Wavenumber-Resolved Spectral Gating for Parameter-Efficient Generative Modeling of Two-Dimensional Turbulence

Anonymous authors
Paper under double-blind review

Abstract

We study unconditional generative modeling of two-dimensional Kolmogorov flow with denoising diffusion. A plain U-Net diffusion model reproduces the point-wise statistics of turbulent vorticity fields but distorts the inertial-range spectrum and the spectral fluxes that carry energy and enstrophy across scales. We add a wavenumber-resolved spectral gate: a small Fourier-domain bottleneck that learns a per-channel multiplicative correction over a coarse radial-wavenumber grid, conditioned on the diffusion noise level and a continuous log-viscosity parameter. The gate is paired with physics losses on enstrophy, modal spectral amplitude, vorticity structure functions, integral length, and spectral flux. On forced 2D Kolmogorov flow at seven viscosity regimes spanning a factor of nine in Reynolds number, and against four unconditional baselines (a plain U-Net, a squeeze-and-excite bottleneck, an FNO-block bottleneck at 2.8× the parameters, and a standalone Fourier neural operator), the gated model (WRSD, 2.51M parameters) reduces the aggregate log-spectral distance to direct numerical simulation (DNS) by 59% relative to the plain U-Net, beats the larger FNO block on that metric, and raises inverse-energy and forward-enstrophy cascade recovery from 10% and 21% to 49% and 89%. Adding an FNO branch to the gate (WRSD+FNO, 7.27M) is best overall, reaching 84% and 110% cascade recovery and the lowest flux error. A factorial ablation separates the gate, which routes spectral power and corrects the large scales and cascade transport, from the losses, which correct small-scale point statistics, and shows a strong positive interaction: the gate alone barely lowers the spectral distance and the losses alone raise it, but together they cut it by more than half. A one-parameter conformal rescaling raises empirical 90% coverage from 0.70–0.84 to 0.85–0.89. We report two negative results: the physics losses worsen the large-scale integral length relative to the gate alone, and no variant fully recovers the DNS cascade peaks. Code is available at `https://anonymous.4open.science/r/wrsd-turbulence`.

1 Introduction

Two-dimensional turbulence is a standard testbed for studying the inverse energy cascade and the forward enstrophy cascade (Kraichnan, 1967; Batchelor, 1969). It is also a demanding testbed for generative models of physical fields. The data is high-dimensional, multi-scale, and constrained by conservation laws whose violation is easy to detect: a generator that fits the mean and variance but distorts the spectrum is exposed by a single log–log plot. A generative model of 2D turbulence can therefore be held to a physical, rather than perceptual, standard.

Denoising diffusion produces high-quality unconditional generators for natural images (Ho et al., 2020; Song et al., 2021; Karras et al., 2022), and recent work applies it to turbulent flows (Shu et al., 2023; Lippe et al., 2023; Kohl et al., 2024). Most of that work studies *conditional* surrogates that predict the next state from the previous one. We study the *unconditional* setting, in which the model draws independent samples from the steady-state distribution given only the flow's viscosity. This setting is harder, because it has no

previous-state input to anchor the prediction, and it is the relevant one for ensemble forecasting and data assimilation (Rozet & Louppe, 2023), which need independent draws from the flow's invariant measure.

In the unconditional setting, the dominant failure mode of a U-Net diffusion model is not amplitude or variance, which are usually correct after enough training. It is the *spectrum*: the distribution of power across wavenumbers, and the spectral fluxes that move energy and enstrophy between scales. We document this failure and address it with two ingredients, evaluated against four unconditional baselines (Section 4.3).

First, we introduce a Fourier-domain bottleneck, the *wavenumber-resolved spectral gate*, which amplifies or attenuates the spectral bands the rest of the network mishandles. Second, we add a set of soft physics losses on enstrophy, modal spectral amplitude, vorticity structure functions, integral length, and spectral flux. A single model conditioned on a continuous log-viscosity scalar covers all seven Reynolds regimes.

Our contributions are:

1. A wavenumber-resolved spectral gate that, with matched physics losses, improves the spectral and cascade fidelity of an unconditional diffusion generator without changing the U-Net's spatial structure. One continuous-viscosity model spans seven regimes.

2. A factorial ablation that separates the gate (spectral routing, large scales, and cascade transport) from the losses (small-scale point statistics), revealing a positive interaction in which the two components are unhelpful or harmful in isolation yet together more than halve the log-spectral distance.

3. A parameter-efficiency comparison in which the gated model (WRSD, 2.51M) beats an FNO-block bottleneck (7.08M) and a standalone FNO operator (4.97M) on spectral and cascade fidelity at 35% of the FNO block's parameter count.

4. A hybrid extension (WRSD+FNO) that adds an FNO branch to the gate and is best overall, together with a one-parameter conformal recalibration of the per-pixel ensemble interval and an analysis of where every variant still fails.

2 Background

Denoising diffusion. Given a clean field x_0, the forward process adds Gaussian noise at scale σ to produce $x_\sigma = x_0 + \sigma\,\epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$. A network learns to denoise x_σ, and sampling starts from pure noise and anneals σ to zero. We use the EDM parameterization of Karras et al. (2022), in which the network output is

$$D(x;\sigma) = c_{\text{skip}}(\sigma)\,x + c_{\text{out}}(\sigma)\,f_\theta(c_{\text{in}}(\sigma)\,x; c_{\text{noise}}(\sigma))\,, \tag{1}$$

where f_θ is the trainable network and the $c_{(\cdot)}(\sigma)$ are fixed preconditioning coefficients that keep the network input and target at unit scale across noise levels. Training minimizes a weighted mean-squared error between $D(x_\sigma;\sigma)$ and the clean field x_0, so the network predicts x_0 directly rather than the noise ϵ. Predicting x_0 lets us apply physics losses to a field with a meaningful spectrum at every training step (Section 4.5). For sampling we use DPM-Solver++ 2M (Lu et al., 2022), a deterministic high-order solver for the reverse process.

Cascades in 2D turbulence. Forced 2D turbulence has two conserved quantities in the inviscid limit, energy and enstrophy (mean-square vorticity). Kraichnan (1967) showed that this produces a dual cascade: energy moves toward large scales (the inverse energy cascade, at wavenumbers $k < k_f$ below the forcing scale k_f), while enstrophy moves toward small scales (the forward enstrophy cascade, at $k > k_f$). The *spectral flux* $\Pi(k)$ measures the net rate at which a conserved quantity crosses wavenumber k; its sign and peak location encode the cascade directions. A generative model can match the spectrum $E(k)$ at each k yet still misrepresent the fluxes, because fluxes depend on nonlinear phase relationships between modes, not on per-mode power alone. We therefore evaluate both per-mode statistics and the fluxes (Section 4.6).

3 Related Work

Generative models for fluid flows. Shu et al. (2023) apply a physics-informed diffusion model to flow-field reconstruction from sparse measurements. Lippe et al. (2023) introduce PDE-Refiner, which uses iterative denoising to stabilize next-step PDE solvers. Kohl et al. (2024) benchmark autoregressive conditional diffusion against deterministic baselines for turbulent flow. These works study conditional generation given a previous state; we study the unconditional setting.

Neural operators. Fourier neural operators (Li et al., 2021; Kovachki et al., 2023) apply learned spectral convolutions over a fixed set of modes and are a standard architectural baseline for grid-based PDE problems. We use them in two roles: as a single FNO block in the U-Net bottleneck, and as a standalone denoiser backbone. The proposed gate differs from an FNO block in three ways: it is multiplicative rather than transform-and-project, it is conditioned on the diffusion noise level and viscosity, and its parameters sit in the conditioning MLP rather than in per-mode learnable weights. This last point makes it more parameter-efficient than the FNO block. We also combine the two, since they act on different parts of the problem (Section 4.3).

