WRONG RAYS: IN-DISTRIBUTION ACCURACY DOES NOT CERTIFY TRANSPORT IN LEARNED PDE PROPAGATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural PDE models are commonly evaluated by prediction error on held-out trajectories, but downstream physical tasks may depend on dynamical responses that are weakly constrained by this metric. We ask when in-distribution prediction accuracy certifies the transport of a localized disturbance. We find that low prediction error can coexist with systematically incorrect packet trajectories, or *wrong rays*, along spectral directions weakly weighted by the standard distribution, from an exact non-identifiability construction to trained models and released PDEBench data. We develop a task-conditioned characterization of this gap: the input distribution determines which operator directions are strongly constrained, whereas transport can be sensitive to different directions. In clean transport settings, this mismatch appears as phase-slope error that accumulates with rollout horizon; on released linear advection, phase-only correction accounts for 81.7% of the improvement achieved by full-complex correction. The error is not removed by the tested representation-consistency, activation-anti-aliasing, or longer-training interventions. An information-presence test further shows that explicit spectral estimation recovers the corrective response from the same broadband observations that the tested end-to-end fine-tuning procedure fails to exploit reliably. These results distinguish predictive accuracy from certification of task-relevant dynamics.

1 INTRODUCTION

Neural operators for time-dependent PDEs are commonly evaluated by field prediction error on held-out trajectories (Li et al., 2021; Lu et al., 2021; Kovachki et al., 2023; Azizzadenesheli et al., 2024). Yet downstream physical tasks can depend on responses that receive little weight under this metric. A localized disturbance, for example, can accumulate substantial position error from a small propagation-speed mismatch. This raises our central question: *when does accurate prediction on the evaluation distribution certify the transport behavior required by a downstream task?*

Figure 1 shows this gap on released PDEBench advection (Takamoto et al., 2022): low standard test error coexists with multi-cell packet-center drift that responds strongly to targeted spectral correction. We call such systematically incorrect packet trajectories *wrong rays*. The separation also appears in an exact non-identifiability construction and trained advection and Schrödinger models, and is not specific to Fourier architectures. It is not universal, however: a periodic U-Net fails on heterogeneous advection but shows little corresponding error on low-viscosity Burgers. Our claim therefore concerns what standard accuracy can *certify*, not an inevitable failure of every learned propagator.

This problem connects spectral bias and frequency-OOD behavior (Rahaman et al., 2019; Xie & Luo, 2026; Mouli et al., 2024; Khodakarami et al., 2026b), distribution-weighted spectral error (Dong & Cheng, 2026), local spectral and Jacobian diagnostics (Gao et al., 2026), and representation consistency (Bartolucci et al., 2023; Gao et al., 2025), while classical numerical analysis relates phase and group-velocity error to wave-packet propagation (Trefethen, 1982; Tam & Webb, 1993; Vichnevetsky & Bowles, 1982; Whitham, 1974). We connect these ideas through a *task-conditioned transport certification* view: prediction risk weights operator error according to the input distribution, whereas a downstream task has its own sensitivity.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

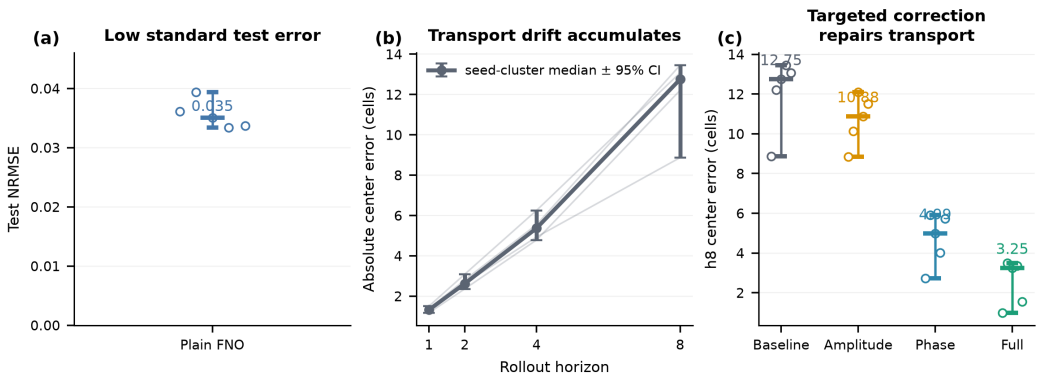


Figure 1: **Accurate in-distribution prediction does not certify packet transport.** **(a)** Five FNOs trained on released PDEBench advection achieve low test NRMSE. **(b)** Localized packets in weakly weighted spectral directions accumulate center error over rollout. **(c)** At horizon 8, phase-only correction removes substantially more drift than amplitude-only correction, while full-complex correction performs best. Intervals are 95% bootstrap intervals across seeds after within-seed aggregation.

For finite-dimensional linear propagators, we show that this mismatch determines the worst-case first-order task error and becomes unbounded when the task depends on an unobserved direction. In Fourier-diagonal transport, the result yields a coverage-dependent certificate for phase-slope error. An exact packet-center identity further shows that adjacent phase differences affect center motion at first order with sensitivity proportional to rollout horizon. Consistent with this mechanism, phase-only correction on released linear advection accounts for 81.7% of the gain achieved by full-complex correction. The remaining gain from full correction, together with tangent amplification in Burgers, limits this phase-dominant interpretation to the clean transport regime.

We further test alternative explanations. The drift persists with zero discretization mismatch, after a support-compatible consistency procedure reduces mismatch by 99.97%, under activation anti-aliasing, and while test NRMSE improves from 0.0361 to 0.0261 over a fourfold training extension. Moreover, the required information is not unique to Fourier probes: explicit Fourier least squares recovers the corrective response from the same six noiseless broadband pairs to 1.63×10^{-7} relative discrepancy, whereas validation-selected end-to-end fine-tuning retains large transport error. This separates information availability from how effectively the tested procedure extracts the task-relevant response. Our contributions are:

- We **demonstrate a transport-certification gap**: low in-distribution prediction error can coexist with large packet-transport error across exact, trained, and released-data settings, with positive and negative architecture cases delimiting its scope.
- We **characterize the gap through task-conditioned sensitivity**, linking weak spectral coverage, phase-slope error, and accumulated packet drift in the clean transport regime.
- We **separate transport identification from representation, optimization, and information availability** through consistency, anti-aliasing, longer-training, and matched-supervision controls.

2 From Prediction Error to Transport Fidelity

2.1 Learned Propagators and Standard Prediction Risk

Let $T : \mathcal{H} \to \mathcal{H}$ denote the true one-step PDE propagator and G_θ its learned approximation. For inputs $u \sim \mu$, squared prediction risk is

$$\mathcal{R}_\mu(G_\theta) = \mathbb{E}_{u\sim\mu}\|G_\theta(u) - T(u)\|_{\mathcal{H}}^2. \tag{1}$$

This is a distribution-weighted notion of operator accuracy: directions with high probability or energy under μ contribute strongly, while weakly weighted directions contribute little. A downstream

physical observable can therefore depend on a response that is only weakly constrained by low held-out field error. We ask when such prediction accuracy provides a certificate for task-relevant transport.

2.2 Packet Transport as a Task-Level Observable

We probe transport with spatially localized, narrowband oscillatory packets. On a periodic grid of N points, define the circular moment, center, and L-step center error by

$$m(u) = \sum_{n=0}^{N-1} |u_n|^2 e^{2\pi i n/N},\ x(u) = \frac{N}{2\pi} \arg m(u),\ m(u) \neq 0,\ e_{\mathrm{ctr}}(L; u_0) = d_{\mathbb{T}}\big(x(G_\theta^L u_0), x(T^L u_0)\big)\,, \tag{2}$$

where $x(u)$ is interpreted modulo the periodic domain and $d_{\mathbb{T}}$ is the shortest periodic distance. The sequences of reference and predicted centers define their respective packet trajectories, or rays; we call a systematic trajectory deviation a *wrong ray*. Packet-center error is a task-specific transport observable, not a replacement for full-field error.

2.3 The Transport Certification Question

We ask whether

$$\mathcal{R}_\mu(G_\theta) \text{ small} \quad \Longrightarrow \quad e_{\mathrm{ctr}}(L; u_0) \text{ small}. \tag{3}$$

We call this *transport certification*. Prediction risk weights operator error according to μ, whereas packet transport can depend strongly on weakly weighted directions. Failure of equation 3 therefore means that standard accuracy does not certify the transport task, not that the prediction metric is invalid.

Empirically, certification asks whether standard evaluation predicts downstream transport reliability. Formally, Proposition 1 gives a first-order certificate for finite-dimensional linear operators and identifies nullspace directions for which no finite first-order certificate exists. It is not a global guarantee for nonlinear neural operators.

This distinction separates predictive accuracy from task-relevant dynamical fidelity and motivates the analysis and interventions that follow.

3 Transport Diagnostics and Structured Interventions

To separate learned response, transport consequence, and information availability, we use a spectral transport diagnostic, component-wise response interventions, and matched-supervision estimation.

3.1 Spectral Response and Transport Diagnostics

For a linear translation-equivariant propagator, a normalized Fourier mode e_k satisfies $Te_k = g(k)e_k$, where the complex multiplier $g(k)$ determines its one-step amplitude and phase response. A nonlinear learned propagator need not be Fourier-diagonal, so we instead measure its projected centered response

$$\widehat{g}_\theta(k; a) = \left\langle e_k,\ \frac{G_\theta(ae_k) - G_\theta(-ae_k)}{2a} \right\rangle. \tag{4}$$

This equals the multiplier for a linear Fourier-diagonal map; for nonlinear models it is the fundamental projection of a finite-amplitude secant response. We use $a = 1$ on released data, so it should not be interpreted as an infinitesimal Jacobian.

For a packet centered at k_0, we use its five-mode band $\mathcal{J}(k_0) = \{k_0 - 2, \ldots, k_0 + 2\}$, unwrap $\phi_\theta(k) = \arg \widehat{g}_\theta(k; a)$, and fit

$$(\widehat{\alpha}, \widehat{\beta}) = \arg\min_{\alpha,\beta} \sum_{k \in \mathcal{J}(k_0)} \big(\phi_\theta(k) - \alpha - \beta k\big)^2. \tag{5}$$

After conversion to the rollout coordinates, the difference between learned and reference slopes defines the transport-velocity diagnostic. Main trained-model results use this fixed unweighted fit; the exact linear construction additionally uses an amplitude-weighted fit, reported in Appendix B.3.

3.2 Phase, Amplitude, and Full-Complex Interventions

To test whether the diagnosed response drives packet drift, let the estimated complex correction be $c(k) = r(k)e^{i\psi(k)}$. At every rollout step we compare

$$\widehat{u}_\theta^{\mathrm{amp}}(k) = r(k)\widehat{u}_\theta^{\mathrm{base}}(k),\ \widehat{u}_\theta^{\mathrm{phase}}(k) = e^{i\psi(k)}\widehat{u}_\theta^{\mathrm{base}}(k),\ \widehat{u}_\theta^{\mathrm{full}}(k) = r(k)e^{i\psi(k)}\widehat{u}_\theta^{\mathrm{base}}(k). \quad (6)$$

The corrected prediction becomes the input to the next step. Thus amplitude-only and phase-only corrections isolate which component contributes to transport, while the full correction measures their combined effect. This decomposition has its cleanest interpretation for constant-speed advection; in dispersive or nonlinear systems, amplitude reweighting and mode coupling can also move the packet center.

3.3 Matched-Supervision Estimation

A successful structured correction may reflect either additional information or a different way of extracting information already present in ordinary fields. We separate these possibilities using the same $M = 6$ noiseless broadband pairs $\{(u^{(m)}, v^{(m)})\}_{m=1}^{M}$, with $v^{(m)} = Tu^{(m)}$, in two procedures. The first performs end-to-end fine-tuning, with strength and stopping selected only by standard validation performance. The second estimates each modal transfer directly,

$$\widetilde{g}(k) = \arg\min_{z\in\mathbb{C}} \sum_{m=1}^{M} \left|\widehat{v}^{(m)}(k) - z\,\widehat{u}^{(m)}(k)\right|^2, \quad (7)$$

and constructs the corresponding spectral correction. We also compare structured pure-mode supervision at the same label budget. The key comparison is therefore between two estimators receiving identical broadband observations: if direct spectral estimation succeeds while end-to-end fine-tuning does not, the difference cannot be attributed to missing information in those labels alone.

4 Standard Accuracy Does Not Certify Transport

We test whether low standard prediction error can coexist with inaccurate transport in weakly constrained directions, from exact constructions to trained and released PDE settings.

Accurate In-Distribution Prediction Can Coexist with Wrong Transport. On released PDEBench advection, five FNOs achieve test NRMSE $0.0351\,[0.0334, 0.0393]$, yet their eight-step packet-center error reaches $12.754\,[8.868, 13.449]$ grid cells (Fig. 1). The error accumulates over rollout and is strongly reduced by targeted spectral correction. Thus, standard in-distribution accuracy does not certify transport in directions weakly weighted by that distribution; we do not require the packets themselves to have low full-field error.

From Exact Non-Identifiability to Learned Models. In the exact construction, linear periodic propagators agree on every observed Fourier mode while differing elsewhere. Validation NRMSE is below 1.5×10^{-14}, yet median eight-step center errors are 19.31 cells for advection and 18.46 for Schrödinger dynamics (Fig. 2a). This establishes non-identifiability: agreement on the observed subspace does not determine off-support transport. The gap also appears after learning. Across trained synthetic convolutional and Fourier-operator models, NRMSE ranges from 0.0012 to 0.0074, while eight-step center error ranges from 9.42 to 18.73 cells (Fig. 2b).

The Gap Persists on Released PDE Data. On released PDEBench advection, six structured spectral observations reduce eight-step center error by 80.7%, improving all five seeds (Fig. 2c). This shows that the transport discrepancy is substantially correctable by information targeted at the weakly constrained response. We do not treat six probes as optimal; budget dependence is reported in Appendix E.1. The result also persists with positive spectral tails, across a broader frequency sweep, and under the naturally non-band-limited PDEBench generator (Appendix C.1–C.3). Raw spectral energy is not a monotone predictor of ray error (Appendix D.6).

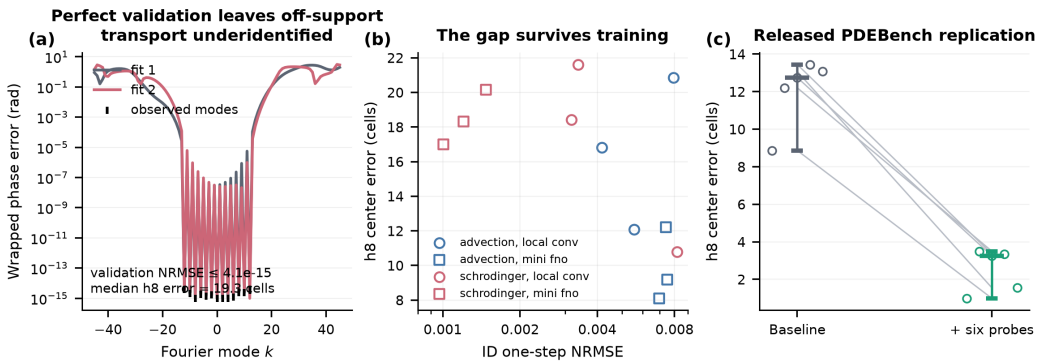


Figure 2: **From non-identifiability to learned wrong rays.** **(a)** Exact propagators match the observed Fourier support yet differ in unseen-band transport. **(b)** Trained advection and Schrödinger models retain large transport error despite low in-distribution NRMSE. **(c)** The same gap appears on released PDEBench advection and is reduced by task-aligned spectral information.

Scope Across Systems and Architectures. The gap is neither FNO-specific nor universal: both FNOs and periodic U-Nets fail on heterogeneous advection, while on low-viscosity Burgers the FNO has 2.90-cell four-step error versus 0.30 for the U-Net. Thus transport error depends on the interaction among dynamics, data, and architecture; see Appendix Table 10 for the full comparison.

5 Task-Conditioned Transport Certification

Low prediction error can coexist with large transport error because prediction risk weights operator error by the input distribution, while a downstream task can depend on different directions. We formalize this mismatch for finite-dimensional linear propagators, specialize it to phase-slope transport, and then connect phase slope to packet-center motion. The nonlinear measurements in Sec. 5.5 are finite-amplitude projected responses and are therefore compared to the theory qualitatively.

5.1 Prediction Risk Induces a Second-Moment Geometry

Let $\mathcal{H} = \mathbb{C}^N$, viewed as a real Hilbert space with state-space inner product $\langle x, y\rangle_{\mathbb{R}} = \mathrm{Re}(x^* y)$. Operator perturbations in $\mathbb{C}^{N\times N}$ use the real Frobenius inner product $\langle X, Y\rangle_{F,\mathbb{R}} = \mathrm{Re}\,\mathrm{tr}(X^* Y)$. For complex-linear propagators T, G, write $A = G - T$ and $M_\mu = \mathbb{E}_{u\sim\mu}[uu^*]$, with $\mathrm{rank}(M_\mu) > 0$.

Lemma 1 (Second-moment geometry of prediction risk)**.**

$$\mathcal{R}_\mu(T + A) = \mathbb{E}\|Au\|_2^2 = \mathrm{tr}(AM_\mu A^*) = \|AM_\mu^{1/2}\|_F^2. \tag{8}$$

Thus prediction risk constrains A in the geometry induced by M_μ: low-second-moment directions are weakly penalized and $\ker M_\mu$ is invisible. Experimental NRMSE is the empirical field metric used in our comparisons; because its normalization need not coincide with the population risk in equation 1, we do not identify the two quantities exactly.

