RADON SPECTRAL NEURAL OPERATOR: LEARNING PDE OPERATORS IN THE RADON–FOURIER DOMAIN

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard Fourier neural operators use frequency-wise multipliers in their linear spectral branches. We introduce the *Radon Spectral Neural Operator* (RSNO), a hybrid neural operator that combines multiscale spatial processing with structured global interactions in Radon–Fourier coordinates. The spatial branch aggregates fine- and coarse-resolution convolutional features, while the spectral branch applies orientation-dependent channel filtering and cross-orientation mixing within radial frequency shells. Their outputs are fused in each residual block, making multiscale aggregation and global spectral processing the two central architectural components. We prove an exact shellwise representation of the continuous branch, establish Sobolev stability, and show that different angular actions range from standard Fourier multipliers to nonlocal Fourier-integral structures. Across four PDE benchmarks spanning Darcy flow, Navier–Stokes forecasting, Helmholtz waves, and Burgers dynamics, RSNO delivers consistently strong performance against a broad set of neural-operator baselines. Ablations show that removing the Radon spectral branch increases mean error across all evaluated benchmark settings, whereas removing cross-orientation coupling produces substantially smaller changes. Together, these results support shellwise Radon–Fourier processing as a principled and analyzable global component within a multiscale neural-operator architecture.

1 INTRODUCTION

Neural operators learn maps from coefficients, initial conditions, or forcing to solutions of partial differential equations (PDEs). In a Fourier Neural Operator (FNO) (Li et al., 2021), each linear Fourier-multiplier branch transforms channels independently at each spatial frequency. We investigate an intermediate parameterization between frequency-diagonal interactions and unrestricted frequency coupling: mixing orientations while preserving frequency magnitude. This restriction concerns the linear spectral branch; nonlinear composition in the complete network can still create cross-frequency interactions.

Radon–Fourier coordinates make this structure explicit. The Radon transform integrates a field over hyperplanes indexed by orientation θ and offset t; Fourier transformation in t yields $\widehat{v}(\rho\theta)$ by the Fourier-slice theorem. Mixing orientations at a fixed frequency radius therefore permits orientation-wise communication while excluding interactions across frequency radii within the branch. This modeling assumption yields an analyzable shellwise structure, without presuming that general PDE solution operators preserve frequency magnitude.

Directional representations also connect naturally to Sobolev regularity. A ridge feature depends only on $\theta \cdot x - t$: θ specifies its direction of variation, while t locates a hyperplane along which the feature is constant. The Sobolev sandwich of He & Tian (2026) shows that the smallest norm of a ridge density representing a function bounds is bounded above and below by Sobolev norms of nearby smoothness orders. These orders coincide in the Hilbert setting, yielding equivalent descriptions of regularity. Thus, organizing a function by orientation and offset captures both its geometry and its smoothness. This connection motivates directional representations, while the Fourier-slice theorem exposes their frequency structure through $\widehat{v}(\rho\theta)$: θ identifies orientation and ρ identifies scale. Radon–Fourier

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.

coordinates therefore provide a natural setting for learning interactions across directions at a common scale, with shellwise mixing remaining a separate architectural choice.

RSNO combines a multiscale spatial branch with a Radon spectral branch. The spectral branch applies orientation-dependent channel filtering and cross-orientation mixing within shells, while the spatial branch preserves local interactions and extends them across coarser resolutions. This multiscale spatial branch complements the global shellwise interaction, and spatial processing together with nonlinear composition can accommodate dependencies beyond the shell-preserving branch. These complementary pathways define the hybrid architecture, while angular mixing is an extension whose marginal empirical effect is assessed separately.

Our contributions are as follows:

- **Hybrid architecture.** We introduce RSNO, which combines explicit fine- and coarse-resolution spatial aggregation with a shell-organized Radon–Fourier spectral branch in each residual block. Their latent updates are fused before activation and residual addition.
- **Spectral structure, analysis and implementation.** We formulate orientation-dependent channel filtering with an additional cross-orientation map, derive an exact continuous shellwise representation, Sobolev bounds, and wavefront characterizations under explicit assumptions. A Fourier-slice implementation uses polar sampling and compensated adjoint gridding without constructing a sinogram.
- **Empirical findings.** Experiments show a favorable parameter–accuracy tradeoff on Darcy, lower errors than the evaluated baselines on Navier–Stokes, and competitive Helmholtz and Burgers results. Ablations examine spatial aggregation, the Radon spectral branch, angular mixing, and radial truncation.

2 Background and Related Work

Operator learning in function spaces. Neural operators learn maps between function spaces with parameters shared across compatible discretizations (Kovachki et al., 2023). DeepONet separates input encoding from evaluation at query coordinates (Lu et al., 2021), while FNO learns Fourier multipliers on regular grids (Li et al., 2021) and factorized FNO uses separable spectral layers (Tran et al., 2023). RSNO extends the linear spectral interaction to directions on a common radial shell; its pointwise angular specialization remains a Fourier multiplier.

Multiscale, geometric, and latent representations. Multiwavelet operators, U-NO, LSM, and MgNO organize computation through wavelet bases, contracting–expanding architectures, latent hierarchies, and multigrid cycles, respectively (Gupta et al., 2021; Rahman et al., 2023; Wu et al., 2023; He et al., 2024). Geo-FNO learns coordinate deformations (Li et al., 2023b), while graph operators support irregular meshes and multiscale connectivity (Migus et al., 2022). RSNO combines a multiscale spatial branch with global interactions organized by orientation and radial frequency.

Attention-based operator learning. Galerkin Transformer, OFormer, and GNOT use attention for operator learning (Cao, 2021; Li et al., 2023a; Hao et al., 2023). Transolver aggregates points into physics-aware slices (Wu et al., 2024); LinearNO develops a linear-attention formulation (Hu et al., 2026). HPM combines domain-level spectral and point-level physical information (Yue et al., 2025), while continuum attention studies attention between function spaces (Calvello et al., 2025). Radon Neural Operator (RNO) (Lu et al., 2025) brings Radon-domain feature processing into a physics-attention backbone through input-dependent angular reweighting, filtered backprojection, and spatial fusion. RSNO instead follows the FNO-style spectral-operator construction: each block parameterizes an input-independent learned interaction between fixed Radon–Fourier analysis and synthesis operators, alongside the multiscale spatial branch. This distinguishes RNO's physics-attention backbone from RSNO's spectral blocks and enables the continuous linear-branch analysis in Appendix A.

Radon transforms and learned reconstruction. Hyperplane projection, Fourier-slice analysis, and filtered backprojection are classical tomography constructions (Deans, 1983; Helgason, 1999). Learned reconstruction methods, such as learned primal–dual algorithms, retain a known forward

model within an inverse solver (Adler & Öktem, 2018). RSNO uses Radon–Fourier coordinates as an internal representation of an available latent field, learning shellwise interactions between fixed analysis and synthesis operators.

3 Radon Spectral Neural Operator

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain, with $d \geq 2$, and let $\mathcal{A}$ and $\mathcal{U}$ be Banach spaces of input and output fields on Ω. Inputs and outputs may be vector-valued. Given training pairs $\{(a_i, u_i)\}_{i=1}^N$, where $u_i = \mathcal{G}(a_i)$ for a target operator $\mathcal{G} : \mathcal{A} \to \mathcal{U}$, we estimate the parameters by solving

$$\min_{\phi} \frac{1}{N} \sum_{i=1}^{N} \ell(\mathcal{G}_\phi(a_i), u_i)\,, \tag{1}$$

where $\ell : \mathcal{U} \times \mathcal{U} \to [0, \infty)$ denotes the training loss, and $\mathcal{G}_\phi$ is the learned neural operator parameterized by ϕ.

To instantiate $\mathcal{G}_\phi$, the network first lifts the input field to a latent field $h_0 : \Omega \to \mathbb{R}^C$, applies a sequence of residual operator blocks, and maps the final latent field to the output space. When a Radon spectral branch is applied, a latent channel is extended by zero outside Ω, yielding a function on $\mathbb{R}^d$. Figure 1 summarizes this realization of $\mathcal{G}_\phi$, including the complete architecture, the two-branch residual block, and the internal sequence of the Radon spectral branch developed below.

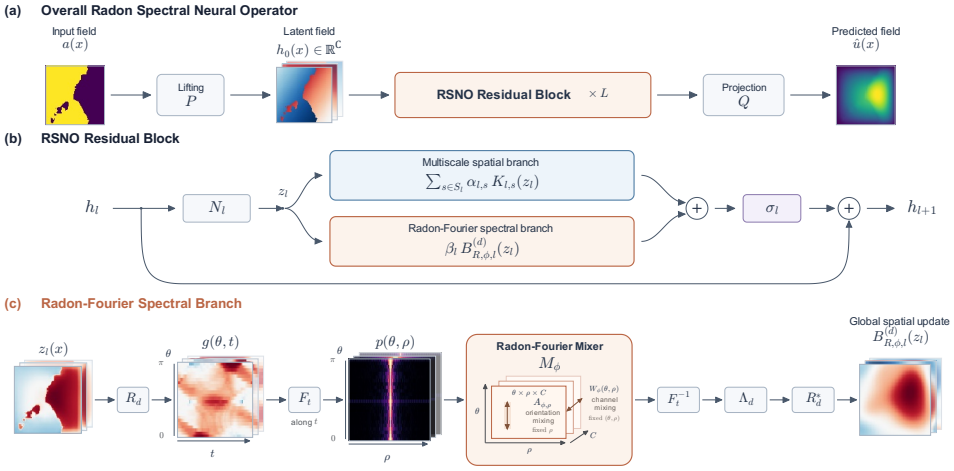

Figure 1: Overview of the Radon Spectral Neural Operator. (a) The input field is lifted to a latent representation, processed by a stack of RSNO residual blocks, and projected to the predicted field. (b) Each residual block combines the multiscale spatial branch and Radon spectral interactions before activation and residual addition. (c) The Radon spectral branch applies the hyperplane Radon transform, a Fourier transform in the signed-offset coordinate, the learned orientation–frequency mixer, inverse synthesis in the signed-offset coordinate, the dimension-aware filter, and adjoint backprojection.

3.1 Hyperplane Radon representation

Let $\mathbb{S}^{d-1}$ denote the set of oriented unit normals. This parametrization is redundant because (θ, t) and $(-\theta, -t)$ represent the same hyperplane. For a sufficiently regular latent channel $v_c : \mathbb{R}^d \to \mathbb{R}$, its hyperplane Radon transform is

$$(\mathcal{R}_d v_c)(\theta, t) = \int_{\mathbb{R}^d} v_c(x)\, \delta(t - x \cdot \theta)\, \mathrm{d}x = \int_{x \cdot \theta = t} v_c(x)\, \mathrm{d}\sigma_{d-1}(x), \qquad (\theta, t) \in \mathbb{S}^{d-1} \times \mathbb{R}, \tag{2}$$

where $\mathrm{d}\sigma_{d-1}$ is the induced measure on a $(d-1)$-dimensional hyperplane. Each coefficient therefore aggregates the field over a complete hyperplane, producing a nonlocal representation whose orientation is indexed by θ and whose signed offset (the detector coordinate) is t.

We apply a one-dimensional Fourier transform only along the signed-offset coordinate t:

$$p_c(\theta, \rho) = (\mathcal{F}_t \mathcal{R}_d v_c)(\theta, \rho) = \int_{\mathbb{R}} (\mathcal{R}_d v_c)(\theta, t) e^{-2\pi \mathrm{i} \rho t} \, \mathrm{d}t. \tag{3}$$

The d-dimensional Fourier-slice theorem gives

$$(\mathcal{F}_t \mathcal{R}_d v_c)(\theta, \rho) = \widehat{v}_c(\rho\theta), \tag{4}$$

so ρ is the signed detector frequency, with $|\rho|$ giving the frequency radius, while θ remains an explicit geometric orientation. The corresponding spectral redundancy is $p(-\theta, -\rho) = p(\theta, \rho)$; positive frequency radii and the full sphere provide independent coordinates.

3.2 Learned orientation–frequency interaction family

Let $p(\theta, \rho) \in \mathbb{C}^C$ denote the vector of latent channel coefficients at (θ, ρ). We define a learned mixer $\mathcal{M}_\phi$ acting on fields over $\mathbb{S}^{d-1} \times \mathbb{R}$ that first mixes channels pointwise in orientation–frequency coordinates and then, in the full model, applies an angular map at each fixed frequency radius. The pointwise-only case is a valid restricted interaction that removes this second map. Let $\mathcal{W}_{\phi,\rho}$ denote the pointwise channel operator, so that $\big(\mathcal{W}_{\phi,\rho} p(\cdot, \rho)\big)(\theta) = W_\phi(\theta, \rho) p(\theta, \rho)$. The learned action is

$$(\mathcal{M}_\phi p)(\theta, \rho) = \Big(\mathcal{A}_{\phi,\rho}\big[W_\phi(\cdot, \rho) p(\cdot, \rho)\big]\Big)(\theta), \tag{5}$$

where $W_\phi(\theta, \rho)$ is complex matrix-valued and $\mathcal{A}_{\phi,\rho}$ couples functions on $\mathbb{S}^{d-1}$. At fixed detector frequency ρ (equivalently, at fixed frequency radius under the redundancy), the features lie in the tensor-product space $\mathbb{C}^{N_\theta} \otimes \mathbb{C}^C$: $\mathcal{A}_{\phi,\rho}$ acts on the orientation factor, whereas $\mathcal{W}_{\phi,\rho}$ acts on the channel factor separately at each orientation. This fixed-frequency-radius action is precisely what preserves the shellwise decomposition: directions can communicate explicitly, while the linear branch does not move information between different radial frequencies. Because $W_\phi(\theta, \rho)$ generally depends on θ, their composition is generally not a single Kronecker product. The continuous angular map may be realized as the identity, an identity plus a dense kernel, a low-rank linear integral operator, or a harmonic parameterization on the sphere. In the finite-dimensional realization considered in this work, each retained nonzero radial shell is associated with a dense complex angular matrix that is shared across channels. The signed-frequency extension preserves the spectral redundancy, and reality compatibility ensures real-valued outputs; both conditions are specified in Appendix A.1.