Physics-informed losses. Soft physics losses appear in physics-informed neural networks (Raissi et al., 2019) and in hybrid solvers (Kochkov et al., 2021; Stachenfeld et al., 2022). Our losses act on enstrophy magnitude, modal spectral amplitude, vorticity structure functions, integral length, and spectral flux (Section 4.5).

Calibrated uncertainty. Our uncertainty estimate is the per-pixel ensemble percentile band, rescaled by a single scalar fit on a held-out split to match a target marginal coverage. This is a one-parameter conformal recalibration in the sense of Vovk et al. (2005) and Angelopoulos & Bates (2021); we make no finite-sample coverage guarantee and report calibration error on a disjoint test split.

4 Setup

4.1 Direct numerical simulation

We solve the 2D incompressible Navier–Stokes equations in vorticity form on a doubly periodic domain $[0, 2\pi]^2$ with Kolmogorov forcing,

$$\partial_t \omega + (\mathbf{u} \cdot \nabla)\omega = \nu \nabla^2 \omega - \alpha \omega + f, \qquad f = k_f^2 \cos(k_f y), \tag{2}$$

where $\omega = \nabla \times \mathbf{u}$ is the scalar vorticity, $\mathbf{u}$ is recovered from ω through the streamfunction, ν is the kinematic viscosity, $\alpha = 0.1$ is a linear (Ekman) drag that removes energy at large scales, and $k_f = 4$ is the forcing wavenumber. We integrate on a 128×128 grid with a pseudo-spectral RK4 scheme at timestep $\Delta t = 10^{-3}$. The convective product is evaluated with full 3/2-rule zero-padding (Orszag, 1971) to remove aliasing of high-wavenumber product modes into the resolved range.

We simulate seven Reynolds regimes at viscosities $\nu \in \{0.005, 0.007, 0.010, 0.013, 0.018, 0.024, 0.032\}$, chosen to span a factor of nine in Reynolds number and so test a viscosity-conditioned generator across regimes. After a 4000-step spinup we save 1000 decorrelated snapshots per regime, one every 200 steps (7000 snapshots in total). From these snapshots, the integral length L (from the radial spectrum) and the 2D Taylor microscale $\lambda = \sqrt{2\langle|\mathbf{u}|^2\rangle / \langle\omega^2\rangle}$ give integral Reynolds numbers $\mathrm{Re}_{\mathrm{int}} = u_{\mathrm{rms}} L/\nu$ from 593 down to 63, and Taylor-microscale Reynolds numbers $\mathrm{Re}_\lambda = u_{\mathrm{rms}} \lambda/\nu$ from 318 down to 33 (Figure 1).

Figure 1 shows the time-averaged vorticity power spectra $|\hat{\omega}(k)|^2$ (the enstrophy spectrum, k^2 times the kinetic-energy spectrum). The fitted slopes over the inertial window $k \in [6, 14]$ run from -2.64 at the least viscous regime to -5.62 at the most viscous, all with $R^2 \geq 0.98$, steepening with viscosity as the dissipation scale moves into the resolved range. These are steeper than the Kraichnan enstrophy-cascade prediction (Kraichnan, 1967; Boffetta & Ecke, 2012) (a k^{-1} vorticity spectrum, equivalently a k^{-3} kinetic-energy spectrum), reflecting the short inertial range and modest Reynolds number at $N = 128$ rather than an asymptotic cascade.

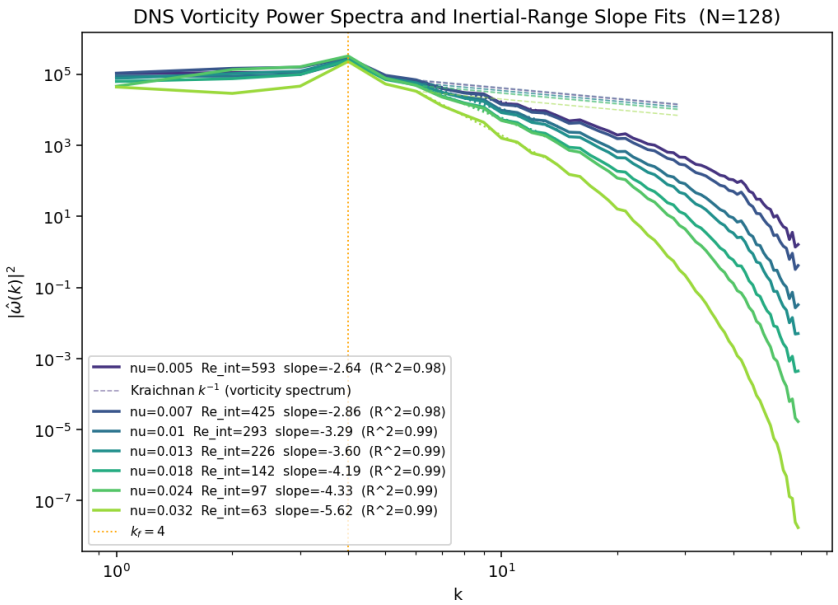

Figure 1: Time-averaged DNS vorticity power spectra $|\hat{\omega}(k)|^2$ at $N = 128$ (k^2 times the kinetic-energy spectrum), one curve per viscosity regime. The inertial-range fit window is $k \in [6, 14]$. The dashed line marks the Kraichnan enstrophy-cascade reference for the vorticity spectrum, $|\hat{\omega}(k)|^2 \sim k^{-1}$ (equivalently a k^{-3} kinetic-energy spectrum). Legend Reynolds numbers are integral-scale, from the time-averaged snapshots.

4.2 Data splits

We hold out 15% of all snapshots (1050 fields) as a fixed test set, with the split fixed by a seed-0 permutation before any training. The same test set evaluates every variant at every training seed. The remaining 5950 snapshots form the training pool, split per seed into inner-train (90%) and inner-val (10%); the inner-val is used only for training-time monitoring, never for reported metrics. Fixing the test split before training, and keeping it independent of the per-seed split, removes the train–test contamination that arises when a per-run training split also sets the per-run evaluation split. The physics metrics of Section 4.6 are computed per seed by comparing 256 unconditionally generated fields against 256 held-out test fields, with the generated set's regime composition matched to that of the test set, so that no training data enters the evaluation.

4.3 Architecture and ablations

The denoiser f_θ in equation 1 maps a noisy single-channel 128^2 vorticity field to a denoised one through a three-level U-Net (Ronneberger et al., 2015) with group normalization. A 3×3 convolution lifts the input to 48 channels; the encoder then applies a residual block at each level while halving the resolution and doubling the width, giving feature maps of $48{\times}128^2$, $96{\times}64^2$, and $192{\times}32^2$, and a symmetric skip-connected decoder inverts this back to a single output channel. (48 base channels was the smallest width at which the plain U-Net reached stable validation loss in pilot runs.) Every residual block is conditioned by a 128-dimensional embedding that fuses sinusoidal embeddings of the noise level $c_{\text{noise}}(\sigma)$ and of a continuous viscosity scalar, the z-scored $\log \nu$, so one model spans all seven regimes on a shared physical scale rather than through a discrete regime index. The bottleneck, acting on the 192-channel 32×32 feature map, is the only component that differs across the U-Net variants. We study a 2×2 factorial design (gate present or absent $\times$ losses present or absent) plus extra baselines and one extension:

- `vanilla`: U-Net, no bottleneck, no physics loss.
- `se`: squeeze-and-excite bottleneck (Hu et al., 2018), no physics loss. Tests whether any conditional bottleneck helps.