5.2 Task-Conditioned Tangent Sensitivity

Let $F : \mathbb{C}^{N\times N} \to \mathbb{R}$ be differentiable at T, with $DF_T[A] = \mathrm{Re}\,\mathrm{tr}(B_F^* A)$, and let P_0 project onto $\ker M_\mu$. Define the task-conditioned tangent sensitivity by

$$\kappa_F(\mu) = \sup_{\|AM_\mu^{1/2}\|_F > 0} \frac{|DF_T[A]|}{\|AM_\mu^{1/2}\|_F},$$

with the convention $\kappa_F(\mu) = +\infty$ whenever a zero-risk perturbation A satisfies $DF_T[A] \neq 0$.

Proposition 1 (Task-conditioned tangent certification)**.**

$$\kappa_F(\mu) = \begin{cases} +\infty, & B_F P_0 \neq 0, \\ \|B_F M_\mu^{\dagger/2}\|_F, & B_F P_0 = 0, \end{cases} \qquad \sup_{\|AM_\mu^{1/2}\|_F \le \varepsilon} |DF_T[A]| = \varepsilon\, \kappa_F(\mu). \tag{9}$$

If the task is sensitive to a direction absent from the input distribution, no finite first-order prediction-risk certificate exists; otherwise the amplification is determined by the task derivative in the inverse second-moment geometry. The result concerns the derivative $DF_T[A]$, not a global bound on $F(T+A)-F(T)$. Proof and a supported local nonlinear bound appear in Appendix A.2 and Appendix A.6.

5.3 A Spectral Transport Certificate

Now assume scalar Fourier-diagonal propagators $T = \mathcal{F}^* \operatorname{diag}(g_k)\mathcal{F}$ and $G = \mathcal{F}^* \operatorname{diag}(\widehat{g}_k)\mathcal{F}$. With $p_k = \mathbb{E}|(\mathcal{F}u)_k|^2$,

$$\mathcal{R}_\mu(G) = \sum_k p_k |\widehat{g}_k - g_k|^2. \tag{10}$$

For packet band J, define the ratio-phase error $\delta\phi_j = \operatorname{Arg}(\widehat{g}_j \overline{g_j})$ on a common phase branch, assume $|\delta\phi_j| \le \eta < \pi$ and $|g_j|, |\widehat{g}_j| \ge a_{\min} > 0$, and write the fixed phase-slope error as $\Delta s = -q^\top \delta\phi$. Let

$$\chi(q,p) = \begin{cases} \|D_p^{\dagger/2} q\|_2, & q_j = 0 \text{ whenever } p_j = 0, \\ +\infty, & \text{otherwise,} \end{cases} \qquad D_p = \operatorname{diag}(p_j)_{j\in J}.$$

Corollary 1 (Spectral transport certificate)**.**

$$|\Delta s| \le \frac{\eta}{2a_{\min}\sin(\eta/2)}\, \chi(q,p)\, \sqrt{\mathcal{R}_\mu(G)}. \tag{11}$$

The certificate separates prediction error from task conditioning. Zero coverage of a mode with nonzero transport weight makes $\chi = +\infty$; weak positive coverage makes the certificate finite but poorly conditioned. The proof is in Appendix A.3. If the induced velocity bias remains approximately persistent over rollout,

$$\Delta x_L = L\Delta s + r_L, \tag{12}$$

so transport error accumulates linearly up to a remainder r_L from deformation, dispersion, changing response, or periodic wrap. This persistence assumption is additional to the one-step certificate.

5.4 Packet-Center Sensitivity to Phase Differences

The packet-center observable makes the role of phase more specific.

Lemma 2 (Circular-moment identity)**.** *With $u_n = N^{-1}\sum_k c_k e^{2\pi i k n/N}$, a scalar Fourier-diagonal propagator satisfies*

$$m_L = \frac{1}{N}\sum_k c_k \overline{c_{k+1}} \big(g_k \overline{g_{k+1}}\big)^L, \qquad x_L = \frac{N}{2\pi}\arg m_L. \tag{13}$$

Hence packet position depends on *adjacent* spectral phase differences; a common phase offset cancels.

Near aligned constant-speed advection, suppose $c_k\overline{c_{k+1}}(g_k\overline{g_{k+1}})^L = w_k e^{i\theta}$, with $w_k \ge 0$, $W = \sum_k w_k > 0$, and perturb $g_k(t) = g_k \exp(t\dot{a}_k + it\dot{\phi}_k)$.

Corollary 2 (Packet-center sensitivity near constant advection)**.**

$$\left.\frac{dx_L(t)}{dt}\right|_{t=0} = \frac{NL}{2\pi W}\sum_k w_k(\dot{\phi}_k - \dot{\phi}_{k+1}). \tag{14}$$

Thus adjacent phase differences affect center position at first order with sensitivity proportional to L, whereas a pure amplitude perturbation $\dot{\phi}_k = 0$ has zero first-order effect. This statement is local to the aligned constant-advection regime; amplitude reweighting, dispersion, mode coupling, and packet deformation may matter outside it. Full derivations are in Appendix A.4 and Appendix A.5.

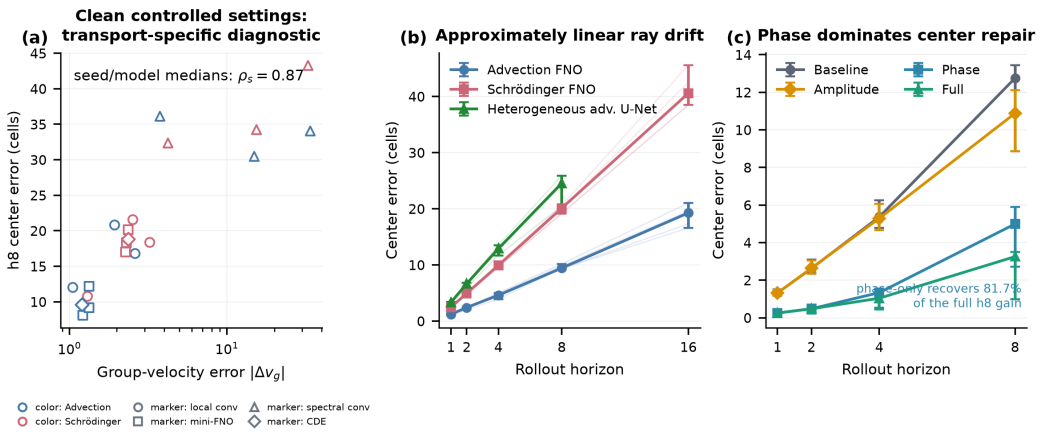


Figure 3: **Empirical signatures of persistent transport mismatch.** **(a)** Across 24 clean system–model–seed aggregates, the phase-slope velocity diagnostic correlates with eight-step center error ($\rho_s = 0.872$). **(b)** Center error grows approximately linearly with rollout horizon. **(c)** On released advection, phase-only correction removes substantially more drift than amplitude-only correction and accounts for 81.7% of the full-complex gain.

5.5 Empirical Signatures of the Transport Mechanism

Figure 3 provides three complementary signatures. The phase-slope diagnostic correlates with center error across clean benchmarks ($\rho_s = 0.872$), while approximately linear horizon growth is consistent with a persistent transport bias but is not phase-specific.

The component intervention is more diagnostic. At $h = 8$, baseline, amplitude-only, phase-only, and full-complex center errors are 12.754, 10.882, 4.992, and 3.254 cells, respectively. Phase-only correction explains 81.7% of the full-complex gain and beats amplitude-only correction in all five seeds. The remaining gap supports a phase-dominant rather than phase-exclusive mechanism. For nonlinear FNOs, these are finite-amplitude projected responses and therefore provide qualitative consistency with the linear theory rather than a direct test of its bounds.

6 Testing Alternative Explanations

We test whether wrong transport is explained by representation mismatch, activation aliasing, or incomplete optimization of the standard objective (Fig. 4).

Representation Consistency Does Not Imply Transport Fidelity. Representation consistency and transport fidelity are distinct. An analytic operator with exactly zero discretization mismatch error (DME) still has a median packet-center error of 12.34 cells (Fig. 4a). Likewise, a support-compatible 256-point procedure reduces DME by 99.97%, from 6.98×10^{-4} to 2.20×10^{-7}, while wrong rays remain (Appendix D.3). Thus discretization consistency alone does not certify transport.

Activation Anti-Aliasing Does Not Remove the Drift. A matched FNO with $2\times$ activation anti-aliasing has nearly unchanged test NRMSE and transport error: the anti-aliased/plain $h = 8$ center-error ratio is 1.006 (Fig. 4b). The tested activation-aliasing mechanism does not explain the drift.

Improved Standard Accuracy Does Not Necessarily Improve Transport. Across three seeds, test NRMSE improves from $0.0361 \to 0.0300 \to 0.0261$ at $400, 800$, and 1600 steps, while $h = 8$ center error changes from $12.205 \to 14.075 \to 13.836$ cells (Fig. 4c,d). Transport therefore does not improve in lockstep with the standard objective over this range. Because validation is still improving at 1600 steps, we do not claim asymptotic saturation. Together with the same-support data-scaling control (Appendix D.1), these results motivate separating missing task-relevant information from failure to extract information already present.

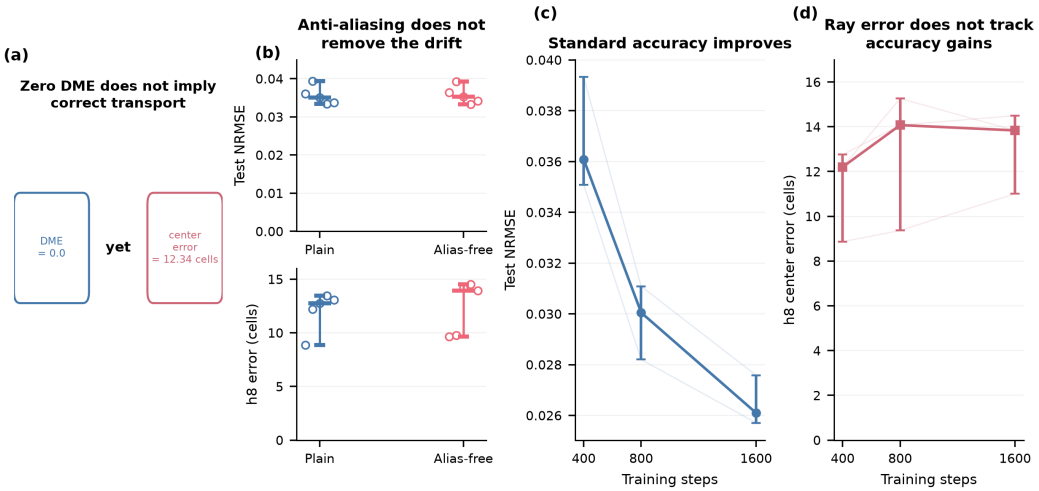


Figure 4: **Improving other error axes does not necessarily repair transport.** **(a)** Zero discretization mismatch can coexist with large center error. **(b)** Activation anti-aliasing leaves transport essentially unchanged. **(c,d)** Longer training improves test NRMSE but not eight-step transport.

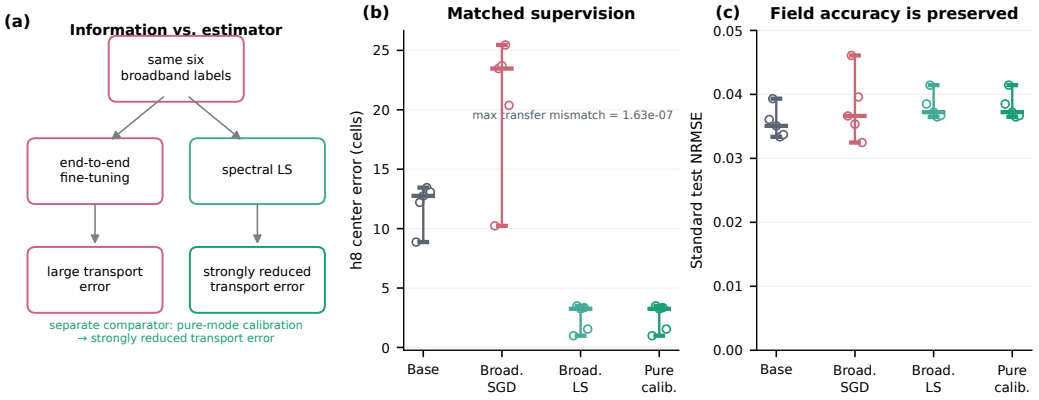


Figure 5: **Broadband observations contain the transport-relevant information.** **(a)** The same six broadband pairs are used for end-to-end fine-tuning and explicit spectral estimation. **(b)** Spectral estimation matches structured correction, whereas fine-tuning retains large $h = 8$ center error. **(c)** Fine-tuning preserves standard test accuracy. This is an information-presence test in linear advection, not an estimator leaderboard.

7 Structured Estimation Reveals Task-Relevant Information

Targeted correction improves transport, but does it add uniquely informative supervision or extract information already present in ordinary fields? We distinguish these possibilities by holding the supervision fixed. Explicit Fourier least squares on the identical observations instead recovers the target modal response to 1.63×10^{-7} relative discrepancy and reduces center error to $3.254\,[0.988, 3.499]$, matching structured six-mode correction (Fig. 5). Thus Fourier probes are not informationally unique: the required response is already present in the broadband data, but the tested end-to-end procedure does not reliably express it. This establishes an information–estimation distinction without identifying why fine-tuning fails or implying general superiority of Fourier least squares. The correction also minimally changes ordinary $h = 8$ rollout NRMSE (0.2320 baseline, 0.2323 phase-only, 0.2350 full-complex); see Appendix E.3.

8 Related Work

Operator learning and learned simulation. Neural operators, including FNOs and DeepONets (Li et al., 2021; Lu et al., 2021; Kovachki et al., 2023), admit distribution-dependent approximation and generalization analyses (Kovachki et al., 2021; Lanthaler et al., 2022) and diverse graph, transformer, convolutional, geometric, and spectral parameterizations (Li et al., 2020; Cao, 2021; Hao et al., 2023; Raonic et al., 2023; Li et al., 2023; Tran et al., 2023). Related learned simulators and physics-informed methods target physical prediction and PDE-constrained learning (Sanchez-Gonzalez et al., 2020; Pfaff et al., 2021; Brandstetter et al., 2022; Gupta & Brandstetter, 2023; Raissi et al., 2019; Wang et al., 2021; Li et al., 2024). We instead ask what standard prediction accuracy certifies about transport.

Spectral bias and distribution shift. Networks learn frequencies unevenly (Rahaman et al., 2019; Xu et al., 2020), motivating Fourier and periodic representations (Tancik et al., 2020; Sitzmann et al., 2020). Neural-operator work studies spectral bias, frequency-OOD behavior, and uncertainty (Khodakarami et al., 2026a;b; Xie & Luo, 2026; Mouli et al., 2024; Lara Benitez et al., 2024), while spectral-error and long-rollout analyses study distribution weighting and stability (Dong & Cheng, 2026; Lippe et al., 2023; Brandstetter et al., 2022; Li et al., 2022). Our focus is whether low distribution-weighted error certifies packet transport.

Spectral diagnostics and dynamics. Spectral and Jacobian analyses expose gain, phase, coupling, and tangent errors (Yang et al., 2023; Gao et al., 2026); related spectral descriptions arise in DMD, Koopman, and generator-based models (Schmid, 2010; Rowley et al., 2009; Williams et al., 2015; Takeishi et al., 2017; Lusch et al., 2018; Jamil & Kapadia, 2026). We connect such operator-level errors to task-conditioned identifiability and a ray-level physical consequence.

Representation consistency and dispersion. Representation-equivalence methods study discretization consistency (Bartolucci et al., 2023; Gao et al., 2025), whereas classical analysis links phase and group-velocity errors to wave-packet propagation (Trefethen, 1982; Tam & Webb, 1993; Vichnevetsky & Bowles, 1982; Whitham, 1974; LeVeque, 2007; Strikwerda, 2004). Our framework separates representation consistency from certification of the transport response.

Structured spectral adaptation. Spectral and phase-aware adaptation targets difficult frequency regimes (Cao et al., 2025; Khodakarami et al., 2026a; Sun et al., 2026). Our corrections instead serve as mechanism interventions, and matched supervision shows that broadband observations can contain the same corrective information under explicit modal estimation. The contribution is therefore task-conditioned certification and an information–estimation separation, rather than a generic spectral correction method.

9 Discussion and Conclusion

Our results distinguish in-distribution accuracy from certification of a task-relevant response: distribution-weighted field error need not constrain the directions governing packet transport. The phase-slope mechanism is narrower, applying most cleanly to linear transport; nonlinear FNO responses are finite-amplitude measurements, and Burgers introduces substantial amplitude and mode-coupling effects. The phenomenon is also architecture- and PDE-dependent, while our spectral corrections are diagnostic rather than optimal. Probe benefit is non-monotone, longer training is not proven saturated, and matched broadband supervision isolates an information–estimation gap without identifying its cause.

Overall, accurate prediction on the evaluation distribution need not certify the task-relevant dynamics of a learned PDE propagator. In clean linear settings, mismatch between prediction-risk geometry and downstream sensitivity appears as weakly constrained phase-slope error that accumulates into wrong rays. Matched supervision further shows that the required transport information can already be present in broadband observations even when end-to-end fine-tuning does not reliably recover it.

