WHITENING THE SPECTRUM: PRECONDITIONED FLOW MATCHING FOR TURBULENCE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models of turbulence transport a spectrally flat source onto fields whose energy follows a steep power law, and the flow matching objective inherits that spectrum through the regression target and the interpolant's signal-to-noise ratio, so energy-containing scales dominate the gradient and dissipative scales are learned last. Our method, WSFM, instead solves transport in a whitened Fourier representation, an invertible, mode-wise affine transform estimated from the training ensemble that equalizes the latent's second-order statistics across wavenumbers, preserves the relative Fourier phases carrying coherent vortical structure, and leaves the modeled distribution unchanged. It therefore improves the conditioning of the optimization rather than altering the target, and we prove the whitened transport problem is uniformly conditioned. At matched backbone and parameter count, WSFM lowers relative L_2 and high-wavenumber spectral error against raw Fourier flow matching and retains the small-scale structure the raw coordinate smooths away. Across benchmark flows the gains concentrate in the physics, with structure-function exponents, enstrophy and dissipation statistics reproduced more faithfully than by the raw coordinate or the published surrogates, while pointwise accuracy stays comparable.

1 INTRODUCTION

The energy spectrum of a developed turbulent velocity field follows $E(k) \sim k^{-5/3}$ across an inertial range that widens with Reynolds number before falling off steeply at dissipative scales (Kolmogorov, 1941; Frisch, 1995). The variance of the Fourier modes therefore spans several orders of magnitude within one field, energy-containing large scales carrying the bulk of the kinetic energy and dissipative scales, which control vorticity, strain rate and enstrophy, a vanishing fraction.

Generative models of turbulence inherit a design choice from image generation, where natural-image spectra are likewise far from flat (van der Schaaf & van Hateren, 1996), that sits awkwardly with these statistics, transporting a white-noise source whose Fourier modes all have equal variance onto data whose modal variances are far from uniform. Wang et al. (2025b) trace the consequences to a frequency-dependent degradation of the signal-to-noise ratio (SNR) and propose frequency-aware architectures, Sambamurthy & Chattopadhyay (2025) analyze the same effect as spectral collapse and prescribe frequency-aware noise schedules, and Ruscio & Rusch (2026) make the source carry the data spectrum. Each adapts the model, schedule or endpoint to a fixed spectrum; we ask instead why the network should confront a spectrally biased coordinate system.

The difficulty is a property of the optimization problem rather than a lack of model capacity. Posed directly on Fourier coefficients, both the regression target and the SNR of the interpolant inherit the energy spectrum, so part of the modeling task reproduces an energy hierarchy already known analytically from the training ensemble, leaving less signal for the non-Gaussian, cross-mode and phase-coherent organization of turbulence.

WSFM removes that task analytically, applying an invertible, mode-wise affine transform to the Fourier coefficients of the training ensemble and running flow matching in the transformed coordinates. The transform is diagonal and constant there, and preserves the relative phases that carry the coherent vortical structure of the field. It rescales every wavenumber mode rather than every radial shell, leaving neither target nor SNR with wavenumber dependence, which is the sense in which whitening acts as a preconditioner. One distinction matters throughout. Whitening aligns the

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.

second-order statistics of the latent with the Gaussian source without making the latent field Gaussian. We instantiate the principle with SPECTRALDIT, a transformer whose positional encoding is the frequency coordinate of each token.

- **Spectrally whitened generative coordinates.** An invertible, parameter-free affine transform of the Fourier coefficients that equalizes the modal statistics of the data and matches them to the white-noise source, applied per wavenumber vector rather than per radial shell, whitening every mode exactly rather than on shell averages.

- **A conditioning account of why it works.** The raw coordinate has $\kappa(D) = (k_{\max}/k_{\min})^{\alpha/2}$, with the regression target and the SNR both carrying the data spectrum; whitening gives $\kappa(D^{\xi}) = 1$, $\mathrm{Var}(v_k) = 2$ and $\mathrm{SNR}_k(t) = ((1-t)/t)^2$ for every k (Propositions 1–2). Being diagonal, invertible and phase-preserving it removes covariance structure and nothing else, leaving the non-Gaussian, cross-mode and phase-coherent structure to be learned (Proposition 3).

- **Empirical validation of the conditioning account.** At matched backbone, parameter count (within 0.04%) and schedule, whitening reduces relative L_2 error by 26% and high-wavenumber spectral error by 99% over raw Fourier flow matching ($p < 10^{-4}$ over 19 paired trajectories). Against the published surrogates pointwise accuracy is comparable rather than superior while the physics leads, WSFM never last on MAE, RMSE or relative L_2 and within $1.5\times$ of the strongest baseline, with the structure-function, enstrophy and dissipation errors lowest on cylinder and foil.

2 RELATED WORK

Generative and deterministic turbulence models. Diffusion and flow matching are standard tools for probabilistic fluid simulation (Kim et al., 2019; Du et al., 2024; Hu et al., 2025; Lienen et al., 2024) and for spatiotemporal generation more broadly (Yin et al., 2026), in latent (Zhou et al., 2024; Cachay et al., 2023) and PDE-informed (Rosu et al., 2025; Mardani et al., 2024) forms. A separate line keeps the map deterministic and learns a solution operator instead, from Fourier neural operators (Li et al., 2021), graph networks (Pfaff et al., 2020; Sanchez-Gonzalez et al., 2020), physics-informed networks (Raissi et al., 2019) and transformer solvers (Wu et al., 2024; Li et al., 2023) to large-scale pre-training (Hao et al., 2024; Zhang et al., 2025) and benchmark collections (Ohana et al., 2024; Mao et al., 2026). Such surrogates are accurate pointwise but single-valued, so cannot represent the ensemble spread of a turbulent flow, and their rollouts drift as errors accumulate (Parthipan et al., 2024); the generative line has already recovered Kolmogorov scaling (Whittaker et al., 2024; Oommen et al., 2025) and near-wall statistics through conditional flow matching (Parikh et al., 2026). The question is therefore not whether a model family can generate turbulence, but in which representation to pose transport.

Frequency-aware generative modeling. A growing line of work identifies frequency imbalance as the mechanism behind degraded small-scale statistics while treating the spectrum as fixed, whether through the spectral bias of coordinate networks (Rahaman et al., 2019), frequency-aware generative backbones (Wang et al., 2025a), spectrally regularized flow matching (Rafiq & Nair, 2026) or frequency-domain physics-informed models (Chen et al., 2026). We instead change the coordinate system, absorbing the data spectrum into the transform so that every mode of the latent carries unit variance by construction.

Learned versus fixed representations. Latent diffusion compresses fields into a learned autoencoder prior (Rombach et al., 2022; Zhou et al., 2023), and self-supervised field encoders learn representations for PDE data (Zhou & Farimani, 2024), an idea carried into scientific generation through neural-field and implicit representations (Mildenhall et al., 2021; Serrano et al., 2023; Yin et al., 2023; Liu et al., 2025; Li et al., 2026); Rozet et al. (2025) find that latent compression systematically damages small-scale physics. Our representation takes the opposite design choice, fixed and invertible rather than learned and lossy, so that physics survives by construction.

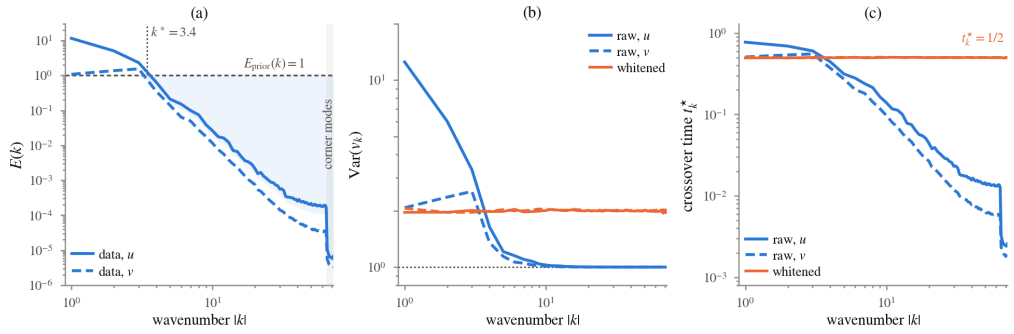


Figure 1: A white-noise prior is spectrally misaligned with turbulent fields. (a) The ensemble energy spectrum of the cylinder-wake data spans several orders of magnitude across the resolved band while the prior is flat. (b) The per-mode target-velocity variance inherits that spectrum in raw coordinates, against a constant 2 after whitening. (c) The time at which a mode's SNR reaches unity is $1/2$ for every mode after whitening and falls with wavenumber before it.

3 Background

We consider a velocity field u with C components on a uniform $N_1 \times \cdots \times N_d$ grid, $d \in \{2, 3\}$, represented by its one-sided orthonormal real FFT $q = \text{rFFT}_d(u)$, a linear bijection onto the Hermitian coefficient array. Modes are indexed by the wavenumber vector k, (k_x, k_y) in two dimensions and (k_x, k_y, k_z) in three, in the discrete sense of the basis function $e^{\mathrm{i}k\cdot x}$ carried by the grid. Write

$$\mu_c(k) = \mathbb{E}[q_c(k)], \qquad D_c(k) = \sqrt{\mathbb{E}|q_c(k)|^2 - |\mathbb{E}q_c(k)|^2}, \tag{1}$$

for the per-mode mean and standard deviation over the data distribution, so the per-mode energy is $\mathbb{E}|q_c(k)|^2 = |\mu_c(k)|^2 + D_c(k)^2$, and averaging over shells of constant $|k|$ weighted by the shell measure gives the discrete energy spectrum $E(k) \sim k^{-\alpha}$ with $\alpha \approx 5/3$ in the inertial range (Kolmogorov, 1941; Frisch, 1995; Pope, 2000). Because white noise is white in any orthonormal basis, the flow-matching source $p_{\text{src}} = \mathcal{N}(0, I)$ is $\mathcal{CN}(0, I)$ in these coordinates, with independent real and imaginary parts of variance $1/2$ per mode. Real flows are also anisotropic, and the modes carrying the coherent streaks, those with vanishing transverse wavenumber, hold energies far above the radial average. Flow matching, and the sense in which the choice of coordinate system fixes the regression problem while leaving the target distribution unchanged, are recalled in Appendix A.3.