3.3 Dimension-aware synthesis and filtered adjoint

For the hyperplane Radon transform in d dimensions, define the detector filter

$$\Lambda_d g = \mathcal{F}_t^{-1}\left[|\rho|^{d-1} \mathcal{F}_t g\right]. \tag{6}$$

The factor $|\rho|^{d-1}$ accounts for the radial Jacobian in d-dimensional Fourier coordinates and is therefore required for a formulation valid in arbitrary spatial dimension. The adjoint Radon transform is

$$(\mathcal{R}_d^* g)(x) = \int_{\mathbb{S}^{d-1}} g(\theta, x \cdot \theta) \, \mathrm{d}\theta, \tag{7}$$

with standard surface measure on $\mathbb{S}^{d-1}$. Combining the fixed transforms with the learned mixer gives the continuous Radon spectral branch

$$\boxed{\mathcal{B}^{(d)}_{\mathrm{R},\phi}(v) = c_d \mathcal{R}_d^* \Lambda_d \mathcal{F}_t^{-1} \mathcal{M}_\phi \mathcal{F}_t \mathcal{R}_d v,} \tag{8}$$

where $c_d = 1/2$ for the Fourier convention above and ordinary surface measure on the full sphere, as derived in Appendix A.1. The branch therefore follows hyperplane projection, Fourier analysis in the signed-offset coordinate, learned orientation–frequency interaction, inverse signed-offset synthesis, dimension-aware filtering, and adjoint backprojection. All trainable parameters are confined to $\mathcal{M}_\phi$, whereas $\mathcal{R}_d$, $\mathcal{F}_t$, $\mathcal{F}_t^{-1}$, Λ_d, and $\mathcal{R}_d^*$ are fixed operators determined by hyperplane geometry and the d-dimensional Radon inversion formula.

3.4 MULTISCALE RESIDUAL OPERATOR BLOCK

Each residual block combines a multiscale spatial branch with a Radon spectral branch. Let $z_\ell = \mathcal{N}_\ell h_\ell$ be the normalized latent input. The spatial branch aggregates convolutional interactions over a set of spatial scales $\mathcal{S}_\ell$:

$$\mathcal{B}_{\mathrm{S},\ell}(z_\ell) = \sum_{s\in\mathcal{S}_\ell} \alpha_{\ell,s}\mathcal{K}_{\ell,s}(z_\ell). \tag{9}$$

Here $\mathcal{K}_{\ell,s}$ denotes the complete scale-s process of downsampling, convolutional processing, and upsampling back to the input resolution. All spatial operations are contained in $\mathcal{K}_{\ell,s}$, and $\alpha_{\ell,s}$ weights its contribution. This aggregation is the model's explicit multiscale mechanism for local structure and coarse-scale dependencies, while the Radon spectral branch communicates global information through orientation–frequency coordinates.

The two branches return compatible latent fields and are combined as

$$h_{\ell+1} = h_\ell + \sigma_\ell\left(\mathcal{B}_{\mathrm{S},l}(z_l) + \beta_\ell \mathcal{B}^{(d)}_{\mathrm{R},\phi,\ell} z_\ell\right), \tag{10}$$

where β_ℓ is a learnable scalar coefficient weighting the Radon spectral branch. The spatial coefficients use $\alpha_{\ell,1} = 1$ and learnable $\alpha_{\ell,s}$ for $s > 1$. Here, σ_ℓ denotes the activation function in block ℓ. An identity skip connection carries h_ℓ directly past both branches and the activation σ_ℓ, adding it to the activated update to form a residual block. Thus the block combines two complementary mechanisms: the multiscale spatial branch and shell-organized global spectral interaction.

3.5 CONTINUUM PROPERTIES OF THE RADON SPECTRAL BRANCH

The construction in Equation (8) can be represented exactly in polar Fourier coordinates. Define $(\mathcal{P}v)(\theta, r) = \widehat{v}(r\theta)$ for $r > 0$, and let $\mathcal{M}_r$ denote the mixer acting at fixed frequency radius. For a scalar or vector z, write $\langle z\rangle = (1 + |z|^2)^{1/2}$. The principal consequences are summarized below; their proofs and full microlocal statements are given in Appendix A.

Theorem 3.1 (Shellwise structure and stability)**.** *Let $\mathcal{M}_r$ act on $L^2(\mathbb{S}^{d-1};\mathbb{C}^C)$. Assume that $r \mapsto \mathcal{M}_r$ is strongly measurable and $\|\mathcal{M}_r\| \le C_M\langle r\rangle^m$ for almost every $r > 0$. The signed-frequency action is understood to be extended from $r > 0$ by the Radon redundancy $q(\theta, -r) = q(-\theta, r)$, so redundancy-compatible fields remain redundancy-compatible. Then the continuous Radon spectral branch satisfies*

$$\mathcal{B}^{(d)}_{\mathrm{R},\phi} = \mathcal{P}^{-1}\mathcal{M}\mathcal{P}, \qquad \widehat{\mathcal{B}^{(d)}_{\mathrm{R},\phi} v}(r\theta) = \mathcal{M}_r[\widehat{v}(r\,\cdot)](\theta). \tag{11}$$

Consequently, the branch preserves radial spectral bands and commutes with every bounded radial multiplier $\psi(|D|)$. For each $s \in \mathbb{R}$,

$$\|\mathcal{B}^{(d)}_{\mathrm{R},\phi} v\|_{H^{s-m}} \le C_M\|v\|_{H^s}.$$

Pointwise multiplication in the angular variable, without cross-orientation mixing, gives a Fourier multiplier. Under the symbol assumptions in Appendix A, its wavefront relation lies in the diagonal, whereas angular transport can produce a non-diagonal canonical graph. Smooth angular kernels yield an equal-radius conormal relation away from the spatial origin. Non-pseudodifferential behavior requires a nonvanishing high-frequency amplitude at an off-diagonal point. These classifications concern specified continuum angular actions.

For a uniformly bounded mixer, the zero-extended Radon spectral branch on Ω inherits the corresponding L^2 bound. Radial-band invariance is a whole-space property and need not survive domain restriction or discrete gridding. These statements concern the continuous Radon spectral branch, which is linear; the complete network additionally contains nonlinearities and its implementation is a finite-dimensional approximation. The following experiments evaluate that finite realization.

4 EXPERIMENTS

We evaluate RSNO on four PDE benchmarks, including Darcy flow, Navier–Stokes (NS), Helmholtz, and Burgers, using relative L^2 error and root-mean-square error (RMSE). The benchmark suite

includes both parameter-to-solution and initial-state-to-trajectory maps. The continuous construction is stated for d-dimensional field domains; task-specific discretizations and temporal representations are described below.

4.1 Evaluation metrics

Let $\widehat{u}_i$ and u_i denote a prediction and its reference in physical units, and let $\mathcal{I}$ contain all evaluated output entries. For time-dependent tasks, $\mathcal{I}$ includes every predicted time and spatial point. We report the per-sample relative L^2 error and global RMSE:

$$e_i = \frac{\|\widehat{u}_i - u_i\|_2}{\|u_i\|_2}, \qquad \mathrm{RMSE} = \left[\frac{1}{N|\mathcal{I}|}\sum_{i=1}^{N}\|\widehat{u}_i - u_i\|_2^2\right]^{1/2}, \tag{12}$$

where N is the number of test samples. Relative L^2 is summarized as the mean $\pm$ standard deviation in percent; RMSE retains the physical output scale. Relative L^2 measures prediction error relative to the magnitude of each reference solution, whereas RMSE measures the absolute error magnitude in physical units across all evaluated output entries.

4.2 Datasets and experimental setup

Darcy flow. The piecewise-constant Darcy benchmark released with FNO (Li et al., 2021) maps a spatially varying permeability field to the steady pressure solution of an elliptic equation. We use an 85×85 grid with 1,000 training, 24 validation, and 200 test samples.

Navier–Stokes. The FNO vorticity dataset (Li et al., 2021) contains solutions of the forced incompressible Navier–Stokes equations on a two-dimensional periodic domain with viscosity $\nu = 10^{-5}$. The task predicts ten future 64×64 vorticity fields from ten observed frames, using 1,000 training and 200 test trajectories. We evaluate two autoregressive RSNO formulations: history-conditioned autoregressive RSNO (RSNO-H10), which uses a sliding ten-frame input window and a growing forecast horizon during training, and single-state autoregressive RSNO (RSNO-H1), which uses adjacent single-frame pairs. Specifically,

$$\widehat{\omega}_{t+1} = G_\phi^{\mathrm{H10}}(\omega_{t-9}, \ldots, \omega_t), \qquad \widehat{\omega}_{t+1} = G_\phi^{\mathrm{H1}}(\omega_t). \tag{13}$$

Both models are evaluated over the complete ten-frame forecast.

Helmholtz. The public wave-propagation dataset of de Hoop et al. (2022) maps a heterogeneous wave-speed field to the solution of a Helmholtz equation with prescribed boundary excitation. We retain the 101×101 grid and use 4,000 training, 200 validation, and 200 test pairs.

Burgers. The one-dimensional PDEBench dataset (Takamoto et al., 2022) describes viscous nonlinear transport with dataset parameter $\nu = 10^{-3}$ and diffusion coefficient ν/π. Each initial field is mapped to 40 future frames on 256 spatial points over $t \in [0, 2]$. We use 9,000 training and 1,000 test trajectories, and RSNO uses direct full-field prediction. Although the physical PDE is one-dimensional in space, RSNO applies its two-dimensional branch to the joint time–space representation. The initial field is replicated over the query-time positions before the operator blocks are applied.

All main benchmark results are from runs on a single NVIDIA RTX 5090 GPU. Within each task, methods use the same task-specific data split and evaluation protocol. We compare twelve baselines on Darcy and Navier–Stokes and seven on each of Helmholtz and Burgers. The ablations in Section 4.4 retain the corresponding main model's data and evaluation protocol. Detailed problem definitions and preprocessing are provided in Appendices C.1, D.1, E.1 and F.1; RSNO training details and baseline size/cost summaries are provided in Appendices C to F.

Hyperparameter policy. For Darcy and Navier–Stokes, the baseline papers report results on these benchmarks and provide public code. We implement the baselines using their released configurations, with the configurations and data-protocol adaptations retained in the experiment code. For Helmholtz

and Burgers, we use fixed configurations for every method, including RSNO, and perform no dataset-specific hyperparameter tuning. When a baseline has published settings for one of these latter datasets, we retain them where applicable; otherwise, we use a fixed configuration adapted from its public implementation to our data and prediction protocols.

4.3 BENCHMARK RESULTS

Darcy and Navier–Stokes results. As shown in Table 1, on Darcy, RSNO achieves the lowest observed relative L^2 error and RMSE, with approximately one ninth of HPM's trainable parameters. The result is therefore primarily a parameter–accuracy tradeoff rather than a large absolute accuracy gap; HPM remains the second-most accurate method. On Navier–Stokes, RSNO-H1 ranks first and RSNO-H10 ranks second on both error metrics, while LinearNO is the strongest baseline. The stronger H1 result may reflect that its single-step update more closely follows the local time evolution of the underlying dynamics. Parameter–error plots and qualitative Darcy predictions are provided in Appendix C.4, while qualitative NS predictions are shown in Figure 4.

Table 1: Darcy and Navier–Stokes comparison under the shared protocols. Bold and underlined errors mark the best and second-best entries within each benchmark.

	Darcy			Navier–Stokes		
Method	Params	Rel. L^2 (%)	RMSE (10^{-5})	Params	Rel. L^2 (%)	RMSE
FNO (2021)	2,368,001	0.9074 ± 0.4829	7.2416	926,517	14.3736 ± 5.3794	0.2219
Galerkin (2021)	2,220,829	0.9546 ± 0.4701	7.5133	862,049	14.6124 ± 5.1964	0.2244
MWT (2021)	9,807,873	0.7022 ± 0.3685	5.5558	71,970,809	18.6230 ± 6.3268	0.2845
GNOT (2023)	675,268	2.9456 ± 1.1177	21.6219	1,391,882	22.6613 ± 7.8134	0.3470
LSM (2023)	19,187,457	0.7022 ± 0.4302	5.9282	19,188,033	24.2856 ± 5.3750	0.3607
OFormer (2023)	2,556,321	1.3607 ± 0.5331	10.0227	1,851,361	18.8958 ± 7.5213	0.2949
U-NO (2023)	16,393,665	0.8297 ± 0.3735	6.3319	30,478,033	13.7822 ± 6.3162	0.2186
MgNO (2024)	678,093	0.4736 ± 0.2318	3.6469	2,922,508	12.5206 ± 5.0523	0.1950
Transolver (2024)	2,826,945	0.5461 ± 0.2530	4.1853	11,232,321	8.9482 ± 3.3339	0.1385
HPM (2025)	2,864,321	$\underline{0.4553 \pm 0.1947}$	$\underline{3.4216}$	11,306,817	7.9408 ± 2.9445	0.1232
RNO-code (2025)	2,815,731	0.7813 ± 0.3957	6.1413	11,205,619	19.9753 ± 7.3238	0.3080
RNO-paper (2025)[1]	—	0.510	—	—	8.94	—
LinearNO (2026)	1,766,145	0.5404 ± 0.2228	4.0277	3,377,921	$\underline{7.3088 \pm 3.6512}$	$\underline{0.1180}$
RSNO (ours)	**317,757**	$\mathbf{0.4426 \pm 0.1870}$	**3.3440**	RSNO-H10: **3,190,137** RSNO-H1: **3,189,849**	$\mathbf{6.3473 \pm 3.2344}$ $\mathbf{3.7623 \pm 2.1732}$	**0.1025** **0.0627**

Parameter counts use trainable real scalars and correspond to each benchmark-specific configuration; a complex scalar counts as two.

