labels also report selective coverage.

Table 11: One- versus two-grid decisions for each check resolution. False acceptance is measured on anchored stress models and false rejection on stationary models. These direction-sensitivity entries are cell-pooled.

	anchored FA (%)		stationary FR (%)	
N_{check}	one	two	one	two
$N/4$	69.9	0.0	2.8	14.3
$N/2$	69.0	2.5	1.2	3.5
$2N$	70.2	5.1	1.0	1.0

At the primary $N/2$ check, Table 12 instead treats each PDE–training-seed combination as one independent cluster. There are nine clusters: three training seeds for each of advection–diffusion, Burgers, and Kuramoto–Sivashinsky. Architectures, regimes, grids, and bands are aggregated within a cluster. The PDE–seed macro anchored FA falls from 68.6% to 2.8%; the mean absolute reduction is 65.8 percentage points with cluster-bootstrap 95% CI $[57.2, 74.2]$ and Student-t interval $[55.3, 76.2]$, all in percentage points. The reduction is positive for every PDE and every leave-one-cluster-out analysis, whose minimum is 63.7 percentage points. This comparative FA conclusion is therefore cluster-robust.

Stationary FR rises from 1.3% to 3.7%. The cell-pooled values in the direction table (69.0%, 2.5%, 1.2%, and 3.5%) agree closely with the corresponding cluster macros, but are not used as replicates. Both two-grid macro point estimates are below 5%. Across nine clusters, the Student-t upper endpoints narrow to 6.0% for FA and 6.3% for FR; we therefore claim a large cluster-robust FA reduction and low observed error rates, not a strict population-level 5% upper-bound guarantee.

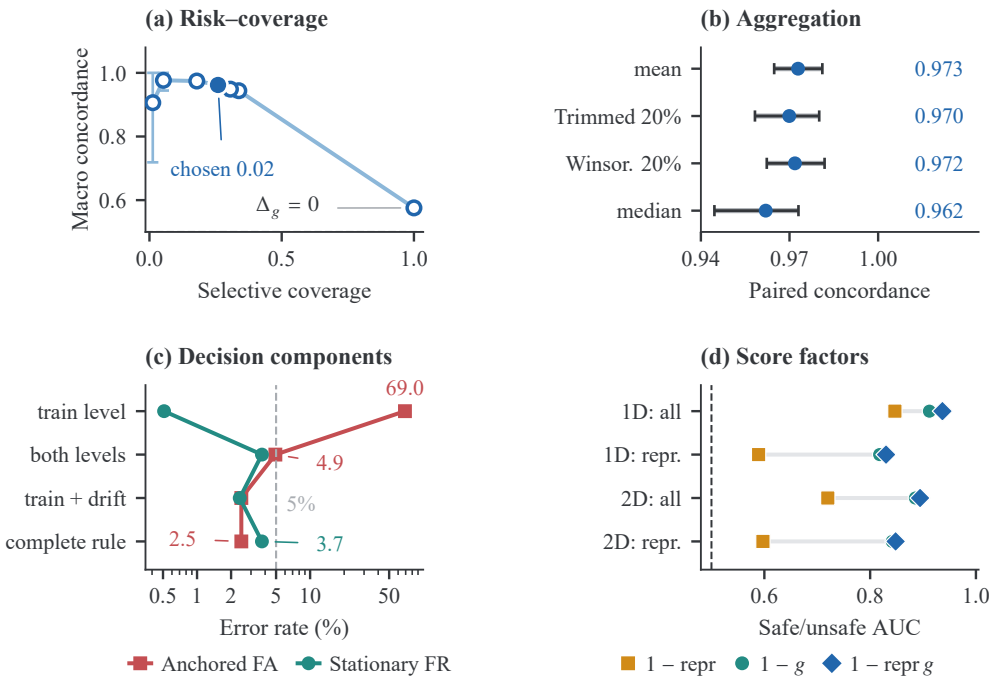


Figure 5: Component ablations of the decision object. (a) Concordance–coverage frontier over seven margins at eight probes. (b) Four aggregators applied to identical probe errors. (c) False acceptance on anchored models and false rejection on stationary models after ablating the check-grid level and drift conditions. (d) Safe/unsafe AUC for target-grid representability, learned response, and their coefficient-free product. Intervals preserve the PDE–seed-cluster hierarchy and all comparisons use unchanged target cells.