4 Method

4.1 Problem formulation

Take the coordinate system of Section 3, in which transport runs directly on the Fourier coefficients, $x_0 = q$ and $x_1 \sim \mathcal{CN}(0, I)$. The conditioning of the regression problem there is fixed by the energy spectrum. Since x_0 and x_1 are independent, the per-mode target velocity $v_k = x_{1,k} - x_{0,k}$ and the mode-wise signal-to-noise ratio of the interpolant satisfy

$$\text{Var}(v_k) = \underbrace{\mathbb{E}|q_k|^2}_{\text{data term}} + \underbrace{1}_{\text{source term}}, \qquad \text{SNR}_k(t) = \frac{(1-t)^2\,\mathbb{E}|q_k|^2}{t^2\,\mathbb{E}|x_{1,k}|^2} = \frac{(1-t)^2}{t^2}\,\mathbb{E}|q_k|^2. \tag{2}$$

Both inherit the energy spectrum exactly. Writing $\mathbb{E}|q_k|^2 \sim k^{-\alpha}$ across a resolved band $[k_{\min}, k_{\max}]$, the ratio between the best- and worst-conditioned modes,

$$\kappa_{\text{spectral}} = \frac{\max_k \mathbb{E}|q_k|^2}{\min_k \mathbb{E}|q_k|^2} \sim \left(\frac{k_{\max}}{k_{\min}}\right)^{\alpha}, \tag{3}$$

grows without bound as the inertial range widens, the spectral condition number κ_{spectral} of the representation.

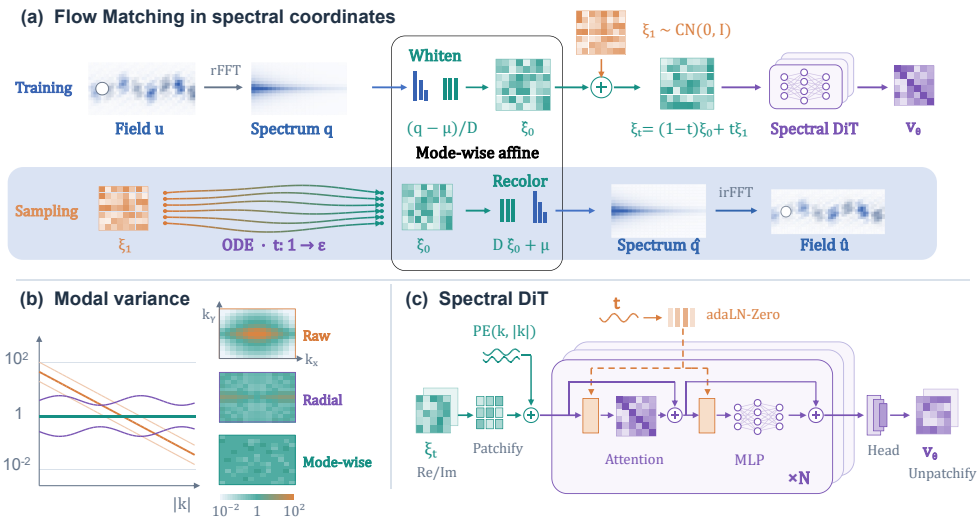


Figure 2: WSFM overview. (a) Training transports in whitened Fourier coordinates and sampling inverts the change of representation before the inverse transform; the change is invertible and parameter-free, the only learned component being the drift network, which acts on a latent whose per-mode second-order statistics match the Gaussian prior. (b) Modal variance under the raw coordinate, radial whitening and mode-wise whitening, as a function of $|k|$ and over (k_x, k_y); the line profiles are schematic. Spectral condition numbers measured on the training ensemble are in Table 4. (c) The SPECTRALDIT backbone, whose tokens carry their frequency coordinate.

Two consequences follow. First, the gradient is dominated by energy-containing modes, so dissipative scales see an effective learning rate smaller by $\kappa(D) = \kappa_{\text{spectral}}^{1/2}$, and high-$k$ modes enter the noise-dominated regime first and are learned last (Sambamurthy & Chattopadhyay, 2025). Second, the white-noise source must build the entire energy hierarchy from a flat spectrum, a task whose optimal solution is known analytically from equation 1, consuming capacity the network should spend on the non-Gaussian organization of the field instead (Figure 1).

4.2 SPECTRALLY WHITENED GENERATIVE COORDINATES

WSFM transports in a coordinate system in which the second-order spectral statistics of the data match the prior. We define the mode-wise affine whitening

$$\xi_c(k) \;=\; \frac{q_c(k) - \mu_c(k)}{D_c(k)} \tag{4}$$

with μ_c, D_c as in equation 1, estimated from the training ensemble, and the inverse (*recoloring*) is $\hat{q}_c(k) = D_c(k)\,\xi_c(k) + \mu_c(k)$, followed by $u = \mathrm{irFFT}_d(\hat{q})$.

Three properties matter. (i) *Invertibility.* The map is a diagonal affine bijection with constant Jacobian, so no information is discarded and no encoder is learned. (ii) *Mode-wise, not radial.* The scale is indexed by the full wavenumber vector k rather than by $|k|$. Radial whitening $q/\sqrt{E(|k|)}$ equalizes shell averages only, whereas the modes carrying the coherent streaks hold far more energy than the radial mean, so those remain under-whitened while the rest of the band is over-whitened (Figure 2b). (iii) *Affine, with the mean removed.* Min-max normalization to $[-1, 1]$ leaves the DC coefficient at a large constant, whereas subtracting the per-mode ensemble mean and dividing by the per-mode standard deviation places every mode, DC included, at zero mean and unit variance. μ and D are estimated over a 200-batch warm-up and then frozen (Figure 2a).

4.3 FLOW MATCHING IN WHITENED COORDINATES

Given the whitened representation, we run physics-aligned flow matching directly in ξ-space:

$$\xi_t = (1-t)\,\xi_0 + t\,\xi_1, \qquad \xi_0 = \mathrm{Whiten}(q_0), \quad \xi_1 \sim \mathcal{CN}(0, I), \tag{5}$$

with the drift network trained against the constant target velocity $\xi_1 - \xi_0$. The pipeline is $u \xrightarrow{\mathrm{rFFT}_d} q \xrightarrow{\text{whiten}} \xi \xrightarrow{\text{FM / ODE}} \hat{\xi}_0 \xrightarrow{\text{recolor}} \hat{q} \xrightarrow{\mathrm{irFFT}_d} \hat{u}$. The $k^{-\alpha}$ factor in equation 2 is removed analytically rather than approximated by a schedule or absorbed into an architecture, leaving the non-Gaussian, cross-mode, phase-coherent organization of the field as the modeling task.

A frequency-coordinate transformer. WSFM is agnostic to the backbone; we instantiate it with SPECTRALDIT, a diffusion transformer (Peebles & Xie, 2023) on the whitened one-sided spectrum. Patches are formed on the wavenumber grid, so a token's coordinate is the frequency it represents, and we encode $\mathrm{PE}(k, |k|)$ by expanding each coordinate in log-spaced sinusoids and projecting the concatenation through a small MLP, in the spirit of the sinusoidal and rotary encodings used for sequence and spatial position (Raffel et al., 2020; Su et al., 2021; Press et al., 2022). The one-sided transform halves only the streamwise axis k_x, so the transverse coordinates k_y and k_z keep their full signed range and the encoding retains those signs, letting the model distinguish a wavenumber vector from its reflection, two dynamically distinct modes in a shear flow. Complex latents are stacked on the channel axis as $[\mathrm{Re}, \mathrm{Im}]$, and timestep conditioning uses a Fourier embedding of $t \in [0, 1]$ with adaLN-Zero modulation (Figure 2c).

Structure-preserving sampling. The one-sided latent describes a real field only on the Hermitian manifold, the subset of the complex array satisfying the conjugate symmetry of a real signal, and successive Euler steps drift off it, so we project onto the manifold after every step; forcing the whole $k_x = 0$ slice real would destroy the transverse asymmetry of the streamwise-averaged profile. Sampling integrates from $t = 1$ to $t = \varepsilon$ with a fixed-step Euler solver from a draw of $\mathcal{CN}(0, I)$. Appendix A.6 gives the projection and the Nyquist checkerboard artifact it prevents.

4.4 WHAT WHITENING PRESERVES

Proposition 1 (Spectral conditioning). For $\mathbb{E}|q_k|^2 = E(k) = c\,k^{-\alpha}$ on $k \in [k_{\min}, k_{\max}]$ with $D_k = \sqrt{E(k)}$ as in equation 4, the raw representation has $\kappa(D) = (k_{\max}/k_{\min})^{\alpha/2}$, whereas the whitened representation has $\kappa(D^\xi) = 1$ and $\mathbb{E}|\xi_k|^2 = 1$ for all k.

Proposition 2 (Transport conditioning). In raw Fourier coordinates with a white source, both $\mathrm{Var}(v_k)$ and $\mathrm{SNR}_k(t)$ carry the $k^{-\alpha}$ spectrum, whereas in whitened coordinates $\mathrm{Var}(v_k) = 2$ and $\mathrm{SNR}_k(t) = ((1-t)/t)^2$ for every k.

Proposition 3 (Structure preservation). The whitening map equation 4 is a deterministic, diagonal, invertible affine bijection, and on the centered field it preserves the relative phase of every pair of modes exactly, so that the bispectrum of the fluctuation field is carried over up to a positive real factor, its phase unchanged, and the induced field map is a bijection. For a Gaussian field with per-mode variance σ_k^2, $D_k = \sigma_k$ makes ξ exactly $\mathcal{CN}(0, I)$.