Helmholtz and Burgers results. Table 2 shows that RSNO ranks second on both error metrics for Helmholtz and Burgers. On Helmholtz, MgNO gives the lowest errors, while RSNO achieves competitive accuracy with fewer trainable parameters than MgNO. On Burgers, F-FNO achieves the lowest errors and RSNO ranks second on both metrics, using direct full-field prediction. These results show that RSNO remains competitive on both oscillatory elliptic and nonlinear transport benchmarks. RSNO training procedures, baseline network sizes and training durations, checkpoint rules, and timing scope are provided in Appendices E and F. Figures 5 and 6 show the highest-error RSNO predictions.

Table 2: Helmholtz and Burgers results under the protocols.

Helmholtz			Burgers		
Method	Rel. L^2 (%)	RMSE (10^{-4})	Method	Rel. L^2 (%)	RMSE
PCA-Net (2021)	3.6359 ± 1.7364	5.3638	FNO (2021)	6.3138 ± 4.4661	0.040887
DeepONet (2021)	15.2521 ± 4.6171	20.2781	U-NO (2023)	4.0748 ± 3.1075	0.032053
FNO (2021)	2.7809 ± 1.3283	4.3507	F-FNO (2023)	$\mathbf{2.8838 \pm 2.3778}$	**0.023087**
PARA-Net (2022)	15.6407 ± 4.1535	21.0343	VCNeF (2024)	4.8549 ± 2.0707	0.031147
U-NO (2023)	1.4757 ± 1.1410	3.1471	LNO (2024)	11.3850 ± 5.8909	0.062426
MgNO (2024)	$\mathbf{0.8315 \pm 0.7481}$	**1.8957**	Transolver (2024)	9.2095 ± 3.4328	0.057704
LinearNO (2026)	1.6237 ± 1.4035	2.6955	LinearNO (2026)	7.9638 ± 2.8522	0.051262
RSNO (ours)	$\underline{0.8878 \pm 0.7616}$	$\underline{1.9492}$	**RSNO (ours)**	$\underline{3.6489 \pm 1.4926}$	$\underline{0.030239}$

[1]RNO-code denotes our evaluation of the authors' released implementation under the shared protocol. RNO-paper reports the corresponding results quoted from Lu et al. (2025). The latter is shown for reference.

4.4 ABLATION STUDIES

We study three aspects of RSNO: sensitivity to architectural choices within NS, run-to-run sensitivity to random seeds for the reference NS configuration, and the relative roles of the Radon spectral branch and its interaction choices across benchmark formulations. Table 3 focuses on the multiscale spatial branch and related capacity choices, Table 4 reports the ten-seed reference-model study, and Table 5 compares the full Radon spectral branch with component removals across Darcy, NS, Helmholtz, and Burgers.

Architectural sensitivity in NS autoregressive prediction. Table 3 examines the sensitivity of the history-conditioned RSNO-H10 model to spatial scales, width, depth, angular resolution, and radial truncation, with and without global angular mixing.

Table 3: Architectural ablations for NS autoregressive prediction.

	Global angular mixing		No angular mixing	
Variant	Real params	Rel. L^2 (%) $\downarrow$	Real params	Rel. L^2 (%) $\downarrow$
Spatial scales $\{1, 2, 4\}$ (full)	3,190,137	6.3473 ± 3.2344	3,125,625	6.3811 ± 3.2213
Spatial scales $\{1\}$	2,894,185	11.2242 ± 5.6192	2,829,673	11.4857 ± 5.6401
Spatial scales $\{1, 2\}$	3,042,161	6.9311 ± 3.8560	2,977,649	6.9381 ± 3.4428
Width $32 \to 24$	1,823,137	7.0129 ± 3.5368	1,758,625	7.1720 ± 3.4250
Depth $8 \to 6$	2,393,267	7.2466 ± 3.5903	2,344,883	7.2929 ± 3.6191
Angles $24 \to 16$	2,236,793	6.6404 ± 3.5687	2,208,121	6.6694 ± 3.5518
Radial locations $8 \to 6$	2,385,273	6.0113 ± 2.9501	2,339,193	6.1564 ± 3.1977

The reported ablations show a larger improvement from expanding the multiscale spatial branch across scales than from adding cross-orientation mixing. Angular mixing consistently lowers the mean error in these runs. Reducing the radial locations from eight to six improves the mean error under both settings, whereas reducing width, depth, or the number of angles increases it. Appendix D.4 specifies the configuration changes.

Random-seed sensitivity. We repeated the reference RSNO-H10 run with seeds 1–9 alongside the original seed 0, keeping all training settings fixed. Table 4 reports means over 200 test trajectories.

Table 4: Random-seed sensitivity of the reference RSNO-H10 configuration on NS.

Seed	0	1	2	3	4	5	6	7	8	9	Mean $\pm$ SD
Rel. L^2 (%) $\downarrow$	6.3473	6.6782	6.5189	6.3904	6.4796	6.5692	6.3110	6.4569	6.5203	6.5892	6.4861 ± 0.1139
RMSE $\downarrow$	0.1025	0.1073	0.1073	0.1030	0.1049	0.1059	0.1030	0.1054	0.1068	0.1059	0.1052 ± 0.0018

Radon-branch and angular-mixing contributions. Table 5 compares the four benchmark configurations in Tables 1 and 2. No Radon spectral branch fixes the spectral residual coefficient to zero, while retaining the spatial branches. No angular mixing retains the Radon spectral branch and its orientation-wise channel mixing but removes the learned cross-orientation map. Parameter counts include trainable real scalars and exclude frozen parameters in the disabled branch, so these component-removal comparisons are not matched-capacity causal estimates.

Table 5: Radon-branch and angular-mixing ablations across RSNO settings.

	Full RSNO		No angular mixing		No Radon spectral branch	
Formulation	Real params	Rel. L^2 (%)	Real params	Rel. L^2 (%)	Real params	Rel. L^2 (%)
Darcy	317,757	0.4426 ± 0.1870	303,933	0.4455 ± 0.2033	73,529	0.6679 ± 0.2709
NS H10	3,190,137	6.3473 ± 3.2344	3,125,625	6.3811 ± 3.2213	373,105	7.3771 ± 3.6310
Helmholtz	5,306,905	0.8878 ± 0.7616	5,260,825	0.9647 ± 0.8937	837,137	3.0684 ± 1.7797
Burgers	939,259	3.6489 ± 1.4926	877,819	3.7705 ± 1.4918	109,813	10.1247 ± 5.0767

Table 5 separates the effects of the two removals. Removing only global angular mixing causes a modest increase in the mean relative L^2 error, whereas removing the entire Radon spectral branch causes a much larger degradation in every formulation. The full spectral branch therefore contributes substantially more to accuracy than the cross-orientation mixing component alone, subject to the parameter-count differences above.

These ablations provide an empirical counterpart to the proposed division of labor. The penalty from removing the Radon spectral branch supports a global representation that organizes orientation information by frequency radius, while the NS scale sweep shows a substantially larger change when the multiscale spatial branch is expanded across resolutions. They assess the representation and architectural components within the tested PDE operator-learning tasks.

4.5 Computational cost

The computational cost depends on the spatial dimension and resolution as well as on the latent width and spectral discretization. Let d denote the field dimension, let $H_1 \times \cdots \times H_d$ be the latent grid with volume $M = \prod_{j=1}^{d} H_j$, let C be the latent width, and let N_θ and N_r denote the retained orientations and radial locations. One RSNO spectral branch requires d-dimensional FFTs and inverse FFTs, costing $O(MC \log M)$, together with channel and optional cross-orientation mixing. Ignoring interpolation constants, its per-block spectral work is

$$O\big(MC \log M + (N_r - 1)N_\theta C^2 + (N_r - 1)N_\theta^2 C\big). \tag{14}$$

The last term is absent without angular mixing; d also affects the orientation discretization on $\mathbb{S}^{d-1}$.

For comparison, an FNO layer retaining K_j Fourier modes along coordinate j has spectral cost approximately $O(MC \log M + KC^2)$, where $K = \prod_j K_j$. Both methods share the FFT term. RSNO replaces FNO's mode-wise channel map with shellwise mixing over $N_r N_\theta$ samples and, when enabled, adds the quadratic angular term.

The multiscale spatial branch adds approximately $O(MC^2 \sum_{s \in \mathcal{S}_\ell} s^{-d})$ per block for scale factors s, up to resampling constants. The total cost scales linearly with the number of residual blocks.

Appendices B and C.2 give parameter counts and Tables 7 to 10 report measured training times.

5 Conclusion

RSNO is a two-branch neural operator that combines multiscale spatial aggregation with structured global processing in Radon–Fourier coordinates. Fine- and coarse-resolution convolutional paths provide explicit spatial aggregation, while the spectral pathway supplies orientation-dependent global filtering with an additional cross-orientation interaction. The continuous analysis establishes a shellwise representation and stability bounds, and identifies the wavefront geometry of several admissible angular actions under explicit assumptions.

Across the reported experiments, RSNO achieves the lowest observed Darcy errors, highlighting a parameter–accuracy tradeoff rather than a large absolute accuracy gap. On Navier–Stokes, both RSNO temporal formulations outperform the evaluated baselines under the common ten-frame forecast metric. RSNO achieves the second-lowest relative L^2 error and RMSE on both Helmholtz and Burgers. Across all four branch comparisons, disabling the Radon spectral branch increases the mean error, whereas removing only global angular mixing produces smaller changes. Together with the Navier–Stokes scale sweep, these component-removal results support the complementary roles of shell-organized global processing and multiscale spatial aggregation in the full RSNO architecture.

Limitations and future work. The current FFT and convolution implementation assumes regular Cartesian grids and has not been validated on irregular meshes. Existing methods offer several routes toward this extension: learned coordinate deformations (Li et al., 2023b), graph-based mappings between irregular meshes and regular latent grids (Li et al., 2023c), and direct spectral evaluation on nonuniform points (Lingsch et al., 2024). These approaches could be adapted to RSNO, but require consistent geometry handling across analysis, synthesis, and the multiscale spatial branch. Future work includes evaluations on additional PDE tasks, and irregular meshes.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and assistance with writing the paper, and for assistance with implementing the baseline code. We reviewed all AI-assisted text and code; the authors checked the implementations and the reported results for correctness. We take responsibility for the final content of this work, including all text, claims, and code produced with the aid of generative AI.

ETHICS STATEMENT

This work uses public numerical PDE benchmarks and involves no human subjects or personal data.

REPRODUCIBILITY STATEMENT

Our code is publicly available at `https://anonymous.4open.science/r/rsno-code/`. The anonymous repository includes all trained RSNO models used in this paper. Section 4 defines the benchmark tasks, splits, metrics, and evaluation protocols. RSNO architecture and training procedures are specified in Appendices B to F; each benchmark appendix also summarizes baseline network sizes, parameter counts, run times, and epochs. Baseline configurations are retained in the experiment code. Reported results are computed from saved test predictions or metrics for the complete training schedules.

REFERENCES

Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE Transactions on Medical Imaging*, 37(6):1322–1332, 2018.

Kaushik Bhattacharya, Bamdad Hosseini, Nikola B. Kovachki, and Andrew M. Stuart. Model reduction and neural networks for parametric PDEs. *The SMAI Journal of computational mathematics*, 7:121–157, 2021.

Edoardo Calvello, Nikola B. Kovachki, Matthew E. Levine, and Andrew M. Stuart. Continuum attention for neural operators. *Journal of Machine Learning Research*, 26(300):1–52, 2025.

Shuhao Cao. Choose a transformer: Fourier or Galerkin. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24924–24940, 2021.

Maarten V. de Hoop, Daniel Zhengyu Huang, Elizabeth Qian, and Andrew M. Stuart. The cost-accuracy trade-off in operator learning with neural networks. *Journal of Machine Learning*, 1(3): 299–341, 2022.

Stanley R. Deans. *The Radon Transform and Some of Its Applications*. John Wiley & Sons, New York, 1983.

Gaurav Gupta, Xiongye Xiao, and Paul Bogdan. Multiwavelet-based operator learning for differential equations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24048–24062, 2021.

Jan Hagnberger, Marimuthu Kalimuthu, Daniel Musekamp, and Mathias Niepert. Vectorized conditional neural fields: A framework for solving time-dependent parametric partial differential equations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 17189–17223, 2024.

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, pp. 12556–12569, 2023.

Juncai He and Zitong Tian. Sharp Sobolev sandwich and approximation rates of Radon-domain L^p ridge integral spaces for ReLU^k networks. *arXiv:2606.24795*, 2026.

Juncai He, Xinliang Liu, and Jinchao Xu. MgNO: Efficient parameterization of linear operators via multigrid. In *International Conference on Learning Representations*, 2024.

Sigurdur Helgason. *The Radon Transform*, volume 5 of *Progress in Mathematics*. Birkhäuser Boston, Boston, MA, 2nd edition, 1999.

Lars Hörmander. *The Analysis of Linear Partial Differential Operators IV: Fourier Integral Operators*, volume 275 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin, 1985.

Wenjie Hu, Sidun Liu, Peng Qiao, Zhenglun Sun, and Yong Dou. Transolver is a linear transformer: Revisiting physics-attention through the lens of linear attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 408–416, 2026.