- `fno`: a single FNO block (Li et al., 2021) with 8 modes per axis as bottleneck, no physics loss. Tests whether a spectral architecture helps inside the U-Net.
- `fno_operator`: a standalone Fourier neural operator denoiser (12 modes, width 64, four spectral layers), trained with the same EDM objective and no physics loss. An external architecture baseline: an FNO used as the denoiser backbone rather than as a U-Net bottleneck.
- `wrsd_gate`: the gate, no physics loss. Isolates the gate.
- `wrsd_phys`: no bottleneck, with physics losses. Isolates the losses.
- `wrsd`: gate plus losses. The main model (wavenumber-resolved spectral diffusion, WRSD).
- `wrsd_fno`: gate plus an FNO branch plus losses (WRSD+FNO). The hybrid extension.

The factorial design holds capacity fixed across the comparisons that matter: `vanilla` and `wrsd_phys` share an architecture (2.32M parameters), and `wrsd` and `wrsd_gate` share an architecture (2.51M parameters). A comparison within a row of Table 1 therefore isolates the losses, and a comparison within a column isolates the gate.

4.4 The wavenumber-resolved gate

Let $h \in \mathbb{R}^{B\times C\times H\times W}$ be the U-Net bottleneck feature map, and let $\hat{h} = \mathcal{F}(h)$ denote its 2D orthonormal Fourier transform. The gate multiplies the Fourier coefficients by a learned per-band factor,

$$\hat{h}_{\text{out}}(k_x, k_y) = \hat{h}(k_x, k_y) \odot \big(1 + s \cdot \big[2\,\sigma(g_{c,b(k)}) - 1\big]\big)\,, \tag{3}$$

and the bottleneck adds the gated signal back to its input as a residual, returning $h + \mathcal{F}^{-1}(\hat{h}_{\text{out}})$ to the spatial domain. In the Fourier domain the bottleneck therefore applies the single per-band multiplier $2 + s \cdot [\,2\,\sigma(g_{c,b(k)}) - 1\,]$, centered at 2 rather than 1. Here σ is the logistic sigmoid, $b(k)$ bins the radial wavenumber $k = \sqrt{k_x^2 + k_y^2}$ into 16 uniformly spaced bins, $g_{c,b} \in \mathbb{R}^{C\times 16}$ is the per-channel, per-bin gate logit, and $s \in \mathbb{R}$ is a learnable scalar initialized to zero. A small MLP maps the conditioning embedding to a low-rank factorization of g, $g_{c,b} = \sum_{r=1}^{R} a_{r,c}\,\beta_{r,b} + d_c$ with rank $R = 2$ and a per-channel bias d_c, so the gate adds $R(C + 16) + C$ parameters rather than the $16C$ of a dense map.

The form of equation 3 is the minimal correction consistent with the symmetries of the statistics it targets, not a free-form spectral transform. The spectrum and the fluxes are isotropic functions of $|k|$, so the correction is indexed by the radial wavenumber rather than by (k_x, k_y); this also makes the parameter count independent of grid size. The correction is multiplicative because spectral bias is a multiplicative distortion of the spectrum, an under- or over-prediction of power within a band, rather than an additive offset. The map $2\sigma(\cdot) - 1 \in (-1, 1)$ centers the correction at zero, and the scale s is initialized to zero, so the gate's correction starts inactive and grows only as needed; the centering also lets the gate raise ($g > 0$) or lower ($g < 0$) the power in a band, redistributing the spurious high-k energy we document in Section 5. Because the gate is applied as a residual, the bottleneck is not the identity at initialization but a uniform doubling of its input ($s = 0$ gives the per-band multiplier 2); this fixed factor is immaterial to the factorial comparison, since the gate's learnable correction still starts at zero, the downstream decoder is trained end-to-end, and the no-gate variants carry no bottleneck at all. Because the bin index depends only on $|k|$, the gate map keeps conjugate symmetry and the inverse transform stays real. We use 16 bins because the gate operates on the 32×32 bottleneck feature map, whose radial spectrum spans about 22 shells (the diagonal Nyquist wavenumber of a 32^2 grid); 16 bins coarsen these enough to pool statistics within a band while still separating the inertial and dissipation ranges. Sensitivity to this choice is discussed in Section 6.

The hybrid bottleneck. The gate controls per-shell amplitude but does not mix Fourier modes, and inter-scale transfer (the fluxes) is a cross-mode effect. The WRSD+FNO bottleneck adds an FNO spectral convolution branch alongside the gate, returning $\text{gate}(h) + s_{\text{F}}\,\text{FNO}(h)$ with a learnable branch weight s_{F} initialized to one. The FNO branch supplies the cross-mode mixing the gate lacks; Section 5 shows it closes most of the remaining cascade-flux gap, at the cost of the FNO block's parameters.

4.5 Diffusion schedule and losses

We standardize the data to zero mean and unit variance with a single global mean and standard deviation, which fixes the EDM data scale at $\sigma_{\text{data}} = 1$ and calibrates the preconditioning coefficients in equation 1. We set $\sigma_{\max} = 20$ rather than the EDM default of 80 because at $\sigma = 80$ the first sampling step is dominated by the raw network output, which our dataset cannot train well; in pilot runs $\sigma_{\max} = 80$ produced near-zero samples. The training noise distribution $\log \sigma \sim \mathcal{N}(-1.0, 1.2^2)$ is recentered from the EDM default to match this reduced range. The remaining diffusion settings ($\sigma_{\min} = 0.002$, DPM-Solver++ 2M with 50 steps) follow the EDM defaults.

We train with AdamW (Loshchilov & Hutter, 2019) at learning rate 2×10^{-4}, weight decay 10^{-5}, and batch size 32 for 160 epochs ($N_{\text{steps}} = 26{,}880$), with cosine annealing after an eight-epoch warmup. We keep an EMA of the weights with decay $d = \min\bigl(1 - 1/(0.2\, N_{\text{steps}}),\, 0.999\bigr)$; at our run length the 0.999 cap binds, giving a window of about 10^3 steps ($\approx 4\%$ of training). The conventional fixed $d = 0.9999$ ($\approx 10^4$ steps) over-smoothed the weights and degraded samples in pilot runs.

For the physics-loss variants (`wrsd_phys`, `wrsd`, `wrsd_fno`) we add soft penalties on physical diagnostics of the predicted clean field x_0, recovered from x_σ through the EDM preconditioning,

$$\begin{aligned}\mathcal{L}_{\text{phys}} = \lambda_E \left\| \langle x_0^2 \rangle_{\text{pred}} - \langle x_0^2 \rangle_{\text{true}} \right\|_1 + \lambda_S \left\| |\hat{x}_0|_{\text{pred}} - |\hat{x}_0|_{\text{true}} \right\|_1 + \lambda_C\, \mathcal{D}_{\text{struct}} \\ + \mathbb{1}[\sigma < 1] \left(\lambda_L\, \mathcal{D}_{\text{len}} + \lambda_\Pi\, \mathcal{D}_{\text{flux}} \right),\end{aligned} \tag{4}$$

with weights $\lambda_E = \lambda_S = \lambda_C = 0.05$, $\lambda_L = 0.10$, $\lambda_\Pi = 0.07$. The first term penalizes enstrophy-magnitude error and the second modal spectral-amplitude error; $\mathcal{D}_{\text{struct}}$ is the log-space error of the order-2 and order-3 vorticity structure functions over separations $r = 1, \ldots, 8$, which targets the small-scale increments the radial gate tends to distort; $\mathcal{D}_{\text{len}}$ is the integral-length error, a large-scale constraint that counteracts the small-scale pull of the other terms; and $\mathcal{D}_{\text{flux}}$ is the peak-normalized energy- and enstrophy-flux error. The flux variant `wrsd_fno` and the main model `wrsd` use all five terms; `wrsd_phys` uses the first four (no flux term). The integral-length and flux terms are applied only at low noise ($\sigma < 1$), where the predicted clean field is informative enough about the large scales and the nonlinear transport for these higher-order diagnostics to give a useful gradient; at high noise the x_0 estimate is too smooth for them to be meaningful. We set the amplitude weights to 0.05 so that $\mathcal{L}_{\text{phys}}$ is roughly one tenth of the EDM denoising loss at initialization, large enough to shape the spectrum but small enough not to override the denoising objective.