Proposition 3 locates what the network has left to learn, the higher-order and phase structure that whitening leaves in place, and why the residual non-Gaussianity is measured rather than assumed away; Appendix A.1 defines the diagnostics and Appendix A.5 the derivations. The condition number of Proposition 1 grows with the resolved bandwidth and the steepness of the spectrum, the basis for Section 5.5.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets and baselines. Our primary benchmark is the two-dimensional RealPDEBench cylinder-wake dataset (Hu et al., 2026), measured velocity fields (u, v) on a 64×128 grid with $\mathrm{Re} \in [3093, 9656]$ in distribution and the ends of the $\mathrm{Re} \in [1781, 11812]$ range held out;

Table 1: Main comparison, all baselines retrained under one protocol (Section 5.3). Lower is better; entries are means to three significant figures with the column exponent applied. Rows are blocked by dataset and values in different blocks are not comparable; within a block the best value in a column is **bold** and the second best underlined. A dash marks a quantity the block does not have.

Benchmark	Method	MAE ↓ $\times 10^{-2}$	RMSE ↓ $\times 10^{-2}$	Rel. L_2 ↓	$E(k)$ ↓ $\times 10^{-2}$	Vort. ↓	ζ_2 ↓	Enstr. ↓	Dissip. ↓
Cylinder	FNO	1.79	2.99	0.196	1.48	**0.617**	0.132	0.509	0.511
	Transolver	2.97	4.70	0.293	4.90	0.830	0.116	0.854	0.835
	DPOT	**1.45**	**2.55**	**0.163**	1.72	0.651	0.223	0.449	0.499
	WDNO	3.42	5.46	0.355	2.52	1.53	0.640	3.36	5.69
	CoNFiLD	2.77	4.62	0.323	1.50	0.785	0.292	0.580	0.597
	FOURIERFLOW	3.12	4.73	0.292	1.79	0.927	0.247	0.413	0.613
	Ours	1.71	3.45	0.215	**1.46**	0.664	**0.0589**	**0.219**	**0.210**
Foil	FNO	**0.299**	**0.464**	**0.0222**	**0.500**	0.899	0.0752	0.284	0.403
	Transolver	0.347	0.552	0.0264	2.00	**0.796**	0.103	0.797	0.791
	DPOT	0.799	1.04	0.0497	1.01	5.67	0.620	38.3	19.7
	WDNO	1.35	1.89	1.75	8.98	8.17	0.859	2.61	5.52
	CoNFiLD	0.511	0.812	0.1358	1.42	0.912	0.201	0.615	0.388
	FOURIERFLOW	0.395	0.588	0.0395	1.09	0.988	0.289	0.902	0.531
	Ours	0.426	0.636	0.0303	0.670	0.803	**0.0628**	**0.134**	**0.0640**
Combustion	FNO	1.83	4.27	1.24	**1740**	–	0.457	–	–
	Transolver	6.59	10.6	2.93	3290	–	**0.135**	–	–
	DPOT	1.34	3.93	0.941	1850	–	0.208	–	–
	WDNO	3.58	5.50	1.46	10500	–	0.657	–	–
	CoNFiLD	2.21	5.10	1.31	2220	–	0.331	–	–
	FOURIERFLOW	1.87	4.95	1.18	2080	–	0.297	–	–
	Ours	**1.04**	**3.82**	**0.922**	1930	–	0.251	–	–
PoolFire	FNO	19.4	35.1	0.912	388	0.972	0.412	1.36	1.21
	Transolver	22.7	40.6	1.08	456	1.24	0.487	2.05	1.87
	DPOT	15.6	29.8	**0.549**	356	**0.581**	0.318	0.944	0.812
	WDNO	61.3	118	2.94	1840	9.15	1.27	48.6	39.2
	CoNFiLD	17.2	32.5	0.866	361	0.940	0.352	1.12	0.983
	FOURIERFLOW	14.9	30.1	0.842	318	0.912	0.341	1.07	0.996
	Ours	**10.8**	**22.3**	0.803	**237**	0.871	**0.145**	**0.398**	**0.374**

$(\mathrm{Re}, \text{angle of attack})$ index the foil data, $(\mathrm{NH_3}$ fraction, equivalence ratio$)$ the combustion scalars, and the three-dimensional PoolFire benchmark the extensions of Section 5.5. Against these we run the deterministic surrogates FNO (Li et al., 2021), Transolver (Wu et al., 2024), DPOT (Hao et al., 2024) and WDNO (Hu et al., 2025), the generative models CoNFiLD (Du et al., 2024) and FOURIERFLOW (Wang et al., 2025b), and three *representation* baselines sharing our backbone exactly, so any difference is attributable to the coordinate system rather than to architecture. These are (B1) physical-space flow matching, (B2) raw Fourier flow matching, (B3) radially whitened Fourier flow matching and (B4) WSFM. Metrics, training and evaluation protocol are in Appendix A.4.

5.2 WHITENING REMOVES THE SPECTRAL CONDITIONING DEFECT

Whitening acts on the conditioning of the transport problem rather than on the scale of its inputs. Measured on the training ensemble, the gain is exact for the target variance and partial for the learned objective. The ratio $\kappa_{\mathrm{spectral}}$ of maximum to minimum per-mode target-velocity variance falls from 527 under the raw coordinate to 91.7 under radial whitening and to 1.00 under ours, the last by construction rather than measurement, and the Spearman correlation between per-mode loss and wavenumber magnitude at the end of training falls in magnitude from 0.972 to 0.505 and 0.468, reversing sign (Table 4). The learned loss profile is much flatter than the raw one without being flat, and the uniformity we claim is one of target variance rather than of the learned loss.

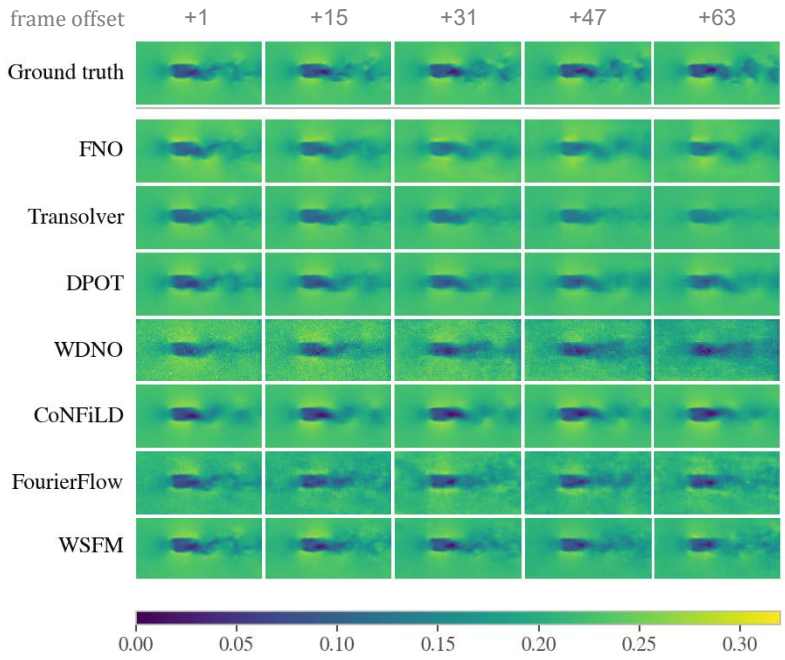


Figure 3: Predictions across the cylinder chunk (Section 5.3). One row for each model of the cylinder block of Table 1 beside the reference, and five columns at frame offsets $0, 16, 32, 48, 64$ from $t^\star$, all at the same Reynolds number and on a shared color scale. Every row advances from the same conditioning frame $t^\star$, so the columns are frame-aligned down the figure and a difference at a given column is a difference in the model.

5.3 Main results

Table 1 reports the main comparison on the same held-out trajectories under the protocol of Appendix A.4. On pointwise accuracy WSFM is never last on MAE, RMSE or relative L_2 and is within $1.5\times$ of the best row; the structure-function, enstrophy and dissipation errors separate more sharply, lowest for WSFM on cylinder by 1.9–$2.4\times$ and on foil by 1.2–$6.1\times$. The energy-spectrum column is the exception, lowest on cylinder by only 1.4% over FNO and third on foil. The controlled comparison shows the mechanism. B2 and B4 of Table 2 share backbone, budget and data and differ only in the coordinate system; WSFM reduces relative L_2 by 26% ($0.292 \to 0.215$) and high-wavenumber energy-spectrum error by 99% ($424 \to 2.97$) at Wilcoxon signed-rank $p < 10^{-4}$. The raw-coordinate model represents low-wavenumber structure but not the small scales carrying almost all of the high-k error. FourierFlow is the architectural answer to the same frequency imbalance, and reimplemented on our skeleton, data and budget it moves part of the small-scale physics away from the raw baseline (6.8% on vorticity error, 26.0% on dissipation) but does not close the conditioning gap, its relative L_2 being identical to B2's ($p = 0.86$). Against WSFM it is behind on every column ($p \leq 3 \times 10^{-2}$), by 26% on relative L_2, 18% on $E(k)$, 28% on vorticity and 76% on ζ_2. The FourierFlow row is our reimplementation and carries 2.0% *more* parameters than ours (82.23M against 80.62M). Only the whitened coordinate was trained on foil, combustion and PoolFire, so those blocks show the representation transferring, and PoolFire is $n = 2$. Figure 3 samples the scored window that Table 1 averages. FNO and Transolver smooth the wake as the rollout lengthens and WDNO carries spurious high-frequency content, whereas WSFM keeps small-scale detail to the end.