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.

Zijie Li, Kazem Meidani, and Amir Barati Farimani. Transformer for partial differential equations' operator learning. *Transactions on Machine Learning Research*, 2023a.

Zongyi Li, Nikola Borislavov 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.

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, 2023b.

Zongyi Li, Nikola Kovachki, Chris Choy, Boyi Li, Jean Kossaifi, Shourya Otta, Mohammad Amin Nabian, Maximilian Stadler, Christian Hundt, Kamyar Azizzadenesheli, and Animashree Anandkumar. Geometry-informed neural operator for large-scale 3D PDEs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 35836–35854, 2023c.

Levi E. Lingsch, Mike Yan Michelis, Emmanuel De Bezenac, Sirani M. Perera, Robert K. Katzschmann, and Siddhartha Mishra. Beyond regular grids: Fourier-based neural operators on arbitrary domains. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 30610–30629, 2024.

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.

Wenbin Lu, Yihan Chen, Junnan Xu, Wei Li, Junwei Zhu, and Jianwei Zheng. Solving partial differential equations via Radon neural operator. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Leon Migus, Yuan Yin, Jocelyn Ahmed Mazari, and Patrick Gallinari. Multi-scale physical representations for approximating PDE solutions with graph neural operators. In *Proceedings of Topological, Algebraic, and Geometric Learning Workshops 2022*, volume 196, pp. 332–340, 2022.

Md Ashiqur Rahman, Zachary E. Ross, and Kamyar Azizzadenesheli. U-NO: U-shaped neural operators. *Transactions on Machine Learning Research*, 2023.

Christopher D. Sogge. *Fourier Integrals in Classical Analysis*, volume 210 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 2nd edition, 2017.

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

Alasdair Tran, Alexander Mathews, Lexing Xie, and Cheng Soon Ong. Factorized fourier neural operators. In *International Conference on Learning Representations*, 2023.

Tian Wang and Chuang Wang. Latent neural operator for solving forward and inverse PDE problems. In *Advances in Neural Information Processing Systems*, volume 37, pp. 33085–33107, 2024.

Haixu Wu, Tengge Hu, Huakun Luo, Jianmin Wang, and Mingsheng Long. Solving high-dimensional PDEs with latent spectral models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 37417–37438, 2023.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for PDEs on general geometries. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 53681–53705, 2024.

Xihang Yue, Yi Yang, and Linchao Zhu. Holistic physics solver: Learning PDEs in a unified spectral-physical space. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 73877–73896, 2025.

A Analysis of the Continuous Radon Spectral Branch

This appendix analyzes the continuous linear Radon spectral branch underlying Theorem 3.1. We establish its shellwise representation and stability properties, characterize the wavefront relations of the angular actions considered here, and state the scope of the bounded-domain and discretized constructions. We use the Fourier-slice inversion formula (Deans, 1983; Helgason, 1999) and standard local pseudodifferential and Fourier integral operator (FIO) calculus (Hörmander, 1985; Sogge, 2017; He & Tian, 2026).

A.1 Polar representation and real-valued fields

Let $d \geq 2$ be the spatial dimension and C the number of channels. We write $x, \xi \in \mathbb{R}^d$ for position and Fourier frequency, and $v : \mathbb{R}^d \to \mathbb{C}^C$ for a channel-valued field. Use the convention

$$\widehat{v}(\xi) = \int_{\mathbb{R}^d} v(x) e^{-2\pi \mathrm{i} x \cdot \xi} \, \mathrm{d}x, \qquad d \geq 2. \tag{15}$$

Here $\mathrm{i}^2 = -1$, $|\cdot|$ is the Euclidean norm, and $\mathbb{S}^{d-1} = \{\theta \in \mathbb{R}^d : |\theta| = 1\}$ is the unit sphere. The unit normal θ and signed offset t specify the hyperplane $x \cdot \theta = t$, with the representation $(\theta, t) \sim (-\theta, -t)$ on $\mathbb{S}^{d-1} \times \mathbb{R}$. Write $\mathcal{R}_d$ for the channelwise Radon transform and $\mathcal{F}_t$ for Fourier transformation in t, as in Equations (2) and (3). The detector frequency ρ is dual to t, and $p = \mathcal{F}_t \mathcal{R}_d v$ satisfies $p(-\theta, -\rho) = p(\theta, \rho)$. We use the independent coordinates $(\theta, r) \in \mathbb{S}^{d-1} \times (0, \infty)$ with $\xi = r\theta$, where $r = |\xi| > 0$ is the frequency radius; negative detector frequencies are recovered by this redundancy.

Let $\mathscr{H} = L^2(\mathbb{S}^{d-1}; \mathbb{C}^C)$ with ordinary surface measure. A polar field $q(\theta, r)$ assigns an angular field $q(\cdot, r) \in \mathscr{H}$ to each radius. Define

$$(\mathcal{P}v)(\theta, r) = \widehat{v}(r\theta), \qquad \mathscr{H}_{\mathrm{pol}} = L^2\big((0, \infty), r^{d-1}\mathrm{d}r; \mathscr{H}\big). \tag{16}$$

Polar integration and Plancherel give

$$\|v\|_{L^2}^2 = \int_0^\infty \|\mathcal{P}v(\cdot, r)\|_{\mathscr{H}}^2 r^{d-1} \, \mathrm{d}r. \tag{17}$$

Thus $\mathcal{P} : L^2(\mathbb{R}^d; \mathbb{C}^C) \to \mathscr{H}_{\mathrm{pol}}$ is unitary and, initially on smooth functions supported in a radial annulus,

$$(\mathcal{P}^{-1} q)(x) = \int_0^\infty \int_{\mathbb{S}^{d-1}} e^{2\pi \mathrm{i} r x \cdot \theta} q(\theta, r) r^{d-1} \, \mathrm{d}\theta \, \mathrm{d}r. \tag{18}$$

Let $\mathcal{L}(\mathscr{H})$ denote the bounded operators on $\mathscr{H}$, and let I and T^* denote the identity and Hilbert-space adjoint. Norms are induced by the corresponding Euclidean or L^2 norms, and $\operatorname{ess\,sup}$ is taken modulo null sets. The mixer is a strongly measurable family $\mathcal{M}_r \in \mathcal{L}(\mathscr{H})$, locally essentially bounded on $(0, \infty)$, with

$$(\mathcal{M}q)(\cdot, r) = \mathcal{M}_r q(\cdot, r), \qquad \mathcal{M}_r = \mathcal{A}_r \mathcal{W}_r, \qquad (\mathcal{W}_r q)(\theta) = W(\theta, r) q(\theta). \tag{19}$$

Here $W(\theta, r) \in \mathbb{C}^{C \times C}$ mixes channels pointwise before the input-independent angular action $\mathcal{A}_r \in \mathcal{L}(\mathscr{H})$. All identities involving radius-dependent families hold for almost every $r > 0$.

Theorem A.1 (Exact shellwise form)**.** *Write $\mathcal{B}$ for the continuous branch $\mathcal{B}^{(d)}_{\mathrm{R},\phi}$ in Equation (8), suppressing its dimension and learned parameters. With the signed-frequency mixer extended so that $q(\theta,-r)=q(-\theta,r)$ is preserved, the normalized continuous branch is*

$$\boxed{\mathcal{B}=\mathcal{P}^{-1}\mathcal{M}\mathcal{P}}, \qquad \widehat{\mathcal{B}v}(r\theta)=\mathcal{M}_r[\widehat{v}(r\,\cdot)](\theta). \tag{20}$$

It is defined on the natural dense domain

$$\mathcal{D}(\mathcal{B})=\{v\in L^2:\mathcal{M}\mathcal{P}v\in\mathscr{H}_{\mathrm{pol}}\}. \tag{21}$$

If $\operatorname{ess\,sup}_{r>0}\|\mathcal{M}_r\|<\infty$, it is bounded on all of L^2.

Proof. Fourier-slice analysis gives $\mathcal{F}_t\mathcal{R}_d v=\widehat{v}(\rho\theta)$. For any smooth annular q, extend it to negative frequencies by the stated redundancy. With ordinary surface measure and the Fourier convention above, the filtered synthesis is

$$\frac{1}{2}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}} e^{2\pi\mathrm{i}\rho x\cdot\theta}|\rho|^{d-1}q(\theta,\rho)\,\mathrm{d}\rho\,\mathrm{d}\theta=\int_{\mathbb{S}^{d-1}}\int_0^\infty e^{2\pi\mathrm{i}rx\cdot\theta}q(\theta,r)r^{d-1}\,\mathrm{d}r\,\mathrm{d}\theta. \tag{22}$$

Changing $\theta\mapsto-\theta$ identifies the negative- and positive-frequency terms, so the normalization in Equation (8) is $c_d=1/2$ and filtered synthesis equals $\mathcal{P}^{-1}$ for annular data, without requiring a smooth Radon-range condition on the mixed output. Density and annular truncation extend the identity to the natural domain in Equation (21), and this domain is dense by local essential boundedness; the norm bound follows from Equation (17). □

Real-valued outputs. Let $Jq(\theta)=\overline{q(-\theta)}$, where the bar denotes complex conjugation. Real-valued spatial fields correspond to $Jq=q$ for almost every radius. A complex-linear mixer preserves this real subspace if and only if

$$\mathcal{M}_rJ=J\mathcal{M}_r \quad \text{for almost every } r. \tag{23}$$

Every q admits a decomposition $q=q_1+\mathrm{i}q_2$ with $Jq_j=q_j$, so preservation of the fixed subspace is equivalent to commutation with J. For the pointwise matrix W and the composed angular kernel K in Equation (37), sufficient reality-compatibility conditions are $W(-\theta,r)=\overline{W(\theta,r)}$ and $K(-\theta,-\theta',r)=\overline{K(\theta,\theta',r)}$, in addition to the hyperplane redundancy. The discrete implementation constructs conjugate frequency partners and takes the real part after inverse Fourier synthesis; Equation (23) is not imposed as a separate training constraint.

A.2 Stability, radial invariance, and inversion

For a scalar or vector z, write $\langle z\rangle=(1+|z|^2)^{1/2}$. For $s\in\mathbb{R}$, the Sobolev space $H^s(\mathbb{R}^d;\mathbb{C}^C)$ consists of tempered distributions with finite norm $\|v\|_{H^s}^2=\int_{\mathbb{R}^d}\langle\xi\rangle^{2s}\|\widehat{v}(\xi)\|^2\,\mathrm{d}\xi$. We use $[T,S]=TS-ST$ for the commutator and $D=(2\pi\mathrm{i})^{-1}\nabla$, so $|D|$ has Fourier symbol $|\xi|$.

Theorem A.2 (Shellwise stability and radial commutation)**.** *Suppose there are constants $m\in\mathbb{R}$ and $C_M>0$ such that*

$$\|\mathcal{M}_r\|\le C_M\langle r\rangle^m \quad \textit{for almost every } r>0. \tag{24}$$

Then $\mathcal{B}$ extends uniquely to a bounded map

$$\mathcal{B}:H^s(\mathbb{R}^d;\mathbb{C}^C)\to H^{s-m}(\mathbb{R}^d;\mathbb{C}^C), \qquad \|\mathcal{B}v\|_{H^{s-m}}\le C_M\|v\|_{H^s} \tag{25}$$

for every real s. For every bounded Borel function $\psi:[0,\infty)\to\mathbb{C}$, the radial Fourier multiplier $\psi(|D|)$ satisfies

$$[\mathcal{B},\psi(|D|)]=0, \qquad \widehat{\psi(|D|)v}(\xi)=\psi(|\xi|)\widehat{v}(\xi). \tag{26}$$

In particular, radial spectral bands are invariant.

Proof. For annularly supported Fourier data, Equation (20) gives

$$\|\mathcal{B}v\|_{H^{s-m}}^2=\int_0^\infty\langle r\rangle^{2(s-m)}\|\mathcal{M}_r\mathcal{P}v(\cdot,r)\|_{\mathscr{H}}^2r^{d-1}\,\mathrm{d}r \tag{27}$$

$$\le C_M^2\int_0^\infty\langle r\rangle^{2s}\|\mathcal{P}v(\cdot,r)\|_{\mathscr{H}}^2r^{d-1}\,\mathrm{d}r=C_M^2\|v\|_{H^s}^2. \tag{28}$$

Density gives Equation (25). In polar coordinates, $\psi(|D|)$ is scalar multiplication by $\psi(r)$ and commutes with $\mathcal{M}_r$. Taking $\psi=\mathbf{1}_E$ for a Borel set $E\subset[0,\infty)$ gives invariance of the band $\{\xi:|\xi|\in E\}$; for unbounded multipliers the identity holds where both compositions are defined. □

Schur bound for angular mixing. Let $K_A(\theta, \theta', r)$ be the matrix-valued integral kernel of $\mathcal{A}_r - I$, so that $(\mathcal{A}_r - I)q(\theta) = \int_{\mathbb{S}^{d-1}} K_A(\theta, \theta', r)q(\theta')\,\mathrm{d}\theta'$. When the two angular integrals below are finite, the matrix-kernel Schur estimate gives

$$\|\mathcal{M}_r\| \le \big(1 + \sqrt{a(r)b(r)}\big)\operatorname*{ess\,sup}_{\theta} \|W(\theta, r)\|, \tag{29}$$

where $a(r) = \operatorname{ess\,sup}_\theta \int \|K_A(\theta, \theta', r)\|\,\mathrm{d}\theta'$ and $b(r) = \operatorname{ess\,sup}_{\theta'} \int \|K_A(\theta, \theta', r)\|\,\mathrm{d}\theta$. Apply the scalar Schur test to the pointwise bound $\|(\mathcal{A}_r - I)q(\theta)\| \le \int \|K_A(\theta, \theta', r)\|\|q(\theta')\|\,\mathrm{d}\theta'$ and then use $\mathcal{M}_r = \mathcal{A}_r\mathcal{W}_r$.