Reliability is therefore task-conditioned: it depends on how the data distribution constrains the response used by the physical observable. Extending this view to nonlinear and heterogeneous dynamics with coupled phase, amplitude, and state-dependent responses is a natural next step.

AI USE STATEMENT

We used generative AI tools to assist with literature retrieval and reference organization, manuscript editing, organization of experimental results, and research and mathematical reasoning. All AI-assisted text, citations, code, results, and proofs were reviewed and verified by the authors, who made all final methodological and scientific decisions. We take responsibility for the complete content of this work.

ETHICS STATEMENT

This work studies statistical properties of language-model responses and does not involve human participants or the collection of personal data. The experiments use existing benchmark tasks and pretrained language models. The certificates studied here characterize response distributions under registered representation groups; they should not be interpreted as guarantees of factual correctness, fairness, or safety.

REPRODUCIBILITY STATEMENT

We provide an anonymized supplementary archive containing the implementation, experiment configurations, definitions, random seeds, result artifacts used in the reported analyses, and scripts for reproducing the tables and figures. The supplement also includes environment information and integrity manifests for the experimental artifacts. Experimental settings and evaluation metrics are specified in the main paper and Appendix, and complete proofs of the theoretical results are provided in the Appendix.

REFERENCES

Kamyar Azizzadenesheli, Nikola Kovachki, Zongyi Li, Miguel Liu-Schiaffini, Jean Kossaifi, and Anima Anandkumar. Neural operators for accelerating scientific simulations and design. *Nature Reviews Physics*, 6(5):320–328, 2024. doi: 10.1038/s42254-024-00712-5.

Francesca Bartolucci, Emmanuel de Bézenac, Bogdan Raonic, 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, 2023.

Johannes Brandstetter, Daniel E. Worrall, and Max Welling. Message passing neural PDE solvers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=vSix3HPYKSU.

Shuhao Cao. Choose a transformer: Fourier or galerkin. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24924–24940, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/d0921d442ee91b896ad95059d13df618-Abstract.html.

Shuhao Cao, Francesco Brarda, Rui Peng Li, and Yuanzhe Xi. Spectral-refiner: Accurate fine-tuning of spatiotemporal fourier neural operator for turbulent flows. In *International Conference on Learning Representations*, 2025.

Hang-Cheng Dong and Pengcheng Cheng. Predicting out-of-distribution generalization of neural operators via observable spectral error decomposition, 2026.

Wenhan Gao, Ruichen Xu, Yuefan Deng, and Yi Liu. Discretization-invariance? on the discretization mismatch errors in neural operators. In *International Conference on Learning Representations*, 2025.

Zhiwei Gao, Liu Yang, and George Em Karniadakis. Spectral audit of in-context operator networks, 2026.

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.

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. GNOT: A general neural operator transformer for operator learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12556–12569. PMLR, 2023. URL https://proceedings.mlr.press/v202/hao23c.html.

Shafayeth Jamil and Rehan Kapadia. Lie generator networks for nonlinear partial differential equations, 2026.

Siavash Khodakarami, Vivek Oommen, Aniruddha Bora, and George Em Karniadakis. Mitigating spectral bias in neural operators via high-frequency scaling for physical systems. *Neural Networks*, 193:108027, 2026a. doi: 10.1016/j.neunet.2025.108027.

Siavash Khodakarami, Vivek Oommen, Nazanin Ahmadi Daryakenari, Maxim Beekenkamp, and George Em Karniadakis. Spectral bias in physics-informed and operator learning: Analysis and mitigation guidelines. *Computer Methods in Applied Mechanics and Engineering*, 461:119156, 2026b. doi: 10.1016/j.cma.2026.119156.

Nikola Kovachki, Samuel Lanthaler, and Siddhartha Mishra. On universal approximation and error bounds for fourier neural operators. *Journal of Machine Learning Research*, 22(290):1–76, 2021. URL https://jmlr.org/papers/v22/21-0806.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.

Samuel Lanthaler, Siddhartha Mishra, and George Em Karniadakis. Error estimates for DeepONets: a deep learning framework in infinite dimensions. *Transactions of Mathematics and Its Applications*, 6(1):tnac001, 2022. doi: 10.1093/imatrm/tnac001.

Jose Antonio Lara Benitez, Takashi Furuya, Florian Faucher, Anastasis Kratsios, Xavier Tricoche, and Maarten V. de Hoop. Out-of-distributional risk bounds for neural operators with applications to the Helmholtz equation. *Journal of Computational Physics*, 513:113168, 2024. doi: 10.1016/j.jcp.2024.113168.

Randall J. LeVeque. *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems*. Society for Industrial and Applied Mathematics, 2007. ISBN 978-0-89871-629-0. doi: 10.1137/1.9780898717839.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, 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.

Zongyi Li, Miguel Liu-Schiaffini, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Learning chaotic dynamics in dissipative systems. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16768–16781, 2022.

Zongyi Li, Daniel Zhengyu Huang, Burigede Liu, and Anima Anandkumar. Fourier neural operator with learned deformations for PDEs on general geometries. *Journal of Machine Learning Research*, 24(388):1–26, 2023. URL https://jmlr.org/papers/v24/23-0064.html.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM/IMS Journal of Data Science*, 1(3):1–27, 2024. doi: 10.1145/3648506.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. PDE-Refiner: Achieving accurate long rollouts with neural PDE solvers. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67398–67433, 2023.

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.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9:4950, 2018. doi: 10.1038/s41467-018-07210-0.

S Chandra Mouli, Danielle C. Maddix, Shima Alizadeh, Gaurav Gupta, Andrew Stuart, Michael W. Mahoney, and Bernie Wang. Using uncertainty quantification to characterize and improve out-of-domain learning for PDEs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36372–36418. PMLR, 2024. URL https://proceedings.mlr.press/v235/mouli24a.html.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=roNqYL0_XP.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred 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.

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

Bogdan Raonic, 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, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f3c1951b34f7f55ffaecada7fde6bd5a-Abstract-Conference.html.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Clarence W. Rowley, Igor Mezić, Shervin Bagheri, Philipp Schlatter, and Dan S. Henningson. Spectral analysis of nonlinear flows. *Journal of Fluid Mechanics*, 641:115–127, 2009. doi: 10.1017/S0022112009992059.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8459–8468. PMLR, 2020. URL https://proceedings.mlr.press/v119/sanchez-gonzalez20a.html.

Peter J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010. doi: 10.1017/S0022112010001217.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

John C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations*. Society for Industrial and Applied Mathematics, 2 edition, 2004. ISBN 978-0-89871-567-5.

Yifan Sun, Lei Cheng, Sijie Chen, Ting Zhang, Jianlong Li, and Shikai Fang. APEX: Amplitude anchors and phase priors for target-scarce higher-frequency wave prediction, 2026.

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, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0a9747136d411fb83f0cf81820d44afb-Abstract-Datasets_and_Benchmarks.html.

Naoya Takeishi, Yoshinobu Kawahara, and Takehisa Yairi. Learning koopman invariant subspaces for dynamic mode decomposition. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Christopher K. W. Tam and Jay C. Webb. Dispersion-relation-preserving finite difference schemes for computational acoustics. *Journal of Computational Physics*, 107(2):262–281, 1993. doi: 10.1006/jcph.1993.1142.

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, 2020.

Alasdair Tran, Alexander Mathews, Lexing Xie, and Cheng Soon Ong. Factorized fourier neural operators. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tmIiMPl4IPa.

Lloyd N. Trefethen. Group velocity in finite difference schemes. *SIAM Review*, 24(2):113–136, 1982. doi: 10.1137/1024038.

Robert Vichnevetsky and John B. Bowles. *Fourier Analysis of Numerical Approximations of Hyperbolic Equations*, volume 5 of *Studies in Applied and Numerical Mathematics*. Society for Industrial and Applied Mathematics, 1982. ISBN 978-0-89871-181-3.

Sifan Wang, Hanwen Wang, and Paris Perdikaris. Learning the solution operator of parametric partial differential equations with physics-informed DeepONets. *Science Advances*, 7(40):eabi8605, 2021. doi: 10.1126/sciadv.abi8605.

G. B. Whitham. *Linear and Nonlinear Waves*. Wiley-Interscience, New York, 1974. ISBN 0-471-94090-9.

Matthew O. Williams, Ioannis G. Kevrekidis, and Clarence W. Rowley. A data-driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015. doi: 10.1007/s00332-015-9258-5.

Runlong Xie and An Luo. Frequency bias and OOD generalization in neural operators under a variable-coefficient wave equation, 2026.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, 2020. doi: 10.4208/cicp.OA-2020-0085.

Liu Yang, Siting Liu, Tingwei Meng, and Stanley J. Osher. In-context operator learning with data prompts for differential equation problems. *Proceedings of the National Academy of Sciences*, 120(39):e2310142120, 2023. doi: 10.1073/pnas.2310142120.

A Proofs and Theoretical Details

This appendix provides the proofs and technical conditions for the results in Sec. 5. Throughout, finite-dimensional complex vector and matrix spaces are regarded as real Hilbert spaces. For matrices X, Y, we use the real Frobenius inner product

$$\langle X, Y \rangle_{\mathbb{R}} = \mathrm{Re}\,\mathrm{tr}(X^* Y).$$

All square roots and pseudoinverses of positive-semidefinite matrices are the canonical Hermitian ones.

A.1 Proof of Lemma 1

We first prove the second-moment representation of population prediction risk.

Proof of Lemma 1. Let

$$A = G - T, \qquad M_\mu = \mathbb{E}_{u\sim\mu}[uu^*].$$

Then

$$\begin{aligned}\mathcal{R}_\mu(T+A) &= \mathbb{E}_{u\sim\mu}\|Au\|_2^2 \\ &= \mathbb{E}_{u\sim\mu}\left[u^*A^*Au\right] \\ &= \mathbb{E}_{u\sim\mu}\,\mathrm{tr}\left(Auu^*A^*\right) \\ &= \mathrm{tr}\left(A\,\mathbb{E}[uu^*]\,A^*\right) \\ &= \mathrm{tr}\left(AM_\mu A^*\right). \end{aligned} \tag{15}$$

Since $M_\mu \succeq 0$,

$$AM_\mu A^* = (AM_\mu^{1/2})(AM_\mu^{1/2})^*.$$

Therefore

$$\mathcal{R}_\mu(T+A) = \mathrm{tr}\left[(AM_\mu^{1/2})(AM_\mu^{1/2})^*\right] = \|AM_\mu^{1/2}\|_F^2,$$

which proves Eq. equation 8. □

The result uses the uncentered second moment $M_\mu = \mathbb{E}[uu^*]$, not the centered covariance. No zero-mean assumption on the input distribution is required.

A.2 Proof of Proposition 1

We now prove the task-conditioned tangent characterization. Let P_0 denote the orthogonal projector onto $\ker M_\mu$, and let

$$P = I - P_0$$

be the projector onto $\mathrm{ran}\, M_\mu$.

Recall that

$$DF_T[A] = \mathrm{Re}\,\mathrm{tr}(B_F^*A).$$

Proof of Proposition 1. We consider separately whether the task derivative is sensitive to the training-null subspace.

Case 1: $B_FP_0 \neq 0$**.** Define

$$A_0 = B_FP_0.$$

Because

$$P_0M_\mu^{1/2} = 0,$$

we have

$$A_0M_\mu^{1/2} = 0.$$

Hence A_0 incurs zero prediction risk. Its task derivative is nevertheless

$$\begin{aligned}DF_T[A_0] &= \mathrm{Re}\,\mathrm{tr}\left(B_F^*B_FP_0\right) \\ &= \mathrm{tr}\left(P_0B_F^*B_FP_0\right) \\ &= \|B_FP_0\|_F^2 > 0. \end{aligned} \tag{16}$$

For any scalar $\alpha > 0$,

$$\|\alpha A_0M_\mu^{1/2}\|_F = 0,$$

whereas

$$|DF_T[\alpha A_0]| = \alpha\|B_FP_0\|_F^2.$$

The task derivative can therefore be made arbitrarily large without increasing prediction risk. Thus

$$\kappa_F(\mu) = +\infty,$$

and, for every $\varepsilon > 0$,

$$\sup_{\|AM_\mu^{1/2}\|_F \le \varepsilon} |DF_T[A]| = +\infty.$$

Case 2: $B_F P_0 = 0$**.** In this case

$$B_F = B_F P.$$

Let

$$C = AM_\mu^{1/2}.$$

Since

$$M_\mu^{1/2} M_\mu^{\dagger/2} = P,$$

we have

$$CM_\mu^{\dagger/2} = AP. \tag{17}$$

The equality is only on the supported input subspace; in general, $A \neq CM_\mu^{\dagger/2}$.

Because $B_F P_0 = 0$,

$$\begin{aligned} DF_T[A] &= \operatorname{Re}\operatorname{tr}(B_F^* A) \\ &= \operatorname{Re}\operatorname{tr}(B_F^* AP) \\ &= \operatorname{Re}\operatorname{tr}\left(B_F^* CM_\mu^{\dagger/2}\right) \\ &= \operatorname{Re}\operatorname{tr}\left[(B_F M_\mu^{\dagger/2})^* (AM_\mu^{1/2})\right]. \end{aligned} \tag{18}$$

Frobenius Cauchy–Schwarz therefore gives

$$|DF_T[A]| \le \|B_F M_\mu^{\dagger/2}\|_F \|AM_\mu^{1/2}\|_F. \tag{19}$$

Hence

$$\kappa_F(\mu) \le \|B_F M_\mu^{\dagger/2}\|_F.$$

It remains to show that the bound is attained. Define

$$D = B_F M_\mu^{\dagger/2}.$$

If $D = 0$, Eq. equation 19 implies $DF_T[A] = 0$ for every admissible A, and the result is immediate.

Otherwise, for any $\varepsilon > 0$, choose

$$A_\star = \frac{\varepsilon}{\|D\|_F} DM_\mu^{\dagger/2}. \tag{20}$$

Since $M_\mu^{\dagger/2} M_\mu^{1/2} = P$ and $DP = D$,

$$A_\star M_\mu^{1/2} = \frac{\varepsilon}{\|D\|_F} D,$$

so

$$\|A_\star M_\mu^{1/2}\|_F = \varepsilon.$$

Using Eq. equation 18,

$$\begin{aligned} |DF_T[A_\star]| &= \frac{\varepsilon}{\|D\|_F} \operatorname{Re}\operatorname{tr}(D^* D) \\ &= \varepsilon \|D\|_F. \end{aligned} \tag{21}$$

Thus the Cauchy–Schwarz upper bound is attained, and

$$\sup_{\|AM_\mu^{1/2}\|_F \le \varepsilon} |DF_T[A]| = \varepsilon \|B_F M_\mu^{\dagger/2}\|_F.$$

Combining the two cases proves Eq. equation 9. □

Proposition 1 is an exact statement about the tangent functional DF_T. The prediction-risk constraint does not generally imply that $\|A\|_F$ is small when M_μ is poorly conditioned, so the proposition should not be interpreted as a global bound on the nonlinear quantity $F(T + A) - F(T)$. A local bound on the latter is given in Sec. A.6.

A.3 Proof of the Spectral Transport Certificate

We next specialize the risk geometry to scalar Fourier-diagonal transport.

Let $\mathcal{F}$ denote a unitary discrete Fourier transform and write

$$\widehat{u} = \mathcal{F}u.$$

Suppose

$$T = \mathcal{F}^* \operatorname{diag}(g_k)\mathcal{F}, \qquad G = \mathcal{F}^* \operatorname{diag}(\widehat{g}_k)\mathcal{F}.$$

Define

$$\delta g_k = \widehat{g}_k - g_k, \qquad p_k = \mathbb{E}|\widehat{u}_k|^2.$$

Prediction risk in Fourier coordinates. By unitarity of $\mathcal{F}$,

$$\begin{aligned}\mathcal{R}_\mu(G) &= \mathbb{E}\|(G-T)u\|_2^2 \\ &= \mathbb{E}\,\|\operatorname{diag}(\delta g_k)\widehat{u}\|_2^2 \\ &= \sum_k |\delta g_k|^2 \mathbb{E}|\widehat{u}_k|^2 \\ &= \sum_k p_k |\widehat{g}_k - g_k|^2. \end{aligned} \tag{22}$$

No assumption that the full Fourier covariance of u is diagonal is needed: scalar Fourier diagonality of $G - T$ makes the output norm a sum of modal contributions.

For a packet band J, define

$$R_{\mathrm{sym},J} = \sum_{j\in J} p_j |\widehat{g}_j - g_j|^2.$$

Since every term in Eq. equation 22 is nonnegative,

$$R_{\mathrm{sym},J} \le \mathcal{R}_\mu(G). \tag{23}$$

Coverage-weighted phase error. For $j \in J$, define the ratio phase

$$\delta\phi_j = \operatorname{Arg}(\widehat{g}_j \overline{g_j})$$

using a common local lift satisfying

$$|\delta\phi_j| \le \eta, \qquad 0 < \eta < \pi.$$

Assume

$$|g_j|,\ |\widehat{g}_j| \ge a_{\min} > 0.$$

Define

$$D_p = \operatorname{diag}(p_j)_{j\in J}, \qquad R_\phi = \delta\phi^\top D_p \delta\phi.$$

Recall the coverage factor

$$\chi(q,p) = \begin{cases} \|D_p^{\dagger/2} q\|_2, & q_j = 0 \text{ whenever } p_j = 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

We first show

$$|q^\top \delta\phi| \le \chi(q,p)\sqrt{R_\phi}. \tag{24}$$

If there exists $j \in J$ such that $p_j = 0$ and $q_j \neq 0$, then a nonzero phase perturbation supported only on that coordinate has $R_\phi = 0$ but $q^\top\delta\phi \neq 0$. Hence no finite constant can control the diagnostic by R_ϕ, exactly as encoded by $\chi(q,p) = +\infty$.