4.6 Evaluation metrics

We report eight physics metrics. Four are point statistics:

- **Aggregate log-spectral distance (LSD)**, $\frac{1}{|K|} \sum_k |\log_{10}\langle E_{\text{pred}}(k)\rangle - \log_{10}\langle E_{\text{true}}(k)\rangle|$, where $E(k)$ is the dealiased *vorticity power spectrum* $|\hat{\omega}(k)|^2$ and $\langle\cdot\rangle$ averages over the field set. This is our single summary of spectral fidelity.
- **Vorticity structure functions** of order 2 and 3, as log-RMSE over a range of separations. We use the absolute-moment form $S_p^\omega(r) = \langle |\omega(\mathbf{x} + r\mathbf{e}) - \omega(\mathbf{x})|^p \rangle$ averaged over the two axis-aligned directions; the absolute form keeps S_3 positive and stable to estimate from small samples, unlike the signed velocity structure functions of Frisch (1995).
- **Integral length relative error**. The length $L = \pi \sum_k k^{-1} |\hat{\omega}(k)|^2 / \sum_k |\hat{\omega}(k)|^2$ is computed from the full (non-dealiased) vorticity power spectrum; it is therefore an enstrophy-weighted length scale, used here as a consistent relative-fidelity metric rather than the classical energy-based integral scale.

Four are spectral-transport metrics:

- **Energy flux RMSE**, normalized by $\max_k |\Pi_E^{\text{DNS}}(k)|$, with Π_E in the sign convention of Boffetta & Ecke (2012): $\Pi_E(k) < 0$ for $k < k_f$ marks the inverse energy cascade.

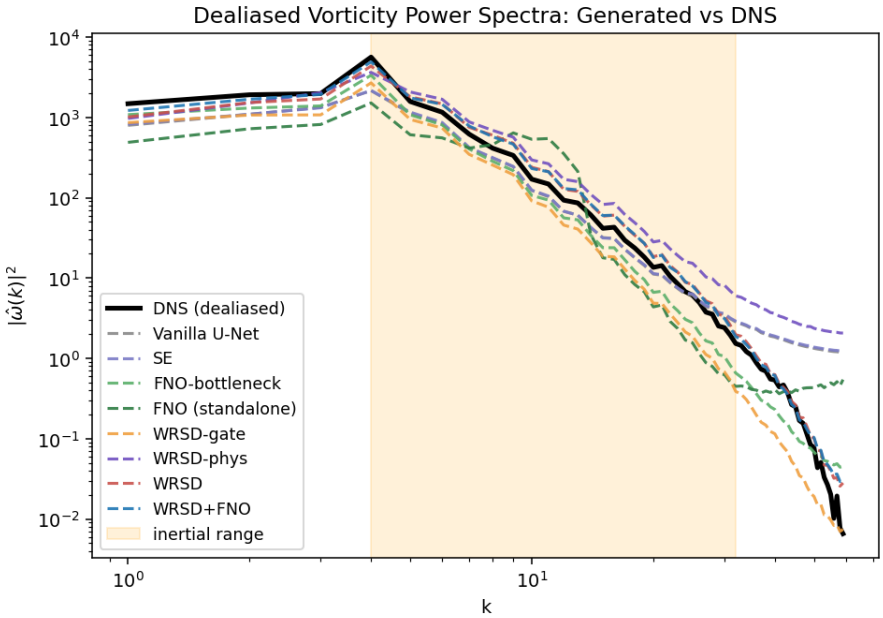

Figure 2: Dealiased vorticity power spectra $|\hat{\omega}(k)|^2$, generated vs DNS, averaged over the test set. The shaded band marks the inertial range, from the forcing wavenumber $k_f = 4$ to $k = 32$. WRSD and WRSD+FNO track the DNS curve across the resolved range. WRSD-phys (losses without the gate) overpredicts power at the smallest scales; the standalone FNO and the plain baselines sit above DNS in the dissipation tail; WRSD-gate (gate without losses) underpredicts the high-k tail.

- **Enstrophy flux RMSE**, defined the same way for Π_Z.
- **Inverse energy cascade recovery**, $100 \cdot \sum_{0<k<k_f} \max(-\Pi_E^{\text{pred}}(k), 0) / \sum_{0<k<k_f} \max(-\Pi_E^{\text{DNS}}(k), 0)$, clipped to $[0, 200]\%$, reporting the fraction of the inverse-cascade flux the model recovers.
- **Forward enstrophy cascade recovery**, defined the same way for Π_Z at $k > k_f$.

For each metric we run five seeds $(29, 47, 89, 101, 149)$, report the seed mean with 95% bootstrap confidence intervals over 2000 resamples (Efron, 1979), and run paired Wilcoxon signed-rank tests (Wilcoxon, 1945) against each baseline.

4.7 Uncertainty quantification

We draw a 32-member ensemble for each of 64 test conditions. For each (pixel, condition) the ensemble gives a median and a 5th–95th percentile band. We fit a single scalar α on a 50% calibration split to match 90% marginal coverage, then report empirical coverage and the continuous ranked probability score (Matheson & Winkler, 1976; Gneiting & Raftery, 2007) on the held-out 50%.

5 Results

5.1 Spectral fidelity

Figure 2 shows the time-averaged generated spectra against DNS. WRSD and WRSD+FNO track the DNS curve across the resolved range, including the forcing peak at $k = k_f = 4$, the inertial slope, and the dissipation tail. The other variants fail at different scales: the plain baselines and the standalone FNO place excess power in the dissipation tail, the losses without the gate (WRSD-phys) overpredict the smallest scales,

Table 1: Headline physics metrics, mean over five seeds. Lower is better for LSD, structure functions, length error, and flux RMSEs; closer to 100% is better for cascade recovery. Best per metric in bold, second-best underlined. Column headers: `van` (vanilla), `se`, `fno` (FNO block), `f-op` (standalone FNO), `gate` (WRSD-gate), `phys` (WRSD-phys), `WRSD`, `+FNO` (WRSD+FNO). 95% bootstrap confidence intervals are shown in Figure 3 and tabulated in Appendix A.

Metric	van	se	fno	f-op	gate	phys	WRSD	+FNO
LSD aggregate ↓	0.744	0.745	0.553	0.757	0.688	0.925	<u>0.304</u>	**0.293**
Vort. S_2 log-RMSE ↓	0.260	0.254	0.224	0.280	0.302	0.054	<u>0.044</u>	**0.033**
Vort. S_3 log-RMSE ↓	0.354	0.348	0.329	0.417	0.424	0.131	<u>0.071</u>	**0.061**
Integral length rel. err. ↓	0.039	0.037	<u>0.033</u>	0.228	**0.013**	0.150	0.112	0.073
Energy flux RMSE ↓	0.138	0.137	0.086	0.139	0.119	0.120	<u>0.083</u>	**0.038**
Enstrophy flux RMSE ↓	0.163	0.162	0.109	0.177	0.144	0.122	<u>0.081</u>	**0.046**
Inv. energy cascade rec. % ↑	10.0	10.3	45.9	8.8	21.4	24.8	<u>48.6</u>	**83.8**
Fwd. enstrophy cascade rec. % ↑	20.7	20.9	41.3	5.8	25.9	71.2	<u>89.0</u>	**110.0**

and the gate without the losses (WRSD-gate) underpredicts the high-k tail. Each failure has a distinct scale signature, which the ablation in Section 5.3 explains.