Corollary A.3 (Fiberwise properties and the inverse)**.** *Here "fiberwise" refers to the angular space at almost every fixed radius. In particular, self-adjointness means $\mathcal{M}_r^* = \mathcal{M}_r$ and positivity means $\langle \mathcal{M}_r h, h\rangle_{\mathscr{H}} \ge 0$ for all $h \in \mathscr{H}$. Under uniform boundedness, these properties transfer to $\mathcal{B}$ on L^2, while unitarity preserves every H^s norm. If, in addition, each $\mathcal{M}_r$ is bijective almost everywhere and $\operatorname{ess\,sup}_r \|\mathcal{M}_r^{-1}\| \le C_I$ for some $C_I > 0$, then*

$$\mathcal{B}^{-1} = \mathcal{P}^{-1}\mathcal{M}^{-1}\mathcal{P}, \qquad \|\mathcal{B}^{-1}v\|_{H^s} \le C_I\|v\|_{H^s}. \tag{30}$$

Proof. Unitary conjugacy gives the first claims. Under the stated uniform bounds, a uniformly convergent Neumann series for $(\mathcal{M}_r^*\mathcal{M}_r)^{-1}\mathcal{M}_r^*$ gives strong measurability of the inverse family. Applying the preceding estimate with $m = 0$ proves Equation (30). □

A.3 Pointwise mixing and angular transport

The wavefront statements below assume the indicated frequency regularity in addition to shellwise boundedness. We use a smooth radial cutoff that vanishes near zero and equals one for large r; bounded low-frequency terms have smooth physical kernels and do not change the local relations.

Symbol and wavefront notation. Multi-indices α, β, γ count derivatives, and j is a nonnegative integer. In local charts on the sphere, ∇_θ^α denotes angular derivatives. We use the standard symbol class $S^m_{1,0}$ of smooth matrix-valued functions $b(x, \xi)$ satisfying

$$\|\partial_x^\alpha \partial_\xi^\gamma b(x, \xi)\| \le C_{\alpha\gamma}\langle\xi\rangle^{m-|\gamma|},$$

locally uniformly in x, and $\Psi^m = \Psi^m_{1,0}$ for the corresponding pseudodifferential operators. Radial symbols satisfy analogous estimates with each r-derivative lowering the order by one. A classical radial symbol has an asymptotic expansion in powers r^{m-j} with smooth angular coefficients and a remainder of order $m - N$ after N terms.

For a distribution u, $\mathrm{WF}(u)$ identifies the position–covector pairs at which u is not microlocally smooth. For a kernel $\mathcal{K}_T(x, y)$, with x and y the output and input positions, define its twisted wavefront relation by

$$\mathrm{WF}'(\mathcal{K}_T) = \{(x, \xi; y, \eta) : (x, y; \xi, -\eta) \in \mathrm{WF}(\mathcal{K}_T)\}.$$

Here ξ and η are output and input covectors. For a nondegenerate homogeneous phase $\Phi(x, y, \zeta)$, the associated FIO relation is generated by $(x, \partial_x\Phi; y, -\partial_y\Phi)$ on $\partial_\zeta\Phi = 0$. A canonical graph gives a local correspondence between input and output covectors; "modulo smoothing" means up to an operator with a smooth kernel.

Proposition A.4 (Pointwise mixing)**.** *If $\mathcal{A}_r = I$, denote the resulting branch by $\mathcal{B}_W$. It is the Fourier multiplier*

$$\widehat{\mathcal{B}_W v}(\xi) = w(\xi)\widehat{v}(\xi), \qquad w(\xi) = W(\xi/|\xi|, |\xi|), \quad \xi \ne 0. \tag{31}$$

If, for $r \ge 1$ and all α, j,

$$\|\partial_r^j \nabla_\theta^\alpha W(\theta, r)\| \le C_{\alpha j} r^{m-j}, \tag{32}$$

its high-frequency part lies in Ψ^m and its wavefront relation is contained in

$$C_{\mathrm{diag}} = \{(x, \xi; x, \xi) : \xi \ne 0\}. \tag{33}$$

Proof. The multiplier formula is Equation (20) with no angular coupling. Every Cartesian frequency derivative is a combination of a radial derivative and r^{-1} times angular derivatives. Thus the cut-off symbol satisfies $\|\partial_\xi^\gamma w\| \lesssim \langle\xi\rangle^{m-|\gamma|}$, which is the $S^m_{1,0}$ condition. The phase $(x - y)\cdot\xi$ gives the wavefront inclusion in Equation (33). □

For a non-pointwise example, let $\Gamma : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ be a smooth diffeomorphism and set

$$\kappa(\xi) = |\xi|\Gamma(\xi/|\xi|), \qquad (\mathcal{M}_r q)(\theta) = a(\theta, r)q(\Gamma(\theta)). \tag{34}$$

Here $a(\theta, r)$ is a scalar or $C \times C$ matrix amplitude. Since Γ is a smooth diffeomorphism of $\mathbb{S}^{d-1}$, κ extends to a smooth degree-one homogeneous diffeomorphism of $\mathbb{R}^d \setminus \{0\}$ that preserves $|\xi|$; consequently, $D\kappa(\xi)$ is invertible for $\xi \neq 0$. Write $\mathcal{B}_{\Gamma,a}$ for the resulting spatial branch and $D\kappa(\xi)$ for the Jacobian matrix of κ; the superscript T denotes transpose. Composing angular transport after $\mathcal{W}_r$ replaces $a(\theta, r)$ by $a(\theta, r)W(\Gamma(\theta), r)$.

Theorem A.5 (Angular transport)**.** *Suppose a is a classical radial symbol of order m, smooth in θ and cut off near zero. Then $\mathcal{B}_{\Gamma,a}$ is an FIO associated with the canonical graph*

$$C_\Gamma = \{(x, \xi; y, \eta) : \xi \neq 0,\ \eta = \kappa(\xi),\ x = D\kappa(\xi)^T y\}. \tag{35}$$

Its wavefront relation is contained in this graph, and $|\eta| = |\xi|$. If $\Gamma(\theta_0) \neq \theta_0$ and the leading homogeneous amplitude is nonzero at θ_0, it is not pseudodifferential modulo smoothing near the corresponding off-diagonal canonical points.

Proof. Fourier inversion yields the oscillatory kernel

$$\mathcal{K}_{\Gamma,a}(x, y) = \int_{\mathbb{R}^d} e^{2\pi \mathrm{i}(x\cdot\xi - y\cdot\kappa(\xi))} a(\xi/|\xi|, |\xi|)\, \mathrm{d}\xi. \tag{36}$$

The phase $\Phi = x \cdot \xi - y \cdot \kappa(\xi)$ is homogeneous, and its critical equations are $x - D\kappa(\xi)^T y = 0$. Their differentials are independent because their derivatives with respect to x form the identity. On the critical set, $(\partial_x \Phi, -\partial_y \Phi) = (\xi, \kappa(\xi))$, which gives Equation (35). Invertibility of $D\kappa$ makes it a canonical graph. A nonzero leading amplitude gives a nonzero microlocal kernel there; if $\Gamma(\theta_0) \neq \theta_0$, then $\eta \neq \xi$, whereas a pseudodifferential kernel has relation contained in the diagonal. □

For example, $\Gamma(\theta) = Q^T\theta$ with $Q \in SO(d) = \{Q \in \mathbb{R}^{d\times d} : Q^TQ = I,\ \det Q = 1\}$, the rotation group, and $a = 1$ (the channel identity) gives $\mathcal{B}v(x) = v(Q^T x)$ exactly when the cutoff is omitted. With the high-frequency cutoff, the difference has a smooth kernel. For real fields, sufficient compatibility conditions for this subclass are $\Gamma(-\theta) = -\Gamma(\theta)$ and $a(-\theta, r) = \overline{a(\theta, r)}$.

A.4 Smooth angular kernels and the equal-radius relation

Consider an identity-plus-integral angular map applied after channel mixing:

$$\begin{aligned}(\mathcal{M}_r q)(\theta) &= W(\theta, r)q(\theta) + \int_{\mathbb{S}^{d-1}} K(\theta, \theta', r)q(\theta')\, \mathrm{d}\theta',\\ K(\theta, \theta', r) &= K_A(\theta, \theta', r)W(\theta', r).\end{aligned} \tag{37}$$

Write $\mathcal{B}_K$ for the integral contribution. We assume a smooth continuum angular kernel K, vanishing near $r = 0$, with the radial symbol estimates

$$\|\partial_r^j \nabla_\theta^\alpha \nabla_{\theta'}^\beta K(\theta, \theta', r)\| \le C_{\alpha\beta j} r^{m-j}, \qquad r \ge 1. \tag{38}$$

In the classical case, the expansion takes the form

$$K(\theta, \theta', r) \sim \sum_{j=0}^{\infty} r^{m-j} K_{m-j}(\theta, \theta'), \tag{39}$$

where K_m is the leading angular coefficient. This expansion is used for the classicality and principal-amplitude conclusions.

Theorem A.6 (Equal-radius conormal structure)**.** *Under Equation* (38)*, the kernel of $\mathcal{B}_K$ is locally conormal to $Z = \{(x, y) : |x| = |y| > 0\}$ on $(\mathbb{R}^d \setminus \{0\}) \times (\mathbb{R}^d \setminus \{0\})$: its singularities are described locally by oscillatory integrals in the normal direction to Z. Its twisted wavefront relation is contained in*

$$C_{\mathrm{sph}} = \left\{ \left(x, \lambda \frac{x}{|x|}; y, \lambda \frac{y}{|y|} \right) : |x| = |y| > 0,\ \lambda \neq 0 \right\}. \tag{40}$$

Here $\lambda \in \mathbb{R} \setminus \{0\}$ is the common signed frequency of the two covectors. Under Equation (39)*, the kernel is a classical FIO kernel on this region.*

Proof. Substitute Fourier inversion into Equation (37):

$$\mathcal{K}_K(x,y) = \int_0^\infty \int_{\mathbb{S}^{d-1}} \int_{\mathbb{S}^{d-1}} e^{2\pi \mathrm{i} r(x\cdot\theta - y\cdot\theta')} K(\theta,\theta',r) r^{d-1}\,\mathrm{d}\theta'\,\mathrm{d}\theta\,\mathrm{d}r. \tag{41}$$

On a compact set with $x, y \neq 0$, the angular stationary points are $\theta = \sigma x/|x|$, $\theta' = \sigma' y/|y|$, with $\sigma, \sigma' \in \{\pm 1\}$. Their Hessians are $-\sigma|x| I_{d-1}$ and $\sigma'|y| I_{d-1}$, where I_{d-1} is the $(d-1)\times(d-1)$ identity matrix, hence nondegenerate. Parameter-dependent stationary phase reduces the kernel, modulo a smooth remainder, to

$$\sum_{\sigma,\sigma'\in\{\pm 1\}} \int_0^\infty e^{2\pi \mathrm{i} r(\sigma|x| - \sigma'|y|)} b_{\sigma\sigma'}(x,y,r)\,\mathrm{d}r. \tag{42}$$

The amplitudes $b_{\sigma\sigma'}$ include the smooth low-frequency cutoff. Each is a symbol of order m: the two factors $r^{-(d-1)/2}$ from stationary phase cancel r^{d-1}. They are classical when K is classical. For equal signs their leading coefficients are

$$b_{\sigma\sigma}(x,y,r) = r^m (|x||y|)^{-(d-1)/2} K_m\left(\sigma\frac{x}{|x|}, \sigma\frac{y}{|y|}\right) + O_S(r^{m-1}), \tag{43}$$

where $O_S(r^{m-1})$ denotes a remainder whose jth radial derivative is $O(r^{m-1-j})$, also after any spatial derivatives on compact sets away from $x = 0$ and $y = 0$. For opposite signs, $|\sigma|x| - \sigma'|y|| = |x| + |y|$ is bounded away from zero, so repeated integration by parts in r gives a smooth kernel. For equal signs the phase is $\sigma r(|x| - |y|)$, the standard conormal phase for Z. Differentiating $F(x,y) = |x| - |y|$ and reversing the sign of the input covector gives Equation (40). □

Corollary A.7 (Off-diagonal wavefront points)**.** *Assume Equation* (39)*. For directions $\theta_0, \theta_1 \in \mathbb{S}^{d-1}$, if $K_m(\theta_0,\theta_1) \neq 0$ for $\theta_0 \neq \theta_1$, then the wavefront relation contains*

$$(R\theta_0, \lambda\theta_0; R\theta_1, \lambda\theta_1), \qquad R > 0, \quad \lambda > 0. \tag{44}$$

Here R is the common spatial radius. Hence $\mathcal{B}_K$ is not pseudodifferential modulo smoothing near these points. The negative ray has the same property when $K_m(-\theta_0,-\theta_1) \neq 0$, in particular under the reality-compatibility conditions.

Proof. The positive conormal amplitude in Equation (43) is nonzero at $(x,y) = (R\theta_0, R\theta_1)$. At least one matrix entry therefore has a nonzero principal conormal symbol. The corresponding point belongs to its wavefront set and lies outside C_{diag}. The negative ray uses the antipodal leading coefficient. □

The relation C_{sph} couples equal-radius base points with radial covectors of the same signed magnitude. For $\mathcal{B}_W + \mathcal{B}_K$, the wavefront relation is contained in $C_{\mathrm{diag}} \cup C_{\mathrm{sph}}$ on the stated region; this union need not be a single smooth canonical relation.