Otherwise $q_j = 0$ on every zero-power coordinate. Componentwise,

$$q^\top\delta\phi = (D_p^{\dagger/2} q)^\top (D_p^{1/2}\delta\phi).$$

Cauchy–Schwarz gives

$$\begin{aligned} |q^\top \delta\phi| &\le \|D_p^{\dagger/2} q\|_2 \|D_p^{1/2} \delta\phi\|_2 \\ &= \chi(q,p)\sqrt{R_\phi}, \end{aligned} \tag{25}$$

which proves Eq. equation 24.

When all relevant $p_j > 0$, the finite factor reduces to

$$\chi(q,p)^2 = \sum_{j \in J} \frac{q_j^2}{p_j}.$$

The weighted Cauchy–Schwarz step is tight over unconstrained local phase-error vectors: equality is attained when, on the positive-support coordinates,

$$\delta\phi \propto D_p^\dagger q.$$

This statement concerns the phase-vector inequality itself and does not assert that the maximizing direction belongs to a particular neural architecture or satisfies additional stability constraints.

Relating phase error to complex symbol error. Write

$$g_j = a_j e^{i\theta_j}, \qquad \widehat{g}_j = \widehat{a}_j e^{i(\theta_j + \delta\phi_j)},$$

with

$$a_j, \widehat{a}_j \ge a_{\min}.$$

Then

$$\begin{aligned} |\widehat{g}_j - g_j|^2 &= (\widehat{a}_j - a_j)^2 + 4 a_j \widehat{a}_j \sin^2\left(\frac{\delta\phi_j}{2}\right) \\ &\ge 4 a_{\min}^2 \sin^2\left(\frac{|\delta\phi_j|}{2}\right). \end{aligned} \tag{26}$$

For $0 \le |\delta\phi_j| \le \eta < \pi$, concavity of $\sin x$ on $[0, \pi/2]$ gives

$$\sin\left(\frac{|\delta\phi_j|}{2}\right) \ge \frac{\sin(\eta/2)}{\eta} |\delta\phi_j|.$$

Combining with Eq. equation 26,

$$|\delta\phi_j| \le C_{\eta, a_{\min}} |\widehat{g}_j - g_j|,$$

where

$$C_{\eta, a_{\min}} = \frac{\eta}{2 a_{\min} \sin(\eta/2)}.$$

Multiplying by p_j, squaring, and summing over J yields

$$R_\phi \le C_{\eta, a_{\min}}^2 R_{\mathrm{sym}, J}. \tag{27}$$

Finally, Eqs. equation 24, equation 27, and equation 23 give

$$\begin{aligned} |\Delta s| = |q^\top \delta\phi| &\le \chi(q,p)\sqrt{R_\phi} \\ &\le C_{\eta, a_{\min}} \chi(q,p)\sqrt{R_{\mathrm{sym}, J}} \\ &\le C_{\eta, a_{\min}} \chi(q,p)\sqrt{\mathcal{R}_\mu(G)}, \end{aligned} \tag{28}$$

which proves Corollary 1.

Relation to the five-mode estimator. In the main trained-model experiments, q is the fixed coefficient vector of the unweighted least-squares slope fit over the five modes $\{k_0 - 2, \ldots, k_0 + 2\}$. The exact construction uses an amplitude-weighted spectral fit; because its regression weights can vary with the multiplier amplitudes, it is not represented exactly by the fixed-q specialization above.

A.4 Circular-Moment Identity

We now prove the exact Fourier representation of the circular center used in Sec. 5.4.

Proof of Lemma 2. We use the discrete Fourier convention

$$u_n^{(L)} = \frac{1}{N} \sum_k c_k g_k^L e^{2\pi i k n/N}.$$

Therefore

$$|u_n^{(L)}|^2 = \frac{1}{N^2} \sum_{k,\ell} c_k \overline{c_\ell} g_k^L \overline{g_\ell^L} e^{2\pi i n(k-\ell)/N}. \tag{29}$$

Multiplying by $e^{2\pi i n/N}$ and summing over n,

$$\begin{aligned} m_L &= \sum_{n=0}^{N-1} |u_n^{(L)}|^2 e^{2\pi i n/N} \\ &= \frac{1}{N^2} \sum_{k,\ell} c_k \overline{c_\ell} g_k^L \overline{g_\ell^L} \sum_{n=0}^{N-1} e^{2\pi i n(k-\ell+1)/N}. \end{aligned} \tag{30}$$

Discrete Fourier orthogonality gives

$$\sum_{n=0}^{N-1} e^{2\pi i n r/N} = \begin{cases} N, & r = 0 \pmod N, \\ 0, & \text{otherwise.} \end{cases}$$

Hence only terms satisfying

$$\ell = k + 1 \pmod N$$

remain in Eq. equation 30, yielding

$$m_L = \frac{1}{N} \sum_k c_k \overline{c_{k+1}} \left(g_k \overline{g_{k+1}}\right)^L. \tag{31}$$

This is Eq. equation 13.

Whenever $m_L \neq 0$, its argument is well defined locally, and the circular center is

$$x_L = \frac{N}{2\pi} \arg m_L$$

modulo the periodic domain. □

Equation equation 31 makes explicit why packet position is sensitive to differences between neighboring spectral phases. A common phase factor applied to every g_k cancels from $g_k \overline{g_{k+1}}$ and therefore cannot translate the packet through this mechanism.

A.5 Packet-Center Sensitivity near Constant Advection

We next derive Corollary 2. The result is a first-order statement around an aligned constant-advection reference.

Assume

$$c_k \overline{c_{k+1}} \left(g_k \overline{g_{k+1}}\right)^L = w_k e^{i\theta}, \qquad w_k \geq 0, \tag{32}$$

and let

$$W = \sum_k w_k > 0.$$

Consider the one-parameter perturbation

$$g_k(t) = g_k \exp\left(t\dot{a}_k + it\dot{\phi}_k\right).$$

Then

$$g_k(t)\overline{g_{k+1}(t)} = g_k\overline{g_{k+1}} \exp\left\{t(\dot{a}_k + \dot{a}_{k+1}) + it(\dot{\phi}_k - \dot{\phi}_{k+1})\right\}. \tag{33}$$

Using Lemma 2 and Eq. equation 32,

$$m_L(t) = \frac{e^{i\theta}}{N} \sum_k w_k \exp\left\{Lt(\dot{a}_k + \dot{a}_{k+1}) + iLt(\dot{\phi}_k - \dot{\phi}_{k+1})\right\}. \tag{34}$$

At $t = 0$,

$$m_L(0) = \frac{We^{i\theta}}{N} \neq 0.$$

Differentiating Eq. equation 34,

$$m_L'(0) = \frac{Le^{i\theta}}{N} \sum_k w_k \left[\dot{a}_k + \dot{a}_{k+1} + i(\dot{\phi}_k - \dot{\phi}_{k+1})\right]. \tag{35}$$

For a differentiable complex-valued function $m(t) \neq 0$,

$$\frac{d}{dt} \arg m(t) = \mathrm{Im}\left(\frac{m'(t)}{m(t)}\right).$$

Therefore

$$\frac{d}{dt} \arg m_L(t)\Big|_{t=0} = \frac{L}{W} \sum_k w_k(\dot{\phi}_k - \dot{\phi}_{k+1}), \tag{36}$$

and hence

$$\boxed{\frac{dx_L(t)}{dt}\Big|_{t=0} = \frac{NL}{2\pi W} \sum_k w_k(\dot{\phi}_k - \dot{\phi}_{k+1})}. \tag{37}$$

This proves Corollary 2.

Two aspects of Eq. equation 37 are worth making explicit. First, the center responds to *adjacent phase differences*; a perturbation with

$$\dot{\phi}_k = c \quad \text{for all } k$$

has zero first-order effect. Second, pure amplitude perturbations satisfy

$$\dot{\phi}_k = 0 \quad \text{for all } k,$$

and therefore

$$\frac{dx_L(t)}{dt}\Big|_{t=0} = 0.$$

Under the stronger condition that the phase remains unchanged and all amplitude factors remain positive, the aligned reference contributions in Eq. equation 34 remain on the same complex ray $e^{i\theta}$. In that special case, amplitude changes preserve the center angle exactly, not only to first order.

Finite perturbations and horizon dependence. The derivative above is exact for every fixed L. A finite-perturbation Taylor expansion requires additional care because the exponent itself scales with L. If a perturbation has characteristic size $\|\delta\|$, the second-order terms scale as

$$O\big(L^2\|\delta\|^2\big),$$

with additional constants depending on the conditioning of the circular moment. Thus the first-order approximation is naturally associated with a regime such as

$$L\|\delta\| \ll 1,$$

and with a perturbation path on which

$$|m_L(t)| \geq m_* > 0.$$

The main-text empirical horizon curves are therefore interpreted as being consistent with an approximately persistent transport-speed bias, rather than as a uniform-in-horizon quantitative validation of Eq. equation 37.

A.6 Supported Local Task-Error Bound

Proposition 1 gives an exact characterization of the task derivative $DF_T[A]$. To translate this into a bound on the actual task difference $F(T+A) - F(T)$, prediction risk must also control the norm of the perturbation. We state one sufficient local result.

Let

$$P_0$$

denote the projector onto $\ker M_\mu$, and restrict attention to perturbations that act only on the supported input subspace:

$$AP_0 = 0.$$

Define the supported task sensitivity

$$\kappa_{F,\mathrm{supp}}(\mu) = \sup_{\substack{AP_0=0 \\ \|AM_\mu^{1/2}\|_F>0}} \frac{|DF_T[A]|}{\|AM_\mu^{1/2}\|_F}. \tag{38}$$

Proposition 2 (Supported local task-error bound)**.** *Suppose that*

$$M_\mu \succeq m(I - P_0)$$

for some $m > 0$ on the supported subspace. Assume that F is twice Fréchet differentiable in the Frobenius ball

$$\{T + A : \|A\|_F \le \delta\}$$

and that

$$\|D^2 F_{T+A}\|_{\mathrm{op}} \le H_F$$

throughout this ball. Then

$$\kappa_{F,\mathrm{supp}}(\mu) = \|B_F M_\mu^{\dagger/2}\|_F. \tag{39}$$

Moreover, for every perturbation satisfying

$$AP_0 = 0, \qquad \|AM_\mu^{1/2}\|_F \le \varepsilon, \qquad \varepsilon \le \delta\sqrt{m},$$

we have

$$|F(T+A) - F(T)| \le \varepsilon \kappa_{F,\mathrm{supp}}(\mu) + \frac{H_F}{2m}\varepsilon^2. \tag{40}$$

Proof. Because $AP_0 = 0$, we have

$$A = A(I - P_0).$$

The task derivative therefore depends only on the component of B_F acting on the supported subspace. Repeating the whitening argument from Proposition 1 gives

$$|DF_T[A]| \le \|B_F M_\mu^{\dagger/2}\|_F \|AM_\mu^{1/2}\|_F.$$

The same aligned construction as in Eq. equation 20 lies in the supported perturbation class, so the bound is tight for the tangent problem. This proves Eq. equation 39.

Next, the support lower bound on M_μ implies

$$\begin{aligned} \|AM_\mu^{1/2}\|_F^2 &= \mathrm{tr}(AM_\mu A^*) \\ &\ge m \ \mathrm{tr}\,(A(I - P_0)A^*) \\ &= m\|A\|_F^2, \end{aligned} \tag{41}$$

where the last equality uses $AP_0 = 0$. Hence

$$\|A\|_F \le \frac{\varepsilon}{\sqrt{m}} \le \delta. \tag{42}$$

The segment $T + tA$, $0 \le t \le 1$, therefore remains inside the region in which the Hessian is bounded.

Taylor's theorem on the real Hilbert space of complex matrices gives

$$F(T+A) - F(T) = DF_T[A] + R_2(A),$$

with

$$|R_2(A)| \le \frac{H_F}{2}\|A\|_F^2.$$

Using the tangent bound and Eq. equation 42,

$$|F(T+A) - F(T)| \le \varepsilon \kappa_{F,\mathrm{supp}}(\mu) + \frac{H_F}{2}\frac{\varepsilon^2}{m}, \tag{43}$$

which proves Eq. equation 40. □

The supported bound clarifies the distinction between tangent certification and a finite task-error guarantee. Proposition 1 allows the global task sensitivity to be infinite when the task reads a training-null direction. Proposition 2 instead conditions on perturbations that exclude that nullspace and adds the regularity needed to control the nonlinear Taylor remainder. Neither result requires identifying the finite- amplitude nonlinear FNO response with a scalar Fourier multiplier.

B Experimental Protocol and Reproducibility

This section specifies the experimental objects underlying the main results. The protocol is designed to distinguish three quantities that are otherwise easy to conflate: standard field prediction on the evaluation distribution, the local response of the learned propagator to a transport-relevant perturbation, and the resulting finite-horizon packet motion. Unless stated otherwise, model seeds are the independent statistical units; packet frequencies, backgrounds, and related probe cases provide repeated measurements within a seed rather than additional independent replicates.

B.1 Datasets and Data Provenance

We use a combination of analytically generated systems, trained synthetic benchmarks, and released PDE datasets. Their roles are deliberately different: the analytic settings isolate identifiability and mechanism, whereas the released datasets test whether the same transport discrepancy occurs outside a deliberately constructed spectral support.

Exact linear constructions. The exact counterexamples use periodic linear propagators for advection and Schrödinger-type dispersive dynamics. The learned propagator is represented by a 25-tap linear convolution and is constrained to agree with the reference dynamics on 13 observed Fourier modes. Multiple extensions outside this observed support are then possible without changing the observed-mode prediction error. These constructions are used only as identifiability examples; they are not treated as evidence for the prevalence of a particular learned solution.

Synthetic trained benchmarks. The controlled learned-model experiments use synthetic periodic advection and Schrödinger dynamics. The formal one-dimensional benchmark is trained on 20% of a fixed pool of 1000 generated examples and evaluated over three independent model seeds. These data provide a setting in which spectral support, transport response, and reference propagation are all directly available. The positive-tail and continuous-frequency extensions modify the spectral coverage while preserving the same basic transport task; their complete results are reported in Appendix C.

We additionally use analytically generated heterogeneous advection to test transport along spatially varying characteristics, and a synthetic two-dimensional acoustic system as a controlled dimensionality extension. The latter is reported in Appendix C.4.

Released one-dimensional PDE data. The primary released-data experiment uses the $\beta = 1$ advection dataset from PDEBench. The source file is checksum-verified, and a fixed 20% training subset is used across the formal five-seed comparison. The released one-dimensional data are represented as complex fields by pairing real-valued trajectories into real and imaginary components and applying the same per-sample normalization used by the experimental pipeline.

We also evaluate released Burgers data at viscosities $\nu = 0.01$ and $\nu = 0.004$. The higher-viscosity setting exposes a physical observability issue because the reference dynamics can strongly attenuate

the high-frequency perturbation itself. The lower-viscosity replication restores reference observability but still exhibits substantial learned tangent-amplitude variation. We therefore use Burgers primarily to establish the boundary of the clean phase-slope interpretation; full results appear in Appendix F.1.

Released two-dimensional data. The public two-dimensional stress test uses a dataset of 1200 Navier–Stokes trajectories. A fixed 20% subset of trajectories is used for training, and the fields are projected to a 32×32 spectral representation for the reported experiments. Both an FNO and a periodic U-Net are trained over five seeds. Because their field NRMSE is moderate rather than near zero, this experiment is interpreted as a cross-dimensional and cross-architectural stress test rather than as another near-perfect-accuracy counterexample; additional details and threshold sensitivity are given in Appendix C.5.

All released-data inputs used in the experiments are associated with recorded file hashes. Training subsets, selected trajectory indices, model checkpoints, and the raw quantitative outputs used for the paper figures are retained together with their data provenance so that the reported subsets can be reconstructed without resampling.

B.2 Model Architectures

The experiments intentionally include both spectral and convolutional architectures so that the transport phenomenon is not identified with a particular Fourier-layer implementation.

Linear and controlled models. The exact construction uses finite-support linear convolutions. The synthetic trained benchmark additionally contains a trainable local convolutional model and a small Fourier neural operator (mini-FNO). These models provide the closest empirical connection to the linear Fourier-diagonal analysis in Sec. 5.

One-dimensional FNO. The released-data and extended one-dimensional experiments use a four-block residual FNO, denoted FNO4 in the main text, following the Fourier-layer construction of Li et al. (2021). Complex fields are represented by two real channels. Each block combines spectral and pointwise transformations with nonlinear activation and normalization, and the model includes residual composition across blocks. Consequently, the full network is nonlinear and a pure Fourier mode need not remain confined to a single output mode. This is why Eq. equation 4 is treated as a projected finite-amplitude response rather than as an exact eigenvalue of the network.

The activation anti-aliasing comparison uses a matched FNO variant in which the nonlinear activation path is evaluated with $2\times$ anti-aliasing. Architecture, data, and evaluation protocol are otherwise paired with the corresponding standard FNO experiment. This comparison targets activation-induced aliasing specifically and is not intended as a test of every possible aliasing mechanism; implementation-level results are reported in Appendix D.4.

