# MoSE: Mixture of Spectral Experts in Diffusion Models

**Anonymous authors**
Paper under double-blind review

## Abstract

Diffusion models are usually trained with pixel-space denoising objectives, which constrain frequency content only indirectly. As a result, generated samples may appear realistic visually, while failing to match global spectral structure or localized details at multiple scales, leading to over-smoothed and detail-poor outputs. We introduce *Mixture-of-Spectral-Experts* (MoSE), a training-time regularizer that combines four spectral experts with diffusion models. Fourier and wavelet energy experts match dataset-level spectral statistics, while per-sample Fourier and wavelet experts preserve sample-specific structure. A timestep-conditioned hierarchical router adjusts the expert weights over the denoising process as the noise level changes. Across four image benchmarks, MoSE improves generation quality and achieves strong FID results with the strongest gains on CIFAR-10, ImageNet and CelebA-HQ. For physical-system tasks, it improves both Fourier-space and solution-error metrics on RealPDEBench dataset, demonstrating its generalizability to different datasets and tasks. These results show that introducing and balancing global and local spectral objectives improve the fidelity of generated samples in diffusion models for both image and scientific data generation.

## 1 Introduction

Diffusion models (DMs) have emerged as one of the most powerful classes of generative models, achieving remarkable success across diverse data modalities including image- (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022), video- (Ho et al., 2022; Blattmann et al., 2023; Brooks et al., 2024) and audio-generation (Kong et al., 2021; Lam et al., 2022; Liu et al., 2023). Despite remarkable advances in perceptual quality, generated samples do not necessarily reproduce statistical properties of the real data. A common and persistent discrepancy arises in the spectral domain. Recent AI-generated image detection methods demonstrate that such spectral discrepancies persist across modern generative models, and can provide sufficiently consistent evidence to distinguish synthetic images from real ones (Karageorgiou et al., 2025; Chu et al., 2025), reflecting a broader challenge in generative models: generating visually realistic samples does not imply faithful recovery of the underlying spectral structure.

Diffusion models learn to reverse a gradual Gaussian corruption process (Ho et al., 2020; Nichol & Dhariwal, 2021). The spectral discrepancies can be traced to the denoising objectives used for training. Standard diffusion models are typically trained using denoising objectives in pixel space. While these objectives are well motivated from both empirical and theoretical perspectives, they do not explicitly account for the structured spectral properties of various datasets. Specifically, natural images and scientific data often exhibit data-specific frequency statistics as well as dependencies across spatial locations and scales, which are characterized differently by Fourier and wavelet representations (Floryan & Graham, 2021). Therefore, there is a growing interest in developing new training strategies that

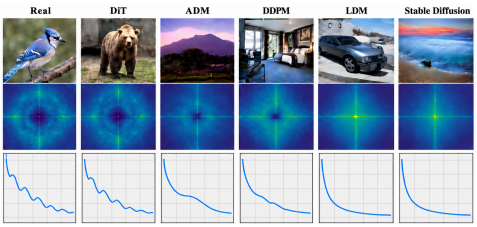



Figure 1: Real (Column 1) and generated images (Columns 2-5) exhibit distinct spectral patterns despite visual realism. First row: images in pixel space. Second row: corresponding log-magnitude Fourier spectra (low frequencies at the center; high frequencies at the edges). Third row: corresponding radial averages of the magnitude Fourier spectra.



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.

explicitly incorporate spectral information. Falck et al. (2025) shows that the standard forward process adds spectrally uniform Gaussian noise into data with nonuniform spectral energy, high-frequency components can reach low signal-to-noise ratios earlier than low-frequency components, inducing a frequency-dependent corruption and reconstruction process. Theoretical analyses of diffusion training have identified a spectral bias under which high-variance modes associated with coarse structures can be learned more rapidly than lower-variance modes corresponding to finer details (Wang & Pehlevan, 2025). Complementarily, analyses of diffusion inference show that different spectral components evolve differently throughout the denoising trajectory and that their behavior is closely related to the noise schedule (Benita et al., 2025). Recent work on diffusion inversion further demonstrates a concrete consequence of spectral deficiency: when the predicted noise lacks sufficient high-frequency variance, the reverse process can produce over-smoothed and texture-poor outputs, while naively restoring this variance may compromise structural fidelity (Bourriez et al., 2026). Consequently, accurate optimization of the conventional denoising objective in pixel space does not necessarily guarantee faithful recovery of the underlying spectral structure.

Recent analyses provide direct evidence that such spectral discrepancies can arise from the diffusion process itself. In particular, Bourriez et al. (2026) demonstrates that deterministic diffusion inversion can hallucinate poor high-frequency details when reconstructing from a spectrally sparse source to a richer target domain, such as in super-resolution and sketch-to-image generation. The recovered images can preserve low-frequency coarse structure while lacking the high-frequency details, which subsequently degrades the performance of diffusion models to generate over-smoothed samples with unrealistic high-frequency information. Importantly, restoring high-frequency information only can degrade structural fidelity, revealing an intrinsic trade-off between matching realistic spectral statistics and preserving sample-specific structure (Bourriez et al., 2026). These observations indicate that spectral fidelity is closely coupled to the representation and dynamics of the diffusion process.

Addressing this problem, however, requires more than introducing a single frequency-domain penalty. On one hand, different spectral representations characterize complementary properties of the data. Fourier representations provide a global description of frequency information, whereas wavelets provide spatial localization and multiscale variations, making them particularly suitable for representing localized textures and fine-detailed structures. On the other hand, spectral fidelity can be considered as a trade-off between matching dataset-level statistics and preserving sample-specific spectral structures. Sample-wise matching encourages the preservation of the spectral structure of its corresponding clean target, whereas dataset-level matching encourages generated samples to reproduce the spectral statistics of the overall data distribution. In addition, the weight of these constraints should not remain constant throughout the denoising process, since the reliability and evolution of different spectral components vary with the diffusion noise level (Falck et al., 2025; Benita et al., 2025). Existing approaches based on spectral regularization, frequency-aware diffusion, or wavelet-domain diffusion (Chandran et al., 2026; Falck et al., 2025; Hu et al., 2025) only deal with a subset of these complementary aspects. This motivates a unified framework that jointly models global and localized spectral structure, dataset-level and sample-wise consistency, and their time-dependent contributions throughout the diffusion process.

To address these limitations, we propose *Mixture-of-Spectral-Experts* (MoSE), a training-time regularizer that augments the standard denoising objective with a mixture of four spectral experts. MoSE leverages the complementary strengths of Fourier and wavelet representations: the Fourier experts model the global, coarse-scale structure of the data, while the wavelet experts capture localized variations and fine-scale details. To exploit these complementary properties at multiple levels, our framework evaluates four spectral experts on the predicted clean sample. Specifically, the Fourier-energy and wavelet-energy experts align dataset-level spectral statistics, thereby reducing distributional drift and discouraging overfitting to individual samples. In contrast, the sample-wise Fourier and sample-wise wavelet experts enforce instance-level spectral consistency, preserving sample-specific structure and mitigating over-smoothing and mode collapse. A timestep-conditioned hierarchical router allocates weights across the experts, allowing the model to adaptively balance the contributions of each expert based on the noise level and prevent any single expert from dominating the routing decision. Experiments on low- and high-resolution image generation benchmarks demonstrate that MoSE improves generation quality for different scales and datasets. MoSE also enhances physical data prediction tasks, achieving strong results on various simulated and real-world physical systems, which shows the generality of our approach.

Our contributions can be summarized as follows:

1. We introduce four spectral experts that combine global Fourier structure with localized wavelet structure and match both dataset-level statistics and per-sample consistency.
2. We develop a timestep-conditioned hierarchical router that adapts the balance among the four experts over the denoising process, allowing the model to adjust its focus based on the noise level and preventing any single expert from dominating the routing decision.
3. We evaluate MoSE on four image benchmarks and five real-world physical tasks from Hu et al. (2026). MoSE improves image generation quality and gives strong results on spectral and physical error metrics across these settings.

## 2 Related Work

**Frequency-aware diffusion models.** Recent work studies diffusion models directly in the frequency domain. Falck et al. (2025) shows that the standard forward process corrupts high-frequency components earlier than low-frequency components and considers an alternative process that corrupts all frequencies at the same rate. Ziashahabi et al. (2024) instead uses a frequency-dependent schedule to control image scales and improve efficiency. For functional data, Phillips et al. (2022) represents stochastic processes through a truncated set of spectral coefficients. These methods change the diffusion process, which limits their generality, or operate mainly in a single spectral representation. An alternative is to add spectral losses while keeping the original diffusion process. Chandran et al. (2026) augments the denoising objective with a choice of Fourier- or wavelet-domain loss and improves image and audio generation. Its spectral-loss weight is set in advance, and it does not jointly route the two transforms or separate dataset-level spectral statistics from per-sample matching, which leads to limited improvement compared to the original method. In contrast, MoSE makes this distinction explicit and uses a timestep-conditioned router to balance four Fourier and wavelet experts during training.