Table 12: Primary stationarity analysis with PDE–training-seed as the independent unit. Bootstrap resampling and Student-t intervals operate on the nine cluster rates, not on resolution–band cells.

independent-unit estimate	one grid	two grids
anchored FA, PDE–seed macro	68.6%	2.8%
cluster-bootstrap 95% CI	[62.5%, 75.1%]	[0.9%, 5.6%]
Student-t 95% CI	[60.8%, 76.3%]	[0.0%, 6.0%]
stationary FR, PDE–seed macro	1.3%	3.7%
cluster-bootstrap 95% CI	[0.4%, 2.5%]	[1.7%, 5.9%]
Student-t 95% CI	[0.0%, 2.7%]	[1.0%, 6.3%]

Table 13: Per-PDE error rates and pooled cell counts across three training seeds at the primary $N/2$ check.

PDE	one-grid FA	two-grid FA	one-grid FR	two-grid FR
advection–diffusion	66.7% (720/1080)	0.0% (0/1080)	0.4% (3/689)	1.6% (11/689)
Burgers	81.0% (797/984)	0.0% (0/984)	1.4% (8/556)	7.4% (41/556)
Kuramoto–Sivashinsky	58.0% (498/858)	8.4% (72/858)	2.0% (10/504)	2.0% (10/504)

The $N/4$ check is most aggressive but rejects more stationary safe cases. The $2N$ check cannot expose all reverse-resolution failures. One check grid is therefore a falsification of stationarity on a direction and interval, not a proof of discretization invariance.

Table 14: Threshold sensitivity on stationary 1D operators.

τ	0.10	0.25	0.50	1.00
macro AUC$(1-g)$	0.894	0.845	0.795	0.757
scoreable bands	5	6	6	7

Threshold sensitivity. At the primary threshold, precision and specificity are 1.00, recall is 0.655, and balanced accuracy is 0.827. The probe is a conservative selection rule, not a calibrated error regressor. In super-resolution the probe mostly amortizes a measured training-grid response (AUC 0.998), whereas reverse resolution changes representability and is the genuine forecast regime (AUC 0.810 versus 0.766 for a band prior).

Varying the retained FNO modes at the same physical band separates learned bandwidth from grid resolution (Table 15).

Table 15: Same physical band $[16, 32)$, different FNO mode counts.

FNO modes	8	16	32
advection–diffusion: g	0.43	0.91	0.98
advection–diffusion: error	0.93	0.11	0.04
KS: g	0.23	0.96	0.96
KS: error	1.02	0.05	0.05

D.2 Radial-band transfer and score construction

The separate 20-model 2D architecture analysis is not pooled with paired selection. It tests whether the same physical-band construction transfers from 1D Fourier bands to 2D radial shells.

Table 16: Separate 2D radial-band analysis. "Transfer" is $1 - \mathrm{repr}\, g$ on representable bands.

subset	transfer AUC	band prior	within-band AUC
advection–diffusion (15)	0.791	0.769	0.695
Navier–Stokes (5)	0.976	0.934	0.815
pooled (20)	0.843	0.772	0.786

Two preliminary untouched mode–width phases reached paired concordance 0.876 and 0.846, but global training error induced the same ordering. They are reported as replications rather than incremental band-specific evidence. The decisive phase fixes modes, width, depth, data, and parameter count, changing only the radial Fourier-loss tilt; this creates bandwise crossovers that one global error cannot express.

Resolution atlas and discarded support factor. The target-grid score is

$$s(b; N_{\mathrm{eval}}) = 1 - \mathrm{repr}(b; N_{\mathrm{eval}}) g(b). \tag{21}$$

An input-support factor explored during development appeared to improve reverse- resolution AUC, but the gain vanished when spectra were recomputed from an alias-free reference. It primarily encoded coarse-grid aliasing and was excluded before the confirmatory paired phases.

E Expanded Architecture Results and Additional Replication

E.1 Complete architecture-level results

Tables 17–19 expose the complete original 105-model development matrix: 39 PDE–architecture–capacity configurations with two or three training seeds each. They show both lower and higher physical bands at reverse, training, and super-resolution grids. Each value is computed from the saved sample-mean band errors, then averaged across training seeds. The selected $[8, 16)$ and

$[16, 32)$ bands are fully represented on all three grids. In the variant labels, m, w, d, s, p, and h denote retained modes, width, depth, sensors, basis size, and attention heads, respectively; n counts training seeds. These descriptive tables are separate from the independent paired holdouts used for the main claims.