5.4 Representation or architecture?

Any representation change invites the objection that it merely substitutes for a larger model. The ablation does not settle that question, a capacity sweep having not been run, but every row shares backbone, schedule and parameter count to within 0.04%, so the contrast is one of coordinate sys-

Table 2: Ablation on the cylinder benchmark (Section 5.4), varying the coordinate system at fixed backbone and schedule. Lower is better; best per column **bold**, second best underlined. (B2+) applies the Hermitian projection at sampling time, High-k is the energy-spectrum error over the upper third of the resolved wavenumber range, and Params is the video backbone size in millions.

Variant	Params	MAE $\times 10^{-2}$	RMSE $\times 10^{-2}$	Rel. L_2	$E(k)$ $\times 10^{-2}$	Vort.	ζ_2	Enstr.	Dissip.	High-k $\times 10^{-1}$
(B1) Physical	80.59	1.74	3.69	0.230	**1.28**	1.07	0.363	0.975	1.48	159
(B2) Raw Fourier	80.62	3.03	4.71	0.292	1.87	0.995	0.243	0.533	0.827	424
(B2+) Raw + Herm.	80.62	3.03	4.70	0.291	2.07	0.985	0.255	0.532	0.785	324
(B3) Radial	80.62	**1.59**	**3.04**	**0.190**	1.38	**0.642**	**0.0494**	**0.216**	**0.153**	3.16
(B4) Ours	80.62	1.71	3.45	0.215	1.46	0.664	0.0589	0.219	0.210	**2.97**

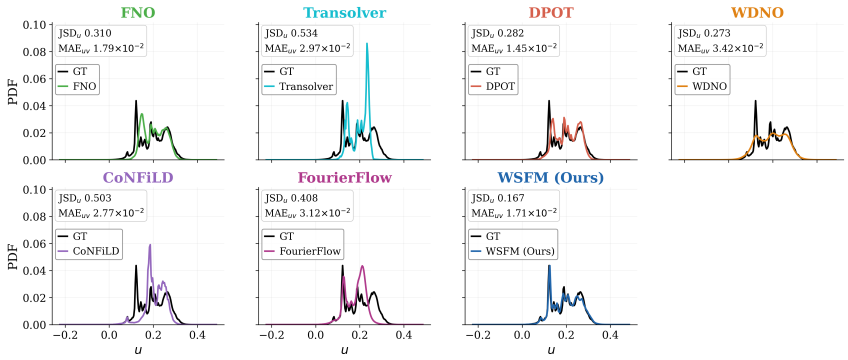


Figure 4: The u marginal behind the first column of Table 3 (Section 5.5). Each panel is the ensemble-mean PDF of u over the scored window, averaged over the 19 held-out trajectories and drawn on shared bins against the same reference curve (black), with JSD_u and MAE_{uv} printed in the panel. CoNFiLD and Transolver shift and broaden the mode around $u \approx 0.2$ that carries most of the mass, DPOT reproduces the two-peak envelope with a displaced first peak, and WSFM tracks the reference throughout.

tem rather than model size. Comparisons are *paired* over the same 19 trajectories, whose standard deviations on relative L_2 are 0.052–0.085; differences below ≈ 0.02 are not resolvable and we read none into them.

Comparing B3 with B4 isolates anisotropy, both whitened, one radially and one per-mode. The predicted signature of the per-mode form, removal of a residual anisotropy in the $E(k_x, k_y)$ of B3 along $k_y \approx 0$ together with better vorticity statistics, does not hold, and the bulk accuracy columns favour the radial variant. What favours WSFM is the high-wavenumber energy-spectrum error, the anisotropic spectrum error of Table 5 and the target variance itself, exactly constant per mode and only equalised on shell averages for the radial form. The two do not separate on the bulk metrics at this budget, so what the per-mode form adds is exactness rather than accuracy (per-column values and p-values in Appendix A.9). Both lie below B2 and B2+ on every column.

5.5 Turbulence physics fidelity

Pointwise accuracy is a weak proxy for turbulence fidelity, an over-smoothed field scoring well while missing the energy-transfer and dissipation structures, so we evaluate the physics directly.

The last four columns of Table 1 report the physics metrics, where the structure-function exponent ζ_2 tests inertial-range scaling rather than merely the spectrum, and Table 3 reports the divergences between generated and reference marginals on cylinder. WSFM holds the lowest value of all six, though the margins are not uniform: $1.6\times$ and $1.1\times$ over the best surrogate on the u and v marginals, and on vorticity only $1.03\times$ over FNO, too narrow to defend, although every other surrogate's vorticity marginal is 1.4–$4.5\times$ WSFM's.

Table 3: Statistical consistency on cylinder against the architectural baselines (Section 5.5). Lower is better in every column; the best value is **bold** and the second best underlined. JSD is the Jensen–Shannon distance in bits and W the Wasserstein-1 distance in physical units, both between generated and reference marginals.

	JSD ($\times 10^{-3}$)			W ($\times 10^{-3}$)		
Method	$u \downarrow$	$v \downarrow$	$\omega \downarrow$	$u \downarrow$	$v \downarrow$	$\omega \downarrow$
FNO	310	166	107	11.0	12.5	1.77
Transolver	534	456	313	20.1	30.2	4.97
DPOT	282	112	151	7.79	8.49	2.42
WDNO	273	314	464	8.48	29.4	9.65
CoNFiLD	503	174	160	22.2	10.6	2.49
FOURIERFLOW	408	167	338	22.3	10.9	4.12
Ours	**167**	**104**	**104**	**4.00**	**6.94**	**1.20**

The coordinate block of Table 6 shows what pointwise accuracy cannot. B1 has the best velocity marginals of the four, both significantly below WSFM's ($p < 10^{-4}$), while its vorticity marginal is more than twice WSFM's. That is the expected trade, the velocity histogram being insensitive to small-scale arrangement, which is what vorticity weights. On the vorticity marginals B3 beats WSFM ($p \leq 4 \times 10^{-6}$), the same ordering as the spectrum and structure-function columns. Foil is the only two-dimensional block WSFM does not lead outright, FNO ahead on five of six divergences by at most $1.8\times$; on combustion, where only u exists, WSFM has the lowest of both. PoolFire is the three-dimensional benchmark and is reported on its own in Table 7.

The conditioning-growth sweeps and the phase and loss-floor diagnostics of Appendix A.1 and A.2 have not been run, so those stand on the propositions alone. Relative to B2 WSFM lowers the error in the energy-containing *and* the dissipative bands at once, and the representation transfers to combustion scalars and to spatiotemporal generation unmodified.

6 Discussion and Limitations

What whitening is and is not. The whitened objective is the raw objective under per-mode reweighting with a corrected source, which yields a falsification criterion: a raw-coordinate model with matched per-mode weights and a spectrally matched source that reproduced these results would show the method reduces to a known technique. **Limitations.** The whitening statistics must be estimated from a training ensemble, so the method presumes a representative set of fields, and a single realization or a strongly non-stationary flow would need a running estimate (Zaki, 2025). The representation is tied to a global Fourier basis on the sampling grid, which represents localized features such as a boundary layer through many coefficients; where the region of interest occupies a small part of the domain the whitening statistics are dominated by the remainder, and a geometry-adapted basis would be the natural choice. Because whitening amplifies dissipative relative to energy-containing scales, it also amplifies measurement noise at high wavenumber, bounding the resolved band over which the equalization usefully extends. **Future work.** The equivalence between whitening and preconditioning applies whenever the data covariance is approximately diagonal in a known basis; combining a whitened representation with frequency-aware schedules is natural, and whether a geometry-aware basis recovers the same benefits in wall-bounded turbulence is the open question.

7 Conclusion

We asked whether generative modeling of turbulence is better conditioned by changing its coordinate system rather than its architecture. Whitening the spectrum removes the energy hierarchy analytically, leaving a transport problem whose per-mode velocity variance and mode-wise SNR are independent of wavenumber while relative phase and higher-order cross-mode structure are preserved exactly. The improvement the ablations measure is therefore one of conditioning rather than model size, and whitening carries it.

AI USE STATEMENT

We used generative AI tools for drafting and editing manuscript text and for assisting with the literature search underlying Section 2. We have not used them to generate experimental results, figures or scientific claims; all quantitative results reported here were produced with code written and verified by the authors, and every cited reference was checked for existence and relevance through a manual literature survey. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work develops generative models of turbulent flow fields from publicly released benchmark datasets. It does not involve human subjects, personal data or practices with direct societal risk, and we are not aware of dual-use concerns specific to this contribution beyond those already general to flow simulation. We report the limitations of the method in Section 6, in particular its dependence on ensemble statistics and its use of a global Fourier basis on the sampling grid.

REPRODUCIBILITY STATEMENT

Section 4 gives the full method, including the whitening transform equation 4, the training interpolation and loss equation 5, the frequency-coordinate positional encoding and the Hermitian projection applied at every sampling step. All hyperparameters, the optimizer and schedule, the training steps, the sampling solver and the whitening warm-up length are listed in Appendix A.4, and proofs of the three propositions in Appendix A.5. Datasets are public benchmark releases (Hu et al., 2026); we will release the training and evaluation code, together with the frozen whitening statistics for each benchmark, upon publication.

REFERENCES

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*. Association for Computing Machinery, 2024. doi: 10.1145/3680528.3687614. URL https://arxiv.org/abs/2401.12945.

Salva Rühling Cachay, Bo Zhao, Hailey Joren, and Rose Yu. Dyffusion: A dynamics-informed diffusion model for spatiotemporal forecasting, 2023. URL https://arxiv.org/abs/2306.01984.

Fangchao Chen, Youhong Xiao, Liang Yu, and Laixu Jiang. Frequency-domain physics-informed neural network for accurate reconstruction of 3d acoustic fields under sparse and multi-frequency measurements. *Neural Networks*, 197:108476, 2026. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2025.108476. URL https://www.sciencedirect.com/science/article/pii/S0893608025013577.

Pan Du, Meet Hemant Parikh, Xiantao Fan, Xin-Yang Liu, and Jian-Xun Wang. Conditional neural field latent diffusion model for generating spatiotemporal turbulence. *Nature Communications*, 15(1), November 2024. doi: 10.1038/s41467-024-54712-1. URL https://doi.org/10.1038/s41467-024-54712-1.