Periodic U-Nets. To test architecture dependence, we also use convolutional U-Nets with periodic boundary handling, following the encoder–decoder lineage of Ronneberger et al. (2015). The one-dimensional model is evaluated on heterogeneous advection and low-viscosity Burgers. Its positive result on heterogeneous advection and negative result on Burgers are both retained in the analysis rather than pooled into a single architecture-level conclusion. The corresponding two-dimensional periodic U-Net is used in the public Navier–Stokes stress test.

Two-dimensional spectral model. The public two-dimensional experiments use a compact FNO2d together with the periodic U-Net2d. These models are intended to test whether the packet-transport observable remains informative in a higher-dimensional setting, not to establish a benchmark architecture comparison.

Exact architectural dimensions, retained spectral modes, optimizer settings, and training durations are fixed by the versioned experiment configurations associated with each checkpoint. No architectural parameter is selected using packet-center performance from the formal evaluation cases.

B.3 Spectral Probes and Phase-Slope Estimation

For a nonlinear learned propagator G_θ, we estimate the fundamental response to Fourier mode e_k through the centered finite-difference probe

$$\delta G_\theta(e_k; a) = \frac{G_\theta(ae_k) - G_\theta(-ae_k)}{2a}, \tag{44}$$

followed by projection onto the same input mode,

$$\widehat{g}_\theta(k; a) = \langle e_k, \delta G_\theta(e_k; a) \rangle . \tag{45}$$

This construction suppresses the even component of the response but does not imply that higher harmonics or channel coupling vanish.

For the released-data and positive-tail FNO experiments, the probe amplitude is $a = 1$. These measurements are therefore finite-amplitude centered secants. The earlier controlled mini-FNO benchmark uses a substantially smaller probe amplitude (10^{-3}), which is closer to a zero-state tangent measurement. We do not merge these two interpretations: the first is an effective finite-amplitude transport response, whereas the second more closely approximates a local linear response.

All principal one-dimensional packets have spectral half-bandwidth two. For a carrier frequency k_0, the transport diagnostic therefore uses

$$\mathcal{J}(k_0) = \{k_0 - 2, k_0 - 1, k_0, k_0 + 1, k_0 + 2\}.$$

The complex responses on these five modes are placed on a common locally unwrapped phase branch. For the trained-model analyses, we fit

$$(\widehat{\alpha}, \widehat{\beta}) = \arg\min_{\alpha,\beta} \sum_{k \in \mathcal{J}(k_0)} \left(\phi(k) - \alpha - \beta k\right)^2 , \tag{46}$$

where

$$\phi(k) = \arg \widehat{g}_\theta(k; a).$$

The same fit is applied to the reference response, and the slope difference is converted to the one-step transport-velocity discrepancy using the Fourier and grid conventions of the corresponding simulation.

The main trained-model diagnostic is therefore a fixed, unweighted five-mode OLS slope. The exact linear construction predates this convention and uses an amplitude-weighted regression with weights proportional to the symbol magnitude. Because those weights can themselves change with the response amplitude, the exact weighted diagnostic is reported separately and is not identified algebraically with the fixed-q estimator used in Corollary 1.

The principal continuous-frequency extension evaluates carrier modes over a broader range rather than only at the calibration frequencies. The full frequency sweep is reported in Appendix C.2; it is used to test whether the observed transport discrepancy is confined to a small hand-selected set of packet centers.

B.4 Packet Construction and Center Metric

The transport probes are narrowband periodic wave packets. In one dimension, a packet centered spectrally at k_0 is constructed from neighboring Fourier components around k_0, with coefficients chosen so that the resulting field is localized in physical space. The principal packet bandwidth contains the five modes in $\mathcal{J}(k_0)$ above. Carrier frequencies, packet locations, background states, and envelope coefficients are fixed by the corresponding experiment configuration before evaluating a model seed.

For a complex field u on an N-point periodic grid, transport is measured through the energy-weighted circular first moment

$$m(u) = \sum_{n=0}^{N-1} |u_n|^2 \exp\left(\frac{2\pi i n}{N}\right), \tag{47}$$

and, whenever $m(u) \neq 0$, the circular center

$$x(u) = \frac{N}{2\pi} \arg m(u). \tag{48}$$

For an L-step rollout, the reported packet-center error is the shortest periodic distance between the predicted and reference centers,

$$e_{\mathrm{ctr}}(L) = d_{\mathbb{T}} \left(x(G_\theta^L u_0), x(T^L u_0) \right). \tag{49}$$

Distances are reported in grid cells.

The official advection component analysis evaluates horizons $L \in \{1, 2, 4, 8\}$. The controlled positive-tail experiments additionally extend the clean horizon analysis to $L = 16$. Horizon curves are interpreted only over the range in which the packet remains observable and the periodic center can be reliably defined.

For tangent experiments in dissipative or nonlinear systems, a packet case is used only when the corresponding reference perturbation retains sufficient energy to define a meaningful transport center according to the pre-specified reliability criterion. This distinction is particularly important for viscous Burgers dynamics, where physical damping can remove the perturbation itself. The consequences of this filter and its threshold sensitivity are analyzed in Appendix F.1 and Appendix D.6.

Packet-center error is intentionally not used as a substitute for full-field error. Standard NRMSE is reported separately, and packet-state NRMSE is retained where it is relevant to interpretation. In particular, the main claim compares strong in-distribution field prediction with transport in directions weakly constrained by standard prediction risk; it does not assume that the same out-of-distribution packet has low full-field error.

B.5 Correction and Estimation Procedures

The spectral interventions modify the response of an already trained base predictor. Let

$$c(k) = r(k) e^{i\psi(k)}$$

denote the estimated complex correction at frequency k. At each rollout step, the base model first produces a field prediction. Its Fourier coefficients are then modified and transformed back to physical space before the next model step is taken. Thus the correction participates in the composed rollout rather than being applied only to the terminal output.

The three component interventions are

$$\widehat{u}^{\,\mathrm{amp}}(k) = r(k) \widehat{u}^{\,\mathrm{base}}(k), \tag{50}$$

$$\widehat{u}^{\,\mathrm{phase}}(k) = e^{i\psi(k)} \widehat{u}^{\,\mathrm{base}}(k), \tag{51}$$

$$\widehat{u}^{\,\mathrm{full}}(k) = r(k) e^{i\psi(k)} \widehat{u}^{\,\mathrm{base}}(k). \tag{52}$$

When response information is available only at selected spectral locations, the correction is interpolated over the frequencies on which it is applied. The official advection probe-budget study considers

$$0,\ 2,\ 4,\ 6,\ \text{and}\ 10$$

structured observations. The main six-observation comparison should not be read as an optimal budget: the measured dependence is slightly non-monotone, with four observations outperforming six and ten in the corresponding sweep. The complete budget dependence is reported in Appendix E.1.

Matched broadband supervision. To distinguish supervision content from estimation procedure, we use exactly the same six noiseless broadband input–target pairs in two pipelines.

The end-to-end pipeline continues optimization of the pretrained propagator using these six pairs. Update strength and stopping are selected using standard validation behavior only; packet-center performance is not used for model selection.

The structured estimator instead solves, independently for each Fourier mode with nonzero input energy,

$$\widetilde{g}(k) = \arg\min_{z \in \mathbb{C}} \sum_{m=1}^{6} \left| \widehat{v}^{(m)}(k) - z \widehat{u}^{(m)}(k) \right|^2. \tag{53}$$

The resulting transfer estimate is converted into the same type of spectral correction used for the pure-mode intervention. Because both procedures receive the identical six broadband fields, differences in transport performance isolate how their information is estimated rather than whether one procedure has access to different labels.

The direct estimator is used only where a scalar linear transfer relation is appropriate. We therefore interpret its near-exact recovery in released linear advection as an information–estimation result, not as a general calibration algorithm for nonlinear PDE dynamics.

B.6 Statistical Aggregation and Uncertainty

The statistical unit for trained-model inference is the independently initialized model seed. Packet centers, carrier frequencies, background states, and related probe cases generated from a single trained model are repeated measurements of that seed and are not treated as independent replicates.

For quantities summarized over multiple packet cases, we first compute the median within each model seed. Let z_s denote the resulting seed-level statistic for seed s. The reported point estimate is the median of the $\{z_s\}$. Uncertainty intervals are obtained by resampling model seeds with replacement and recomputing this median over $10{,}000$ bootstrap draws. Main figures use 95% bootstrap intervals. The bootstrap random-number seed used for the final reported intervals is 92731.

Most released-data comparisons use five independent model seeds. The synthetic trained benchmark uses three formal seeds, and the longer-training comparison uses three independently initialized training trajectories evaluated at the 400, 800, and 1600 checkpoints of each trajectory. Confidence intervals therefore describe variability across model initializations, not uncertainty across individual packet examples.

Correlations used to summarize the transport-specific diagnostic are computed at the aggregation level shown in the corresponding figure. In particular, the main mechanism plot first aggregates packet cases at the model–system–seed level before computing the reported association. We do not use packet-level pseudo-replication to inflate the effective sample size.

Exact constructions and deterministic controls without independent model seeds are reported descriptively rather than assigned seed-bootstrap confidence intervals. Likewise, the same-support data-scaling construction in Appendix D.1 is interpreted as a controlled coverage example rather than as a stochastic sample-complexity estimate.

A complete table of the seed-cluster estimates and intervals used by the primary claims is given in Appendix H.4.

B.7 Compute and Reproducibility

All reported quantitative results are traceable to versioned raw outputs and model checkpoints. The reproducibility archive contains 94 formal checkpoints spanning 15 distinct seed identifiers, together with a checkpoint manifest and SHA-256 hashes. Seed-cluster statistics are stored separately from packet-level outputs; the final archive contains 62 precomputed seed-cluster confidence-interval rows.

External datasets are associated with file hashes, and the recorded software environment includes the package versions used for the experiments. Training and analysis implementations are separated from their command-line entry points, and the reproduction record contains the commands required to regenerate the principal metrics from the stored data and checkpoints. Nine representative compute profiles record the execution characteristics of the principal model families.

Figure generation is also provenance-tracked. Each paper panel is linked to its raw CSV or JSON inputs, and the source manifest records input hashes, file sizes, schemas, and row counts. The paper figures are generated from these quantitative sources rather than by reading values back from previously rendered plots.

Engineering-only configuration searches and smoke tests are retained separately from the formal model seeds and are never pooled into the reported uncertainty estimates. This includes preliminary capacity checks and abandoned architectures. Their role is documented in the negative-result ledger

in Appendix H.2, whereas the quantitative tables use only the corresponding formal experimental units.

At the final recorded state, the experiment and analysis test suite contains 231 passing tests. Additional figure-level provenance and source-to-panel traceability are summarized in Appendix H.3.

C Robustness Across Coverage, Systems, and Dimensions

The main results combine an exact support-based construction with trained models and released PDE data. We next examine whether the transport discrepancy persists when the idealized assumptions behind the clean construction are relaxed. In particular, we replace strict spectral gaps by weak positive coverage, evaluate carrier frequencies beyond the calibration locations, use a naturally non-band-limited advection distribution, and extend the transport test to two-dimensional systems.

These experiments are robustness tests rather than attempts to impose a single mechanism on every setting. The clean phase-slope interpretation is strongest for the linear translation-equivariant systems analyzed in Sec. 5. In the broader settings below, we ask the more limited question of whether standard prediction accuracy can still leave a task-relevant transport response poorly determined, and whether targeted response information can improve that transport.

C.1 Positive Spectral Tails

The exact construction in Sec. 4 contains directions with zero training support. A natural concern is that the resulting ambiguity is therefore a discontinuous artifact: once every relevant Fourier mode receives any nonzero training energy, ordinary prediction risk might be sufficient to determine the transport response.

To test this possibility, we replace the strict spectral cutoff by positive tails of increasing magnitude. The controlled extension considers tail levels

$$0,\ 10^{-4},\ 10^{-3},\ 10^{-2},\ 10^{-1}$$

for both advection and Schrödinger dynamics, using residual FNO4 models. The primary positive-tail setting is 10^{-3}, for which five independently trained models are evaluated across carrier frequencies and rollout horizons.

Wrong-ray behavior remains pronounced despite strictly positive spectral coverage. At the primary 10^{-3} tail, the seed-cluster eight-step packet-center errors are

$$9.45\,[9.34, 10.19] \quad \text{cells for advection,}$$

and

$$20.10\,[19.14, 20.59] \quad \text{cells for Schrödinger dynamics.}$$

Thus exact zero support is sufficient for non-identifiability, but it is not necessary for a practically large transport-certification gap.

The dependence on tail magnitude is also not monotone. Increasing the high-frequency tail does not produce a corresponding monotone reduction in packet-center error, and the largest tested tail can yield worse transport than intermediate tails. This is consistent with the theoretical role of coverage as a conditioning geometry rather than a scalar predictor of realized error. Modal energy affects how strongly a direction is weighted by prediction risk, but the realized learned response also depends on optimization, parameterization, and the orientation of the resulting operator error. A more direct analysis of raw spectral energy is given in Appendix D.6.

C.2 Continuous-Frequency Evaluation

The main packet experiments necessarily evaluate a finite collection of carrier frequencies. To test whether the reported effect depends on a small set of selected locations, we perform a broader carrier-frequency sweep over modes

$$k_0 \in \{14, \dots, 50\}$$

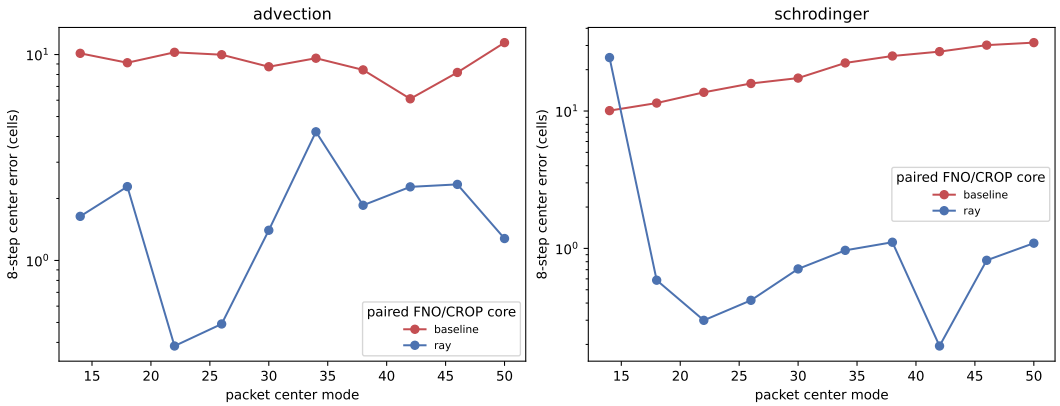


Figure 6: **Wrong transport persists across a broad range of carrier frequencies.** Packet-center error is evaluated across carrier modes 14–50 for the positive-tail advection and Schrödinger experiments. The primary setting uses nonzero spectral tail 10^{-3} and five independently trained FNO4 models. Large transport errors occur at multiple carrier frequencies, including locations that are not themselves calibration anchors. The frequency dependence is structured but non-monotone, so the result supports robustness beyond a hand-selected packet frequency rather than a claim that raw spectral energy alone determines transport error.

in the positive-tail setting. The same trained models are evaluated at frequencies between and beyond the spectral locations used for structured correction.

Figure 6 shows that large center errors are not confined to one isolated packet band. They occur across multiple frequencies and persist at non-anchor locations. This reduces the possibility that the main result is produced by selecting a single anomalous carrier frequency after training.

The sweep also illustrates why the task-conditioned interpretation is more appropriate than a universal frequency threshold. Error magnitude varies substantially across the evaluated band, and increasing spectral coverage does not induce a simple monotone frequency-by-frequency ordering. The relevant statement is therefore that weakly constrained spectral response can leave transport inaccurate over an extended region, not that all high frequencies fail equally or that a single frequency cutoff separates correct from incorrect transport.

C.3 Non-Band-Limited Advection

The positive-tail experiments retain a deliberately parameterized spectral distribution. We therefore perform a complementary test using the official PDEBench advection initial-condition generator, whose inputs have naturally non-band-limited spectra rather than an imposed hard cutoff or constant spectral tail. The protocol uses a fixed deterministic 20% subset and five FNO4 model seeds.

The learned models obtain a standard NRMSE of 0.0471, while the baseline packet evaluation exhibits a median center error of 9.37 grid cells. Applying the structured transport correction reduces this error to 2.32 cells, with improvement in all five model seeds. The associated transport-response discrepancy is reduced by 84.9%.

This experiment shows that the wrong-ray/repair pattern does not require a synthetically band-limited input distribution. At the same time, it provides an important boundary for the scalar mechanism diagnostic. The rank correlation between the group-velocity diagnostic and realized packet-center error is only $\rho = 0.244$ in this setting, substantially weaker than in the clean synthetic benchmarks. The structured intervention remains effective even though a single scalar phase-slope error no longer ranks individual packet errors reliably.

We therefore interpret this experiment as evidence for robustness of the broader transport-identification problem, not as evidence that group-velocity error is a universal scalar predictor for finite-amplitude nonlinear FNO responses. This diagnostic boundary is examined further in Appendix F.4.