**Diffusion models for physical systems.** Diffusion models have also been applied to scientific data whose spectra contain important physical information. Whittaker et al. (2024) generates two-dimensional turbulent vorticity fields and evaluates their energy spectra and scaling statistics. Hu et al. (2025) performs diffusion over full PDE trajectories in the wavelet domain and uses multi-resolution training to handle abrupt changes and resolution shifts. These methods are designed for specific simulation or forecasting settings. MoSE instead provides a general training-time regularizer that applies to both image generation and physical fields without requiring an equation-specific residual loss.

## 3 Preliminaries

### 3.1 Denoising Diffusion Probabilistic Models

DDPMs consist of two main components: a forward (diffusing) process and a backward (denoising) process. The forward process (Ho et al., 2020) is defined by a Markov chain parameterized with Gaussian transitions that gradually add Gaussian noise $\epsilon_{\mathbf{t}}$ to progressively corrupt any image data $\mathbf{x}_0 \sim p_{data}(\mathbf{x})$, creating increasingly noisy data points. For any $t \in [0, T]$, the latent variable $\mathbf{x}_t$ is defined as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \tag{1}$$

where $\epsilon_t \sim \mathcal{N}(\mathbf{0},\ \mathbf{I})$, $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$, and $\alpha_s$ is the scheduler. The denoising framework is then trained to invert this process by applying a reverse Markov chain starting from $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0},\ \mathbf{I})$ with a noise predictor $\epsilon_\theta(\mathbf{x}_t, t)$:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left[\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right] + \sqrt{\beta_t}\mathbf{z}, \tag{2}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\ \mathbf{I})$ and $\beta_t := 1 - \alpha_t$.

### 3.2 Spectral Transforms

#### 3.2.1 Fourier Transform

The spectral decomposition of the input data $\mathbf{x}_t(\mathbf{s}) \in C(\mathbb{R}^2)$ can be defined by a linear combination of spatial basis functions and spectral components (Wikle & Cressie, 1999; Sigrist et al., 2015):

$$\mathbf{x}_t(\mathbf{s}) \approx \sum_{j=1}^{K} \hat{x}_{t,j} f_j(\mathbf{s}) \equiv \boldsymbol{f}^T(\mathbf{s})\hat{\mathbf{x}}_t, \tag{3}$$

where $f_j(\mathbf{s}) = \exp(i\mathbf{k}_j^T\mathbf{s})$ is the corresponding spatial basis function, with $\mathbf{s} \in \mathbb{R}^2$ representing the spatial variable and $\mathbf{k}_j$ representing the spatial wavenumbers (frequency vectors). Collectively, these form the Fourier basis, denoted by $\boldsymbol{f}(\mathbf{s}) = (f_1(\mathbf{s}), \cdots, f_K(\mathbf{s}))^T$, which spans the entire pixel space $C(\mathbb{R}^2)$. Then the Fourier transform $\mathscr{F}$ can be defined as a linear operator that maps the input data $\mathbf{x}_t(\mathbf{s})$ to its corresponding spectral components $\hat{\mathbf{x}}_t = (\hat{x}_{t,1}, \cdots, \hat{x}_{t,K})^T$:

$$\begin{aligned} \mathscr{F} : C(\mathbb{R}^2) &\to \mathbb{C}^K \\ \mathbf{x}_t(\mathbf{s}) &\mapsto \hat{\mathbf{x}}_t, \end{aligned} \tag{4}$$

#### 3.2.2 Wavelet Transform

The discrete wavelet transform (DWT) provides a multi-scale representation of an image by jointly localizing its content in spatial and frequency domains. Unlike the Fourier transform, whose basis functions are globally supported, wavelet basis functions are spatially localized and therefore capture local structures at different resolutions. This property makes wavelet representations particularly suitable for characterizing edges, textures, and other spatially heterogeneous structures.

Given an image at diffusion timestep $t$, $x_t \in \mathbb{R}^{C\times H\times W}$, a $J$-level two-dimensional wavelet transform can be expressed as

$$\mathcal{W}(x_t) = \left\{ A_J, \left\{ D_j^{LH}, D_j^{HL}, D_j^{HH} \right\}_{j=1}^{J} \right\}, \tag{5}$$

where $A_J$ denotes the low-frequency approximation coefficients at the coarsest decomposition level, and $D_j^{LH}$, $D_j^{HL}$, and $D_j^{HH}$ denote the detail coefficients at scale $j$ along the three spatial orientations.

### 3.3 Mixture-of-Experts Routing

Mixture-of-Experts (MoE) (Shazeer et al., 2017) models combine multiple specialized experts through a routing mechanism that determines the contribution of each expert for a given input. Let $\{\mathcal{E}_m\}_{m=1}^M$ denote a set of $M$ experts, where $\mathcal{E}_m(\mathbf{x})$ represents the output of the $m$-th expert. A router $\mathcal{R}$ maps input $\mathbf{x}$ and a conditioning variable $\mathbf{z}$ to a set of routing scores,

$$\mathbf{g} = \mathcal{R}(\mathbf{x}, \mathbf{z}) \in \mathbb{R}^M, \tag{6}$$

where $\mathbf{z}$ may contain information about the input or the current state of the diffusion process. The resulting MoE output is then expressed as

$$\mathcal{E}_{\mathrm{MoE}}(\mathbf{x}; \mathbf{z}) = \sum_{m=1}^{M} g_m(\mathbf{x}, \mathbf{z})\, \mathcal{E}_m(\mathbf{x}). \tag{7}$$

### 3.4 Distribution-Aware Discrepancy

The Mahalanobis distance (Lee et al., 2018) measures the discrepancy between a sample and a reference distribution while incorporating the covariance structure of the feature space. Given a feature vector $\mathbf{x} \in \mathbb{R}^d$, mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$, and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, the squared Mahalanobis distance can be defined as

$$\mathcal{M}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2d}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}), \tag{8}$$

where $d$ denotes the dimensionality of the feature vector.

## 4 Methodology

Our method (shown in Fig. 2) augments the standard diffusion objective with a mixture of spectral experts that constrain the Fourier and wavelet representations of the predicted clean sample $\tilde{\mathbf{x}}_0$. The key distinction is between *dataset-level* spectral statistics and *sample-specific* spectral structure. We introduce four experts: Fourier and wavelet energy experts (FE and WE), which encourage $\tilde{\mathbf{x}}_0$ to follow the spectral statistics of the real data, and Fourier and wavelet per-sample experts (FP and WP), which directly match individual spectral structure of $\tilde{\mathbf{x}}_0$ to its corresponding clean target. The Fourier experts focus on global and coarse structure, whereas the wavelet experts deal with local and fine-scale details. MoSE constrains both dataset-level and sample-wise spectral information that are not explicitly learned or controlled by the conventional denoising objective in pixel space. A timestep-conditioned router then adaptively combines these four experts.

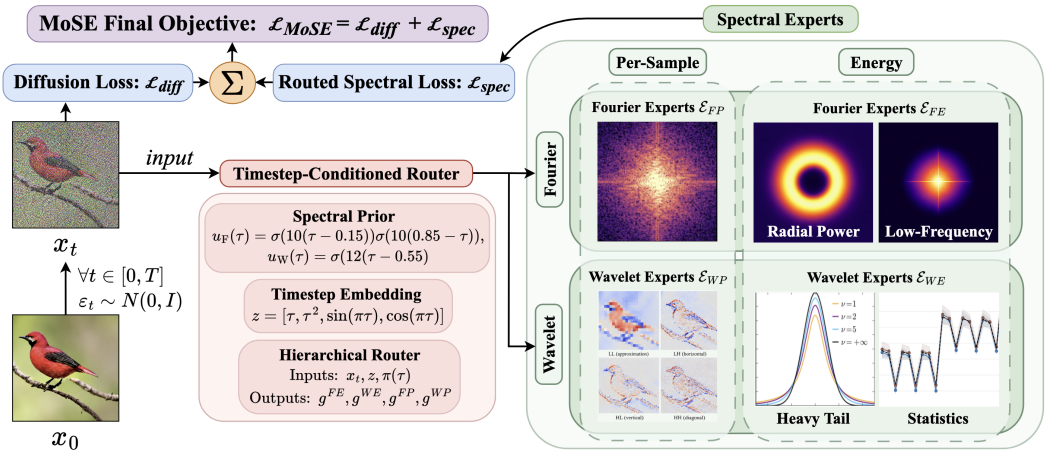



Figure 2: Framework of the proposed method.

### 4.1 Energy Experts

For a $d$-dimensional data or feature vector $\mathbf{x}$, recall that its mean and covariance are denoted by $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. Assume that $\mathbf{x}$ follows a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. From Lee et al. (2018) we know that the Mahalanobis distance can be used to measure the distance between $\mathbf{x}$ and the corresponding Gaussian distribution.