Uriel Frisch. *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge University Press, 1995. ISBN 9780521457132. doi: 10.1017/CBO9781139170666. URL https://doi.org/10.1017/CBO9781139170666.

Zhongkai Hao, Chang Su, Songming Liu, Julius Berner, Chengyang Ying, Hang Su, Anima Anandkumar, Jian Song, and Jun Zhu. Dpot: Auto-regressive denoising operator transformer for large-scale pde pre-training. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 17597–17615, 2024. URL https://proceedings.mlr.press/v235/hao24c.html.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. URL https://arxiv.org/abs/2006.11239.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 8633–8646, 2022. doi: 10.52202/068431-0628. URL https://arxiv.org/abs/2204.03458.

Peiyan Hu, Rui Wang, Xiang Zheng, Tao Zhang, Haodong Feng, Ruiqi Feng, Long Wei, Yue Wang, Zhi-Ming Ma, and Tailin Wu. Wavelet diffusion neural operator. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2412.04833.

Peiyan Hu, Haodong Feng, Hongyuan Liu, Tongtong Yan, Wenhao Deng, Tianrun Gao, Rong Zheng, Haoren Zheng, Chenglei Yu, Chuanrui Wang, Kaiwen Li, Zhi-Ming Ma, Dezhi Zhou, Xingcai Lu, Dixia Fan, and Tailin Wu. Realpdebench: A benchmark for complex physical systems with real-world data. *arXiv preprint arXiv:2601.01829*, 2026. URL https://arxiv.org/abs/2601.01829. ICLR 2026 (Oral).

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24174–24184, 2024. doi: 10.1109/CVPR52733.2024.02282. URL https://arxiv.org/abs/2312.02696.

Byungsoo Kim, Vinicius C Azevedo, Nils Thuerey, Theodore Kim, Markus Gross, and Barbara Solenthaler. Deep fluids: A generative network for parameterized fluid simulations. In *Computer graphics forum*, volume 38, pp. 59–70. Wiley Online Library, 2019.

Andrei N. Kolmogorov. The local structure of turbulence in incompressible viscous fluid for very large reynolds numbers. *Doklady Akademii Nauk SSSR*, 30:301–305, 1941. English translation: Proc. R. Soc. Lond. A 434(1890):9–13, 1991, doi:10.1098/rspa.1991.0075.

Zijie Li, Kazem Meidani, and Amir Barati Farimani. Transformer for partial differential equations' operator learning, 2023. URL https://arxiv.org/abs/2205.13671.

Ziwei Li, Yuhan Duan, Tianyu Xiong, Yi-Tang Chen, Wei-Lun Chao, and Han-Wei Shen. Fa-inr: Adaptive implicit neural representations for interpretable exploration of simulation ensembles, 2026. URL https://arxiv.org/abs/2506.06858.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.08895.

Marten Lienen, David Lüdke, Jan Hansen-Palmus, and Stephan Günnemann. From zero to turbulence: Generative modeling for 3d flow simulation. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2306.01776.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.02747.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024. URL https://arxiv.org/abs/2412.06264.

Ke Liu, MENGXUAN LI, Feng Liu, Jiajun Bu, and Haishuai Wang. INRCT: An end-to-end framework for encoding and generating implicit neural representation, 2025. URL https://openreview.net/forum?id=eajZxJm9gt.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2209.03003.

Runze Mao, Rui Zhang, Xuan Bai, Tianhao Wu, Teng Zhang, Zhenyi Chen, Minqi Lin, Bocheng Zeng, Yangchen Xu, Yingxuan Xiang, Haoze Zhang, Shubham Goswami, Pierre A. Dawe, Yifan Xu, Zhenhua An, Mengtao Yan, Xiaoyi Lu, Yi Wang, Rongbo Bai, Haobu Gao, Xiaohang Fang, Han Li, Hao Sun, and Zhi X. Chen. Benchmarking neural surrogates on realistic spatiotemporal multiphysics flows, 2026. URL https://arxiv.org/abs/2512.18595.

Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, Karthik Kashinath, Jan Kautz, and Mike Pritchard. Residual corrective diffusion modeling for km-scale atmospheric downscaling, 2024. URL https://arxiv.org/abs/2309.15214.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina Agocs, Miguel Beneitez, Marsha Berger, Blakesly Burkhart, Stuart Dalziel, Drummond Fielding, et al. The well: a large-scale collection of diverse physics simulations for machine learning. *Advances in Neural Information Processing Systems*, 37:44989–45037, 2024.

Vivek Oommen, Aniruddha Bora, Zhen Zhang, and George Em Karniadakis. Integrating neural operators with diffusion models improves spectral representation in turbulence modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 481(2309), 2025. doi: 10.1098/rspa.2024.0819. URL https://doi.org/10.1098/rspa.2024.0819.

Meet Hemant Parikh, Xiantao Fan, and Jian-Xun Wang. Conditional flow matching for generative modelling of near-wall turbulence with quantified uncertainty. *Journal of Fluid Mechanics*, 1029, 2026. doi: 10.1017/jfm.2026.11193. URL https://doi.org/10.1017/jfm.2026.11193.

Raghul Parthipan, Mohit Anand, Hannah M. Christensen, J. Scott Hosking, and Damon J. Wischik. Defining error accumulation in ml atmospheric simulators, 2024. URL https://arxiv.org/abs/2405.14714.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023. doi: 10.1109/ICCV51070.2023.00387. URL https://arxiv.org/abs/2212.09748.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Learning mesh-based simulation with graph networks. In *International conference on learning representations*, 2020.

Stephen B. Pope. *Turbulent Flows*. Cambridge University Press, 2000. ISBN 9780511840531. doi: 10.1017/CBO9780511840531. URL https://doi.org/10.1017/CBO9780511840531.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2108.12409.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Khalid Rafiq and Aditya G. Nair. Spectrally regularized latent flow matching for turbulence generation. *arXiv preprint arXiv:2606.11691*, 2026. URL https://arxiv.org/abs/2606.11691. Accepted at the AI4Physics Workshop at ICML 2026.

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

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022. URL https://arxiv.org/abs/2112.10752.

Paul Rosu, Muchang Bahng, Erick Jiang, Rico Zhu, and Vahid Tarokh. A pde-informed latent diffusion model for 2-m temperature downscaling, 2025. URL https://arxiv.org/abs/2510.23866.

François Rozet, Ruben Ohana, Michael McCabe, Gilles Louppe, François Lanusse, and Shirley Ho. Lost in latent space: An empirical study of latent diffusion models for physics emulation. *arXiv preprint arXiv:2507.02608*, 2025. URL https://arxiv.org/abs/2507.02608.

Francesco M. Ruscio and T. Konstantin Rusch. Low-pass flow matching. *arXiv preprint arXiv:2606.02177*, 2026. URL https://arxiv.org/abs/2606.02177. ICLR 2026 Delta Workshop.

Anish Sambamurthy and Ashesh Chattopadhyay. Lazy diffusion: Mitigating spectral collapse in generative diffusion-based stable autoregressive emulation of turbulent flows. *arXiv preprint arXiv:2512.09572*, 2025. URL https://arxiv.org/abs/2512.09572.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pp. 8459–8468. PMLR, 2020.

Louis Serrano, Lise Le Boudec, Armand Kassaï Koupaï, Thomas X. Wang, Yuan Yin, Jean-Noël Vittaut, and Patrick Gallinari. Operator learning with neural fields: Tackling pdes on general geometries. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70581–70611, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d771c14ec4a2df249b54d297f02cbdaf-Abstract-Conference.html.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2011.13456. Oral presentation.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

A. van der Schaaf and J. H. van Hateren. Modelling the power spectra of natural images: Statistics and information. *Vision Research*, 36(17):2759–2770, 1996. doi: 10.1016/0042-6989(96)00002-8. URL https://doi.org/10.1016/0042-6989(96)00002-8.

Haiteng Wang, Lei Ren, Yikang Li, and Yuqing Wang. Metaindux-ts: Frequency-aware aigc foundation model for industrial time series. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):18062–18074, 2025a. doi: 10.1109/TNNLS.2025.3577203. URL https://doi.org/10.1109/TNNLS.2025.3577203.

Haixin Wang, Jiashu Pan, Hao Wu, Fan Zhang, and Tailin Wu. Fourierflow: Frequency-aware flow matching for generative turbulence modeling. *arXiv preprint arXiv:2506.00862*, 2025b. URL https://arxiv.org/abs/2506.00862.

Tim Whittaker, Romuald A. Janik, and Yaron Oz. Turbulence scaling from deep learning diffusion generative models. *Journal of Computational Physics*, 514:113239, 2024. doi: 10.1016/j.jcp.2024.113239. URL https://doi.org/10.1016/j.jcp.2024.113239.

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 (ICML)*, volume 235 of *PMLR*, pp. 53656–53680, 2024. URL https://proceedings.mlr.press/v235/wu24q.html.

Yuan Yin, Matthieu Kirchmeyer, Jean-Yves Franceschi, Alain Rakotomamonjy, and Patrick Gallinari. Continuous pde dynamics forecasting with implicit neural representations, 2023. URL https://arxiv.org/abs/2209.14855.

Zhiyu Yin, Kehai Chen, Xuefeng Bai, Ruili Jiang, Juntao Li, Hongdong Li, Jin Liu, Yang Xiang, Jun Yu, and Min Zhang. A survey: Spatiotemporal consistency in video generation. *ACM Comput. Surv.*, April 2026. ISSN 0360-0300. doi: 10.1145/3802588. URL https://doi.org/10.1145/3802588. Just Accepted.

Tamer A. Zaki. Turbulence from an observer perspective. *Annual Review of Fluid Mechanics*, 57 (1):311–334, 2025. doi: 10.1146/annurev-fluid-030424-114735. URL https://doi.org/10.1146/annurev-fluid-030424-114735.