The archived aggregate error is an unweighted mean of band-relative errors, despite having been labeled as a global relative error. It is not used in these tables and cannot be converted into a true global norm from the aggregate records. The training-error baselines use separately recorded true training relative norms and are unaffected. The additional replication reports true global relative error separately from mean band-relative error; the archived measurements are retained for reproducibility.

Table 17: Advection–diffusion: original in-distribution per-band relative errors. Each entry is mean ± sample SD across the listed independently trained seeds; these are archived experiments, not new runs.

Family	Variant	n	$[8, 16)$			$[16, 32)$		
			$N = 64$	128	256	64	128	256
CNO	w48 d3	3	1.194 ± 0.004	0.009 ± 0.000	2.876 ± 0.001	1.300 ± 0.272	0.117 ± 0.014	7.505 ± 0.009
DEEPONET	s128 p128 w128 d4	2	0.499 ± 0.001	0.499 ± 0.001	0.499 ± 0.001	1.604 ± 0.105	1.573 ± 0.069	1.604 ± 0.053
DEEPONET	s128 p256 w256 d5	3	0.351 ± 0.011	0.351 ± 0.011	0.351 ± 0.011	0.961 ± 0.090	0.959 ± 0.091	0.961 ± 0.091
DEEPONET	s64 p64 w128 d4	2	0.403 ± 0.009	0.403 ± 0.009	0.403 ± 0.009	1.549 ± 0.085	1.488 ± 0.042	1.498 ± 0.029
FNO	m12 w64 d4	3	0.044 ± 0.004	0.044 ± 0.004	0.044 ± 0.004	0.636 ± 0.046	0.620 ± 0.048	0.619 ± 0.048
FNO	m16 w64 d4	3	0.004 ± 0.000	0.004 ± 0.000	0.004 ± 0.000	0.170 ± 0.007	0.168 ± 0.008	0.168 ± 0.007
FNO	m24 w64 d4	3	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000	0.052 ± 0.001	0.047 ± 0.001	0.047 ± 0.001
FNO	m32 w64 d4	3	0.003 ± 0.000	0.003 ± 0.000	0.003 ± 0.000	0.067 ± 0.009	0.048 ± 0.006	0.047 ± 0.006
FNO	m8 w64 d4	3	0.261 ± 0.011	0.261 ± 0.011	0.261 ± 0.011	2.507 ± 0.016	2.479 ± 0.016	2.478 ± 0.016
GALERKIN	w32 h4 d3	2	1.299 ± 0.006	1.294 ± 0.009	1.294 ± 0.009	6.016 ± 0.261	5.780 ± 0.283	5.802 ± 0.218
GALERKIN	w64 h4 d4	3	1.046 ± 0.070	1.043 ± 0.073	1.044 ± 0.073	6.266 ± 0.246	6.141 ± 0.246	6.156 ± 0.243
GALERKIN	w96 h4 d4	2	0.871 ± 0.314	0.869 ± 0.322	0.870 ± 0.323	6.030 ± 1.418	5.801 ± 1.263	5.791 ± 1.234
UNET	w48 d3	3	1.198 ± 0.010	0.009 ± 0.001	2.876 ± 0.000	1.359 ± 0.298	0.123 ± 0.024	7.502 ± 0.006

Table 18: Burgers: original in-distribution per-band relative errors. Each entry is mean ± sample SD across the listed independently trained seeds; these are archived experiments, not new runs.