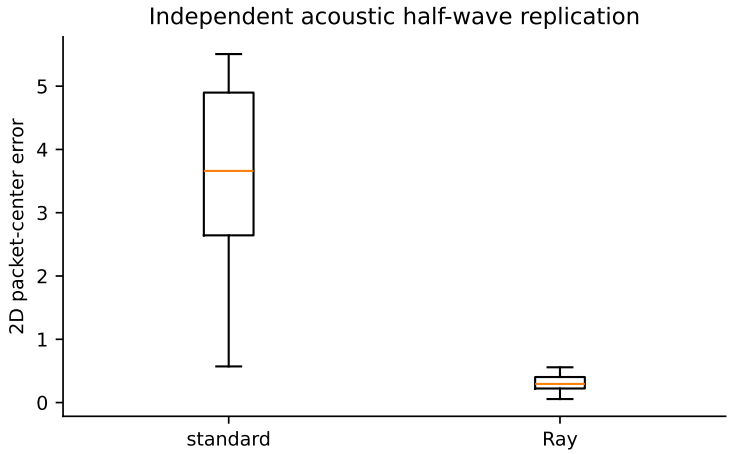


Figure 7: **Controlled two-dimensional acoustic replication.** In a synthetic 2D half-wave system evaluated over three model seeds, structured transport calibration reduces packet-center error from 3.66 to 0.29 grid cells (91.9%) while standard in-distribution NRMSE changes only from 0.0138 to 0.0140. The experiment provides a controlled dimensionality extension of the prediction–transport separation; it does not establish a universal 2D phase-slope mechanism.

C.4 Synthetic Two-Dimensional Acoustics

The one-dimensional experiments make packet trajectories and local phase response particularly transparent. To test whether the same evaluation principle extends beyond one spatial dimension without immediately introducing the additional complexity of released turbulent dynamics, we use a controlled two-dimensional acoustic half-wave system.

The experiment is evaluated over three independently trained models. Before task-aligned correction, the packet-center error is 3.66 grid cells. Structured transport calibration reduces this error to 0.29 cells, corresponding to a 91.9% reduction. In contrast, standard in-distribution prediction changes only slightly: NRMSE is 0.0138 before correction and 0.0140 afterward. Figure 7 summarizes this controlled dimensionality extension.

The result extends the central evaluation distinction to a controlled two-dimensional setting: a response can be important for localized transport while having little effect on the standard distribution-weighted field metric. We do not use this experiment to claim that the one-dimensional scalar phase-slope theory extends unchanged to arbitrary two-dimensional dynamics. Instead, it provides a bridge between the clean one-dimensional mechanism experiments and the released two-dimensional stress test below.

C.5 Public Two-Dimensional Navier–Stokes

We finally evaluate the transport diagnostic on a released two-dimensional Navier–Stokes dataset containing 1200 trajectories. A fixed 20% subset is used for training, fields are represented on a 32×32 grid, and both a compact FNO2d and a periodic U-Net2d are trained over five independent seeds.

The two architectures achieve standard test NRMSEs of 0.198 and 0.173, respectively, compared with 0.428 for the persistence baseline. Their corresponding four-step packet-center errors are

$$6.267\,[5.502, 6.668] \quad \text{cells for FNO2d,}$$

and

$$5.566\,[4.955, 6.733] \quad \text{cells for periodic U-Net2d.}$$

The transport discrepancy therefore appears in both a spectral and a convolutional architecture in this released two-dimensional setting.

The registered perturbation-retention criterion retains 80.6% of the evaluated cases. Because the model field errors are moderate, these results should not be read as evidence of a hidden physical failure under nearly perfect standard prediction. Their role is instead to test external validity: the packet-center observable continues to expose multi-cell transport discrepancies after moving simultaneously from one to two dimensions, from synthetic to released dynamics, and from a Fourier architecture to a periodic U-Net.

The retention threshold does not drive the qualitative conclusion. As shown in Appendix D.6, relaxing the perturbation-energy threshold until all cases are retained leaves multi-cell transport errors for both model families. We nevertheless leave the mechanism status unassigned in this setting: two-dimensional nonlinear Navier–Stokes dynamics do not satisfy the scalar Fourier-diagonal assumptions of Sec. 5, and the present experiment does not isolate phase, amplitude, and mode-coupling contributions separately.

Taken together, the robustness results establish a bounded but useful extension of the main claim. The transport-certification gap is not restricted to strict zero spectral support, a single selected packet frequency, an artificially band-limited input generator, one spatial dimension, or Fourier architectures. At the same time, the broader experiments weaken rather than strengthen the case for a universal scalar mechanism: the clean phase-slope account remains most appropriate in the linear transport regime, while heterogeneous, nonlinear, and two-dimensional settings require a more general task-conditioned response description.

D Alternative Explanations and Additional Controls

The main experiments establish that standard prediction accuracy can coexist with incorrect packet transport, but several simpler explanations could produce a similar empirical pattern. In this section we test whether the observed discrepancy can be accounted for by insufficient sample count on the same support, a small amount of additional IID supervision, discretization mismatch, activation aliasing, incomplete optimization of the standard objective, or a particular choice of spectral-energy or reliability threshold.

These controls address different hypotheses and should not be interpreted as showing that the corresponding mechanisms are irrelevant in general. Their role is narrower: in the settings studied here, improving these quantities does not by itself determine the transport response.

D.1 Increasing Same-Support Training Data

The second-moment view in Sec. 5.1 distinguishes the amount of data from the directions covered by that data. If transport failure were simply a finite-sample effect, increasing the number of examples from the same distribution could be expected to reduce the packet error even without changing spectral support.

We test this possibility in a deterministic controlled setting for advection and Schrödinger dynamics. The fitting procedure is repeated using 10%, 20%, and 100% of the available training pool while keeping the spectral support fixed. Across these fractions, one-step prediction error remains near 10^{-7}, and the packet-center error is unchanged:

$$30.53 \quad \text{cells for advection,} \qquad 34.30 \quad \text{cells for Schrödinger dynamics.}$$

Increasing the number of samples within the same support therefore does not identify the missing transport response in this construction.

By contrast, augmenting the 20% fit with task-aligned spectral information reduces the corresponding packet-center errors to approximately numerical precision, on the order of 10^{-6} grid cells. The contrast is therefore between *additional samples on the same support* and *additional information in a task-relevant direction*.

This experiment should not be read as a general neural scaling law. It is a deterministic controlled example with fixed support and no independent model-seed distribution. Its role is to demonstrate that increasing sample count need not alter the population-level directions constrained by the training data.

Table 1: **Cross-resolution representation consistency and transport.** At 64 points, relevant transport modes are lost above Nyquist, so the comparison is not support compatible and DME decreases only modestly. At 256 points, the relevant support is preserved and the consistency procedure reduces DME by 99.97%, from 6.98×10^{-4} to 2.20×10^{-7}, while the training-grid wrong-ray behavior remains. The comparison therefore separates successful representation-consistency improvement from physical transport identification.

Resolution	Support compatible?	DME change	Transport interpretation
64	no	29.6% reduction	Nyquist-limited comparison
256	yes	$6.98\times10^{-4} \rightarrow 2.20\times10^{-7}$	wrong rays remain

D.2 Additional IID Supervision

The preceding construction changes the size of the training set globally. We also test a more local alternative: whether a small number of additional ordinary in-distribution examples can reproduce the effect of task-aligned calibration.

Starting from the same five pretrained official-advection models used in the matched label-budget study, we provide six additional IID input–target pairs and fine-tune the model through the ordinary prediction objective. At horizon eight, the resulting packet-center error remains approximately 12.17 grid cells, compared with 12.75 cells for the unmodified model. Under the same six-label budget, the task-aligned spectral correction reduces the error to approximately 3.25 cells.

Thus, in this comparison, the benefit of the structured intervention is not reproduced merely by adding the same number of ordinary IID labels. This result does not imply that IID data can never improve transport: sufficiently different data coverage, substantially larger datasets, or different optimization procedures may do so. The supported conclusion is that six additional samples drawn from the ordinary distribution do not provide the same transport correction as six task-aligned observations in the evaluated setting.

Together with Sec. D.1, this control separates sample quantity from task-relevant coverage. The result is also complementary to the matched broadband experiment in Sec. 7, where the required information is shown to be recoverable from generic broadband fields when it is estimated explicitly.

D.3 Cross-Resolution Representation Consistency

A distinct explanation is that wrong transport results from discretization mismatch: the learned operator may represent different effective mappings when evaluated on different grids. The zero-DME construction in Sec. 6 already shows that perfect continuous–discrete consistency is not sufficient for correct transport. We additionally examine a learned representation-consistency procedure across multiple resolutions.

The aggregate cross-resolution comparison includes 64-, 128-, and 256-point grids. Interpreting all resolutions together is complicated by support compatibility: at the 64-point resolution, part of the spectral region relevant to the transport probe lies above the Nyquist limit and therefore cannot be represented. The pooled comparison consequently reduces DME by only 41.35%, below the criterion used for the aggregate experiment.

The resolution-specific analysis separates this support loss from genuine representation inconsistency. Table 1 summarizes the two most informative cases.

On the support-compatible 256-point evaluation, DME is reduced by 99.97%, yet the wrong-ray behavior measured on the training grid remains. This provides the cleaner comparison for the scientific question: a model can become substantially more consistent across discrete representations without acquiring the physical response required by the packet-transport task.

The 256-point decomposition does not replace the aggregate cross-resolution outcome; rather, it explains why the pooled DME result is weaker. The relevant conclusion is therefore conditional on support compatibility. Representation consistency and transport identification are distinct axes once the compared grids are capable of representing the task-relevant modes.

D.4 Activation Anti-Aliasing

Nonlinear activations can generate frequencies beyond the represented band and fold them back into the resolved spectrum. Such activation aliasing is therefore a plausible source of an incorrect high-frequency response even when the underlying operator architecture is nominally spectral.

We test this explanation on released PDEBench advection by comparing the standard FNO4 with a matched variant that applies $2\times$ anti-aliasing around the nonlinear activation path. Both models attain approximately the same standard test accuracy, with NRMSE near 0.035. Their transport errors are also nearly unchanged: the ratio of the anti-aliased to standard eight-step center error is

$$1.006.$$

Thus the anti-aliasing modification does not remove the observed packet drift. This control rules out the particular activation-aliasing mechanism addressed by the tested architecture modification as a sufficient explanation of the result. It does not rule out every form of spectral aliasing, truncation error, or nonlinear mode coupling.

The distinction is important because anti-aliasing and task identification solve different problems. The former changes how nonlinear spectral content is represented inside the network; the latter asks whether the learned response agrees with the physical response along directions important to the downstream transport task.

D.5 Longer Standard Training

Another possibility is that the transport discrepancy is simply a transient optimization error. If so, continued progress on the standard training objective should eventually be accompanied by improved packet transport.

We examine uninterrupted training of the official-advection FNO4 over a fourfold increase in optimization duration. Three independently initialized training trajectories are evaluated at 400, 800, and 1600 steps. Standard test NRMSE improves monotonically:

$$0.0361 \;\to\; 0.0300 \;\to\; 0.0261.$$

Ordinary in-distribution eight-step rollout error likewise improves from

$$0.2365 \;\to\; 0.2012 \;\to\; 0.1811.$$

The corresponding seed-cluster eight-step packet-center errors are

$$12.205 \;\to\; 14.075 \;\to\; 13.836 \quad \text{grid cells.}$$

Thus substantial progress on both one-step/test and ordinary rollout metrics is not accompanied by transport improvement over the evaluated training range.

The comparison should not be interpreted as an asymptotic statement. Validation performance continues to improve from 800 to 1600 steps, including a 19.1% improvement over that interval, so the models have not been shown to reach a complete optimization plateau. The result establishes only that wrong-ray error persists through the tested fourfold training budget while the standard objective continues to improve.

This distinction is stronger than attributing the phenomenon to a completely underfit model but weaker than claiming that ordinary optimization could never recover the required response. Longer schedules, different objectives, or different optimization procedures remain outside the present comparison.

D.6 Spectral-Energy and Threshold Sensitivity

The theory assigns an important role to input coverage, but this should not be confused with a claim that raw spectral energy monotonically predicts realized packet error. We test this stronger interpretation in both the one-dimensional and two-dimensional experiments.

Local spectral energy. For the official advection models, we compare the local training-spectrum energy near each packet band with the corresponding realized ray error. The Spearman correlations are approximately

$$-0.088 \quad \text{for the standard FNO,} \qquad -0.125 \quad \text{for the anti-aliased FNO.}$$

These near-zero associations show that local modal energy alone is not a useful scalar ranker of packet-center error across the evaluated cases.

The positive-tail sweep provides the same warning in a different form. Increasing the imposed tail amplitude does not monotonically reduce transport error. At the largest tested tail, 10^{-1}, the eight-step center errors rise to approximately 24.39 cells for advection and 29.23 cells for Schrödinger dynamics. Coverage therefore enters as part of a conditioning geometry; the realized error also depends on the learned extension, phase structure, architecture, and optimization.

Two-dimensional reliability threshold. The public Navier–Stokes analysis uses a reference-perturbation energy threshold to exclude cases in which the perturbation is too weak to define a reliable transport center. Under the registered threshold, 80.6% of cases are retained. To test whether this filtering creates the observed result, we sweep the threshold over

$$0.0025, \ldots, 0.08.$$

At the lowest threshold, all cases are retained, yet the four-step center errors remain approximately

$$6.66 \quad \text{cells for FNO2d,} \qquad 6.46 \quad \text{cells for periodic U-Net2d.}$$

The two-dimensional transport discrepancy therefore does not depend on selecting a narrow subset through the registered reliability cutoff.

Taken together, these sensitivity analyses refine the interpretation of "coverage" used throughout the paper. Low or zero coverage can make a task-conditioned certificate weak or vacuous, but realized transport error is not a monotone function of raw spectral energy. Likewise, the public two-dimensional result is not an artifact of one perturbation-retention threshold. These controls support a geometric conditioning interpretation while ruling out simpler scalar energy and filtering explanations.

E Additional Calibration Ablations

The main text uses structured spectral correction as a diagnostic intervention: it tests whether modifying the response identified by the transport analysis changes the downstream packet trajectory. Here we examine three additional questions about that intervention. First, how does transport depend on the number of structured spectral observations? Second, does correcting packet center also improve the full packet state, or merely the center statistic? Third, does the correction substantially alter ordinary in-distribution rollouts?

These analyses refine the interpretation of the calibration result rather than introducing a separate prediction method. In particular, they show that a small number of probes can be effective without implying monotone improvement with probe count, that the phase/full corrections alter the packet state as well as its measured center, and that the same intervention has little effect on standard in-distribution rollout error.

E.1 Probe-Budget Dependence

The main released-advection experiments use six structured spectral observations to demonstrate that the diagnosed transport response is correctable. The choice of six should not be interpreted as an optimal budget. We therefore evaluate the same correction procedure using

$$0,\ 2,\ 4,\ 6,\ \text{and } 10$$

spectral observations while holding the trained predictor and packet evaluation protocol fixed.

For the plain FNO, the descriptive eight-step packet-center errors in the budget sweep are

$$12.17, \quad 2.86, \quad 1.96, \quad 2.35, \quad 2.47$$

Table 2: **Dependence of packet transport on the number of structured spectral observations.** Values are descriptive $h = 8$ center-error summaries from the official-advection probe-budget sweep for the plain FNO and should not be confused with the seed-cluster estimates used for the main inferential figures. A small number of probes produces a large reduction in transport error, but the improvement is not monotone: four probes perform best in this sweep, while six and ten are slightly worse.

Number of probes	$h = 8$ **center error**
0	12.17
2	2.86
4	1.96
6	2.35
10	2.47

Table 3: **Packet-state NRMSE under phase and full-complex correction on released advection.** Both interventions strongly improve the packet state at short horizon. At $h = 8$, substantial residual error remains despite the large reduction in packet-center drift, showing that sparse spectral correction does not reconstruct the entire out-of-distribution rollout.

Correction	$h = 1$ **NRMSE**	$h = 8$ **NRMSE**
Baseline	1.58	1.11
Phase-only	0.119	0.757
Full-complex	0.091	0.555

grid cells for 0, 2, 4, 6, and 10 probes, respectively. Table 2 reports the same descriptive sweep.

The large improvement between zero and a small number of probes supports the claim that the relevant transport response can be strongly constrained with limited task-aligned information. The subsequent non-monotonicity is equally important. Adding more probes does not guarantee further improvement, so these experiments do not support a "more anchors is always better" interpretation or a universal optimal probe count.

Several mechanisms could contribute to this behavior. The corrected quantity is a finite-amplitude effective response of a nonlinear FNO rather than an exact scalar Fourier multiplier; increasing the number of fitted locations can therefore expose irregularities that are not represented by a smooth diagonal correction. Phase unwrapping and interpolation geometry may also influence the result. The present experiments do not distinguish among these explanations. We consequently use the budget sweep only to establish that sparse structured information is sufficient in the tested setting and that its benefit saturates rather than increasing monotonically.

E.2 Packet-State Error under Component Corrections

Packet-center error isolates the physical observable studied in the paper, but a large reduction in center drift would be less informative if it arose from changing the center statistic while leaving the packet state essentially arbitrary. We therefore also measure full-field NRMSE on the out-of-distribution packet rollouts under the same component corrections.

The strongest change occurs at short horizon. At $h = 1$, the baseline packet-state NRMSE is approximately 1.58. Phase-only correction reduces it to 0.119, while full-complex correction reduces it further to 0.091. Thus the early transport repair is accompanied by a substantial improvement of the packet field, not merely by a shift of the circular-center statistic.