A.5 Localized wavefront statements

For smooth compactly supported output and input cutoffs χ_o, χ_i, the operator $T_\chi = \chi_o \mathcal{B} \chi_i$ is properly supported. In the smooth-angular-kernel case, assume both supports avoid the spatial origin. The standard FIO wavefront theorem gives

$$\mathrm{WF}(T_\chi v) \subset \mathcal{C} \circ \mathrm{WF}(\chi_i v), \tag{45}$$

with $\mathcal{C} = C_{\mathrm{diag}}$, C_Γ, or C_{sph} for the three respective cases, restricted to the cutoff supports. For a set E of input position–covector pairs, relational composition means

$$\mathcal{C} \circ E = \{(x,\xi) : \text{some } (y,\eta) \in E \text{ satisfies } (x,\xi;y,\eta) \in \mathcal{C}\}.$$

On these regions, neither component of a kernel wavefront covector vanishes, so distributional composition is defined. For matrix-valued kernels, take the union over entries. The inclusion allows cancellation and vanishing amplitudes; for smooth angular kernels, only input covectors radial to their base point can contribute to output singularities.

A.6 Bounded domains and scope of the theory

Let $\Omega \subset \mathbb{R}^d$ be a bounded open domain. Write $E_0 : L^2(\Omega; \mathbb{C}^C) \to L^2(\mathbb{R}^d; \mathbb{C}^C)$ for extension by zero outside Ω and R_Ω for restriction back to Ω. If $\operatorname{ess\,sup}_r \|\mathcal{M}_r\| \le C_M$, the bounded-domain branch satisfies

$$\mathcal{B}_\Omega = R_\Omega \mathcal{B} E_0, \qquad \|\mathcal{B}_\Omega v\|_{L^2(\Omega)} \le C_M \|v\|_{L^2(\Omega)}. \tag{46}$$

The bound follows because E_0 is an isometry and R_Ω a contraction. Sobolev estimates require $E_0 v \in H^s(\mathbb{R}^d)$; a sufficient whole-space class is $\widetilde{H}^s(\Omega) = \overline{C_c^\infty(\Omega)}^{H^s(\mathbb{R}^d)}$, with test functions extended by zero. Localized wavefront statements must include boundary singularities introduced by this extension.

The exact shellwise identities hold on the whole space; restriction and spatial cutoffs generally destroy radial-band invariance.

Relation to PDE operators. For a constant-coefficient elliptic symbol $p(\xi)$, the multiplier $W(\theta, r) = p(r\theta)^{-1}$ with $\mathcal{A}_r = I$ represents its inverse where $p \neq 0$, with cutoffs as needed. Constant anisotropic diffusion $-\nabla \cdot (A\nabla)$, with $A \in \mathbb{R}^{d\times d}$ symmetric positive definite, has symbol $4\pi^2 r^2 \theta^T A \theta$. This example concerns a constant-coefficient whole-space operator. The canonical relations above preserve Euclidean frequency magnitude, which restricts their applicability to wave or Helmholtz propagators.

Discretization and nonlinear composition. A finite angular matrix does not determine a unique continuum kernel or its high-frequency asymptotics. The microlocal results apply to continuum extensions satisfying the stated symbol assumptions; convergence of Fourier-slice interpolation, density compensation, padding, and cropping requires a separate analysis. Stability across resolutions requires uniform bounds on the learned mixers. The theory concerns an input-independent linear branch. The full network uses nonlinear composition to model maps such as the NS and Burgers solution operators, with interactions across shells.

B RSNO Numerical Implementation

B.1 Fourier-slice realization

The planar Radon spectral branch in Equation (8) samples a Cartesian Fourier grid directly; it does not form a spatial sinogram. For a square $H \times H$ input, each latent channel is zero-padded to an odd $D \times D$ grid with $D \ge \sqrt{2}H$. We apply an orthonormal two-dimensional FFT and shift its zero-frequency coefficient to the grid center. Bilinear interpolation then samples nonnegative integer radii, in FFT-index units, along uniformly spaced angles in $[0, \pi)$. The opposite rays are supplied by Hermitian conjugation. At each retained non-DC radius, the mixer first applies a complex channel map that depends on angle and radius and, when enabled, a dense angular map shared by all channels.

The DC coefficient counts as the first radial location. If N_r locations are retained, only indices $1, \dots, N_r - 1$ are passed to the mixer; the spectral-branch output at DC and beyond the cutoff is zero. The complete residual block can still have a nonzero mean through its input and spatial branch. Here, *radial locations* means the retained sample positions, *radial truncation* their upper cutoff, and *radial resolution* their spacing. The radial-truncation ablations change the cutoff while keeping the unit spacing fixed.

To synthesize the field, the mixed coefficients are multiplied by a discrete ramp, completed with their Hermitian partners, and scattered back to the Cartesian grid using the transpose of the bilinear interpolation weights. On a square padded grid of side D, the ramp weight at index k is $2k/D$. The density correction is

$$d_h(\xi) = \sum_m r_m w_{m\xi}, \qquad \mathcal{D}_h(\xi) = \begin{cases} \operatorname{median}_{d_h(\eta)>\epsilon} d_h(\eta)/d_h(\xi), & d_h(\xi) > \epsilon, \\ 0, & \text{otherwise}, \end{cases} \tag{47}$$

where m indexes polar samples and their Hermitian partners, $w_{m\xi}$ is a bilinear scatter weight, r_m is the ramp weight, and $\epsilon = 10^{-8}$. The density map is precomputed from the full sampling plan before radial truncation. For the reported RSNO configurations, applying $\mathcal{D}_h$ before the inverse FFT

compensates for the nonuniform scatter coverage. We take the real part after the inverse FFT, crop it to the original grid, and multiply by a fixed gain. The gain is calibrated once at initialization from a centered Gaussian g and its reconstruction g_{raw} through the same analysis–synthesis pair:

$$\frac{\langle g_{\mathrm{raw}}, g\rangle}{\max(\|g_{\mathrm{raw}}\|_2^2, 10^{-12})}.$$

These operations correct the numerical scale and gridding density; bilinear gridding remains an approximation to the continuous inverse Radon transform. The rectangular Burgers realization is specified in Appendix F.2.

Multiscale spatial branch. Each residual block begins with one-group GroupNorm. The fine-scale path applies a dilated convolution. At each coarse scale, average pooling uses the scale factor as both kernel size and stride, followed by two convolutions with GELU between them; bilinear upsampling then restores the original grid. The task-specific dilation schedules apply only to the fine-scale path, while coarse-scale convolutions use unit dilation. The local, coarse-scale, and Radon spectral outputs are multiplied by their learned coefficients, summed, passed through GELU, and added to the block input.

At the network level, a pointwise convolution lifts the input (including the coordinate channels) to the latent field. Two pointwise convolutions with GELU between them project the final latent field to the output.

B.2 Initialization and parameter counts

Each complex channel weight is stored as two real scalars, with real and imaginary parts independently initialized from Gaussian draws scaled by $1/C$, where C is the latent width. Each angular map is parameterized as the identity plus a dense complex matrix initialized to zero. Spatial weights $\alpha_{\ell,s}$ for $s > 1$ start at 0.1, with $\alpha_{\ell,1} = 1$ fixed; the learnable Radon spectral branch weights β_ℓ start at 1.0. The no-Radon-branch ablation fixes $\beta_\ell = 0$ and freezes the Radon spectral branch parameters, as described in Section 4.4.

All parameter counts refer to trainable real scalars. With N_θ angles and $N_r - 1$ nonzero shells, each residual block of the full model contains $2(N_r - 1)N_\theta C^2$ channel-mixing and $2(N_r - 1)N_\theta^2$ angular-mixing parameters. The no-angular-mixing variant omits the latter parameters; the no-Radon-branch variant excludes both groups and the fixed branch weights β_ℓ from its trainable parameter count.

B.3 Shared RSNO training procedure

The reference RSNO runs use seed 0, FP32, AdamW, gradient-norm clipping at 1.0, and 500 complete epochs; the NS seed-sensitivity study in Table 4 varies only the random seed. A cosine OneCycle schedule is advanced after each optimizer update, with maximum learning rate 5×10^{-4}, peak fraction 0.2, initial division factor 2, final division factor 100, and momentum cycling enabled. The initial and terminal learning rates are therefore 2.5×10^{-4} and 2.5×10^{-6}.

Scalar normalization statistics for input and output fields are fitted on training data only. Predictions are decoded to physical units before the training loss and evaluation metrics are computed.

Except for Darcy, the training objective is the mean relative L^2 error evaluated in physical units,

$$\mathcal{L}_{L^2} = \frac{1}{B}\sum_{i=1}^{B} \frac{\|\widehat{u}_i - u_i\|_2}{\max(\|u_i\|_2, 10^{-8})}. \tag{48}$$

Here B is the minibatch size. For each sample, the norm covers one frame for RSNO-H1, all frames in the active forecast horizon for RSNO-H10, and the complete output field for Helmholtz and Burgers. Thus the RSNO-H10 loss is a relative norm over the active forecast horizon, rather than an average of separately normalized frame errors.

Table 6 lists batch sizes, weight decay, and checkpoint-selection rules. The following appendices specify the architecture and training protocol for each task.

Table 6: Full-model RSNO training settings. Every run completes the shared 500-epoch schedule. Best validation denotes the checkpoint with the lowest mean validation relative L^2 error evaluated in physical units.

Task	Batch	Weight decay	Objective	Checkpoint
Darcy	4	10^{-4}	relative H^1	best validation (epoch 492)
NS (RSNO-H10)	8	10^{-4}	relative L^2	final
NS (RSNO-H1)	16	10^{-4}	relative L^2	final
Helmholtz	16	10^{-5}	relative L^2	best validation (epoch 500)
Burgers	32	10^{-4}	relative L^2	final

C Darcy Experimental Details

C.1 Problem and data

The piecewise-constant Darcy benchmark (Li et al., 2021) solves the elliptic problem

$$-\nabla\cdot\big(a(x)\nabla u(x)\big) = 1, \qquad x \in \Omega = (0,1)^2, \qquad u\big|_{\partial\Omega} = 0, \tag{49}$$

where a is the permeability field and u is the pressure. Each dataset sample pairs a permeability field taking values in $\{3, 12\}$ with its corresponding pressure solution. Endpoint-preserving stride-5 subsampling converts the released 421×421 fields to 85×85. Using zero-based indices, the training and validation sets contain samples $[0, 1000)$ and $[1000, 1024)$ from the first source subset, respectively. The test set contains samples $[0, 200)$ from the second source subset. All methods use this 1,000/24/200 split and the evaluation metrics in Section 4.1.

C.2 RSNO configuration and optimization

RSNO uses width 20, four residual blocks, 24 uniformly spaced angles over $[0, \pi)$, and four radial locations including DC. Each residual block combines the Radon spectral branch with a multiscale spatial branch using scales $\{1, 2, 4\}$, 3×3 kernels, and fine-scale dilations $(1, 2, 4, 8)$. The spatial convolutions use circular padding as a network implementation choice. Raw Cartesian coordinate channels are appended before lifting. The Radon spectral branch zero-pads the 85×85 field to 121×121 and applies learned mixing on three nonzero radial shells. The complete model contains 317,757 trainable real parameters, including 244,224 in the orientation–frequency mixers.

Training follows Appendix B.3, with batch size 4 and weight decay 10^{-4}. The relative H^1 objective uses the following discrete Fourier-weighted norm:

$$\mathcal{L}_{H^1} = \frac{1}{B}\sum_{i=1}^{B} \frac{\|\widehat{u}_i - u_i\|_{H^1,h}}{\|u_i\|_{H^1,h} + 10^{-12}}, \qquad \|w\|^2_{H^1,h} = \sum_{k_1,k_2} (1 + k_1^2 + k_2^2)|(\mathcal{F}_h w)_{k_1,k_2}|^2. \tag{50}$$

Here $\mathcal{F}_h$ is the orthonormal two-dimensional FFT, and k_1, k_2 are the integer frequency indices for a length-85 transform with sample spacing $1/85$. The weight is exactly $1 + k_1^2 + k_2^2$, with no $4\pi^2$ factor; it defines the implemented loss rather than a finite-difference norm on the endpoint-inclusive PDE grid. The training loss is evaluated on pressure fields decoded to physical units.

C.3 Model configurations and training cost

Table 7 summarizes model configurations, trainable parameter counts, recorded run times, and training epochs. MWT and MgNO use parameter-free grid adapters ($85 \leftrightarrow 128$ and $85 \leftrightarrow 211$); evaluation remains on the common 85×85 grid. Other baseline settings follow the released implementations and the policy in Section 4.2. Throughout Appendices C and D, RNO denotes our evaluation of the authors' released implementation.

C.4 Supplementary results

Figures 2 and 3 show the parameter–accuracy tradeoff, RSNO training history, and the three test samples with the largest RSNO relative L^2 errors.

Table 7: Darcy model configurations, trainable real-parameter counts, recorded run times, and training epochs.

Method	Configuration	Params	Time (min)	Epochs
FNO	width 32; 4 layers; 12 modes	2,368,001	3.4	500
Galerkin	width 128; 6 layers; 4 heads; 29^2 grid	2,220,829	19.7	100
MWT	4 blocks; $\alpha = 12$, $c = 4$, order 4	9,807,873	95.7	500
GNOT	width 96; 3 layers; 4 heads; 1 expert	675,268	52.5	500
LSM	width 64; 12 bases; 4 tokens; 6^2 patches	19,187,457	20.0	501
OFormer	embedding 96; latent 256; 6 layers; 4 heads	2,556,321	365.1	1,588*
U-NO	UNO-9; width 32	16,393,665	9.0	700
MgNO	5 layers; 24 channels; 6 V-cycle levels	678,093	40.2	500
Transolver	width 128; 8 layers; 8 heads; 64 slices	2,826,945	105.5	500
HPM	width 128; 8 layers; 8 heads; 128 frequencies	2,864,321	129.7	500
RNO	width 128; 8 layers; 8 heads; 32 angles; Radon $\times 4$	2,815,731	92.4	500
LinearNO	width 128; 8 layers; 8 heads; key ratio 64	1,766,145	125.4	500
RSNO (ours)	width 20; 4 blocks; 24 angles; 4 radial locations	317,757	57.9	500

*OFormer stops during epoch 1,588 at 100,000 updates. Run times are measured end to end.