Family	Variant	n	$[8, 16)$			$[16, 32)$		
			$N = 64$	128	256	64	128	256
CNO	w48 d3	3	1.187 ± 0.002	0.029 ± 0.002	1.502 ± 0.003	1.243 ± 0.006	0.063 ± 0.005	1.714 ± 0.001
DEEPONET	s128 p128 w128 d4	2	1.126 ± 0.022	1.127 ± 0.022	1.128 ± 0.022	1.155 ± 0.003	1.149 ± 0.010	1.148 ± 0.011
DEEPONET	s128 p256 w256 d5	3	1.018 ± 0.010	1.018 ± 0.010	1.018 ± 0.009	1.208 ± 0.025	1.201 ± 0.023	1.201 ± 0.023
DEEPONET	s64 p64 w128 d4	2	1.165 ± 0.009	1.165 ± 0.009	1.165 ± 0.009	1.150 ± 0.021	1.148 ± 0.020	1.149 ± 0.020
FNO	m12 w64 d4	3	0.104 ± 0.002	0.088 ± 0.003	0.088 ± 0.003	0.344 ± 0.003	0.144 ± 0.008	0.144 ± 0.008
FNO	m16 w64 d4	3	0.077 ± 0.001	0.054 ± 0.002	0.054 ± 0.002	0.319 ± 0.002	0.109 ± 0.004	0.109 ± 0.004
FNO	m24 w64 d4	3	0.078 ± 0.003	0.057 ± 0.004	0.057 ± 0.004	0.321 ± 0.005	0.128 ± 0.007	0.128 ± 0.007
FNO	m32 w64 d4	3	0.077 ± 0.004	0.060 ± 0.003	0.060 ± 0.003	0.322 ± 0.011	0.154 ± 0.004	0.154 ± 0.004
FNO	m8 w64 d4	3	0.225 ± 0.009	0.218 ± 0.011	0.218 ± 0.011	0.421 ± 0.008	0.287 ± 0.010	0.287 ± 0.010
GALERKIN	w32 h4 d3	2	0.747 ± 0.063	0.745 ± 0.061	0.746 ± 0.060	1.020 ± 0.056	0.990 ± 0.065	0.990 ± 0.060
GALERKIN	w64 h4 d4	3	0.679 ± 0.074	0.674 ± 0.075	0.673 ± 0.075	0.978 ± 0.089	0.929 ± 0.089	0.925 ± 0.091
GALERKIN	w96 h4 d4	2	0.740 ± 0.053	0.735 ± 0.055	0.735 ± 0.056	1.109 ± 0.119	1.056 ± 0.110	1.053 ± 0.106
UNET	w48 d3	3	1.190 ± 0.003	0.035 ± 0.002	1.500 ± 0.004	1.251 ± 0.005	0.074 ± 0.004	1.711 ± 0.002

E.2 DIFFUSION-STRENGTH REPLICATION

Purpose and scope. This new experiment varies the physical smoothing strength as well as model bandwidth. It is separate from the original 1D and 2D evidence and uses newly trained models and disjoint data. Its purpose is a rapid, controlled test of the physical-frequency mechanism, with exact

Table 19: Kuramoto–Sivashinsky: original in-distribution per-band relative errors. Each entry is mean ± sample SD across the listed independently trained seeds; these are archived experiments, not new runs.

Family	Variant	n	$[8, 16)$			$[16, 32)$		
			$N = 64$	128	256	64	128	256
CNO	w48 d3	3	0.987 ± 0.001	0.012 ± 0.001	1.001 ± 0.000	2.315 ± 0.461	0.170 ± 0.009	19.613 ± 0.027
DEEPONET	s128 p128 w128 d4	2	0.929 ± 0.022	0.929 ± 0.023	0.929 ± 0.023	1.334 ± 0.055	1.321 ± 0.049	1.326 ± 0.049
DEEPONET	s128 p256 w256 d5	3	0.847 ± 0.006	0.847 ± 0.006	0.847 ± 0.006	1.341 ± 0.036	1.342 ± 0.036	1.343 ± 0.037
DEEPONET	s64 p64 w128 d4	2	0.888 ± 0.029	0.888 ± 0.029	0.888 ± 0.029	1.283 ± 0.010	1.270 ± 0.010	1.275 ± 0.015
FNO	m12 w64 d4	3	0.028 ± 0.001	0.028 ± 0.001	0.028 ± 0.001	0.232 ± 0.006	0.224 ± 0.006	0.223 ± 0.006
FNO	m16 w64 d4	3	0.006 ± 0.000	0.006 ± 0.000	0.006 ± 0.000	0.082 ± 0.007	0.076 ± 0.007	0.076 ± 0.007
FNO	m24 w64 d4	3	0.007 ± 0.001	0.007 ± 0.001	0.007 ± 0.001	0.247 ± 0.019	0.222 ± 0.016	0.221 ± 0.016
FNO	m32 w64 d4	3	0.006 ± 0.001	0.005 ± 0.001	0.005 ± 0.001	0.109 ± 0.014	0.083 ± 0.012	0.082 ± 0.012
FNO	m8 w64 d4	3	0.261 ± 0.002	0.260 ± 0.002	0.260 ± 0.002	1.790 ± 0.062	1.749 ± 0.068	1.749 ± 0.071
GALERKIN	w32 h4 d3	2	0.893 ± 0.012	0.895 ± 0.011	0.896 ± 0.010	14.880 ± 0.356	14.673 ± 0.298	14.638 ± 0.287
GALERKIN	w64 h4 d4	3	1.123 ± 0.030	1.116 ± 0.028	1.115 ± 0.029	18.440 ± 1.121	17.375 ± 0.829	17.301 ± 0.756
GALERKIN	w96 h4 d4	2	1.034 ± 0.034	1.028 ± 0.034	1.026 ± 0.035	17.016 ± 1.468	16.186 ± 1.424	16.165 ± 1.421
UNET	w48 d3	3	0.987 ± 0.001	0.013 ± 0.001	1.001 ± 0.000	2.481 ± 0.569	0.185 ± 0.015	19.607 ± 0.019