**Fourier energy.** We use two Fourier terms to capture different aspects of the Fourier energy. First, the average radial-angular power spectrum summarizes how energy is distributed from low to high frequencies. We partition the Fourier coefficients jointly by radius and angle to capture both the frequency and direction information. Let $B_{r,\theta}$ denote the $r$-th radial and $\theta$-th angular frequency bin. We define

$$P_{r,\theta}(\mathbf{x}) = |B_{r,\theta}|^{-1} \sum_{k \in B_{r,\theta}} |\mathscr{F}(\mathbf{x})_k|^2, \quad l_{r,\theta}(\mathbf{x}) = \log(P_{r,\theta}(\mathbf{x}) + \varepsilon), \tag{9}$$

where $P_{r,\theta}(\mathbf{x})$ is the average power in the $r$-th radial and $\theta$-th angular frequency bin, and $\varepsilon$ is a small constant to avoid logarithms of zero. To retain coarse structure, we additionally consider the low-frequency Fourier coefficients. Let $B_{low}$ denote the set of low-frequency bins and define

$$\mathscr{F}_{low}(\mathbf{x}) = \mathscr{F}(\mathbf{x})_{\mathbf{k} \in B_{low}} \tag{10}$$

These low-frequency coefficients capture large-scale structures, such as overall shapes, global lighting, and gradual color changes. The Fourier energy expert is then defined as the sum of two Mahalanobis distances:

$$\mathcal{E}_{\mathrm{FE}}(\mathbf{x}) = \mathcal{M}(l_{rad}(\mathbf{x}); \boldsymbol{\mu}_{rad}, \boldsymbol{\Sigma}_{rad}) + \mathcal{M}(\mathscr{F}_{low}(\mathbf{x}); \boldsymbol{\mu}_{low}, \boldsymbol{\Sigma}_{low}). \tag{11}$$

where $l_{rad}(\mathbf{x}) = \{l_{r,\theta}(\mathbf{x})\}_{r,\theta}$, $\boldsymbol{\mu}_{rad}, \boldsymbol{\Sigma}_{rad}$ are mean and covariance of the log radial-angular power spectrum, and $\boldsymbol{\mu}_{low}, \boldsymbol{\Sigma}_{low}$ are the mean and covariance of the low-frequency Fourier coefficients. The radial-angular term captures the anisotropic energy across different frequency and direction, while the low-frequency term preserves the coarse coefficient and cross-frequency dependencies.

**Proposition 1** *Let $\mathbf{z}_{rad} = l_{rad}(\mathbf{x})$, $\mathbf{z}_{low} = \mathscr{F}_{low}(\mathbf{x})$, and $j \in \{rad, low\}$. Assume that the Fourier features $\mathbf{z}_j$ are Gaussian with mean $\boldsymbol{\mu}_j$ and positive-definite covariance $\boldsymbol{\Sigma}_j$. Then minimizing $\mathcal{E}_{\mathrm{FE}}(\mathbf{x})$ is equivalent to minimizing the covariance-weighted score discrepancy from the corresponding means.*

The proof is given in Appendix A.1.

**Wavelet energy.** While Fourier statistics capture the global frequency information, they provide limited local spatial information, such as edges and textures. We combine FE with a wavelet energy expert that models local spectral behavior across multiple resolutions. Recall that $D_{jcouv}(\boldsymbol{x})$ denotes the Haar detail coefficient at scale $j$, orientation $o \in \{\mathrm{LH}, \mathrm{HL}, \mathrm{HH}\}$, channel $c$, and spatial location

$(u, v)$. The average power of each subband is defined as

$$p_{jco}(\mathbf{x}) = \frac{1}{H_j W_j} \sum_{u,v} D_{jcouv}(\mathbf{x})^2, \tag{12}$$

Wavelet coefficients of natural data are typically sparse and strongly non-Gaussian. Most coefficients have almost zero values, while a small set of coefficients exhibit large magnitudes. This heavy-tailed behavior breaks the Gaussian assumption of the diffusion models and can lead to poor performance. Kurtosis provides a fourth-order statistic that characterizes the extremity of distributional tails (Westfall, 2014). Higher kurtosis indicates greater tail extremity relative to distributions with comparable lower-order moments. We therefore include subband kurtosis in the wavelet energy expert to capture the heavy-tailed behavior of wavelet coefficients,

$$\kappa_{joc}(\mathbf{x}) = \frac{\mathbb{E}\left[(D_{joc}(\mathbf{x}) - \mu_{joc})^4\right]}{\left(\mathbb{E}[(D_{joc}(\mathbf{x}) - \mu_{joc})^2]\right)^2}. \tag{13}$$

Then we define the wavelet energy expert as:

$$\mathcal{E}_{\mathrm{WE}}(\mathbf{x}) = \sum_{j,o,c} \left|\log\left(\kappa_{joc}(\tilde{\mathbf{x}})\right) - \log\left(\kappa_{joc}(\mathbf{x})\right)\right| + \mathcal{M}(p_{sub}(\mathbf{x}); \boldsymbol{\mu}_{sub}, \boldsymbol{\Sigma}_{sub}) + \mathcal{M}(\mathscr{W}(\mathbf{x}); \boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w), \tag{14}$$

where $p_{sub}(\mathbf{x}) = \{p_{jco}(\mathbf{x})\}_{j,o,c}$, $\mathscr{W}(\mathbf{x}) = \{D_{jco}(\mathbf{x})\}_{j,o,c}$, $\boldsymbol{\mu}_{sub}, \boldsymbol{\Sigma}_{sub}$ are the mean and covariance of the subband powers, and $\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w$ are the mean and covariance of the wavelet coefficients. The wavelet energy expert captures both the heavy-tailed behavior of the wavelet coefficients and the statistical dependencies across different subbands by combining different orders of statistics.

**Proposition 2** *For a wavelet subband with $N$ coefficients and nonzero variance, let $\mathbf{v}$ denote its centered coefficient vector and let $n_{\mathrm{eff}} = \frac{\|\mathbf{v}\|_2^4}{\|\mathbf{v}\|_4^4}$ denote its effective coefficient support size. Then, the log-kurtosis term in $\mathcal{E}_{\mathrm{WE}}$ matches the effective coefficient support sizes of the wavelet coefficients*

The proof is given in Appendix A.2.

Consequently, these two energy experts enforce dataset-level priors: FE constrains global frequency allocation and coarse structure, while WE constrains localized, multiscale, and heavy-tailed spectral behavior. However, matching these statistics alone does not guarantee that the generated sample can match its corresponding clean target in spectral space, which may result in over-smoothing or loss of fine details.

### 4.2 Per-sample Experts

The energy experts described above encourage the generated samples to match the dataset-level spectral statistics, but they do not enforce spectral consistency between an individual predicted sample and its corresponding clean target. We therefore complement the energy experts with two per-sample experts that directly constrain the spectral representation.

Recall that the Fourier and wavelet transforms are denoted by $\mathscr{F}$ and $\mathscr{W}$, respectively. We consider the amplitude and phase of the Fourier coefficients, i.e. $\mathscr{F}(\mathbf{x})_k = A_k e^{i\phi_k}$. Given a predicted clean sample $\tilde{\mathbf{x}}_0$ and its corresponding target $\mathbf{x}_0$, we define the per-sample Fourier expert by:

$$\mathcal{E}_{\mathrm{FP}} = \frac{1}{K} \sum_k \left|\log(A_k) - \log(\tilde{A}_k)\right| + \frac{\sum_k w_k \left[1 - \mathrm{Re}\left(\frac{\mathscr{F}(\tilde{\mathbf{x}})_k \mathscr{F}(\mathbf{x})_k^*}{\tilde{A}_k A_k}\right)\right]}{\sum_k w_k}, \tag{15}$$

where $w_k = \sqrt{\tilde{A}_k A_k}$ is a weighting factor that emphasizes the phase matching of high-amplitude coefficients while de-emphasizing low-amplitude coefficients. The second term measures the phase discrepancy since

$$\mathrm{Re}\left(\frac{\mathscr{F}(\tilde{\mathbf{x}})_k \mathscr{F}(\mathbf{x})_k^*}{\tilde{A}_k A_k}\right) = \mathrm{Re}\left(e^{i(\phi_k - \tilde{\phi}_k)}\right) = \cos(\phi_k - \tilde{\phi}_k). \tag{16}$$

Similarly, the wavelet per-sample expert compares the multiscale wavelet representations:

$$\mathcal{E}_{\mathrm{WP}}(\tilde{\mathbf{x}}_0, \mathbf{x}_0) = \sum_{j,o,c} \|D_{j,o,c}(\tilde{\mathbf{x}}_0) - D_{j,o,c}(\mathbf{x}_0)\|_1, \tag{17}$$

which excludes the low-frequency approximation coefficients $A_J$ and only focuses on the detail coefficients $D_{j,o,c}$ at each scale $j$, orientation $o$, and channel $c$.