Rui Zhang, Qi Meng, Han Wan, Yang Liu, Zhi-Ming Ma, and Hao Sun. Omnifluids: Physics pretrained modeling of fluid dynamics, 2025. URL https://arxiv.org/abs/2506.10862.

Anthony Zhou and Amir Barati Farimani. Masked autoencoders are pde learners, 2024. URL https://arxiv.org/abs/2403.17728.

Anthony Zhou, Zijie Li, Michael Schneier, John R Buchanan Jr, and Amir Barati Farimani. Text2pde: Latent diffusion models for accessible physics simulation. *arXiv preprint arXiv:2410.01153*, 2024.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models, 2023. URL https://arxiv.org/abs/2211.11018.

A Appendix

A.1 Phase diagnostics

The three phase-sensitive quantities of Section 5.5 are the phase-increment distribution $\arg q_k - \arg q_{k'}$ between neighbouring modes, the per-shell phase resultant $R = |\mathbb{E}[e^{\mathrm{i}\arg q_k}]|$, and the bicoherence $b(k_1, k_2) = |B|/\sqrt{\mathbb{E}|q_1 q_2|^2\,\mathbb{E}|q_3|^2}$ with $B = \mathbb{E}[q(k_1)q(k_2)\overline{q(k_1+k_2)}]$, which detects triad interactions.

A.2 Loss-floor diagnostic

We separate the two explanations for a loss plateau with three predictors of increasing power on the whitened latent: the zero predictor $v = 0$, the optimal *linear* predictor $v = a(t)\xi_t$ with $a(t) = (2t-1)/(1-2t+2t^2)$ (Appendix A.5), and a non-parametric per-mode Bayes predictor $\hat{\xi}_0 = \mathbb{E}[\xi_0 \mid \xi_t]$ estimated by kernel regression within each wavenumber mode. The last is an *upper* bound on the true Bayes floor, since it ignores cross-mode coupling, so the gap $L_{\mathrm{model}} - L_{\mathrm{Bayes}}$ bounds the cross-mode structure that remains available to be learned. A model sitting close to the Bayes line in all bands has hit a statistical floor and its headroom lies in cross-mode modelling, whereas one sitting well above the line at high k but matching it at low k has an architectural bottleneck localised to small scales. The comparison has not been run, so the diagnostic stands as a defined procedure rather than a measured result.

A.3 Background and notation

Diffusion (Ho et al., 2020; Song et al., 2021) and flow matching (Lipman et al., 2023; 2024) transport a source $p_{\mathrm{src}} = \mathcal{N}(0, I)$ to data p_{data} along a chosen interpolation. For the straight-line path of rectified flow (Liu et al., 2023), $x_t = (1-t)x_0 + t\,x_1$ with $x_0 \sim p_{\mathrm{data}}$ and $x_1 \sim p_{\mathrm{src}}$, the

conditional target velocity is constant along the path, $\dot{x}_t = x_1 - x_0$, and a network v_θ regressed on it recovers the marginal velocity field that generates p_{data}; sampling integrates $\mathrm{d}x/\mathrm{d}t = v_\theta(x,t)$ from $t = 1$ to $t = \varepsilon$. The construction presumes a state space in which the source, the data and the interpolant are written, and that space is a design decision rather than a fixed part of the framework. Posing transport in a transformed coordinate system is established practice, whether the transform is learned, as in latent diffusion (Rombach et al., 2022), or fixed, as in wavelet-domain diffusion for physical fields (Hu et al., 2025). The choice of coordinates fixes the regression problem that the network is trained on, for a target distribution that the choice itself leaves unchanged.

The transform of Section 3 is the one-sided orthonormal real FFT, $q = \mathrm{rFFT}_d(u) \in \mathbb{C}^{C\times N_1\times\cdots\times N_{d-1}\times N_f}$ with $N_f = \lfloor N_d/2 \rfloor + 1$. It is a linear bijection between the grid values and the Hermitian coefficient array, since the negative-k_d half plane conjugates the positive half. Every quantity defined on q is defined per mode and is therefore independent of the dimension d.

A.4 Experimental protocol

Metrics. We report three families. *Prediction accuracy* covers MAE, RMSE and relative L_2 error. *Physical fidelity* covers the energy-spectrum error $\|\hat{E}(k) - E(k)\|_1/\|E(k)\|_1$, the anisotropic spectrum error on $E(k_x, k_y)$, the vorticity error, the enstrophy $\Omega = \langle\omega^2\rangle$, the dissipation $\epsilon \propto 2\nu\langle S_{ij}S_{ij}\rangle$, and the structure-function exponents ζ_p obtained from $S_p(r) = \langle|\delta u(r)|^p\rangle$; we report $p = 2$ and $p = 4$, the two orders the pipeline stores. *Statistical consistency* covers the Jensen–Shannon divergence between the generated and reference PDFs of velocity, vorticity magnitude and dissipation, together with a bicoherence error that probes nonlinear phase coupling.

Implementation. SpectralDiT uses patch size 4, width 512, depth 12, 8 heads of dimension 64 and MLP ratio 4.0. On the 64×65 one-sided latent of the cylinder benchmark this gives 272 tokens per frame; the image backbone we initialize from has 57.9M parameters, and the video extension used in all experiments of Section 5 has 80.6M. Training uses AdamW at learning rate 10^{-4} with no learning-rate schedule and no gradient clipping, batch size 1 and 60K steps, with EMA decay 0.995 for sampling (Karras et al., 2024). We use the linear interpolant with an L2 drift loss, Euler integration with 100 steps to $\varepsilon = 10^{-3}$, and whitening statistics frozen after a 200-batch warm-up. All models train on a single NVIDIA A40.[1]

Evaluation protocol. One script computes every number reported in Section 5 from the stored fields, and it reproduces the reference pipeline to 0–1×10^{-7} on every cylinder metric the two share. The scored window of each dataset starts at frame $t^\star = T/2$ of the trajectory, so that a real past exists on both sides of it. The deterministic surrogates are given the 20 frames ending at $t^\star$ as genuine input, not a tiled copy, roll out autoregressively in 20-frame blocks from $t^\star + 1$, and are scored on $[t^\star, t^\star + W)$ with W the dataset's window length (64 on cylinder, 32 on foil and combustion, 16 on PoolFire); the generative rows, namely WSFM, FourierFlow and CoNFiLD, are guided on the same single frame $t^\star$ at guidance scale $d_{\mathrm{ps}} = 2$ and generate the rest of the same window. The scored chunk therefore holds exactly one reference frame for every row, the surrogates conditioning directly and the generative rows through guidance, and on cylinder the two families are scored on bit-identical fields ($\max|\Delta| = 0$). Guidance does not transfer across datasets: it is worse on foil, so foil and combustion are reported unguided under one protocol and PoolFire uses first-frame replacement. Every row of Tables 1, 3, 6 and 7 observes the same single frame under the same rule, so a lead there is architectural rather than a difference in available context. Both divergences of Tables 3, 6 and 7 are computed on the flattened field over all frames of a trajectory, with the Wasserstein distance evaluated on a 1-in-20 subsample of that vector for cost. The surrogate rows of those tables were rescored from their stored prediction fields by the same code that scores WSFM, which reproduces the WSFM rows to 1×10^{-6} relative on cylinder and 3×10^{-7} on foil.

[1] These numbers are read from the checkpoints and training logs of the runs behind Tables 1 and 2, not from a reference configuration: the backbone is smaller and the batches thinner than an image-scale recipe, and we report what was actually run. The parameter counts are measured, and the schedule is stated by its absence because the training script supports cosine annealing and gradient clipping but neither was enabled in these runs.

A.5 Additional proofs and derivations

Whitening as loss reweighting. Under a spectrally matched source $\tilde{x}_1 = D\,\xi_1$ the raw-coordinate L1 objective becomes $\sum_k D_k \mathbb{E}|v_k^{\xi} - v_{\theta,k}^{\xi}|$, so the whitened objective is the matched-source objective under per-mode weights D_k^{-1}. Whitening is therefore the change of variables that makes those weights uniform, and it corrects the source as well, which reweighting alone does not. Radial whitening is the same transform with a coarser D, so the gap between those rows of Table 2 isolates spectral anisotropy rather than any architectural effect.

Optimal linear predictor. For the whitened interpolant $\xi_t = (1-t)\xi_0 + t\xi_1$ with $\xi_0 \sim \mathcal{CN}(0, I)$ (to second order) and $\xi_1 \sim \mathcal{CN}(0, I)$ independent, $\mathrm{Cov}(\xi_t) = ((1-t)^2 + t^2)I$ and $\mathrm{Cov}(\xi_t, \xi_0) = (1-t)I$, so that the linear minimum-mean-square predictor of ξ_0 given ξ_t is $\hat{\xi}_0 = \frac{1-t}{(1-t)^2+t^2}\xi_t$. Since $v = \xi_1 - \xi_0$ and $\xi_1 = (\xi_t - (1-t)\xi_0)/t$, the induced velocity predictor is $a(t)\xi_t$ with $a(t) = \frac{2t-1}{1-2t+2t^2}$, which is the expression used in the loss-floor diagnostic of Section 5.5. The per-mode regression used for the Bayes floor replaces this linear rule by the exact posterior mean under a per-mode empirical prior.

Whitening and the Jacobian. The map equation 4 has Jacobian $\mathrm{diag}(D^{-1})$ as a real-linear map on the real degrees of freedom of the Hermitian latent. Its determinant is constant, so that the change of variables contributes only an additive constant to any log-density and the maximum-likelihood problem in whitened coordinates is the raw problem under fixed per-mode weights. This is the formal content of the remark in Section 4.4.