reference solutions. It does not add a new nonlinear PDE family or establish large-scale benchmark superiority. The protocol was fixed before these outcomes were inspected.

Models, data, and exact references. We solve $u_t = -u_x + \nu u_{xx}$ on $[0, 2\pi)$ at $T = 0.5$, with $\nu \in \{0.005, 0.02, 0.08\}$. Each setting has three FNO capacities (8, 16, or 32 retained modes; width 32, depth 3) and three independent training seeds (100–102), totaling 27 models. Each is trained on 512 fields at $N = 128$ for 200 epochs with batch size 64, AdamW learning rate 10^{-3}, weight decay 10^{-5}, and a cosine schedule. Gaussian Fourier coefficients have variance proportional to $k^{-1.5}$, cutoff 16, and unit RMS amplitude. Viscosity-specific random-stream offsets ensure independence across the nine viscosity–seed clusters, while candidate capacities share fields within each cluster. Training, probe, check, and evaluation streams are disjoint.

For coefficient a_k, the exact reference multiplier is $H_\nu(k) = \exp(-ik/2 - \nu k^2/2)$. Inputs are sampled directly as continuous trigonometric polynomials; output references are projected to modes strictly below target Nyquist. Thus coarse-grid input aliasing is retained, while no time-stepping or reference convergence error is introduced. We use 96 fresh evaluation fields per regime, with cutoffs 16, 32, and 48, on seven grids $N \in \{32, 48, 64, 96, 128, 192, 256\}$. This gives 54,432 model–field–grid evaluations; they are correlated measurements, not independent replicates. Table 20 reports the global-norm results for cutoff 16.

Table 20: Diffusion-strength replication: true global relative L^2 error in units of 10^{-3}, mean ± sample SD across three training seeds, averaged over 96 held-out fields at each seed. All rows use the in-distribution cutoff 16.

ν	modes	32	48	64	96	128	192	256
0.005	8	122.6 ± 11.1	131.1 ± 8.6	131.2 ± 8.4	131.2 ± 8.4	131.2 ± 8.4	131.2 ± 8.4	131.2 ± 8.4
0.005	16	9.1 ± 0.6	6.5 ± 0.2	6.5 ± 0.2	6.5 ± 0.2	6.5 ± 0.2	6.5 ± 0.2	6.5 ± 0.2
0.005	32	7.6 ± 0.5	1.4 ± 0.4	1.0 ± 0.1	0.8 ± 0.0	0.8 ± 0.0	0.9 ± 0.0	0.9 ± 0.0
0.02	8	57.1 ± 2.6	56.9 ± 2.5	56.9 ± 2.6	56.9 ± 2.6	56.9 ± 2.6	56.9 ± 2.6	56.9 ± 2.6
0.02	16	3.7 ± 0.7	3.1 ± 0.5	3.1 ± 0.5	3.1 ± 0.5	3.1 ± 0.5	3.1 ± 0.5	3.1 ± 0.5
0.02	32	2.4 ± 0.9	2.2 ± 0.9	1.1 ± 0.2	0.9 ± 0.1	0.9 ± 0.1	1.0 ± 0.1	1.0 ± 0.1
0.08	8	9.0 ± 0.2	8.8 ± 0.2	8.8 ± 0.2	8.8 ± 0.2	8.8 ± 0.2	8.8 ± 0.2	8.8 ± 0.2
0.08	16	1.3 ± 0.2	1.0 ± 0.1	0.9 ± 0.1	0.9 ± 0.1	0.9 ± 0.1	0.9 ± 0.1	0.9 ± 0.1
0.08	32	1.8 ± 0.6	1.4 ± 0.3	0.8 ± 0.0	0.7 ± 0.0	0.7 ± 0.0	0.7 ± 0.0	0.7 ± 0.0