Jointly, the energy and per-sample experts provide two complementary forms of spectral regularization: the former force the generated samples toward the spectral distribution of the real data, whereas the latter preserve the individual spectral structure. However, the relative importance of these experts varies throughout the denoising process, since the spectral reliability of $\tilde{\mathbf{x}}_0$ changes with the noise level, which motivates the timestep-conditioned router introduced in the next section.

### 4.3 Timestep-Conditioned Router

In this section, we introduce a timestep-conditioned router that adaptively allocates regularization mass across the four spectral experts. Let $\tau_t = 1 - t/T$ be a rescaled timestep and $\mathcal{G}_\Phi$ be the router function. The input of $\mathcal{G}_\Phi$ includes a low-dimensional timestep embedding

$$\boldsymbol{z} = [\tau, \tau^2, \sin(\pi\tau), \cos(\pi\tau)]^\top, \tag{18}$$

and its output consists of one scalar $m_E$ and two vectors $\boldsymbol{g}_E \in \mathbb{R}^2$ and $\boldsymbol{g}_P \in \mathbb{R}^2$. $m_E$ controls the total mass allocated to the energy experts, while $m_P = 1 - m_E$ is assigned to the per-sample experts. $\boldsymbol{g}_E$ and $\boldsymbol{g}_P$ control the Fourier-wavelet allocation within the energy and per-sample experts, respectively. We constrain $m_E$ to a fixed range such that both energy and per-sample experts cannot vanish or dominate the loss. The hierarchical parameterization separates the trade-off between dataset-level statistics and per-sample discrepancies from the choice of spectral representation.

We introduce two smooth, nonnegative scheduling priors, $u_{\mathrm{F}}$ and $u_{\mathrm{W}}$, encoding the Fourier and wavelet scheduling constraints along the denoising trajectory. The Fourier scheduler assigns a higher weight to the intermediate denoising process, where the global structure of the sample has emerged but remains under refinement. In contrast, weight assigned by the wavelet scheduler increases toward the later time of denoising process, when local and high-frequency structures are sufficiently recovered to support meaningful fine-scale regularization. Given the rescaled timestep $\tau_t$, we define the normalized prior by

$$\boldsymbol{\pi}(\tau) = \frac{[u_{\mathrm{F}}, u_{\mathrm{W}}]^\top}{u_{\mathrm{F}} + u_{\mathrm{W}}}, \tag{19}$$

where $u_{\mathrm{F}}$ and $u_{\mathrm{W}}$ are given smooth functions of $\tau$ that control the relative importance of Fourier and wavelet experts. The prior encourages Fourier constraints over a broad intermediate range and wavelet constraints at later timesteps, since at early timesteps the predicted sample is noisy and unreliable, while at later timesteps the predicted sample is cleaner and more accurate, allowing for more detailed wavelet constraints. A visual illustration of the scheduling prior is shown in Fig. 3. The normalized prior is applied to the learned logits before obtaining the Fourier and wavelet weights:

$$\mathrm{softmax}(p_j + \log \pi_j(\tau)) \tag{20}$$

where $j \in \{F, W\}$ and $p_j$ is the learned logits. We then define the timestep-conditioned router as:

$$\boldsymbol{g}(\tau) = \mathcal{G}_\Phi(\mathbf{x}_t, \boldsymbol{z}, \boldsymbol{\pi}(\tau)) = [m_E, g_{\mathrm{FE}}, g_{\mathrm{WE}}, g_{\mathrm{FP}}, g_{\mathrm{WP}}]^\top \tag{21}$$

where $\mathbf{x}_t$ is the noisy sample at timestep $t$, $\boldsymbol{z}$ and $\boldsymbol{\pi}(\tau)$ are the timestep embeddings, and $g_{\mathrm{FE}}, g_{\mathrm{WE}}, g_{\mathrm{FP}}, g_{\mathrm{WP}}$ are the routing weights for the four experts. The spectral loss is then defined as the weighted sum of the four expert losses:

$$\mathcal{L}_{\mathrm{spec}} = m_E \left(g_{\mathrm{FE}}(t)\mathcal{E}_{\mathrm{FE}} + g_{\mathrm{WE}}(t)\mathcal{E}_{\mathrm{WE}}\right) + m_P \left(g_{\mathrm{FP}}(t)\mathcal{E}_{\mathrm{FP}} + g_{\mathrm{WP}}(t)\mathcal{E}_{\mathrm{WP}}\right). \tag{22}$$

Our final objective combines the standard diffusion loss with the routed spectral loss:

$$\mathcal{L} = \mathcal{L}_{\mathrm{diff}} + \mathcal{L}_{\mathrm{spec}}, \tag{23}$$

where $\mathcal{L}_{\mathrm{diff}}$ is the standard diffusion loss.

## 5 Experiments

We evaluate MoSE across image generation and real-world PDE forecasting to examine two research questions:

1. Does MoSE consistently improve the generation quality across different scales and resolutions with complicated spectral variations?

2. Can MoSE be generalized to different datasets and tasks, which exhibit different spectral behavior, instead of being limited to natural images?

To this end, we consider image generation at resolutions ranging from $32 \times 32$ to $256 \times 256$, forecasting on five real-world physical systems from RealPDEBench.

### 5.1 Experimental Setup

**Datasets and tasks.** For image generation, we use CIFAR-10 (Krizhevsky et al., 2009), ImageNet (Deng et al., 2009) at $64 \times 64$ resolution, CelebA-HQ (Karras et al., 2017), and FFHQ (Karras et al., 2019). For physical forecasting, we evaluate all five paired simulated-to-real scenarios from RealPDEBench (Hu et al., 2026): Cylinder, Controlled Cylinder, FSI, Foil, and Combustion. We further evaluate physical data generation on Linear Transport and Burgers equations.

**Baselines and implementation details.** For image generation, we compare MoSE with its DDPM-based baseline and other generative models listed in the corresponding tables. On CIFAR-10 and ImageNet, we apply MoSE to VP-EDM (Karras et al., 2022). On CelebA-HQ and FFHQ, we apply MoSE to LDM-4 (Rombach et al., 2022) with pretrained VQ-4 autoencoding model. The spectral experts are applied to the latent space of the autoencoder, which has a resolution of $64 \times 64$ and 3 channels. On RealPDEBench, we compare with U-Net, CNO, DeepONet, FNO, WDNO, MWT, GK-Transformer, and Transolver. We apply MoSE to the conventional DDPM for all five scenarios, using the same U-Net architecture as the baseline WDNO. For one-dimensional generation, we use vanilla DDPM as the baseline.

**Evaluation metrics.** For images, we report Fréchet Inception Distance (FID), Inception Score (IS), precision, and recall. For RealPDEBench, we report root mean squared error (RMSE), relative $L_2$ error (Rel. $L_2$), and Fourier-space RMSE (fRMSE). For the one-dimensional tasks, we report residual errors, distribution distances, solution MSE, and final-time MSE. Arrows in the tables indicate whether higher or lower values are better.

Table 1: Generation results on CIFAR-10 and ImageNet $64 \times 64$. NFE denotes the number of function evaluations; lower FID is better.

| Unconditional CIFAR-10 | | | Class-conditional ImageNet $64 \times 64$ | | |
|---|---|---|---|---|---|
| Model | NFE | FID (↓) | Model | NFE | FID (↓) |
| VP-EDM | 35 | <u>1.97</u> | VP-EDM | 511 | <u>1.36</u> |
| DPM-Solver-3 | 48 | 2.65 | DDPM | 250 | 11.00 |
| DDIM | 100 | 4.16 | ADM | 250 | 2.07 |
| DDPM | 1000 | 3.17 | DiT-L/2 | 250 | 2.91 |
| NCSN++ | 1000 | 2.38 | U-ViT | 50 | 4.26 |
| iDDPM | 4000 | 2.90 | iCT | 1 | 4.02 |
| sCT | 2 | 2.06 | iCT-deep | 1 | 3.25 |
| ECT | 2 | 2.11 | ECT | 1 | 2.49 |
| iCT | 2 | 2.46 | | | |
| Ours | 35 | **1.78** | Ours | 511 | **1.21** |

Table 2: Unconditional generation on CelebA-HQ and FFHQ at $256 \times 256$.

| CelebA-HQ $256 \times 256$ | | | | FFHQ $256 \times 256$ | | | |
|---|---|---|---|---|---|---|---|
| Method | FID (↓) | Prec. (↑) | Recall (↑) | Method | FID (↓) | Prec. (↑) | Recall (↑) |
| DC-VAE | 15.8 | – | – | ImageBART [21] | 9.57 | – | – |
| VQGAN+T. ($k = 400$) | 10.2 | – | – | U-Net GAN (+aug) [77] | 10.9 (7.6) | – | – |
| PGGAN | 8.0 | – | – | UDM [43] | 5.54 | – | – |
| LSGM | 7.22 | – | – | StyleGAN [41] | <u>4.16</u> | 0.71 | 0.46 |
| UDM | 7.16 | – | – | ProjectedGAN [76] | **3.08** | 0.65 | 0.46 |
| LDM-4 | <u>5.11</u> | <u>0.72</u> | <u>0.49</u> | LDM-4 | 4.98 | <u>0.73</u> | <u>0.50</u> |
| *MoSE* (ours) | **4.82** | **0.80** | **0.55** | *MoSE* (ours) | 4.35 | **0.78** | **0.55** |

### 5.2 Image Generation

We compare MoSE with the baseline methods in Tables 1-2. Among the methods listed in each table, MoSE gives the lowest FID on three of the four benchmarks, and a lower FID compared to the baselines with the same architecture. On CIFAR-10 and ImageNet, it obtains FID scores of 1.78 and 1.21, compared to 1.97 and 1.36 for VP-EDM. On CelebA-HQ, MoSE improves over LDM-4 in FID (4.82 vs. 5.11), precision (0.80 vs. 0.72), and recall (0.55 vs. 0.49). On FFHQ, it improves FID from 4.98 for LDM-4 to 4.35 and gives the highest precision and recall in the table.