A.6 The Hermitian manifold and its projection

The one-sided array q is not an arbitrary complex array. It is the image of a real field under rFFT_d, so it lies on the Hermitian manifold, the subspace of arrays whose entries satisfy the conjugate symmetry inherited from the reality of u. Concretely, the last axis k_x is half-range, $k_x \in [0, N_x/2]$ with $N_x^f = \lfloor N_x/2 \rfloor + 1$, while the transverse axes k_y and k_z run over the full signed range and are tied to their reflections by $q(k_x, -k_y, k_z) = \overline{q(k_x, k_y, k_z)}$ and likewise for k_z, with the two reflections combined at the corners. Any array outside this subspace inverts to a complex field, and irFFT_d silently discards the imaginary part, which for a boundary slice that has drifted off the manifold appears as a checkerboard at the Nyquist frequency.

Projection. Let F be the one-sided latent after an Euler step, of shape $N_y \times N_x^f$ in two dimensions and $N_y \times N_z \times N_x^f$ in three, with x the streamwise axis. The projection acts in two stages; the distinction between them is what makes it easy to get wrong. *Self-conjugate modes*, the indices whose transverse coordinates each lie at zero or at the Nyquist value, are their own conjugates and must therefore be real, so F is replaced by $\mathrm{Re}\,F$ there. *Interior conjugate pairs*, $k_y \leftrightarrow N_y - k_y$ and $k_z \leftrightarrow N_z - k_z$, carry genuine information in their antisymmetric part, and replacing both members by their symmetrized average $\frac{1}{2}(F_k + \overline{F_{k'}})$ is the orthogonal projection onto the subspace of arrays that invert to a real field. The $k_x = 0$ slice requires care because it is not uniformly self-conjugate. Only its transverse origin and Nyquist lines are; the rest of the slice is complex and encodes the transverse asymmetry of the streamwise-averaged profile. Forcing the whole slice real discards that asymmetry and biases the mean flow toward mirror symmetry, so the projection forces real only the intersection with the self-conjugate set and symmetrizes the remainder. The projection is applied after every integration step. At $t = 1$ the draw from $\mathcal{CN}(0, I)$ has real self-conjugate entries by construction, so the first projection is a no-op and the operator acts only on drift accumulated along the trajectory.

A.7 Whitening statistics and warm-up

The per-mode mean $\mu_c(k)$ and standard deviation $D_c(k)$ are accumulated over a warm-up of 200 batches drawn from the training loader, using the per-mode energy $\sum_b |q_c^{(b)}(k)|^2$ and the per-mode complex sum $\sum_b q_c^{(b)}(k)$. After the warm-up both grids are frozen and stored as non-trainable parameters, so that exponential moving averages over model parameters keep the sampling copy

synchronized and so that checkpoints carry the spectrum. A one-time diagnostic verifies that the DC mode has mean ≈ 0 and standard deviation ≈ 1 after whitening, as required for it to match the $\mathcal{CN}(0, I)$ prior; under min-max normalization the DC coefficient is a large constant, so that this check fails for scale-only whitening and is the reason the affine form is used.

A.8 BACKBONE CONFIGURATION

For completeness we list the SPECTRALDIT configuration used in all experiments. Patch size is 4, width 512 and depth 12, with 8 attention heads of dimension 64 and an MLP ratio of 4.0. The frequency-coordinate encoding uses 12 log-spaced sinusoids per coordinate spanning 2^0 to 2^4, the timestep is embedded by a Fourier embedding of the continuous $t \in [0, 1]$, and modulation is adaLN-Zero with zero-initialized output projections in every block and at the head. On the 64×65 one-sided latent that our experiments use this yields 272 tokens per frame; the image backbone carries 57.9M parameters and the video backbone of Section 5 carries 80.6M. The video extension, in the manner of video diffusion models (Ho et al., 2022; Bar-Tal et al., 2024), factorizes attention into a per-frame spatial pass and a cross-frame temporal pass on the same tokens, with a zero-initialized sinusoidal frame encoding so that the spatial branch initializes to exactly the image model; the temporal attention adds a band-conditioned learnable lag bias of one free scalar per (head, radial band, signed frame offset).

A.9 ADDITIONAL EXPERIMENTAL RESULTS

The per-column detail behind Section 5.4 is as follows, all paired over the same 19 trajectories. Between (B3) and (B4), the anisotropic spectrum error (4.85 against 4.23) is not resolved ($p = 0.11$); the vorticity marginal divergence is 89.6 against 104 ($p \leq 4 \times 10^{-6}$) and the vorticity error 0.642 against 0.664 ($p = 1.1 \times 10^{-5}$); relative L_2 separates by a paired margin of 0.025 (0.190 against 0.215, $p = 6.5 \times 10^{-4}$) and the dissipation error by 0.057 (0.153 against 0.210, $p \leq 4 \times 10^{-6}$).

Table 4: Measured conditioning under each representation (Section 5.2), computed on the training ensemble before any optimization. Lower is better except in the last column, which is higher-is-better; best among the comparable rows **bold**, second best <u>underlined</u>.

Representation	$\kappa_{\text{spectral}} \downarrow$	$\text{Var}(v_k)$ range	$\lvert r(L, \lvert k \rvert) \rvert \downarrow$	Inertial-range loss $\uparrow$
Physical space	5.78	0.0103–0.0597	–	–
Raw Fourier (B2)	527	1.00–527	0.972	0.550
Radial whitening (B3)	<u>91.7</u>	1.01–92.3	<u>0.505</u>	<u>1.03</u>
Mode-wise whitening (Ours, B4)	**1.00**	2.00–2.00	**0.468**	**1.05**

In Table 4, κ_{spectral} is the ratio of maximum to minimum per-mode target-velocity variance and $r(L, |k|)$ the Spearman correlation between per-mode loss and wavenumber magnitude at the end of training, reported as $|r|$ because the white-coordinate prediction is a *flat* profile, $r = 0$, rather than a negative one; the signed values are given below because $|r|$ alone would hide that reversal. The inertial-range loss ratio is the mean loss over the inertial bands divided by the mean loss over all modes with $|k| \geq 1$, so 1.0 is the uniform profile. Physical space is measured per pixel rather than per mode, so the last two columns have no meaning there, neither its dashes nor its κ_{spectral} are comparable with the three per-mode rows, and it therefore carries no marks. κ_{spectral} is computed on the unrounded per-mode variances and so is not exactly reproducible from the two endpoints printed in the range column. Both whitened representations sit above 1.0, whitening equalising the target variance while the learned loss profile stays mildly tilted, now towards high wavenumbers rather than away from them ($r_{\text{B4}} = +0.468$, $r_{\text{B3}} = +0.505$, $r_{\text{B2}} = -0.972$). The correlations are Spearman coefficients over the per-mode loss at the fixed time $t = 1/2$, estimated from 91 training trajectories (5824 samples per mode), each significant at $p < 10^{-4}$. The B3 and B4 rows differ by 0.037, small enough that we do not read a ranking into it.

Table 5 uses the $n = 19$ held-out cylinder trajectories with DPS guidance ($d_{\text{ps}} = 2$), under the protocol and script of Table 1; its ζ_2 column repeats the corresponding column of Table 2 unchanged. Only the even structure-function orders appear because the pipeline stores $p = 2$ and $p = 4$ only, so

Table 5: Turbulence-statistics comparison (Section 5.5). Extends the coordinate-system variants of Table 2 with the anisotropic spectrum error and the fourth-order structure-function exponent. Every column is an error against the reference, so lower is better; best per column **bold**, second best underlined.

Method	$E(k_x, k_y)$ ($\times 10^{-2}$)	ζ_2	ζ_4	Bicoherence
(B1) Physical-space FM	**1.53**	0.363	0.474	0.166
(B2) Raw Fourier FM	18.0	0.243	0.365	**0.161**
(B3) Radial whitened	4.85	**0.0494**	**0.119**	0.165
(B4) Ours	4.23	0.0589	0.143	0.164

Table 6: Statistical consistency on the coordinate ablation and the two-dimensional benchmarks (Section 5.5). Conventions, exponents and protocols are those of Table 3, and each row is the same run as its counterpart in Table 1. Lower is better; best value in a block is **bold**, second best underlined, and a dash marks a quantity the block does not have.

Dataset	Method	JSD ($\times 10^{-3}$)			W ($\times 10^{-3}$)		
		$u\downarrow$	$v\downarrow$	$\omega\downarrow$	$u\downarrow$	$v\downarrow$	$\omega\downarrow$
Cylinder	(B1) Physical-space FM	**82.6**	**72.0**	222	**1.45**	**3.58**	7.43
	(B2) Raw Fourier FM	417	144	326	23.4	10.8	4.17
	(B2+) Raw + Hermitian proj.	425	145	325	23.5	11.3	4.13
	(B3) Radial whitened	177	95.1	**89.6**	4.48	6.11	**1.08**
	(B4) Ours	167	104	104	4.00	6.94	1.20
Foil	FNO	**420**	**91.9**	**74.3**	**2.50**	**0.69**	**0.12**
	Transolver	611	284	355	3.04	2.53	0.56
	DPOT	718	180	777	8.74	1.26	6.81
	WDNO	903	878	918	290	265	266
	CoNFiLD	507	631	704	42.8	51.3	66.2
	FourierFlow	762	719	837	68.5	74.9	88.1
	Ours	503	164	76.8	3.00	1.02	**0.12**
Combustion	FNO	688	–	–	13.9	–	–
	Transolver	916	–	–	65.5	–	–
	DPOT	657	–	–	7.00	–	–
	WDNO	804	–	–	27.0	–	–
	CoNFiLD	871	–	–	44.2	–	–
	FourierFlow	933	–	–	71.6	–	–
	Ours	**248**	–	–	**5.94**	–	–

ζ_3 and ζ_6 are not computed. The anisotropic error $E(k_x, k_y)$ falls from 18.0 for B2 to 4.23 for Ours ($4.3\times$, paired $p = 6.2 \times 10^{-3}$) and ζ_4 from 0.365 to 0.143 ($2.6\times$, $p = 1.2 \times 10^{-2}$). B3 and B4 do not separate on $E(k_x, k_y)$ ($p = 0.11$), and on ζ_4 they are ordered the other way, B3 being 17% lower (0.119 against 0.143, $p = 0.023$), so across the two columns the pattern is split rather than ordered and at this budget neither difference resolves the pair. The bicoherence column is flat across all four rows (0.161–0.166, a spread under 3%), the expected outcome for a phase-neutral rescaling (Proposition 3); we therefore read it as a consistency check on the whitening rather than as a metric on which the variants separate, and do not treat its bold and underlined entries, which sit inside that spread, as a ranking.