5.2 Headline metrics

Table 1 summarizes the eight metrics across all eight variants. WRSD+FNO is best on seven of the eight metrics and fifth on the remaining one (integral length); WRSD is second on those same seven. The two proposed models lead the four baselines on every metric except integral length, where the gate alone (`wrsd_gate`) is best and the physics losses move the proposed models behind the plain baselines, the trade-off we analyze in Sections 5.3 and 6.

WRSD (2.51M) beats the FNO block (7.08M) on the spectrum, both structure functions, and forward-enstrophy recovery, and ties it on energy flux; Section 5.4 gives the per-metric numbers and the parameter-efficiency comparison. The standalone FNO operator (`fno_operator`) is the weakest variant overall: an FNO used as the denoiser backbone, with no U-Net and no physics loss, does not match even the plain U-Net on the spectrum or the cascade. The spectral architecture helps as a bottleneck inside the U-Net, not as a replacement for it.

Paired tests. WRSD improves over `vanilla` on seven of eight metrics at paired Wilcoxon $p = 0.0625$, the floor for $n = 5$ (every seed agrees on the sign of the effect), with large effect sizes (Cohen's $|d|$ from 7.7 on energy flux to 41.5 on the S_2 structure function); it is worse than `vanilla` only on integral length. WRSD+FNO improves on WRSD on the flux and cascade metrics ($p = 0.0625$; energy flux $\Delta = -0.044$, $d = -7.2$) but is statistically tied on LSD ($\Delta = -0.010$, $p = 0.125$, $d = -1.1$): the FNO branch buys cascade transport, not aggregate spectral fidelity. Full per-metric differences with bootstrap CIs are in Appendix A.

5.3 Ablation: gate, losses, and their interaction

Reading each effect against `vanilla` (Figure 3):

- The gate alone (`wrsd_gate`) cuts energy flux RMSE by 14% and enstrophy flux RMSE by 11%, doubles inverse-energy cascade recovery (21.4% vs 10.0%), and nearly removes the integral-length error (1.3% vs 3.9%, the best of any variant). It worsens the structure functions (S_2 from 0.260 to 0.302). The gate acts as a spectral router: it corrects how power is distributed across scales but does not fit small-scale point statistics on its own.
- The losses alone (`wrsd_phys`) sharply lower the structure-function log-RMSE (S_2 0.054 vs 0.260) and raise forward-enstrophy recovery (71.2% vs 20.7%), but they *raise* LSD to 0.925 (the worst of any variant) and worsen the integral length. Pushed toward small-scale and transport fidelity with no gate to route the spectral correction, the losses distort the aggregate spectrum.

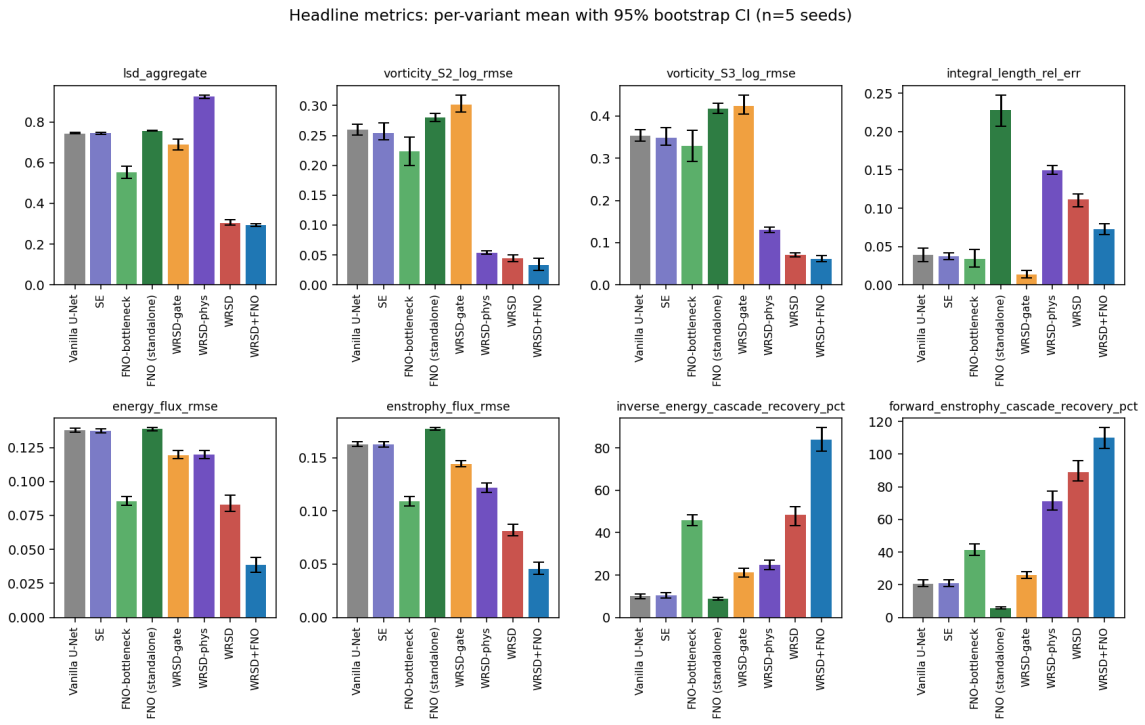

Figure 3: Per-variant mean with 95% bootstrap CI over five seeds, across the eight headline metrics. The gate (WRSD-gate) reduces integral-length error and cascade-flux error but is worst on structure functions. The losses (WRSD-phys) are strong on structure functions and forward-enstrophy recovery but raise LSD and integral-length error. The full model (WRSD) and the hybrid (WRSD+FNO) inherit the gains of both.

The interaction is the main finding of the ablation. The gate alone lowers LSD only slightly (gate 0.688 vs vanilla 0.744), and the losses alone raise it to 0.925; adding the two individual effects would predict no improvement. Yet together they reach 0.304, a 59% reduction. The gate gives the network a place to apply a clean spectral correction, and the losses tell it what correction to make; the losses are harmful without the gate to route them, and the gate does little without the losses to drive it. WRSD+FNO then adds the FNO branch and improves the flux and cascade metrics further (energy flux 0.038, inverse-energy recovery 84%), at the FNO block's parameter cost. Figure 4 shows the flux curves; the structure-function curves are in Appendix E.

5.4 Parameter efficiency

The FNO-block baseline uses 7.08M parameters; WRSD uses 2.51M, or 35% of it. At that size WRSD beats the FNO block on LSD (0.304 vs 0.553; $\Delta = -0.249$, $p = 0.0625$), on both structure functions, and on forward-enstrophy recovery, and matches it on energy flux RMSE (0.083 vs 0.086; $\Delta = -0.003$, $p = 0.625$, not significant). WRSD+FNO (7.27M) is the strongest model and uses about the same parameter count as the FNO block, so the gain over the FNO block on the flux and cascade metrics comes from the gate-plus-loss combination, not from added capacity. Figure 5 shows these fronts.