Probe and statistical protocol. Each model receives the same 32-probe sequence as the other candidates in its cluster, using nested prefixes of 4, 8, 16, and 32. A separate matched check uses grids 64 and 128 and common cutoff $\min(k_{\max}, 31)$. Six octave bands cover $[0, 64)$. Partially represented

bands and per-field reference band energies at most 10^{-10} (Parseval-normalized) are excluded from band-relative inference. Target band errors for this extension are medians over eligible held-out fields. Pairwise comparisons require finite positive errors for both models, an absolute natural-log target-error gap of at least 0.02, and a probe fidelity gap of at least 0.02. Training-grid target cells are excluded. The global training baseline uses the last-epoch relative-norm training loss on the exact same selected cells; this is not the independently confidence-ranked, matched-coverage baseline of the original 2D experiment.

Table 21: New replication: paired selection on independent held-out fields. Macro means and percentile 95% intervals resample nine viscosity–seed clusters (5,000 resamples). Training-error concordance uses the same selected cells, not its own confidence ranking.

n_p	direction	concordance [95% CI]	coverage	training error	selected/eligible
4	reverse	0.984 [0.970,0.996]	0.399	0.882	501/1228
4	super	1.000 [1.000,1.000]	0.387	0.842	320/788
4	all	0.990 [0.981,0.998]	0.395	0.867	821/2016
8	reverse	0.981 [0.962,0.996]	0.391	0.879	491/1228
8	super	1.000 [1.000,1.000]	0.378	0.846	310/788
8	all	0.988 [0.977,0.998]	0.386	0.867	801/2016
16	reverse	0.979 [0.961,0.993]	0.383	0.877	480/1228
16	super	1.000 [1.000,1.000]	0.371	0.844	304/788
16	all	0.987 [0.976,0.996]	0.378	0.864	784/2016
32	reverse	0.979 [0.961,0.993]	0.387	0.877	486/1228
32	super	1.000 [1.000,1.000]	0.375	0.845	308/788
32	all	0.987 [0.976,0.996]	0.383	0.865	794/2016

Observed outcome. At eight probes, concordance is 0.988 (cluster-bootstrap 95% interval $[0.977, 0.998]$), compared with 0.867 for training error on the selected cells. Macro coverage is 0.386; 801 of 2,016 eligible cells are selected when pooled. The pooled fraction differs from macro coverage because clusters receive equal weight in the latter. Reverse-resolution concordance is 0.981 with coverage 0.391, and super-resolution concordance is 1.000 with coverage 0.378. The degenerate super-resolution bootstrap interval reflects no observed discordance in this finite suite; it is not a population guarantee. Additional probes do not improve this already saturated linear setting monotonically: four and 32 probes reach 0.990 and 0.987, respectively. All measured operating points are retained in Table 21.

Table 20 separates grid refinement from learned bandwidth. At weak diffusion, the eight-mode model remains near 0.131 error on every grid from 48 through 256: more target points do not supply missing learned frequencies. Stronger diffusion reduces this global error to approximately 0.009 because high-frequency output energy is attenuated. A low global error therefore need not imply uniformly accurate high-frequency response. These are controlled linear-PDE findings; the nonlinear and 2D claims continue to rest on the original separate experiments.

Computational precision and numerical checks. The models were trained on two RTX 4090 GPUs using PyTorch 2.6.0 with CUDA 12.4. Each worker used one CPU thread. Training and reference generation were batched on GPU; full float32 training was used without automatic mixed precision or TF32, while Fourier field construction used complex128. The 27 training jobs took 318.1 summed worker-seconds; training plus evaluation took 323.0 summed worker-seconds. These are aggregate job times, not a claimed speedup over a separately benchmarked serial run. The replication artifact contains the model and evaluation records.

Reference checks verify the exact solver and common physical modes. Numerical verification covers 2,000 median perturbations, 100 Fourier-multiplier cases across four grids, and 2,000 orthogonal energy decompositions. Maximum multiplier and energy-identity discrepancies are 1.12×10^{-16} and 6.67×10^{-16}, respectively. The untested-grid example has errors $(0, 0, 1)$ at grids $(128, 64, 256)$, illustrating Appendix A.