Overall, MoSE improves image quality across both low- and high-resolution benchmarks. It also shows that MoSE is effective in latent diffusion models, demonstrating that latent representations exhibit spectral structure that can be exploited by MoSE, and that regularizing in latent space can also improve the quality of generated images. A direct comparison with vanilla DDPM is provided in Appendix C.1.

### 5.3 REALPDEBENCH

We next evaluate whether MoSE can be effectively transferred from natural images to physical systems. In contrast to images, the objective of this task is accurate temporal forecasting rather than visual realism, and errors in spatial frequencies can correspond directly to incorrect physical structures or dynamics. Following Hu et al. (2026), we train each method on either simulated or real-world data and evaluate both settings on the real-world test set. The simulated training setting is more challenging because of the distribution shift between the simulated and real-world data. We report RMSE, relative $L_2$ error, and fRMSE. Lower values are better for all three metrics.

Table 3 compares MoSE with the best baseline value for each metric and training setting, excluding DPOT-S-FT and DPOT-L-FT because both are pretrained on external datasets. The baseline entries may therefore come from different methods. Full results for all baselines, including both DPOT models, together with their parameter counts, are provided in Appendix C.3.

Table 3: For each column, we report the best result among the compared baseline methods, excluding DPOT-S-FT and DPOT-L-FT. All models are evaluated on the real-world test set, and lower is better. The better value in each pair is shown in bold, with ties bolded for both entries.

| Scenario | Entry | Simulated training | | | Real-world training | | |
|---|---|---|---|---|---|---|---|
| | | RMSE($\times$100) | Rel. $L_2$($\times$100) | fRMSE($\times$100) | RMSE($\times$100) | Rel. $L_2$($\times$100) | fRMSE($\times$100) |
| Cylinder | Baseline | 6.99 | **18.49** | 1.05 | 4.24 | **7.01** | 0.69 |
| | MoSE | **6.76** | 18.98 | **0.91** | **4.10** | 8.38 | **0.59** |
| Controlled Cyl. | Baseline | 1.67 | **12.09** | 0.22 | **0.80** | **5.55** | 0.09 |
| | MoSE | **1.63** | 12.11 | **0.20** | **0.80** | 5.79 | **0.07** |
| FSI | Baseline | 2.23 | 15.89 | 0.25 | 0.85 | **5.83** | 0.07 |
| | MoSE | **2.09** | **15.77** | **0.19** | **0.70** | 5.91 | **0.06** |
| Foil | Baseline | **2.27** | 4.38 | **0.26** | **1.00** | **1.59** | **0.11** |
| | MoSE | 2.31 | **4.28** | 0.30 | 1.32 | 2.28 | **0.11** |
| Combustion | Baseline | 3.58 | **72.90** | **0.51** | 2.16 | **54.87** | **0.26** |
| | MoSE | **3.43** | 79.88 | 0.52 | **2.12** | 61.12 | 0.27 |

MoSE performs most strongly on Cylinder, Controlled Cylinder and FSI, where it has the best performance on a majority of the metrics. On Foil and Combustion, MoSE is competitive with the best baseline, either outperforming it on some metrics or being close to the best value on most metrics. Notably, MoSE consistently outperforms the WDNO baseline on all five scenarios, which is a DDPM-based model as well. Consequently, MoSE is able to improve the accuracy of physical forecasting on complicated physical systems with strong spectral variations, even when trained on simulated data with distribution shifts. Results on one-dimensional physical data generation with discontinuous initial conditions and direct comparison with conventional DDPM are provided in Appendix C.4.

## 6 CONCLUSION

We introduced Mixture-of-Spectral-Experts (MoSE), a training-time regularizer for diffusion models. MoSE combines Fourier experts for global structure with wavelet experts for local details. It matches both dataset-level spectral statistics and individual sample structure, while a timestep-conditioned router adaptively adjusts the contribution of each expert during denoising process.

MoSE improves image generation across all evaluated datasets and gives the best FID compared with other diffusion-based baselines. It also improves physical data generation, including Cylinder, Controlled Cylinder and FSI in RealPDEBench. Overall, the results show that combining Fourier and wavelet constraints as a mixture of spectral experts can effectively improve the spectral structure learned by diffusion models, and also improves the quality of generated samples.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing and polishing of the manuscript, as well as with the reimplementation of baseline methods. AI-assisted code was reviewed, modified, and tested by the authors, and the baseline implementations were checked against the corresponding papers and publicly available implementations when applicable. Generative AI tools were not used to generate experimental results, formulate research claims, or make scientific conclusions. We have reviewed all AI-assisted content and take responsibility for the final content of this work, including text, code, claims, and artifacts produced with the aid of generative AI.

## REFERENCES

Roi Benita, Miki Elad, and Joseph Keshet. Spectral analysis of diffusion models with application to schedule design. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/031bd3fc62939fcb38ed887c6f66d5c7-Abstract-Conference.html.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 22563–22575. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02161. URL https://doi.org/10.1109/CVPR52729.2023.02161.

Nicolas Bourriez, Alexandre Verine, and Auguste Genovesio. Spectral collapse in diffusion inversion. *CoRR*, abs/2602.13303, 2026. doi: 10.48550/ARXIV.2602.13303. URL https://doi.org/10.48550/arXiv.2602.13303.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Satish Chandran, Nícolas Roque dos Santos, Yunshu Wu, Greg Ver Steeg, and Evangelos Papalexakis. Spectral regularization for diffusion models. *CoRR*, abs/2603.02447, 2026. doi: 10.48550/ARXIV.2603.02447. URL https://doi.org/10.48550/arXiv.2603.02447.

Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. FIRE: robust detection of diffusion-generated images via frequency-guided reconstruction error. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 12830–12839. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01197. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chu_FIRE_Robust_Detection_of_Diffusion-Generated_Images_via_Frequency-Guided_Reconstruction_Error_CVPR_2025_paper.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Fabian Falck, Teodora Pandeva, Kiarash Zahirnia, Rachel Lawrence, Richard E. Turner, Edward Meeds, Javier Zazo, and Sushrut Karmalkar. A fourier space perspective on diffusion models. *CoRR*, abs/2505.11278, 2025. doi: 10.48550/ARXIV.2505.11278. URL https://doi.org/10.48550/arXiv.2505.11278.

Daniel Floryan and Michael D Graham. Discovering multiscale and self-similar structure with data-driven wavelets. *Proceedings of the National Academy of Sciences*, 118(1):e2021299118, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *CoRR*, abs/2210.02303, 2022. doi: 10.48550/ARXIV.2210.02303. URL https://doi.org/10.48550/arXiv.2210.02303.

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 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=FQhDIGuaJ4.

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. *CoRR*, abs/2601.01829, 2026. doi: 10.48550/ARXIV.2601.01829. URL https://doi.org/10.48550/arXiv.2601.01829.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution ai-generated image detection by spectral learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 18706–18717. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01743. URL https://openaccess.thecvf.com/content/CVPR2025/html/Karageorgiou_Any-Resolution_AI-Generated_Image_Detection_by_Spectral_Learning_CVPR_2025_paper.html.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4396–4405. IEEE, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/a98846e9d9cc01cfb87eb694d946ce6b-Abstract-Conference.html.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=a-xFK8Ymz5J.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Max W. Y. Lam, Jun Wang, Dan Su, and Dong Yu. BDDM: bilateral denoising diffusion models for fast and high-quality speech synthesis. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=L7wzpQttNO.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 7167–7177, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21450–21474. PMLR, 2023. URL https://proceedings.mlr.press/v202/liu23f.html.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 2021. URL http://proceedings.mlr.press/v139/nichol21a.html.

Angus Phillips, Thomas Seror, Michael J. Hutchinson, Valentin De Bortoli, Arnaud Doucet, and Emile Mathieu. Spectral diffusion processes. *CoRR*, abs/2209.14125, 2022. doi: 10.48550/ARXIV.2209.14125. URL https://doi.org/10.48550/arXiv.2209.14125.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Fabio Sigrist, Hans R Künsch, and Werner A Stahel. Stochastic partial differential equation based modelling of large space–time data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(1):3–33, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Binxu Wang and Cengiz Pehlevan. An analytical theory of spectral bias in the learning dynamics of diffusion models. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/8a0d3f77bb435817807d463c5dcef1ab-Abstract-Conference.html.