At $h = 8$, the baseline packet-state NRMSE is approximately 1.11. Phase-only and full-complex corrections reduce it to 0.757 and 0.555, respectively. The horizon-specific values are collected in Table 3. The improvement remains substantial, but the residual field error is much larger than at the first step.

Table 4: **Effect of component corrections on ordinary in-distribution rollouts.** At horizon 8, phase-only and full-complex corrections that substantially reduce packet-center drift change standard rollout NRMSE only slightly. The correction therefore acts strongly on the packet-transport task while having limited collateral effect on the distribution-weighted field metric in this released-advection setting.

Correction	**ID $h = 8$ NRMSE**
Baseline	0.2320
Amplitude-only	0.2334
Phase-only	0.2323
Full-complex	0.2350

The horizon dependence helps delimit what the intervention establishes. For phase-only correction, packet-center error is approximately

$$0.23, \quad 0.50, \quad 0.93, \quad 4.97$$

grid cells at horizons $1, 2, 4, 8$, respectively. The correction is therefore particularly effective at repairing the immediate transport response, while residual errors accumulate under repeated application of the nonlinear base model.

This pattern is consistent with the role assigned to the intervention in the main text. The phase correction targets one diagnosed component of the effective response; it is not expected to remove all amplitude errors, harmonic leakage, mode coupling, or compositional error. The remaining gap between phase-only and full-complex correction at long horizon is therefore evidence against a phase-exclusive account, even in the linear-advection task for which phase is the dominant component.

The large baseline packet-state error also reinforces an important scope statement: our main result is not that the same out-of-distribution packet appears accurate under ordinary full-field error while following the wrong trajectory. Rather, in-distribution accuracy fails to certify the packet transport response in the directions weakly constrained by standard prediction risk considered here.

E.3 In-Distribution Rollouts after Correction

A complementary concern is that improved packet transport might be obtained only by substantially changing predictions on the standard data distribution. We therefore apply the same post-hoc component corrections during ordinary in-distribution rollouts and measure the resulting field NRMSE.

At horizon eight, the seed-aggregated NRMSE values are

$$0.2320 \quad \text{(baseline)}, \qquad 0.2334 \quad \text{(amplitude-only)},$$
$$0.2323 \quad \text{(phase-only)}, \qquad 0.2350 \quad \text{(full-complex)}.$$

Table 4 summarizes this matched comparison.

The contrast with packet transport is large. At the same horizon, the corresponding packet-center errors for the baseline, amplitude-only, phase-only, and full-complex conditions are

$$12.754, \quad 10.882, \quad 4.992, \quad 3.254$$

grid cells. Thus phase-only and full-complex correction substantially alter the task-relevant trajectory while leaving the standard rollout metric nearly unchanged.

This result supports two interpretations. First, the observed repair is not obtained by globally changing the model until it behaves differently on ordinary trajectories. Second, improving the task-relevant response need not produce a corresponding gain in the standard evaluation metric when that response is weakly weighted by the in-distribution data.

The conclusion is specific to the released linear-advection experiment. We do not assume that a similar correction would have negligible in-distribution effects for nonlinear equations, broader spectral interventions, or other architectures. Nevertheless, within the setting where the phase mechanism is most clearly isolated, the separation between large transport improvement and small in-distribution change provides an additional empirical manifestation of the task-conditioned geometry developed in Sec. 5.

F NONLINEAR AND ARCHITECTURAL BOUNDARIES

The clean mechanism developed in Sec. 5 is intentionally narrow: it describes transport through a scalar Fourier response in a linear or locally linear regime. The broader experiments show that the transport-certification question extends beyond this setting, but they also reveal cases in which packet motion can no longer be interpreted through a single phase-slope error.

This section records those boundaries explicitly. Burgers dynamics introduce both physical perturbation damping and learned tangent amplification; architecture transfer is strongly equation dependent; and the scalar group-velocity diagnostic that is predictive in the clean benchmarks becomes much less informative under a natural non-band-limited input distribution. These results motivate separating the general certification problem from the specific phase-dominant mechanism established for clean transport.

F.1 BURGERS DYNAMICS AND PERTURBATION OBSERVABILITY

Nonlinear dissipative dynamics introduce a difficulty that is largely absent from constant-speed advection: the physical system can attenuate the perturbation used to define the transport observable. A small or poorly localized reference perturbation can make packet-center measurements unstable even when the learned model itself is not pathological. We therefore treat perturbation observability as part of the evaluation protocol rather than attributing every missing or unreliable packet to the learned operator.

Our first Burgers experiment uses the released PDEBench system at viscosity $\nu = 0.01$. Both the standard and activation-anti-aliased FNOs exhibit substantial four-step packet-center error, approximately

$$4.56 \quad \text{and} \quad 4.66 \quad \text{grid cells},$$

respectively. However, only 72.5% of the registered packet cases satisfy the reference-perturbation retention criterion. Inspection of the reference trajectories shows that viscosity itself strongly damps part of the high-frequency perturbation population.

This limits what can be inferred from the high-viscosity experiment. It supports the existence of transport discrepancies in a nonlinear dissipative system, but it does not provide a clean test of the same mechanism used for linear advection because the reference transport observable is not uniformly well conditioned.

We therefore repeat the experiment prospectively at lower viscosity, $\nu = 0.004$, keeping the relevant protocol choices fixed. Reference observability improves to 100% retention. The FNO and its anti-aliased counterpart retain multi-cell transport errors, with four-step center errors of approximately

$$2.89 \quad \text{and} \quad 2.87 \quad \text{grid cells}.$$

For the standard FNO, the final seed-cluster estimate is

$$2.896\,[2.194, 4.374] \quad \text{cells}.$$

The lower-viscosity result therefore supports the conclusion that the earlier Burgers behavior was not produced solely by loss of the reference perturbation. At the same time, the learned tangent response exhibits an additional failure mode, discussed next, that prevents Burgers from serving as clean evidence for a phase-only transport mechanism.

F.2 TANGENT AMPLIFICATION IN BURGERS

For nonlinear dynamics, a wrong packet center can arise together with incorrect perturbation amplitude, harmonic generation, and state-dependent mode coupling. To distinguish these effects from the approximately unit-modulus transport regime, we compare the energy of the learned finite-difference response with that of the corresponding reference perturbation.

The Burgers models exhibit substantial amplification. At $\nu = 0.004$, the learned-to-reference tangent-energy ratio has a median of approximately

$$1.59,$$

Table 5: **Perturbation observability and tangent amplification in Burgers dynamics.** Lowering the viscosity from $\nu = 0.01$ to $\nu = 0.004$ restores reliable reference-perturbation retention, but substantial learned tangent-amplitude amplification remains. The low-viscosity experiment therefore removes the strongest observability confound without converting Burgers into a clean phase-only transport setting.

Viscosity	Retention	Median energy ratio	90th percentile
$\nu = 0.01$	72.5%	≈ 3.85	≈ 323–413
$\nu = 0.004$	100%	≈ 1.59	≈ 7.64–8.55

while its 90th percentile lies between approximately

$$7.64 \quad \text{and} \quad 8.55$$

for the tested FNO variants. The distribution is heavy-tailed, with substantially larger rare values.

The effect is even more pronounced at $\nu = 0.01$. There the median ratio is approximately 3.85, while the 90th percentile reaches roughly 323–413; rare cases exceed 10^5. These extreme ratios occur in a regime where the physical reference perturbation is itself strongly damped, so the raw magnitude should not be interpreted as a stable architecture-wide amplification factor. Instead, it illustrates why center error alone is insufficient to assign a clean phase mechanism in this regime. Table 5 places the retention and amplification statistics side by side.

This behavior places Burgers outside the clean mechanism used for Corollary 2. In constant advection, the reference is approximately unit modulus and the center is locally sensitive to adjacent phase differences while pure amplitude changes vanish at first order. In Burgers, the learned and physical perturbation amplitudes can evolve very differently, and nonlinear mode coupling can affect both the packet shape and its center.

The supported conclusion is therefore not that Burgers wrong rays are caused by a scalar dispersion error. Rather, the Burgers experiments show that the broader task-certification problem persists in a nonlinear setting while the underlying response error can mix transport, amplification, and mode-coupling effects.

F.3 Architecture Dependence

If wrong-ray behavior were purely a consequence of Fourier layers, replacing the FNO with a spatial convolutional architecture should remove it across systems. Conversely, if the effect were architecture independent, different model families should fail on the same tasks. The periodic U-Net experiments support neither extreme.

On heterogeneous advection, both architectures exhibit large transport errors. The seed-cluster eight-step center errors are

$$20.260\,[17.934, 20.849] \quad \text{for FNO4},$$

and

$$24.526\,[20.508, 25.919] \quad \text{for the periodic U-Net}.$$

The U-Net therefore reproduces the transport discrepancy without any Fourier layer. Its horizon dependence is also approximately linear over the evaluated range, supporting the interpretation of a persistent characteristic-speed mismatch without requiring a global scalar Fourier symbol.

Low-viscosity Burgers gives the opposite result. The FNO has a four-step center error of

$$2.896\,[2.194, 4.374],$$

whereas the periodic U-Net has only

$$0.298\,[0.150, 0.882] \quad \text{cells}.$$

Only one of the five U-Net seeds satisfies the corresponding wrong-ray criterion in the original architecture-transfer evaluation. We therefore retain this result as a genuine negative case rather than averaging it into the positive heterogeneous- advection result.

Table 6: **Architecture dependence of packet transport.** Periodic U-Nets reproduce substantial transport error on heterogeneous advection and public two-dimensional Navier–Stokes, but not on low-viscosity Burgers. The phenomenon is therefore not specific to Fourier layers, yet neither is it architecture independent.

System	Model	Center error
Heterogeneous advection	FNO4	$h = 8:\ 20.260$
Heterogeneous advection	periodic U-Net	$h = 8:\ 24.526$
Burgers $\nu = 0.004$	FNO4	$h = 4:\ 2.896$
Burgers $\nu = 0.004$	periodic U-Net	$h = 4:\ 0.298$
2D Navier–Stokes	FNO2d	$h = 4:\ 6.267$
2D Navier–Stokes	periodic U-Net2d	$h = 4:\ 5.566$

The two-dimensional Navier–Stokes stress test again shows occurrence in both families: FNO2d and periodic U-Net2d have four-step center errors of 6.267 and 5.566 cells, respectively. Because their standard field errors are moderate, this comparison establishes cross-architectural occurrence in that setting rather than a hidden failure under near-perfect prediction. Table 6 summarizes these positive and negative architecture–equation pairs.

These comparisons suggest that architecture controls how an underconstrained response is extended beyond the directions emphasized by training, but the relevant inductive bias depends on the equation. The exact construction shows that under-identification can exist independently of architecture; the trained-model experiments show that whether it becomes a large realized transport error depends on both architecture and dynamics.

F.4 When Group-Velocity Diagnostics Break Down

The phase-slope diagnostic is designed for regimes in which a local scalar spectral response provides a meaningful description of packet transport. Its strongest empirical association therefore occurs in the clean advection and Schrödinger benchmarks.

Using the model–system–seed aggregates displayed in Fig. 3a, the Spearman association between the transport-specific velocity diagnostic and eight-step packet-center error is approximately

$$\rho_s \approx 0.87.$$

Restricting the comparison to the most nearly linear spectral models gives a similar association. Together with the horizon law and the phase/amplitude intervention, this supports the use of phase slope as a mechanism-aligned diagnostic in the clean transport regime.

The relationship becomes much weaker when the input distribution and learned response are less idealized. Under the non-band-limited PDEBench-generator protocol of Appendix C.3, the same type of scalar diagnostic has Spearman correlation only

$$\rho_s = 0.244$$

with realized packet-center error. This occurs even though the structured intervention reduces center error from 9.37 to 2.32 cells and reduces the associated transport-response discrepancy by 84.9%.

Thus a weak cross-case correlation does not imply that the corrected spectral response is irrelevant. Rather, it shows that one scalar summary is no longer sufficient to rank finite-amplitude packet errors when other sources of variation become important. Packet deformation, amplitude reweighting, nonlinear coupling, background dependence, and nonuniform response across the local band can all alter the realized center error without being represented by a single fitted phase slope.

Burgers provides an even clearer boundary. There, large tangent-amplitude variation makes it inappropriate to interpret the packet trajectory using phase slope alone. Similarly, the approximately linear horizon dependence observed for the heterogeneous-advection U-Net supports a persistent characteristic-speed discrepancy but cannot identify Fourier phase as its origin because that model does not admit the global scalar Fourier-diagonal description assumed by the clean theory.

The role of the group-velocity diagnostic is therefore regime dependent. In clean linear transport it is a useful task-aligned predictor and is mechanistically linked to packet-center motion. In broader

nonlinear or naturally distributed settings, it remains one component of the learned response but should not be treated as a universal scalar score of transport fidelity. The more general object is the task-conditioned response itself: whichever combination of phase, amplitude, coupling, and state dependence the downstream transport observable actually reads.

G Qualitative Transport Examples

The quantitative analyses in the main text summarize transport through packet-center error, spectral response, and seed-level statistics. Here we provide complementary examples that make the corresponding failure mode concrete at the level of individual packets and exact spectral responses. These examples are illustrative rather than additional statistical evidence: the inferential conclusions continue to rely on the seed-level analyses reported in the main text and Appendix H.

G.1 Representative Packet Trajectories

Figure 1b shows the horizon dependence of absolute packet-center error for the released PDEBench advection experiment. The underlying packet-level examples reveal a consistent qualitative pattern. At short horizon, many packets remain close to the reference location, but their center discrepancy grows over successive applications of the learned propagator. The resulting error is therefore not well described as an isolated terminal-time displacement: it develops progressively along the rollout, consistent with the persistent transport mismatch analyzed in Sec. 5.5.

For these released-data examples, the stored quantity is the absolute shortest periodic distance between learned and reference packet centers. Signed packet-center trajectories were not retained for the formal experiment, so we do not reconstruct a left- or right-directed displacement from the absolute errors. The qualitative examples should therefore be read as trajectories of *transport-error magnitude*, not as recovered signed space–time rays.

The representative packet-level trajectories used for visualization are drawn from the median model seed of the official-advection experiment. They are not treated as additional independent replicates; frequencies and packet instances belonging to the same trained model remain repeated measurements within that seed. The purpose of the visualization is instead to show the within-seed structure that is compressed by the seed-level summaries in the main figures.

Several features of these trajectories are useful for interpreting the aggregate results. First, large final center errors generally emerge through accumulation over multiple rollout steps rather than through one abrupt terminal failure. Second, different packet bands can accumulate error at different rates, consistent with the frequency dependence documented in Appendix C.2. Third, component correction changes the trajectory over the rollout rather than merely moving the final output: phase-only and full-complex correction reduce the center discrepancy already at early horizons, although residual error can compound again at later times.

This last behavior is also visible in the horizon-wise phase correction reported in Appendix E.2. For released advection, the phase-only center errors are approximately

$$0.23, \qquad 0.50, \qquad 0.93, \qquad 4.97$$

cells at horizons $1, 2, 4, 8$, respectively. Thus the correction substantially repairs the immediate propagation direction while not eliminating all later compositional error. This is consistent with the interpretation of the measured spectral response as one important component of a nonlinear learned propagator, rather than as a complete scalar description of its rollout.

We do not use these packet examples to claim that packet-center error is the only observable revealing out-of-distribution failure. In the official-advection packet experiment, the full packet-state NRMSE is already substantial at long horizon. Their qualitative value is instead to isolate the particular physical consequence studied throughout the paper: where the learned dynamics carry a localized disturbance over time.

G.2 Exact Spectral Counterexample

The exact construction provides the cleanest visual separation between observed-mode agreement and off-support transport. Figure 8 shows a representative construction for which the candidate

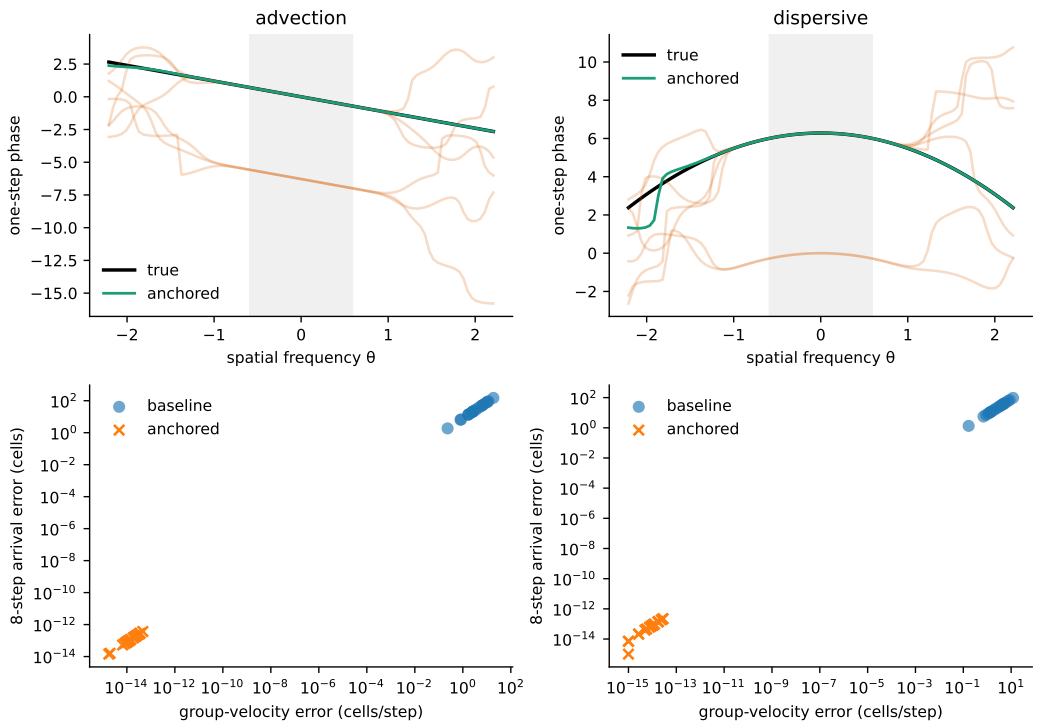


Figure 8: **Exact spectral agreement on the observed support can coexist with a wrong packet ray.** A representative exact linear construction is shown for the controlled transport setting. The candidate and reference propagators agree to numerical precision on the Fourier modes constrained by the observed data, so their validation error on that support is essentially zero. Outside the constrained support, their spectral responses differ, producing a different packet propagation trajectory and a large finite-horizon center displacement. This example establishes non-identifiability under restricted support; it is a constructive possibility result rather than evidence about the frequency with which training selects such an extension.

propagator agrees with the target on every Fourier mode constrained by the observed distribution but differs outside that support.