5.5 Per-regime breakdown

Table 2 shows the LSD improvement over `vanilla` holds across all seven regimes and grows in absolute terms with viscosity.[1] At the most viscous regime ($\nu = 0.032$, $\mathrm{Re}_{\mathrm{int}} = 63$) WRSD cuts LSD from 3.41 to 1.80 (a

[1]The per-regime LSD here pools snapshots within a single regime, whereas the headline LSD in Table 1 pools all seven regimes into one ensemble-mean spectrum before taking the log-spectral distance. Because the LSD is a nonlinear function of

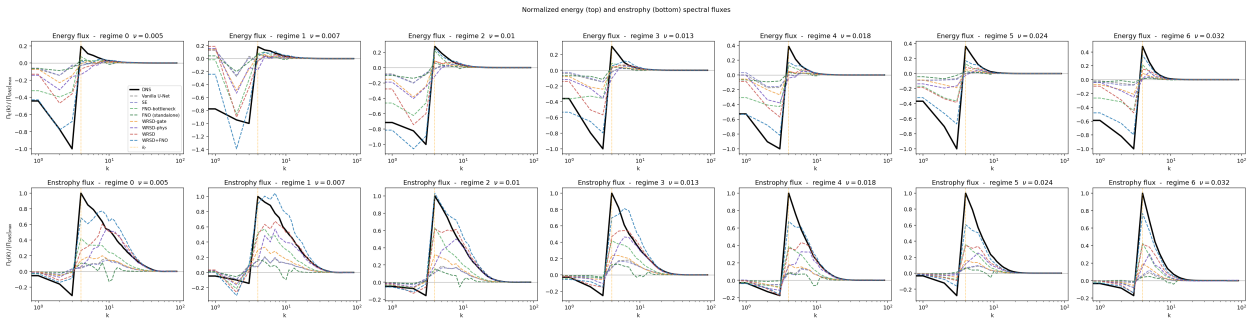

Figure 4: Normalized spectral fluxes by regime. Top row: energy flux $\Pi_E(k)$. Bottom row: enstrophy flux $\Pi_Z(k)$. Curves are normalized by the peak DNS flux magnitude in each panel. WRSD+FNO recovers the largest fraction of the DNS flux; no variant fully matches the peaks. The residual gap is a limitation of current diffusion-based generators (Section 6).

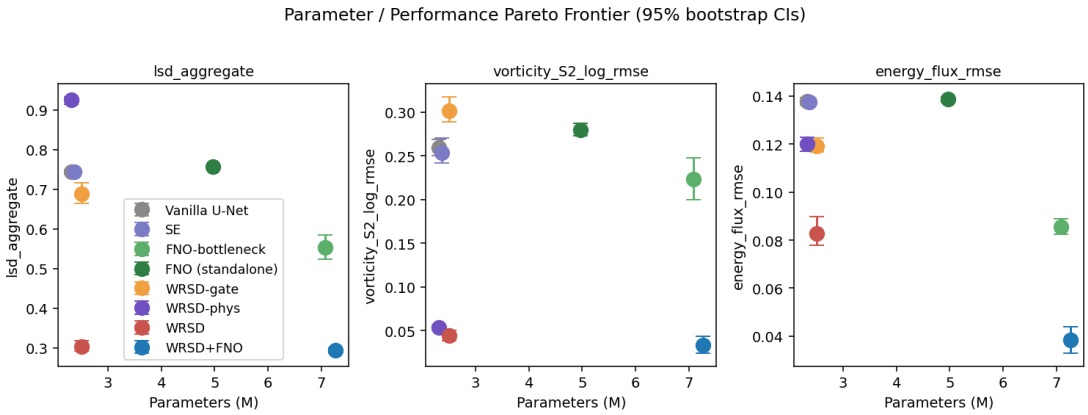

Figure 5: Parameter–performance Pareto fronts for three metrics, with 95% bootstrap CIs. WRSD (2.51M) is on the front for LSD and the structure functions, beating the 2.8×-larger FNO block. WRSD+FNO (7.27M) extends the front on energy flux. The standalone FNO operator is off the front on every metric.

47% reduction); at the least viscous ($\nu = 0.005$) it goes from 0.56 to 0.32 (43%). `vanilla` struggles most at high ν, where the dissipation tail steepens, and the gate compensates by moving capacity away from the low-amplitude high-k modes. WRSD and WRSD+FNO hold first and second place in every regime, with the largest absolute margin over `vanilla` at the most viscous regimes; the FNO block and the gate-only model also keep LSD below 2 throughout but never reach the top two.

5.6 Calibrated uncertainty

The raw per-pixel 5–95 band gives 0.70–0.84 empirical coverage (per-variant means), below the 90% target (Figure 6). The one-parameter conformal rescaling, with $\alpha \in [1.19, 1.66]$ fit on the calibration split, raises coverage to 0.854–0.891 on the disjoint test split: the WRSD models reach 0.887 and the FNO block 0.891 (closest to target), while the standalone FNO operator, with the most under-dispersed ensemble, reaches only 0.854. A single global scale narrows the gap but does not close it, because the per-pixel band is under-dispersed in a way one scalar cannot fully correct. CRPS clusters in $[0.485, 0.520]$, lowest for WRSD+FNO and WRSD (0.485, 0.487).

the ensemble-mean spectrum and the regimes differ by orders of magnitude in spectral power, the headline value is not the average of the per-regime values and can fall below all of them.

Table 2: Per-regime LSD aggregate. The improvement over `vanilla` holds across all seven regimes and grows in absolute terms with viscosity, where the steep dissipation tail makes spectral fitting harder for every method. WRSD and WRSD+FNO are first or second in every regime.

Variant	ν=.005	.007	.010	.013	.018	.024	.032
vanilla	0.555	0.701	0.994	1.343	1.890	2.287	3.408
se	0.557	0.704	0.995	1.339	1.893	2.297	3.415
fno	0.528	0.578	0.634	0.795	1.004	1.292	1.990
fno_operator	0.727	0.775	0.852	1.028	1.258	1.492	2.282
wrsd_gate	0.677	0.709	0.724	0.823	0.896	1.180	1.874
wrsd_phys	0.701	0.837	1.159	1.511	2.076	2.426	3.552
wrsd	0.317	0.313	0.386	**0.534**	**0.791**	**1.036**	**1.801**
wrsd_fno	**0.308**	**0.306**	**0.375**	0.549	0.803	1.045	1.823

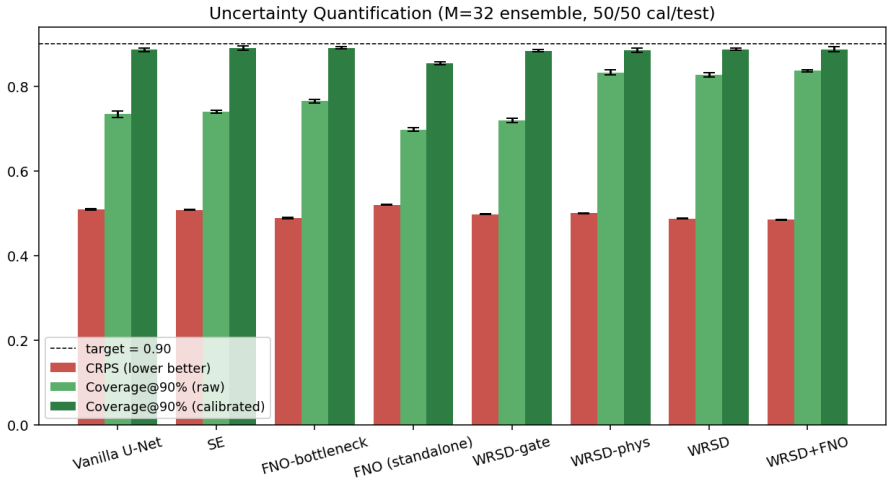

Figure 6: Uncertainty quantification. The raw 90% per-pixel band gives 0.70–0.84 coverage; the one-parameter conformal rescaling raises it to 0.85–0.89, short of the 0.90 target. CRPS is tight (0.485–0.520), best for WRSD+FNO.