Peter H Westfall. Kurtosis as peakedness, 1905–2014. rip. *The American Statistician*, 68(3):191–195, 2014.

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

Christopher K Wikle and Noel Cressie. A dimension-reduced approach to space-time kalman filtering. *Biometrika*, 86(4):815–829, 1999.

Amir Ziashahabi, Baturalp Buyukates, Artan Sheshmani, Yi-Zhuang You, and Salman Avestimehr. Frequency domain diffusion model with scale-dependent noise schedule. In *IEEE International Symposium on Information Theory, ISIT 2024, Athens, Greece, July 7-12, 2024*, pp. 19–24. IEEE, 2024. doi: 10.1109/ISIT57864.2024.10619452. URL https://doi.org/10.1109/ISIT57864.2024.10619452.

# A PROOFS OF THE PROPOSITIONS

## A.1 PROOF OF PROPOSITION 1

Recall that $\mathbf{z}_{\mathrm{rad}} = l_{\mathrm{rad}}(\mathbf{x})$, $\mathbf{z}_{\mathrm{low}} = \mathscr{F}_{\mathrm{low}}(\mathbf{x})$, and $j \in \{\mathrm{rad}, \mathrm{low}\}$. Let $d_j = \dim \mathbf{z}_j$ and assume $\mathbf{\Sigma}_j \succ 0$. The score of $\mathbf{z}_j$ is

$$\boldsymbol{s}(\mathbf{z}_j) = \nabla_{\mathbf{z}_j} \log q_j(\mathbf{z}_j) = -\mathbf{\Sigma}_j^{-1}(\mathbf{z}_j - \boldsymbol{\mu}_j). \tag{24}$$

Writing $\|\boldsymbol{v}\|_{\mathbf{\Sigma}}^2 = \boldsymbol{v}^\top \mathbf{\Sigma} \boldsymbol{v}$, Equation 8 gives

$$\begin{aligned}\mathcal{E}_{\mathrm{FE}}(\mathbf{x}) &= \sum_j \frac{1}{2d_j}(\mathbf{z}_j - \boldsymbol{\mu}_j)^\top \mathbf{\Sigma}_j^{-1}(\mathbf{z}_j - \boldsymbol{\mu}_j) \\ &= \sum_j \frac{1}{2d_j}\|\boldsymbol{s}(\mathbf{z}_j) - \boldsymbol{s}(\boldsymbol{\mu}_j)\|_{\mathbf{\Sigma}_j}^2. \end{aligned} \tag{25}$$

Since $\boldsymbol{s}(\boldsymbol{\mu}_j) = \mathbf{0}$, minimizing $\mathcal{E}_{\mathrm{FE}}(\mathbf{x})$ is equivalent to minimizing the covariance-weighted score discrepancy from the corresponding means.

## A.2 PROOF OF PROPOSITION 2

Fix subband $(j, o, c)$. Let the centered coefficients be:

$$\mathbf{v} = (v_1, \cdots, v_N), \quad N = H_j W_j, \tag{26}$$

where $v_i = D_i - \mu$. The kurtosis is:

$$\kappa_{joc}(\mathbf{x}) = \frac{\frac{1}{N}\sum_i v_i^4}{(\frac{1}{N}\sum_i v_i^2)^2} = N\frac{\|\mathbf{v}\|_4^4}{\|\mathbf{v}\|_2^4}. \tag{27}$$

Define the effective coefficient support size as

$$n_{\mathrm{eff}}(\mathbf{x}) := \frac{\|\mathbf{v}\|_2^4}{\|\mathbf{v}\|_4^4}. \tag{28}$$

Then

$$\kappa_{joc}(\mathbf{x}) = \frac{N}{n_{\mathrm{eff}}(\mathbf{x})}. \tag{29}$$

By using Cauchy-Schwarz inequality, we have

$$\left(\sum_i v_i^2\right)^2 \le \|\mathbf{v}\|_0 \sum_i v_i^4. \tag{30}$$

Therefore,

$$\|\mathbf{v}\|_2^4 \le \|\mathbf{v}\|_0 \|\mathbf{v}\|_4^4, \tag{31}$$

which implies that

$$n_{\mathrm{eff}}(\mathbf{x}) = \frac{\|\mathbf{v}\|_2^4}{\|\mathbf{v}\|_4^4} \le \|\mathbf{v}\|_0. \tag{32}$$

Consider a predicted subband $\tilde{\mathbf{x}}$ and a target subband $\mathbf{x}$. We have

$$\begin{aligned}|\log \kappa_{joc}(\tilde{\mathbf{x}}) - \log \kappa_{joc}(\mathbf{x})| &= \left|\log \frac{N}{n_{\mathrm{eff}}(\tilde{\mathbf{x}})} - \log \frac{N}{n_{\mathrm{eff}}(\mathbf{x})}\right| \\ &= \left|\log n_{\mathrm{eff}}(\tilde{\mathbf{x}}) - \log n_{\mathrm{eff}}(\mathbf{x})\right|. \end{aligned} \tag{33}$$

Hence, the kurtosis term in $\mathcal{E}_{\mathrm{WE}}$ matches the effective coefficient support sizes in log space. It reaches its minimum of zero when

$$n_{\mathrm{eff}}(\tilde{\mathbf{x}}) = n_{\mathrm{eff}}(\mathbf{x}) \tag{34}$$

# B Wavelet Transform Details

Recall that given an image at diffusion timestep $t$, $x_t \in \mathbb{R}^{C\times H\times W}$, a $J$-level two-dimensional wavelet transform can be expressed as

$$\mathcal{W}(x_t) = \left\{ A_J, \left\{ D_j^{LH}, D_j^{HL}, D_j^{HH} \right\}_{j=1}^{J} \right\}, \tag{35}$$

where $A_J$ denotes the low-frequency approximation coefficients at the coarsest decomposition level, and $D_j^{LH}$, $D_j^{HL}$, and $D_j^{HH}$ denote the detail coefficients at scale $j$ along the three spatial orientations.

One of the most commonly used wavelets is the Haar wavelet. The one-dimensional Haar scaling function $\phi$ and mother wavelet $\psi$ are defined as

$$\phi(s) = \begin{cases} 1, & 0 \le s < 1, \\ 0, & \text{otherwise}, \end{cases} \qquad \psi(s) = \begin{cases} 1, & 0 \le s < \frac{1}{2}, \\ -1, & \frac{1}{2} \le s < 1, \\ 0, & \text{otherwise}. \end{cases} \tag{36}$$

At scale $j$ and translation $k$, the corresponding basis functions are

$$\phi_{j,k}(s) = 2^{j/2}\phi(2^j s - k), \qquad \psi_{j,k}(s) = 2^{j/2}\psi(2^j s - k). \tag{37}$$

For two-dimensional inputs, the Haar basis is constructed separably from the one-dimensional scaling and wavelet functions. Let $\mathbf{k} = (k_1, k_2)$ denote the spatial translation index. The three detail basis functions are

$$\Psi_{j,\mathbf{k}}^{LH}(u,v) = \phi_{j,k_1}(u)\psi_{j,k_2}(v),\ \Psi_{j,\mathbf{k}}^{HL}(u,v) = \psi_{j,k_1}(u)\phi_{j,k_2}(v),\ \Psi_{j,\mathbf{k}}^{HH}(u,v) = \psi_{j,k_1}(u)\psi_{j,k_2}(v), \tag{38}$$

and the low-frequency approximation basis is given by

$$\Phi_{j,\mathbf{k}}(u,v) = \phi_{j,k_1}(u)\phi_{j,k_2}(v). \tag{39}$$

The wavelet coefficient of channel $c$ at scale $j$, orientation $o \in \{LH, HL, HH\}$, and spatial location $\mathbf{k}$ can be defined as

$$D_{j,c,o,\mathbf{k}}(x_t) = \left\langle x_{t,c}, \Psi_{j,\mathbf{k}}^{o} \right\rangle, \tag{40}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product over the spatial dimensions.

# C Additional Results

## C.1 Comparison with Vanilla DDPM

Table 4 compares MoSE with vanilla DDPM. Both methods are trained on the same datasets with the same architecture and smaller training budget. MoSE reduces FID by 5.3% on CIFAR-10, 38.5% on ImageNet, and 35.6% on CelebA. It also improves IS on CIFAR-10 and ImageNet, with a small decrease on CelebA (from 2.52 to 2.49).