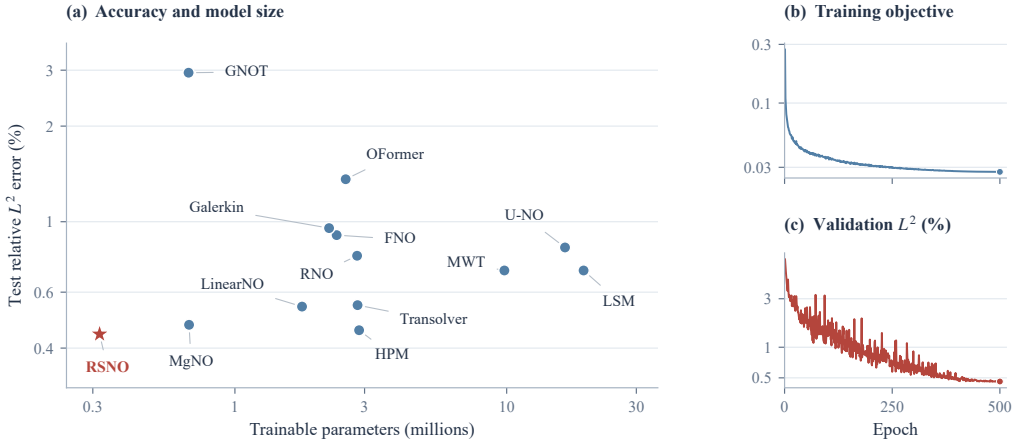

Figure 2: Darcy parameter–accuracy tradeoff and RSNO training history. (a) Test relative L^2 error versus trainable real parameters for our runs in Table 1 with both axes logarithmic. (b, c) Recorded RSNO training objective and validation relative L^2 error over 500 epochs, with logarithmic vertical axes.

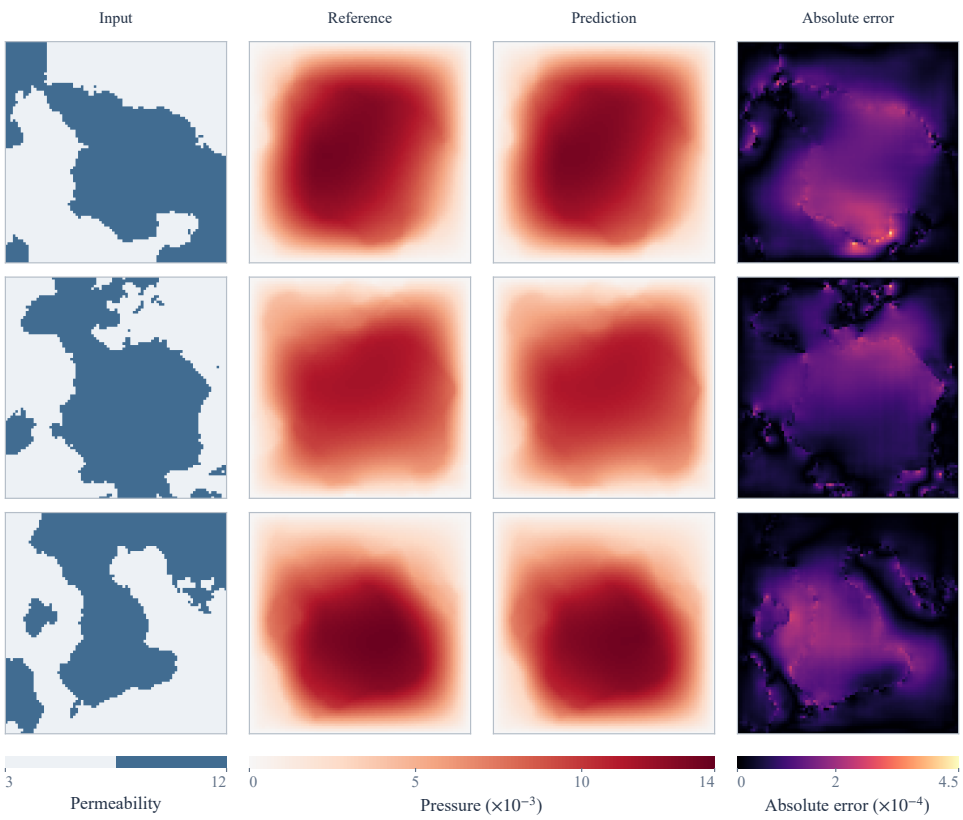

Figure 3: RSNO predictions for the three Darcy test samples with the largest relative L^2 errors.

D NAVIER–STOKES EXPERIMENTAL DETAILS

D.1 PROBLEM AND DATA

The FNO benchmark (Li et al., 2021) solves the vorticity equation on the periodic unit torus,

$$\partial_t \omega + u \cdot \nabla \omega = \nu \Delta \omega + f, \qquad \nabla \cdot u = 0, \qquad \omega(\cdot, 0) = \omega_0, \tag{51}$$

with $\nu = 10^{-5}$, $\omega_0 \sim \mathcal{N}(0, 7^{3/2}(-\Delta + 49I)^{-2.5})$, and $f(x) = 0.1[\sin(2\pi(x_1 + x_2)) + \cos(2\pi(x_1 + x_2))]$.

The common evaluation target is $(\omega_{11}, \ldots, \omega_{20})$, with observations available through ω_{10}. RSNO-H10 uses the full ten-frame history, while RSNO-H1 uses only ω_{10}. Using zero-based indices, the training and test sets contain trajectories $[0, 1000)$ and $[1000, 1200)$, respectively, with no validation set. All methods are evaluated using their final checkpoints over the complete ten-frame forecast.

D.2 RSNO CONFIGURATION AND TEMPORAL TRAINING

Both RSNO-H10 and RSNO-H1 use width 32, eight residual blocks, 24 angles, and eight radial locations including DC. Their multiscale spatial branches use scales $\{1, 2, 4\}$, 3×3 kernels, fine-scale dilations $(1, 2, 4, 4, 8, 8, 16, 16)$, and circular padding. Coordinate channels use periodic sine/cosine embeddings with one harmonic per axis. Before coordinate channels are appended, RSNO-H10 has ten input-field channels and RSNO-H1 has one. Both predict one output-field channel. Training follows Appendix B.3, with weight decay 10^{-4} and batch sizes 8 and 16, respectively.

History-conditioned autoregressive training (RSNO-H10). RSNO-H10 predicts one frame from a sliding ten-frame input window. During training, each step drops the oldest frame and appends the ground-truth next frame (teacher forcing); the appended frame is decoded to physical units with the output normalizer and then re-encoded with the input normalizer. The horizon curriculum uses 1, 2, 4, 6, 8, and 10 forecast steps starting at epochs 1, 51, 101, 151, 201, and 251, respectively. For each minibatch, predictions over the active forecast horizon are concatenated. The relative L^2 loss in Equation (48) is then computed over this target, followed by one optimizer update. At test time, the model starts from the observed ten-frame history and appends its own predictions, using the same normalization conversion, for ten autoregressive steps.

Single-state autoregressive training (RSNO-H1). All 19 adjacent-frame pairs from each training trajectory form the training set. Each update minimizes the one-frame relative L^2 loss evaluated in physical units, without a training rollout or horizon curriculum. At test time, RSNO-H1 starts from ω_{10} and recursively predicts the same ten future frames, decoding each predicted frame to physical units before re-encoding it with the input normalizer.

Both formulations use the final checkpoint after 500 epochs, but an RSNO-H1 epoch traverses the adjacent-frame pairs, whereas an RSNO-H10 epoch traverses the trajectories.

D.3 MODEL CONFIGURATIONS AND TRAINING COST

Table 8 summarizes model configurations and training costs; Figure 4 compares predictions at the final forecast frame.

D.4 AUTOREGRESSIVE ABLATIONS

All fourteen configurations in Table 3 use seed 0, batch size 8, and the RSNO-H10 data split, training loss, and horizon curriculum. Each uses its final checkpoint after 500 epochs and ten-step autoregressive inference. Six-block variants use the first six entries of the fine-scale dilation schedule. Table 5 reuses the results for the full RSNO-H10 model and its no-angular-mixing variant, and adds the no-Radon-branch variant trained under the same protocol. Component removal and trainable parameter counts follow Section 4.4.

Table 8: Navier–Stokes model configurations, trainable real-parameter counts, recorded run times, and training epochs.

Method	Configuration	Params	Time (min)	Epochs
FNO	width 20; 4 Fourier layers; 12 modes	926,517	33.6	500
Galerkin	width 48; 4 layers; 1 head; 12 modes	862,049	71.0	100
MWT	4 blocks; $\alpha = 12$; 4 channels; order 3	71,970,809	109.3	500
GNOT	CGPT; width 128; 3 layers; 1 head	1,391,882	34.0	500
LSM	width 64; 12 bases; 4 tokens; 4^2 patches	19,188,033	119.4	501
OFormer	enc. $96 \to 192$; 5 layers; 1 head; dec. 384	1,851,361	292.8	2,032*
U-NO	width 32; 7 operator levels; factor 0.75	30,478,033	68.4	500
MgNO	5 layers; 32 channels; periodic V-cycle	2,922,508	98.6	500
Transolver	width 256; 8 layers; 8 heads; 32 slices	11,232,321	1042.5	500
HPM	width 256; 8 layers; 8 heads; 128 frequencies	11,306,817	879.4	500
RNO	width 256; 8 layers; 8 heads; 32 angles; Radon $\times 2$	11,205,619	933.0	500
LinearNO	width 256; 8 layers; 8 heads; key ratio 32	3,377,921	718.0	500
RSNO-H10	width 32; 8 blocks; 24 angles; 8 radial locations	3,190,137	337.1	500
RSNO-H1	width 32; 8 blocks; 24 angles; 8 radial locations	3,189,849	610.3	500

*OFormer stops during epoch 2,032 at 128,000 updates. Run times are measured end to end.

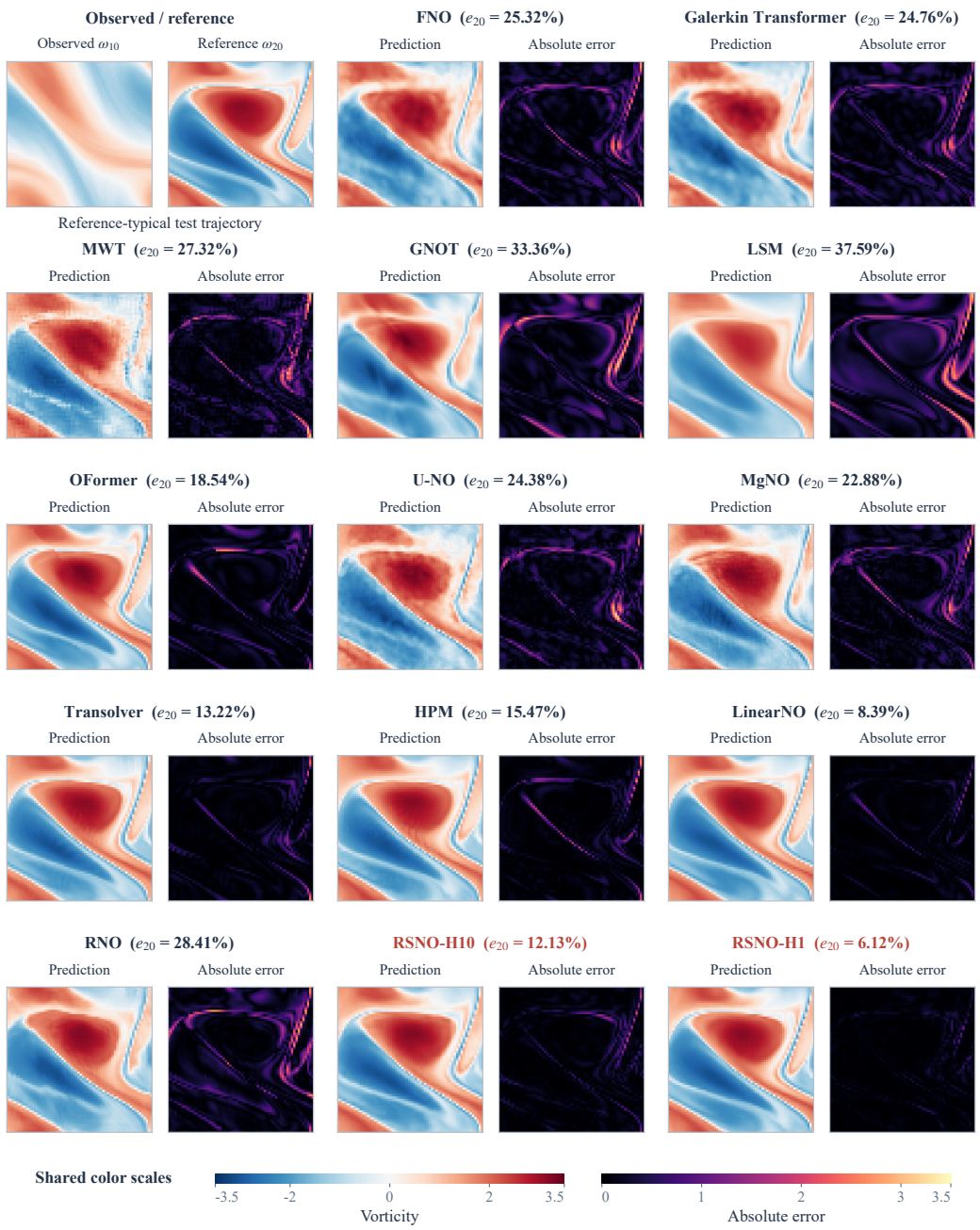

Figure 4: Navier–Stokes predictions on a test trajectory selected using only the reference solutions, with near-median energy and periodic-gradient energy over the ten target frames. The first pair shows the last observation ω_{10} and reference solution ω_{20}; each remaining pair shows a prediction $\widehat{\omega}_{20}$ and absolute error $|\widehat{\omega}_{20} - \omega_{20}|$ on the periodic unit square. All methods use their final checkpoints without ground-truth feedback. Labels give $e_{20} = \|\widehat{\omega}_{20} - \omega_{20}\|_2 / \|\omega_{20}\|_2$ in percent.