Across the exact advection and Schrödinger constructions, observed-support validation NRMSE is below 1.5×10^{-14}, while median eight-step packet-center errors are 19.31 and 18.46 grid cells, respectively. The visual example illustrates how these two statements can hold simultaneously: prediction error constrains the symbol only where the evaluation distribution supplies information, whereas a localized packet can depend on neighboring frequencies outside that set.

The transport difference is particularly interpretable in the linear setting because each Fourier component evolves independently. Agreement of the multipliers at the observed frequencies does not determine their extension to the remaining modes, and the packet trajectory depends on variation of spectral phase across the packet band. Two propagators can therefore be indistinguishable under the observed-support prediction metric while inducing different adjacent phase differences and hence different packet motion.

The construction is also exactly correctable once the missing response is specified. A small number of task-aligned spectral constraints determines the relevant part of the extension and collapses the packet-center discrepancy to numerical precision. A representative visualization of this intervention is shown in Fig. 9.

The exact example therefore serves a narrower role than the trained-model results in Sec. 4. It shows that the certification gap is mathematically possible even under essentially perfect observed-support prediction and that task-aligned information can resolve the ambiguity. The trained synthetic and

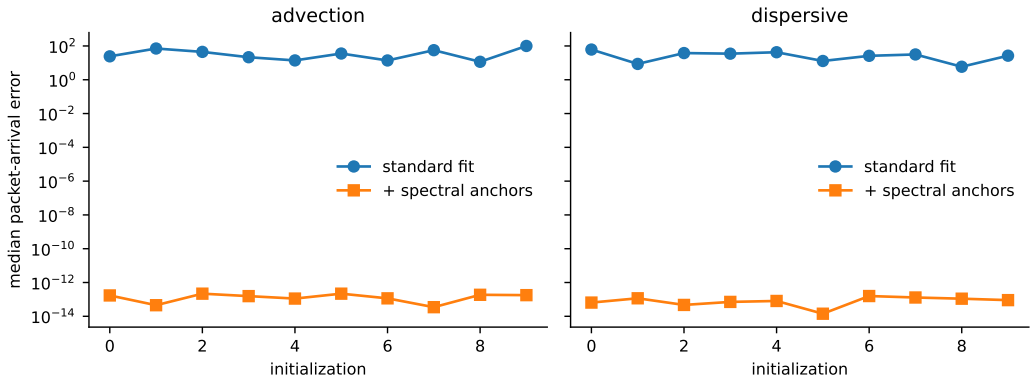


Figure 9: **Task-aligned spectral information resolves the exact ambiguity.** Starting from the exact support-limited construction in Fig. 8, additional spectral constraints are supplied in the transport-relevant region. The resulting propagator retains agreement on the originally observed support while its packet trajectory is brought into agreement with the reference to numerical precision. The intervention demonstrates correctability of this exact ambiguity; it does not imply that the same number or placement of observations is optimal for trained nonlinear models.

released-data experiments are needed separately to establish that consequential transport errors can remain after optimization rather than only under an explicitly constructed extension.

H Additional Quantitative Results

This section consolidates the quantitative estimates underlying the main empirical claims, together with the negative outcomes and provenance information needed to interpret them. The purpose is not to introduce additional evidence after the main analysis, but to make the statistical units, uncertainty estimates, boundary cases, and figure-level data sources explicit in one place.

H.1 Seed-Cluster Confidence Intervals

For trained-model experiments, the independent statistical unit is the model seed. Packet frequencies, backgrounds, horizons, and related probe cases are first reduced within seed, and uncertainty is then computed across seed-level summaries as described in Appendix B.6. Table 7 collects the principal seed-cluster estimates used throughout the paper.

Several distinctions in Table 7 are important for interpreting the main text. First, the component-correction comparison uses the same five official-advection checkpoints, so the reduction from baseline to phase-only and full-complex correction is evaluated on matched trained models. Second, the broadband-estimation comparison holds the additional supervision fixed: stabilized end-to-end fine-tuning retains large transport error, whereas direct spectral least squares reaches the same seed-cluster $h = 8$ error as the structured full-complex correction.

Third, the training-duration experiment uses only three independent training trajectories and follows each trajectory through the 400, 800, and 1600-step checkpoints. Its reported values are therefore seed-cluster estimates from those three trajectories. They should not be mixed with older descriptive pooled-case medians from the same raw outputs.

Finally, the Burgers U-Net interval is retained as an important negative boundary case rather than omitted from the quantitative summary. Its small center error contrasts with both the FNO on the same equation and the U-Net on heterogeneous advection, demonstrating that transport failure is not universal across architecture–equation pairs.

Table 7: **Seed-cluster estimates for the principal quantitative results.** Packet or trajectory cases are first aggregated within each independently trained model seed. The point estimate is the median across seed-level summaries, and 95% confidence intervals are obtained from $10{,}000$ bootstrap resamples of the model seeds. The intervals therefore quantify variation across independently initialized models, not across individual packet cases.

Setting	Quantity	Estimate [95% CI]	Seeds
Official advection	Baseline test NRMSE	0.0351 $[0.0334, 0.0393]$	5
Official advection	Baseline $h = 8$ center error	12.754 $[8.868, 13.449]$	5
Official advection	Amplitude-only $h = 8$ center error	10.882 $[8.852, 12.105]$	5
Official advection	Phase-only $h = 8$ center error	4.992 $[2.727, 5.898]$	5
Official advection	Full-complex $h = 8$ center error	3.254 $[0.988, 3.499]$	5
Official advection	Stabilized broadband test NRMSE	0.0366 $[0.0325, 0.0461]$	5
Official advection	Stabilized broadband $h = 8$ center error	23.468 $[10.233, 25.454]$	5
Official advection	Direct broadband LS $h = 8$ center error	3.254 $[0.988, 3.499]$	5
Aliasing control	Alias-free $h = 8$ center error	13.931 $[9.652, 14.506]$	5
Training audit	400-step test NRMSE	0.0361 $[0.0351, 0.0393]$	3
Training audit	800-step test NRMSE	0.0300 $[0.0282, 0.0311]$	3
Training audit	1600-step test NRMSE	0.0261 $[0.0257, 0.0276]$	3
Training audit	400-step $h = 8$ center error	12.205 $[8.868, 12.754]$	3
Training audit	800-step $h = 8$ center error	14.075 $[9.367, 15.262]$	3
Training audit	1600-step $h = 8$ center error	13.836 $[11.009, 14.497]$	3
Heterogeneous advection	FNO4 $h = 8$ center error	20.260 $[17.934, 20.849]$	5
Heterogeneous advection	periodic U-Net $h = 8$ center error	24.526 $[20.508, 25.919]$	5
Burgers $\nu = 0.004$	FNO4 $h = 4$ center error	2.896 $[2.194, 4.374]$	5
Burgers $\nu = 0.004$	periodic U-Net $h = 4$ center error	0.298 $[0.150, 0.882]$	5
Public 2D Navier–Stokes	FNO2d $h = 4$ center error	6.267 $[5.502, 6.668]$	5
Public 2D Navier–Stokes	periodic U-Net2d $h = 4$ center error	5.566 $[4.955, 6.733]$	5

H.2 Negative and Boundary Results

The experimental program contains several outcomes that constrain the final claims. We retain them explicitly because removing experiments that failed a secondary gate, produced a negative result, or were superseded by a cleaner comparison would give an incorrect impression of uniform support.

Table 8 summarizes the principal positive, negative, boundary, and superseded outcomes. A *HOLD* decision does not imply that every signal in the experiment was absent; rather, at least one prespecified criterion failed, and the result is therefore used only for the narrower conclusions supported by the observed measurements.

The negative outcomes are scientifically useful because they rule out several stronger versions of the story. In particular, the U-Net Burgers result directly contradicts a universal cross-architecture failure claim. The PDEBench-generator result contradicts treating group-velocity error as a universal scalar ranker. The non-monotone probe-budget result prevents interpreting structured calibration as a simple monotone function of label count. Finally, the continued validation improvement at 1600 steps prevents the longer-training experiment from being described as an asymptotic saturation test.

The superseded broadband experiment is retained for a different reason. Its original unit-weight fine-tuning comparator degraded ordinary field accuracy and therefore could not distinguish transport-specific failure from global model degradation. The subsequent validation-stabilized comparison removes that confound, and the direct Fourier least-squares estimator on the same six broadband observations further shows that the relevant linear information was present in those labels. Accordingly, only the later matched-supervision comparison is used for the main information-versus-estimation conclusion.

H.3 Figure and Data Provenance

All figures and tables in the final analysis are generated from retained raw quantitative outputs rather than from values digitized or reconstructed from earlier plots. Table 9 summarizes the principal data sources for the five main figures and the quantitative tables.

The figure-generation implementation and its reproduction entry point are maintained separately from the underlying experiment code. A source manifest records the SHA-256 hash, file size, schema, and row count of every quantitative input used by the paper-figure layer, while panel-

Table 8: **Negative, boundary, and superseded experimental outcomes.** The table preserves experiments that failed a pre-specified criterion or produced mixed evidence. These results delimit the claims in the main text: wrong rays are not universal across architectures or equations, group-velocity error is not a universal scalar diagnostic, sparse-calibration benefit is not monotone in probe count, and continued standard training was not shown to reach an optimization plateau.

Experiment	Outcome	Observed result	Interpretive boundary
Exact null-space construction	GO	Observed-support error at numerical precision with $19.31/18.46$-cell packet errors	Constructive non-identifiability result; not prevalence evidence
Paper benchmark smoke	HOLD	Sparse-repair criterion failed	Engineering-only seed; excluded from formal aggregation
Deep FNO without global residual	Engineering failure	NRMSE 0.0864; calibration changed center error $24.26 \to 25.34$	Underfit configuration; retained as an abandoned design rather than confirmatory evidence
Positive-tail / CROP extension	HOLD	Six of seven criteria passed; wrong-ray, horizon, and repair signals remained	Pooled DME reduction was 41.4%, below the aggregate representation-consistency criterion
PDEBench-generator protocol	HOLD	Center error $9.37 \to 2.32$; all seeds improved	Group-velocity/center correlation only $\rho = 0.244$; scalar diagnostic is not universal
Official PDEBench advection	HOLD	Wrong rays and 75.5–80.7% six-probe repair	Probe-budget monotonicity failed; four probes outperform six and ten
High-viscosity Burgers	HOLD	Wrong-ray criterion passed in all five seeds	Only 72.5% of cases retained because physical damping reduces perturbation observability
Low-viscosity Burgers	GO	Full reference retention and multi-cell FNO transport error	Tangent amplification remains heavy-tailed, preventing a clean phase-only interpretation
Periodic U-Net transfer	HOLD	Positive on heterogeneous advection but negative on Burgers	Establishes architecture–equation dependence rather than architecture universality
Original matched-label repair	HOLD / superseded	Structured probes outperform IID and initial broadband fine-tuning	Initial broadband comparator degraded standard test accuracy and is not used as definitive estimator evidence
Stabilized broadband fine-tuning	GO	Standard test accuracy preserved, but $h = 8$ center error remains 23.47	Supports failure of the tested fine-tuning procedure, not an asymptotic impossibility result
Public 2D Navier–Stokes	GO	FNO2d and U-Net2d both exhibit multi-cell $h = 4$ transport error	Standard NRMSE is moderate; this is an external-validity stress test, not a near-perfect-fit example
Phase/amplitude decomposition	GO	Phase-only accounts for 81.7% of the full center-error gain	Full-complex correction remains better; mechanism is phase-dominant, not phase-exclusive
Direct broadband spectral estimator	Structured calibration	Maximum target-transfer discrepancy 1.63×10^{-7}; same transport repair	Shows that pure-mode labels are not informationally unique in the noiseless linear setting
400/800/1600-step training audit	HOLD	Test and ID-rollout errors improve while ray error remains large	Validation is still improving; no claim of complete optimization saturation

level provenance records which source contributes to each visual element. Checkpoint hashes and external-data hashes are stored independently in the reproducibility archive.

Several provenance choices are important for interpreting the displayed results. First, packet/background/frequency cases are reduced within model seed before inferential uncertainty is computed. The figure pipeline therefore does not treat the large number of packet cases as independent replicates. Second, the training-duration panels use the seed-cluster summaries

$$12.205, \quad 14.075, \quad 13.836$$

rather than the older pooled-case medians from the same raw CSV. This keeps the statistical unit consistent with the rest of the main evidence.

Third, no signed one-dimensional packet-center trajectories were reconstructed after the fact. Where only absolute periodic center errors were retained, the figures and qualitative supplement use those stored quantities directly. Likewise, the direct broadband estimator has a frozen scalar maximum transfer discrepancy of 1.63×10^{-7}, but no independent per-mode transfer plot was fabricated when the corresponding per-mode visualization was not retained.

Finally, preliminary smoke runs and engineering-only configurations remain in the repository but are separated from the formal seed clusters. Their outputs are used only to document development history or negative outcomes and are never pooled into the uncertainty intervals in Table 7. This separation, together with the raw-source manifest and checkpoint hashes, ensures that the quantitative claims in the paper can be traced to the specific retained observations from which they were computed.

H.4 Seed-Cluster Confidence Intervals

Table 9: **Primary figure and table provenance.** Each paper artifact is linked to retained quantitative sources. The figure-generation manifest additionally records source hashes, byte sizes, schemas, and row counts.

Artifact	Primary source	Content
Fig. 1	`public_rescue_results/metrics.csv`; `reviewer_rescue_results/phase_component_packet_metrics.csv`	Official PDEBench advection accuracy, horizon-dependent packet error, and phase/amplitude/full-complex intervention
Fig. 2	`results/ray_metrics.csv`; `paper_results/model_metrics_1d.csv`; `public_rescue_results/metrics.csv`	Exact non-identifiability, trained synthetic models, and released-advection occurrence
Fig. 3	`paper_results/model_metrics_1d.csv`; `extension_results/frequency_horizon_metrics.csv`; `architecture_transfer_results/ray_metrics.csv`; component-correction packet outputs	Phase-slope diagnostic, horizon accumulation, heterogeneous-advection U-Net, and component intervention
Fig. 4	`paper_results/cross_resolution.csv`; `public_rescue_results/metrics.csv`; training-audit scalar and ray outputs	Representation-consistency control, activation anti-aliasing, and $400/800/1600$-step training comparison
Fig. 5	`stabilized_broadband_results/metrics.csv`; direct-estimator outputs	Validation-stabilized broadband fine-tuning and direct Fourier least-squares estimation on identical supervision
Table 10	Frozen decision records; `final_audit_results/seed_cluster_ci.csv`	Cross-system and cross-architecture evidence matrix
Table 7	`final_audit_results/seed_cluster_ci.csv`	Seed-level estimates and 95% bootstrap intervals
Table 8	Frozen decision records and engineering logs	Positive, HOLD, negative, failed, and superseded experimental outcomes

Table 10: **Transport fidelity across systems and architectures.** Standard accuracy is field NRMSE; transport is periodic packet-center error. "Clean" denotes a scalar phase-slope interpretation, while "mixed" indicates additional amplitude, heterogeneity, or mode-coupling effects.

System / data	Model	Standard acc.	Transport	Mechanism	Repair
Exact advection	25-tap linear conv.	val. $\leq 9.3 \times 10^{-15}$	$h = 8 : 19.31$	clean	yes
Trained advection	mini-FNO	NRMSE 0.0074	$h = 8 : 9.42$	clean	yes
Schrödinger	mini-FNO	NRMSE 0.0012	$h = 8 : 18.73$	clean	yes
Advection $\beta = 1$	FNO4	NRMSE 0.0351	$h = 8 : 12.75$	phase-dominant	yes
Heterogeneous advection	FNO4	NRMSE 0.0304	$h = 8 : 20.26$	mixed	—
Heterogeneous advection	periodic U-Net	NRMSE 0.0092	$h = 8 : 24.53$	mixed	—
Burgers $\nu = 0.004$	FNO4	NRMSE 0.0261	$h = 4 : 2.90$	mixed / ampl.	—
Burgers $\nu = 0.004$	periodic U-Net	NRMSE 0.0093	$h = 4 : 0.30$	not assessed	—
2D Navier–Stokes	FNO2d	NRMSE 0.198	$h = 4 : 6.27$	not assessed	—
2D Navier–Stokes	periodic U-Net2d	NRMSE 0.173	$h = 4 : 5.57$	not assessed	—