Table 4: Comparison with DDPM on image generation. Best results for each dataset are in bold, and second-best results are underlined.

| Metric | CIFAR-10 | | ImageNet $64 \times 64$ | | CelebA | |
|---|---|---|---|---|---|---|
| | DDPM | Ours | DDPM | Ours | DDPM | Ours |
| FID (↓) | <u>9.55</u> | **9.04** | <u>30.74</u> | **18.91** | <u>63.63</u> | **40.99** |
| IS (↑) | <u>8.11</u> | **8.30** | <u>12.38</u> | **13.36** | **2.52** | <u>2.49</u> |

## C.2 Ablation Studies

We design the ablations to answer two questions:

1. Whether Fourier and wavelet (energy and per-sample) experts provide complementary information?
2. Whether adaptive routing is preferable to a fixed combination?

Same denoising backbone, data split, optimizer, training budget, and sampling schedule are used for all variants. We report the mean and standard deviation over at least three random seeds and using the same checkpoint-selection rule for all variants.

The image ablation is conducted on CIFAR-10, while the physical system ablation uses one-dimensional Burger's equation. In addition to FID, RMSE, and fRMSE, we add two spectral metrics to evaluate the spectral fidelity of the generated samples. Fourier error (F-Err) is the mean absolute difference between the mean log radial power spectra of generated samples and corresponding ground-truth targets. Wavelet error (W-Err) is the corresponding mean absolute difference between log wavelet-subband energies, averaged over scales, orientations, and channels. Lower is better for all the metrics in the ablation tables.

Table 5 separates the two axes of the method. The Fourier-only and wavelet-only variants test global versus localized spectral representations, whereas the energy-only and per-sample-only variants test dataset-level statistics versus sample-specific matching. The four single-expert rows evaluate whether the gain from a family is dominated by one loss, and the following rows measure each expert's marginal contribution in the complete framework. Comparing uniform weighting with adaptive routing tests whether the timestep-conditioned router is necessary to balance the four experts. The final row is the full MoSE model with all four experts and adaptive routing.

Table 5: Results for the expert component ablation (mean $\pm$ standard deviation). "Y" means that an expert is enabled. Uniform routing assigns equal weight to the enabled experts, and adaptive denotes timestep-conditioned routing over the enabled experts. Lower is better for every metric, and the best value is in bold.

| | Enabled experts | | | | | CIFAR-10 | | | Burgers | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variant | FE | WE | FP | WP | Routing | FID | F-Err | W-Err | RMSE | fRMSE |
| DDPM backbone | – | – | – | – | – | $9.55 \pm 0.12$ | $0.184 \pm 0.006$ | $0.231 \pm 0.008$ | $2.78 \pm 0.09$ | $0.392 \pm 0.014$ |
| FE only | Y | – | – | – | – | $9.38 \pm 0.10$ | $0.155 \pm 0.005$ | $0.218 \pm 0.007$ | $2.51 \pm 0.08$ | $0.341 \pm 0.012$ |
| WE only | – | Y | – | – | – | $9.36 \pm 0.11$ | $0.173 \pm 0.006$ | $0.191 \pm 0.006$ | $2.47 \pm 0.08$ | $0.326 \pm 0.011$ |
| FP only | – | – | Y | – | – | $9.30 \pm 0.10$ | $0.146 \pm 0.004$ | $0.210 \pm 0.007$ | $2.34 \pm 0.07$ | $0.304 \pm 0.010$ |
| WP only | – | – | – | Y | – | $9.28 \pm 0.09$ | $0.166 \pm 0.005$ | $0.182 \pm 0.006$ | $2.29 \pm 0.07$ | $0.291 \pm 0.009$ |
| Energy only | Y | Y | – | – | Uniform | $9.23 \pm 0.09$ | $0.142 \pm 0.004$ | $0.178 \pm 0.005$ | $2.18 \pm 0.06$ | $0.270 \pm 0.009$ |
| Per-sample only | – | – | Y | Y | Uniform | $9.18 \pm 0.09$ | $0.134 \pm 0.004$ | $0.169 \pm 0.005$ | $1.97 \pm 0.06$ | $0.240 \pm 0.008$ |
| Fourier only | Y | – | Y | – | Uniform | $9.16 \pm 0.08$ | $0.121 \pm 0.003$ | $0.195 \pm 0.006$ | $1.91 \pm 0.05$ | $0.228 \pm 0.007$ |
| Wavelet only | – | Y | – | Y | Uniform | $9.14 \pm 0.09$ | $0.149 \pm 0.005$ | $0.151 \pm 0.004$ | $1.84 \pm 0.05$ | $0.216 \pm 0.007$ |
| All experts, fixed | Y | Y | Y | Y | Uniform | $9.12 \pm 0.08$ | $0.116 \pm 0.003$ | $0.143 \pm 0.004$ | $1.61 \pm 0.05$ | $0.190 \pm 0.006$ |
| Full w/o FE | – | Y | Y | Y | Adaptive | $9.11 \pm 0.07$ | $0.122 \pm 0.004$ | $0.139 \pm 0.004$ | $1.55 \pm 0.04$ | $0.181 \pm 0.006$ |
| Full w/o WE | Y | – | Y | Y | Adaptive | $9.13 \pm 0.08$ | $0.109 \pm 0.003$ | $0.154 \pm 0.005$ | $1.58 \pm 0.05$ | $0.186 \pm 0.006$ |
| Full w/o FP | Y | Y | – | Y | Adaptive | $9.15 \pm 0.09$ | $0.124 \pm 0.004$ | $0.145 \pm 0.004$ | $1.64 \pm 0.05$ | $0.194 \pm 0.007$ |
| Full w/o WP | Y | Y | Y | – | Adaptive | $9.14 \pm 0.08$ | $0.116 \pm 0.004$ | $0.158 \pm 0.005$ | $1.60 \pm 0.05$ | $0.189 \pm 0.007$ |
| Full MoSE | Y | Y | Y | Y | Adaptive | $\mathbf{9.04 \pm 0.07}$ | $\mathbf{0.102 \pm 0.003}$ | $\mathbf{0.131 \pm 0.004}$ | $\mathbf{1.37 \pm 0.04}$ | $\mathbf{0.164 \pm 0.005}$ |

All rows in Table 6 retain the same four expert losses and vary only the rule that combines them. Uniform routing is a parameter-free reference. Validation-tuned static routing tests whether a single global mixture is sufficient. The remaining variants remove the spectral scheduling prior, the noisy-sample input, or learned routing from the full router in Equation 21.

Table 6: Results for the routing ablation with all four experts enabled. The input columns show the information available to each routing rule.

| Routing rule | Router information | | | CIFAR-10 | | | Burgers | |
|---|---|---|---|---|---|---|---|---|
| | $\mathbf{x}_t$ | Time embedding $\boldsymbol{z}$ | Prior $\boldsymbol{\pi}(t)$ | FID | F-Err | W-Err | RMSE | fRMSE |
| Uniform static | – | – | – | 9.12 | 0.116 | 0.143 | 1.61 | 0.190 |
| Validation-tuned static | – | – | – | 9.10 | 0.112 | 0.140 | 1.55 | 0.183 |
| Scheduling prior only | – | – | Y | 9.09 | 0.109 | 0.138 | 1.49 | 0.177 |
| Learned, no prior | Y | Y | – | 9.08 | 0.108 | 0.135 | 1.46 | 0.173 |
| Learned, no sample input | – | Y | Y | 9.07 | 0.105 | 0.136 | 1.43 | 0.170 |
| Full router | Y | Y | Y | **9.04** | **0.102** | **0.131** | **1.37** | **0.164** |

### C.3 FULL REALPDEBENCH RESULTS

Tables 7-11 report the complete RealPDEBench comparisons. All methods are evaluated on the real-world test set after training on either simulated or real-world data. Lower values are better; all three error metrics are multiplied by 100 and reported to two decimal places. The bold and underlined entries mark the best and second-best results in each column, respectively.

Overall, our model achieves competitive performance across both simulated- and real-data training, with particularly strong results on Cylinder, Controlled Cylinder and FSI, where it achieves the best or second-best performance across nearly all metrics. Notably, our method consistently improves upon WDNO, the diffusion-based baseline, demonstrating that the proposed framework is capable of improving the accuracy of diffusion-based PDE forecasting. On several individual settings, specialized deterministic or pretrained models achieve lower errors, particularly on Foil and Combustion. However, no single baseline uniformly dominates across all PDE systems, while our method maintains strong performance across heterogeneous dynamics with a single framework. These results indicate that our method provides a strong balance between predictive accuracy and robustness across different physical fields.