Table 7: Statistical consistency on the three-dimensional PoolFire benchmark (Section 5.5). Conventions, exponents and protocol are those of Table 3, and each row is the same run as its counterpart in Table 1. Velocity has three components on this benchmark, so the divergences are reported for u, v, w and the vorticity magnitude ω; lower is better in every column, and the best value is **bold** and the second best underlined.

	JSD ($\times 10^{-3}$)				W ($\times 10^{-3}$)			
Method	$u \downarrow$	$v \downarrow$	$w \downarrow$	$\omega \downarrow$	$u \downarrow$	$v \downarrow$	$w \downarrow$	$\omega \downarrow$
FNO	402	371	438	**187**	18.6	31.8	34.2	**16.4**
Transolver	586	524	611	356	52.8	27.9	58.9	44.1
DPOT	486	351	421	471	**15.2**	28.7	42.1	39.4
WDNO	734	698	597	663	74.2	69.5	81.6	64.8
CoNFiLD	905	891	868	812	92.4	87.6	96.3	84.1
FourierFlow	942	918	946	857	118	104	121	103
Ours	**110**	**107**	**196**	203	19.9	**18.5**	**19.1**	18.2

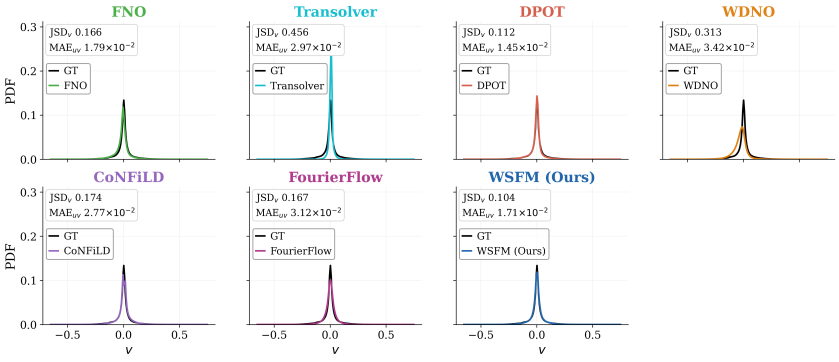


Figure 5: The v marginal behind the second column of Table 3 (Section 5.5). Panels, reference curve, bins and averaging are those of Figure 4, with JSD_v printed in place of JSD_u; the cross-trajectory envelope is $2.0\times$ the mean peak here, so it is again omitted. The transverse component is far more concentrated than the streamwise one and the differences are correspondingly narrower, which is why the margin over the best surrogate in this column is the smaller of the two.

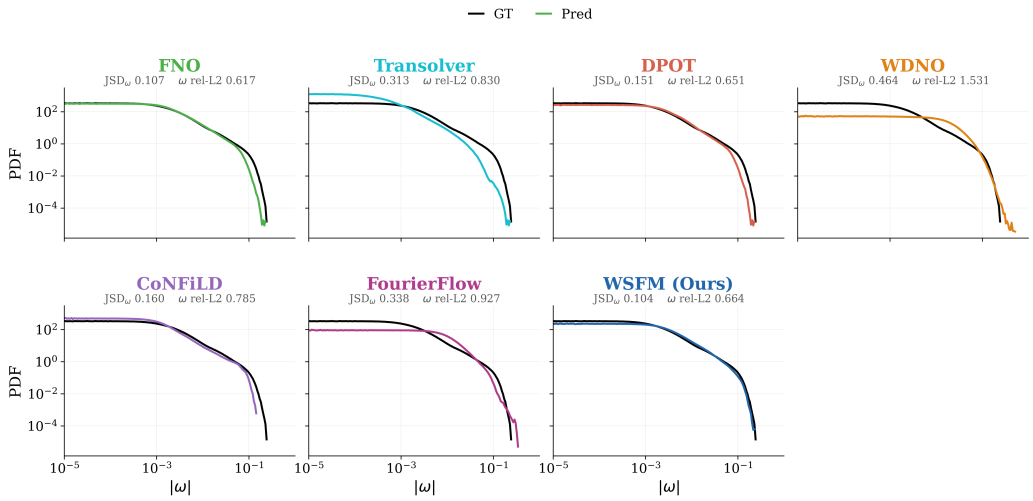


Figure 6: The vorticity-magnitude marginal behind the third column of Table 3 (Section 5.5), on log–log axes. Both axes are logarithmic because $|\omega|$ spans about ten decades and on a linear axis 54% of the reference mass would sit in the first bin; the plotted window is $[10^{-5}, 10^{0}]$, below which the flow is quiescent and $|\omega|$ sits at the finite-difference noise floor, and the excluded mass is at most 1.2% for every model. Each panel carries the divergence JSD_ω and the vorticity relative L_2 of that table. WDNO and FourierFlow have not reached the dissipative tail at the plotted cutoff, which is the spurious-high-frequency failure mode Figure 3 shows in the fields.

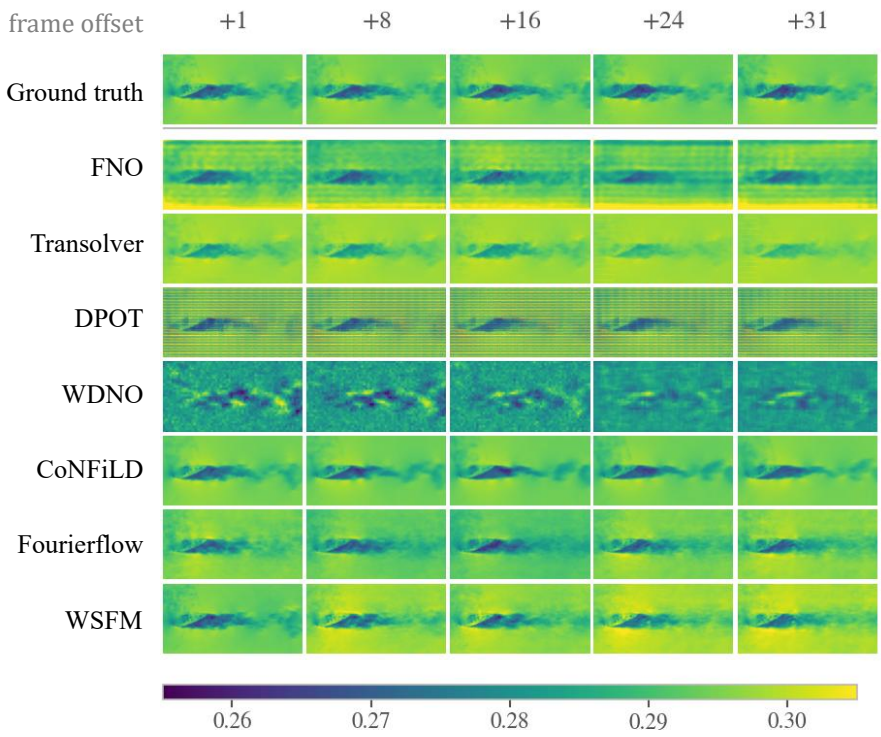


Figure 7: Predictions across the foil chunk (Section 5.5). One row per model and a common set of frames sampling the 32-frame scored window at offsets $+1, +8, +16, +24, +31$, all at the same Reynolds number and angle of attack, on the shared color scale of the bar. The rows are the ground truth, FNO, Transolver, DPOT, WDNO, CoNFiLD, FOURIERFLOW and WSFM; the coordinate ablation is run on cylinder only and does not appear here. Foil is reported unguided (Appendix A.4), so every row sees the same conditioning and the contrast between rows is the model rather than the available context.

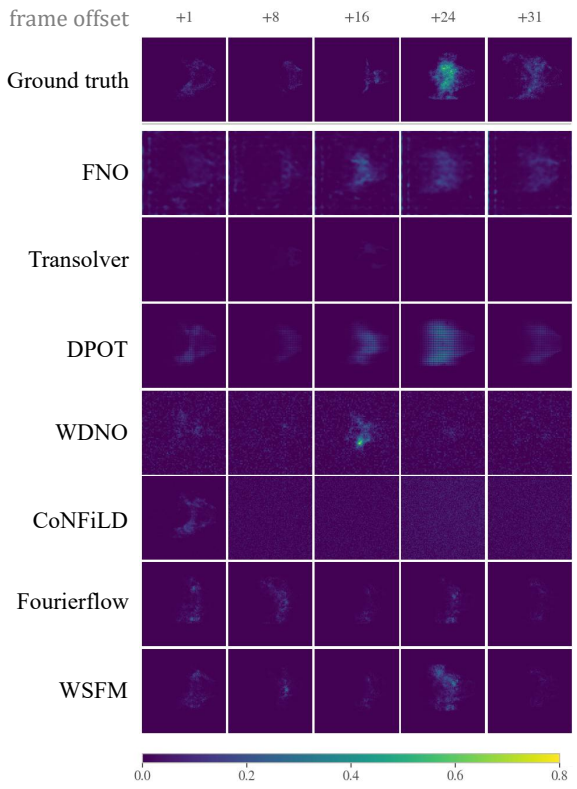


Figure 8: Predictions across the combustion chunk (Section 5.5). Only the u component exists on this benchmark, so each cell is that field alone; rows and frame offsets are those of Figure 7, at the same (NH_3 fraction, equivalence ratio) and on the shared color scale of the bar. Transolver collapses to a near-uniform field and CoNFiLD loses the plume after the first frame, whereas WSFM keeps a localized structure over the whole window.