E HELMHOLTZ EXPERIMENTAL DETAILS

E.1 PROBLEM AND DATA

The public Helmholtz problem of de Hoop et al. (2022) maps the wave-speed field c to the real-valued solution u on $\Omega = (0,1)^2$:

$$\left(-\Delta - \frac{\omega^2}{c(x)^2}\right) u(x) = 0, \quad x \in \Omega, \quad \omega = 10^3, \tag{52}$$

$$\partial_n u(x) = \begin{cases} \mathbf{1}_{[0.35,0.65]}(x_1), & x_2 = 1, \\ 0, & \text{on the other three edges.} \end{cases} \tag{53}$$

Here ∂_n denotes the outward normal derivative. The wave-speed field is $c = 20 + \tanh(\widetilde{c})$, where $\widetilde{c} \sim \mathcal{N}(0, (-\Delta + 9I)^{-2})$ on the spatial-mean-zero subspace with homogeneous Neumann conditions.

We use the CaltechDATA archive without spatial resampling. Zero-based source intervals $[0, 4000)$, $[8200, 8400)$, and $[8000, 8200)$ define the training, validation, and test sets. All methods, including RSNO, use fixed configurations without dataset-specific hyperparameter tuning, following Section 4.2.

E.2 RSNO CONFIGURATION AND OPTIMIZATION

RSNO uses width 48, eight residual blocks, 24 angles, and six radial locations including DC, totaling 5,306,905 trainable real parameters. The multiscale spatial branch uses scales $\{1, 2, 4\}$, 3×3 kernels, fine-scale dilations $(1, 2, 4, 4, 8, 8, 16, 16)$, and zero padding. Raw Cartesian coordinate channels are appended before lifting. The Radon spectral branch follows Appendix B.1.

Training follows the shared procedure in Appendix B.3, with batch size 16, weight decay 10^{-5}, and the relative L^2 objective evaluated in physical units. The 500-epoch schedule contains 125,000 optimizer updates. The checkpoint with the lowest validation relative L^2 error occurs at epoch 500 and is evaluated on all 200 test samples.

E.3 MODEL CONFIGURATIONS AND TRAINING COST

Table 9 summarizes model configurations and training costs. PCA-Net, DeepONet, PARA-Net, and FNO follow the public implementation of de Hoop et al. (2022); the remaining baselines use fixed adaptations to the shared grid. U-NO and MgNO select their checkpoints by validation, while the other baselines use their final checkpoints.

PCA-Net uses input and output PCA ranks of 101 and 322, respectively. DeepONet's branch network takes 101 input PCA coefficients, while its trunk takes two query coordinates. PARA-Net takes the concatenation of these PCA coefficients and query coordinates as input.

Table 9: Helmholtz model configurations, trainable real-parameter counts, recorded run times, and training epochs.

Method	Configuration	Params	Time (min)	Epochs
PCA-Net	$(101, 128, 128, 128, 322)$	87,618	6.7	1,000
DeepONet	branch $(101, 128, 128, 128, 128)$; trunk $(2, 128, 128, 128)$	96,001	11.0	1,000
FNO	width 16; 3 spectral blocks; 12×12 modes; padding 9	443,233	13.0	500
PARA-Net	$(103, 128, 128, 128, 1)$	46,465	46.9	200
U-NO	UNO-9; width 8; padding 12	1,024,785	45.4	700
MgNO	3 layers; level widths $[18, 36, 72, 144, 256]$; reflect padding	7,539,792	23.2	100
LinearNO	width 96; 8 conv. attention blocks; 8 heads; key ratio 64	998,625	516.4	500
RSNO (ours)	width 48; 8 blocks; 24 angles; 6 radial locations	5,306,905	162.2	500

E.4 Qualitative results

Figure 5 shows RSNO predictions for the three test samples with the largest relative L^2 errors.

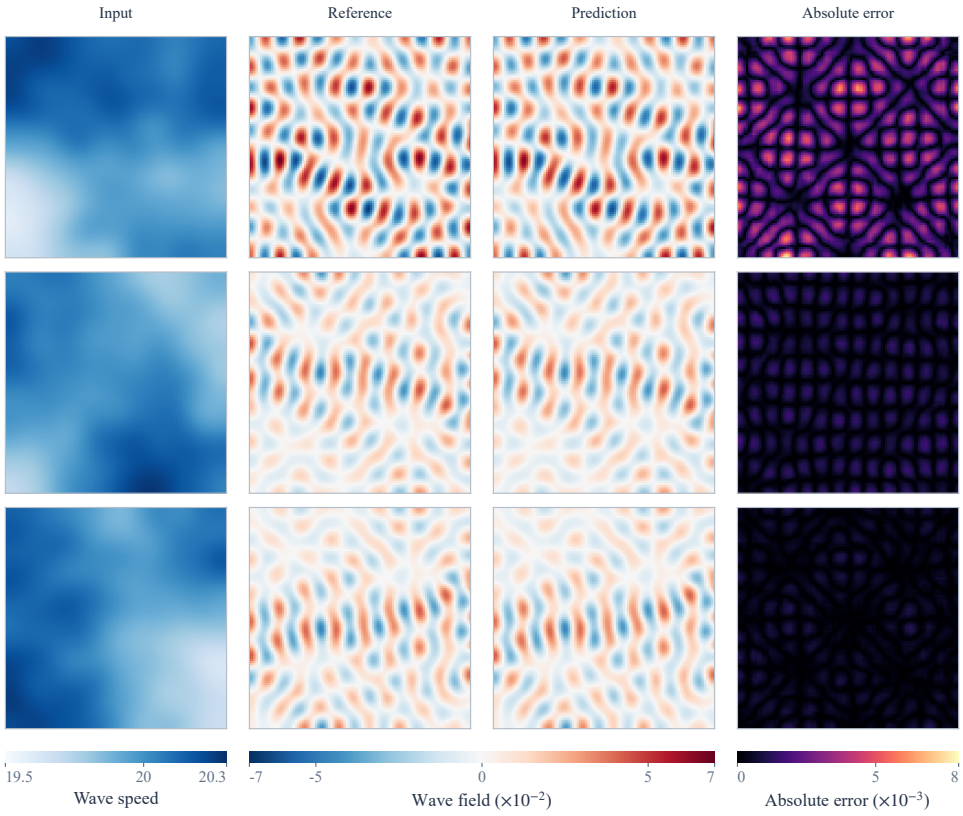

Figure 5: RSNO predictions for the three Helmholtz test samples with the largest relative L^2 errors.

F Burgers Experimental Details

F.1 Problem and data

The PDEBench task (Takamoto et al., 2022) uses the periodic one-dimensional viscous Burgers equation,

$$\partial_t u + \partial_x\left(\frac{u^2}{2}\right) = \frac{\nu}{\pi}\partial_{xx} u, \qquad \nu = 10^{-3}. \tag{54}$$

Subsampling with strides of 5 in time and 4 in space converts the source tensor from $10{,}000 \times 201 \times 1{,}024$ to $10{,}000 \times 41 \times 256$. The stored times are $0, 0.05, \ldots, 2$. Using zero-based indices, the training and test sets contain trajectories $[0, 9000)$ and $[9000, 10000)$, respectively, with no validation set.

All methods predict the same 40 future frames from $u(0, x)$. They use seed 0 and scalar normalization statistics fitted on training data only, and are evaluated in physical units using their final checkpoints. Configurations are fixed without dataset-specific hyperparameter tuning for any method, including RSNO.

F.2 RSNO representation and training

Although the Burgers equation is one-dimensional in space, the two-dimensional branch acts on the joint time–space representation. RSNO replicates the initial field over 40 query-time positions,

appends two coordinate channels, and lifts the result to a latent 40×256 field. Each coordinate channel is evenly spaced from 0 to 1 along its respective axis; these encode normalized positions along the time and space axes rather than physical coordinates. Six width-20 residual blocks operate jointly over time and space. The final projection returns one scalar value at each time–space position, producing the complete future trajectory. The Radon spectral branch uses 32 angles and six radial locations including DC. The multiscale spatial branch uses scales $\{1, 2, 4\}$, 3×3 kernels, and fine-scale dilations $(1, 2, 4, 4, 8, 8)$. Circular convolutional padding is applied along both axes as an architectural convention, without imposing temporal periodicity on the PDE. The complete model contains 939,259 trainable real parameters.

Fourier gridding pads the two axes independently to 41×257 and samples polar rays at radial indices $k = 0, \ldots, 5$ in centered FFT-index coordinates. Thus the shell radius is defined in index space, without equating physical temporal and spatial frequencies. The ramp weight is $2k/257$. Density compensation and gain calibration follow Appendix B.1, but use only this retained sampling plan. Before Hermitian completion, the shared mixer projects the coefficient at the last retained radial location onto its real part at each angle. The rectangular grid is preserved throughout. All 40 future frames are produced in one forward pass, as required by the direct full-field prediction formulation.

Training follows Appendix B.3, with batch size 32, weight decay 10^{-4}, and no gradient accumulation. The loss in Equation (48) covers the complete 40×256 target for each trajectory. After 500 epochs, the final checkpoint is evaluated once on all 1,000 test trajectories.

F.3 Model configurations and training cost

Table 10 summarizes model configurations and training costs. FNO, U-NO, and F-FNO train on all 40 adjacent-frame pairs per training trajectory and use 40-step autoregressive inference; the remaining methods use direct full-field prediction. An epoch therefore traverses adjacent-frame pairs for the autoregressive methods and complete trajectories for the direct full-field methods.

Table 10: Burgers model configurations, trainable real-parameter counts, recorded run times, and training epochs.

Method	Configuration	Params	Time (min)	Epochs
FNO	width 64; 4 spectral layers; 16 modes	549,569	5.5	9^*
U-NO	width 32; 1-D spectral encoder–decoder	530,881	12.9	9^*
F-FNO	width 64; 4 factorized Fourier blocks; 16 modes	616,897	10.6	9^*
VCNeF	width 96; 8 heads; 3 encoder and 3 modulation layers	855,841	161.8	500
LNO	width 96; 8 heads; 128 tokens; 4 latent layers	484,552	25.6	500
Transolver	width 128; 4 blocks; 8 heads; 32 slices	701,608	24.3	500
LinearNO	width 128; 4 blocks; 8 heads; key dimension 32	573,160	25.8	500
RSNO (ours)	width 20; 6 blocks; 32 angles; 6 radial locations	939,259	111.1	500

*The first three methods complete eight epochs and part of the ninth, stopping at 50,000 updates.

F.4 QUALITATIVE RESULTS

Figure 6 shows RSNO predictions for the three test trajectories with the largest relative L^2 errors, using direct full-field prediction over the time–space grid.

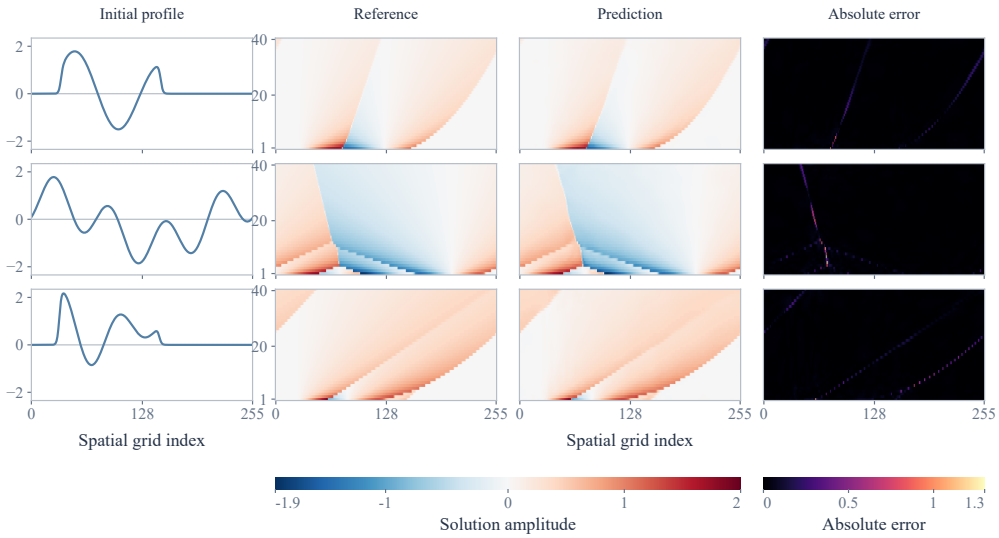

Figure 6: RSNO predictions for the three Burgers test trajectories with the largest relative L^2 errors. Columns show the initial profile, reference solution, direct full-field prediction, and absolute error. Heatmap rows run from steps 1 to 40, bottom to top.