Qualitative samples (Appendix C) and training curves (Appendix D) are provided in the appendix: all U-Net variants produce realistic vorticity fields, with the trade-offs already present in the metrics, and the denoising MSE is nearly identical across the U-Net variants, so the architectural differences in Table 1 are visible only through the distribution-level metrics of Section 4.6.

6 Limitations

The losses worsen the integral length. The physics losses sharpen the small scales and the cascade but degrade the large-scale integral length: WRSD-gate reaches an integral-length error of 0.013, while adding the losses raises it to 0.112 for WRSD and 0.073 for WRSD+FNO, both worse than the plain U-Net (0.039). The integral-length term in equation 4 reduces this regression but does not remove it. The structure-function and flux terms pull model capacity toward the small scales and the transport range, at the cost of the lowest-wavenumber modes that set the integral length. A practitioner who needs the large-scale length above all should use the gate without the losses.

Residual cascade gap. No variant fully matches the DNS flux peaks (Figure 4). WRSD+FNO recovers 84% of the inverse energy cascade flux but *overshoots* the forward enstrophy cascade to 110%; WRSD reaches 49% and 89%. Matching the fluxes, which depend on nonlinear transport, with a model trained only on snapshots and no hard conservation constraint likely needs an explicit conservation loss or a hybrid

generator–solver. We tempered the forward overshoot with a smaller FNO-branch weight in pilot runs, but that gave back the inverse-cascade recovery the branch provides, so we report the neutral setting and the overshoot as a trade-off.

Calibration does not reach the target. The conformal rescaling raises 90% coverage to 0.85–0.89, not to 0.90. The per-pixel ensemble is under-dispersed, and a single global scale cannot fully correct it; a per-wavenumber or per-pixel calibration would likely close the gap, at the cost of the one-parameter simplicity.

Bin count and grid. The gate uses 16 radial bins on the 32^2 bottleneck. Coarser binning raised LSD in pilot runs and finer binning gave no measurable gain while adding parameters, so 16 is near the useful range for this 128^2 setup; a different grid resolution would need re-tuning.

Sample size and scope. Five seeds give a Wilcoxon floor of $p = 0.0625$, so significance bottoms out before the effect sizes are exhausted; the large Cohen's d values indicate the floor is reached because every seed agrees on direction, not because the per-seed effect is marginal. All runs are at 128×128 in 2D, where the DNS is exact and the spectral diagnostics are well defined. The radial binning of $|k|$ is dimension-agnostic, so the gate extends to 3D in principle, but we have not tested this.

7 Discussion

The factorial decomposition (Section 5.3) refines the usual "physics losses help" message of the PINN literature (Raissi et al., 2019): here neither the additive losses nor the multiplicative gate improves the aggregate spectrum on its own, and the gain is their interaction. The FNO branch adds a third, separable effect: cross-mode mixing that the per-shell gate cannot provide, closing most of the cascade-flux gap.

We compare against the four unconditional baselines of Section 4.3 at comparable or larger capacity, the like-for-like comparison for the unconditional task. We do not benchmark against conditional next-step solvers such as PDE-Refiner (Lippe et al., 2023) or autoregressive conditional diffusion (Kohl et al., 2024): they solve a different problem, predicting the next state from the current one, and their metrics are not comparable to ours, since a conditional model can reach near-zero pointwise error while an unconditional model can only match the distribution. Porting the gate to a conditional setting is left for future work.

8 Conclusion

We presented a wavenumber-resolved spectral gate for diffusion-based unconditional generation of 2D turbulence, paired with five soft physics losses, and a hybrid extension that adds an FNO branch. Across seven Reynolds regimes with a single viscosity-conditioned model, the gated model improves over a plain U-Net on the spectrum and the cascade and is parameter-efficient against an FNO block of nearly three times the size; the hybrid is best overall. A factorial ablation shows the gate and the losses correct distinct errors and interact, and the conformal recalibration narrows the coverage gap. The residual cascade gap and the integral-length trade-off (Section 6) remain the main open problems. The gate is a self-contained bottleneck that can be added to an existing U-Net, and the suite of physics metrics gives a reference point for future work on generative turbulence modeling.

Broader Impact Statement

This work studies generative modeling of an idealized fluid system (forced 2D Kolmogorov flow) at modest resolution. We see no direct dual-use concern. A plausible application is ensemble generation for uncertainty quantification in geophysical or atmospheric modeling, where calibrated uncertainty is important. Any deployment of a generative surrogate for physical simulation requires validation against the target real-world distribution, which differs from the idealized training distribution used here.

Reproducibility Statement

Code, training configurations, and seed values are available at `https://anonymous.4open.science/r/wrsd-turbulence`. Sampling at evaluation time is seeded, so the reported tables reproduce exactly from the released checkpoints. The full sweep (eight variants, five seeds, seven regimes, 160 epochs at 128^2) takes about 100 GPU-hours. Per-variant parameter counts and training times are in Appendix B. The DNS solver, data splits (Section 4), architecture and hyperparameters (Sections 4.3–4.5), and evaluation protocol (Section 4.6) are specified in full.

References

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

George K. Batchelor. Computation of the energy spectrum in homogeneous two-dimensional turbulence. *The Physics of Fluids*, 12(12):II–233–II–239, 1969.

Guido Boffetta and Robert E. Ecke. Two-dimensional turbulence. *Annual Review of Fluid Mechanics*, 44: 427–451, 2012.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2nd edition, 1988.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

Uriel Frisch. *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge University Press, 1995.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, 2018.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577, 2022.

Dmitrii Kochkov, Jamie A. Smith, Ayya Alieva, Qing Wang, Michael P. Brenner, and Stephan Hoyer. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21):e2101784118, 2021.

Georg Kohl, Li-Wei Chen, and Nils Thuerey. Benchmarking autoregressive conditional diffusion models for turbulent flow simulation. *arXiv preprint arXiv:2309.01745*, 2024.

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

Robert H. Kraichnan. Inertial ranges in two-dimensional turbulence. *The Physics of Fluids*, 10(7):1417–1423, 1967.

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

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.

James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976.

Steven A. Orszag. On the elimination of aliasing in finite-difference schemes by filtering high-wavenumber components. *Journal of the Atmospheric Sciences*, 28(6):1074–1074, 1971.

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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241. Springer, 2015.

François Rozet and Gilles Louppe. Score-based data assimilation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Dule Shu, Zijie Li, and Amir Barati Farimani. A physics-informed diffusion model for high-fidelity flow field reconstruction. *Journal of Computational Physics*, 478:111972, 2023.

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*, 2021.

Kimberly Stachenfeld, Drummond B. Fielding, Dmitrii Kochkov, Miles Cranmer, Tobias Pfaff, Jonathan Godwin, Can Cui, Shirley Ho, Peter Battaglia, and Alvaro Sanchez-Gonzalez. Learned coarse models for efficient turbulence simulation. In *International Conference on Learning Representations*, 2022.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

A Detailed Statistical Tests

For each headline metric we run a paired test of a proposed model against each baseline at matched seeds, reporting the mean difference with bootstrap CI, the Wilcoxon signed-rank p-value, and Cohen's d (Cohen, 1988) on the seed-wise differences. With $n = 5$ seeds the Wilcoxon floor is $p = 0.0625$, reached whenever all five seeds agree on the sign. All differences, intervals, and effect sizes are computed from the unrounded per-seed metrics, so a difference recomputed from the rounded means in Table 1 may differ in the last digit. We tabulate the main model (WRSD) against each baseline on three representative metrics below; the per-seed values for all eight metrics and both proposed models are released with the code.

Table 3: Paired tests of WRSD on the LSD aggregate metric. Negative differences favor WRSD.

Baseline	Δ mean	95% CI	Wilcoxon p	Cohen's d
vanilla	-0.441	$[-0.455, -0.426]$	0.0625	-21.9
fno	-0.249	$[-0.291, -0.210]$	0.0625	-4.7
fno_operator	-0.453	$[-0.465, -0.439]$	0.0625	-26.6
wrsd_gate	-0.385	$[-0.423, -0.348]$	0.0625	-7.9
wrsd_phys	-0.621	$[-0.634, -0.610]$	0.0625	-40.9

Table 4: Paired tests of WRSD on energy flux RMSE. Negative differences favor WRSD; WRSD ties the FNO block.

Baseline	Δ mean	95% CI	Wilcoxon p	Cohen's d
vanilla	-0.055	$[-0.060, -0.049]$	0.0625	-7.7
fno	-0.003	$[-0.008, +0.003]$	0.6250	-0.4
fno_operator	-0.056	$[-0.061, -0.050]$	0.0625	-7.7
wrsd_gate	-0.036	$[-0.040, -0.032]$	0.0625	-7.4
wrsd_phys	-0.037	$[-0.042, -0.032]$	0.0625	-5.6

Table 5: Paired tests of WRSD on forward enstrophy cascade recovery (positive favors WRSD).

Baseline	Δ mean	95% CI	Wilcoxon p	Cohen's d
vanilla	$+68.3$	$[+64.2, +73.9]$	0.0625	$+10.9$
fno	$+47.6$	$[+44.7, +51.9]$	0.0625	$+9.8$
fno_operator	$+83.2$	$[+78.3, +89.8]$	0.0625	$+10.7$
wrsd_gate	$+63.0$	$[+59.3, +68.6]$	0.0625	$+10.3$
wrsd_phys	$+17.7$	$[+15.1, +20.3]$	0.0625	$+5.3$

The hybrid extension improves on WRSD on the transport metrics (energy flux $\Delta = -0.044$, $[-0.049, -0.040]$, $p = 0.0625$, $d = -7.2$; forward enstrophy recovery $\Delta = +21.0$, $p = 0.0625$, $d = +4.6$) but is tied with WRSD on LSD ($\Delta = -0.010$, $p = 0.125$, $d = -1.1$).

B Compute Profile

Table 6: Per-variant parameter count and wall-clock training time. Training time and peak memory are means over five seeds, on one V100 GPU, for 160 epochs at 128^2. The physics-loss variants are slower because the structure-function and flux losses add per-step FFT work; `fno` and `wrsd_fno` carry the FNO block's parameters.

Variant	Parameters (M)	Mean train time (s)	Mean peak memory (MB)
vanilla	2.321	6647	2960
se	2.371	6890	2985
fno	7.077	8161	3178
fno_operator	4.970	8387	2132
wrsd_gate	2.511	8738	3069
wrsd_phys	2.321	8789	2960
wrsd	2.511	11329	3087
wrsd_fno	7.266	12628	3243

C Qualitative Samples

Figure 7 shows paired samples from the same initial noise across the eight variants, with a DNS reference. The U-Net variants produce realistic vorticity fields with coherent cores and filaments; the standalone FNO operator is visibly grainier and less coherent, in line with its weaker physics metrics. The zoom shows the trade-offs already in the metrics: the plain baselines are slightly smoother, the physics-loss variants produce sharper small-scale detail, and WRSD and WRSD+FNO look alike despite the parameter gap.

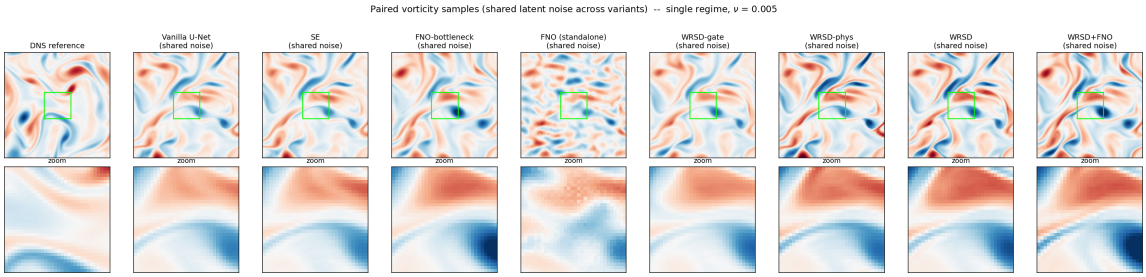

Figure 7: Paired vorticity samples from the same initial noise across variants, at $\nu = 0.005$. Top row: full 128×128 fields. Bottom row: the centered 32×32 patch. The DNS reference is one test draw, for orientation; the model panels are paired by latent noise.

D Training Dynamics

Figure 8 shows training and validation MSE. The seven U-Net variants converge to nearly the same MSE on the diffusion objective, so standard MSE validation does not separate architectures that produce visibly different samples. The standalone FNO operator sits above this cluster in both training loss and validation MSE, consistent with its being the weakest model on the physics metrics. Within the U-Net family the denoising MSE is flat while the spectral and cascade metrics in Table 1 vary widely, so we evaluate with the distribution-level metrics of Section 4.6.

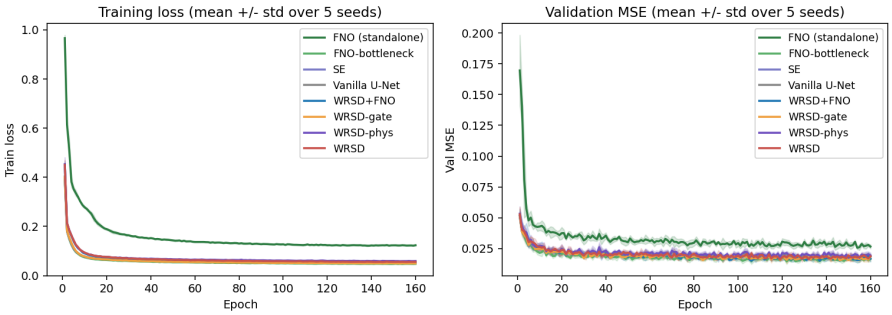

Figure 8: Training loss (left) and inner-validation MSE (right), mean $\pm$ standard deviation over five seeds. The seven U-Net variants converge to nearly the same denoising MSE despite their different sample statistics; the standalone FNO operator sits above the cluster.

E Structure Functions

Figure 9 shows the order-2 and order-3 vorticity structure functions against DNS. The physics-loss variants (WRSD-phys, WRSD, WRSD+FNO) match DNS most closely across separations; the gate without the losses (WRSD-gate) sits furthest off, consistent with the log-RMSE values in Table 1.

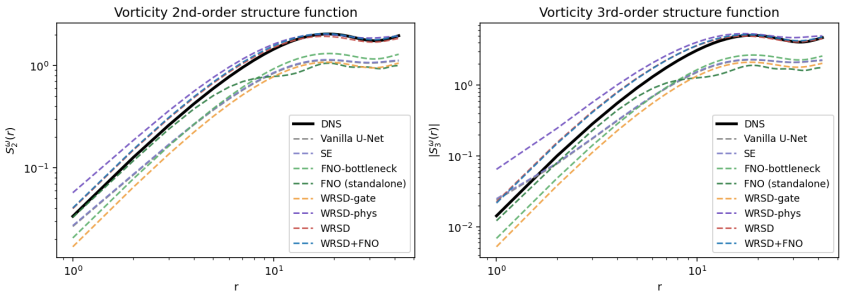

Figure 9: Vorticity structure functions, averaged over the test set. The physics-loss variants (WRSD-phys, WRSD, WRSD+FNO) match DNS most closely; the gate without the losses (WRSD-gate) sits furthest off.