Table 7: RealPDEBench results on Cylinder. Models are trained on simulated or real-world data and evaluated on the real-world test set. Lower is better for all metrics. **Bold** and underlined values indicate the best and second-best results, respectively, excluding DPOT-S-FT and DPOT-L-FT.

| Method | Params | Simulated training | | | Real-world training | | |
|---|---|---|---|---|---|---|---|
| | | RMSE ($\times 100$) | Rel. $L_2$ ($\times 100$) | fRMSE ($\times 100$) | RMSE ($\times 100$) | Rel. $L_2$ ($\times 100$) | fRMSE ($\times 100$) |
| U-Net | 23.0M | 7.58 | 21.65 | 1.22 | 7.00 | **7.01** | 1.03 |
| CNO | 8.0M | 7.29 | **18.49** | 1.13 | 4.24 | 8.97 | 0.69 |
| DeepONet | 3.6M | 8.63 | 35.92 | 1.32 | 7.13 | 15.28 | 1.08 |
| FNO | 50.4M | 7.39 | 25.75 | 1.15 | 5.85 | 8.55 | 0.87 |
| WDNO | 104.3M | 6.99 | 20.53 | 1.05 | 6.89 | 16.66 | 0.91 |
| MWT | 2.9M | 7.33 | 22.40 | 1.07 | 5.40 | 8.32 | 0.81 |
| GK-Transformer | 84.4M | 8.76 | 29.41 | 1.39 | 11.96 | 8.98 | 1.66 |
| Transolver | 4.3M | 11.21 | 32.24 | 1.74 | 10.93 | 18.87 | 1.60 |
| DPOT-S-FT | 30.8M | 5.13 | 14.74 | 0.76 | 5.86 | 9.83 | 0.86 |
| DPOT-L-FT | 649.8M | 4.86 | 14.46 | 0.70 | 3.90 | 8.12 | 0.56 |
| MLAverage | – | 7.52 | 23.56 | 1.15 | 6.92 | 11.06 | 1.01 |
| Ours | 13.0M | **6.76** | 18.98 | **0.91** | **4.10** | 8.38 | **0.59** |

Table 8: RealPDEBench results on Controlled Cylinder. All metrics are evaluated on the real-world test set. **Bold** and underlined values indicate the best and second-best results, respectively, excluding DPOT-S-FT and DPOT-L-FT.

| Method | Params | Simulated training | | | Real-world training | | |
|---|---|---|---|---|---|---|---|
| | | RMSE ($\times 100$) | Rel. $L_2$ ($\times 100$) | fRMSE ($\times 100$) | RMSE ($\times 100$) | Rel. $L_2$ ($\times 100$) | fRMSE ($\times 100$) |
| U-Net | 23.0M | 1.95 | 13.60 | 0.29 | **0.80** | **5.55** | 0.10 |
| CNO | 8.0M | 1.67 | **12.09** | 0.22 | 0.81 | 5.83 | 0.09 |
| DeepONet | 3.6M | 5.40 | 42.56 | 1.34 | 3.09 | 23.99 | 0.58 |
| FNO | 50.4M | 2.85 | 20.07 | 0.59 | 0.97 | 7.23 | 0.12 |
| WDNO | 359.8M | 2.40 | 18.29 | 0.37 | 1.15 | 9.27 | 0.14 |
| MWT | 2.9M | 2.20 | 15.94 | 0.37 | 1.02 | 7.47 | 0.11 |
| GK-Transformer | 50.8M | 2.61 | 18.16 | 0.50 | 1.03 | 7.67 | 0.12 |
| Transolver | 4.3M | 2.94 | 18.94 | 0.69 | 1.71 | 12.00 | 0.24 |
| DPOT-S-FT | 30.8M | 2.33 | 17.22 | 0.39 | 0.84 | 5.98 | 0.10 |
| DPOT-L-FT | 649.8M | 2.48 | 17.84 | 0.43 | 0.84 | 6.03 | 0.10 |
| MLAverage | – | 2.68 | 19.47 | 0.52 | 1.23 | 9.10 | 0.17 |
| Ours | 13.0M | **1.63** | 12.11 | **0.20** | **0.80** | 5.79 | **0.07** |

Table 9: RealPDEBench results on FSI. All metrics are evaluated on the real-world test set. **Bold** and underlined values indicate the best and second-best results, respectively, excluding DPOT-S-FT and DPOT-L-FT.

| Method | Params | Simulated training | | | Real-world training | | |
|---|---|---|---|---|---|---|---|
| | | RMSE ($\times 100$) | Rel. $L_2$ ($\times 100$) | fRMSE ($\times 100$) | RMSE ($\times 100$) | Rel. $L_2$ ($\times 100$) | fRMSE ($\times 100$) |
| U-Net | 23.0M | 2.23 | 15.89 | 0.25 | 0.85 | **5.83** | 0.07 |
| CNO | 8.0M | 2.41 | 17.24 | 0.29 | 1.05 | 7.41 | 0.09 |
| DeepONet | 3.4M | 6.37 | 46.06 | 0.97 | 3.50 | 25.02 | 0.51 |
| FNO | 268.5M | 4.26 | 30.95 | 0.59 | 1.29 | 8.92 | 0.12 |
| WDNO | 91.7M | 3.69 | 27.00 | 0.50 | 1.17 | 8.17 | 0.11 |
| MWT | 2.9M | 3.39 | 24.87 | 0.46 | 1.28 | 9.10 | 0.10 |
| GK-Transformer | 67.6M | 3.07 | 22.41 | 0.39 | 1.27 | 9.16 | 0.11 |
| Transolver | 4.3M | 3.05 | 21.19 | 0.43 | 2.36 | 15.67 | 0.27 |
| DPOT-S-FT | 41.3M | 2.60 | 18.86 | 0.31 | 1.05 | 7.46 | 0.08 |
| DPOT-L-FT | 673.5M | 2.62 | 18.98 | 0.30 | 0.99 | 6.87 | 0.08 |
| MLAverage | – | 3.37 | 24.34 | 0.45 | 1.48 | 10.36 | 0.15 |
| Ours | 89.8M | **2.09** | **15.77** | **0.19** | **0.70** | 5.91 | **0.06** |

Table 10: RealPDEBench results on Foil. All metrics are evaluated on the real-world test set. **Bold** and underlined values indicate the best and second-best results, respectively, excluding DPOT-S-FT and DPOT-L-FT.

| Method | Params | Simulated training | | | Real-world training | | |
|---|---|---|---|---|---|---|---|
| | | RMSE ($\times 100$) | Rel. $L_2$ ($\times 100$) | fRMSE ($\times 100$) | RMSE ($\times 100$) | Rel. $L_2$ ($\times 100$) | fRMSE ($\times 100$) |
| U-Net | 23.0M | 2.72 | 7.45 | 0.39 | **1.00** | **1.59** | **0.11** |
| CNO | 8.0M | **2.27** | 4.38 | 0.28 | 1.36 | 2.53 | 0.18 |
| DeepONet | 3.6M | 3.39 | 5.65 | 0.51 | 2.22 | 3.63 | 0.28 |
| FNO | 50.4M | 2.74 | 5.40 | 0.37 | 1.30 | 2.28 | 0.15 |
| WDNO | 358.4M | 2.42 | 4.90 | **0.26** | 1.62 | 4.44 | 0.18 |
| MWT | 2.9M | 2.64 | 4.92 | 0.36 | 1.33 | 2.27 | 0.15 |
| GK-Transformer | 50.8M | 2.77 | 5.12 | 0.38 | 1.42 | 2.45 | 0.18 |
| Transolver | 4.3M | 3.45 | 4.65 | 0.50 | 2.20 | 3.70 | 0.27 |
| DPOT-S-FT | 41.3M | 2.21 | 3.97 | 0.27 | 1.06 | 1.66 | 0.10 |
| DPOT-L-FT | 673.5M | 2.29 | 4.02 | 0.29 | 1.05 | 1.59 | 0.11 |
| MLAverage | – | 2.69 | 5.05 | 0.36 | 1.46 | 2.61 | 0.17 |
| Ours | 100.64M | 2.31 | **4.28** | 0.30 | 1.32 | 2.28 | **0.11** |

Table 11: RealPDEBench results on Combustion. All metrics are evaluated on the real-world test set. **Bold** and underlined values indicate the best and second-best results, respectively, excluding DPOT-S-FT and DPOT-L-FT.

| Method | Params | Simulated training | | | Real-world training | | |
|---|---|---|---|---|---|---|---|
| | | RMSE ($\times 100$) | Rel. $L_2$ ($\times 100$) | fRMSE ($\times 100$) | RMSE ($\times 100$) | Rel. $L_2$ ($\times 100$) | fRMSE ($\times 100$) |
| U-Net | 23.3M | 3.58 | **72.90** | **0.51** | 2.16 | **54.87** | **0.26** |
| CNO | 8.0M | 4.01 | 82.74 | 0.57 | 2.48 | 60.30 | 0.32 |
| DeepONet | 3.5M | 4.03 | 85.65 | 0.56 | 2.29 | 57.51 | 0.28 |
| FNO | 67.1M | 3.63 | 76.06 | 0.52 | 2.26 | 56.64 | 0.27 |
| WDNO | 122.7M | 4.39 | 110.16 | 0.61 | 3.80 | 86.05 | 0.55 |
| MWT | 2.9M | 3.67 | 75.36 | 0.52 | 2.21 | 55.60 | 0.27 |
| GK-Transformer | 67.6M | 4.00 | 91.43 | 0.56 | 2.47 | 60.83 | 0.31 |
| Transolver | 4.3M | 4.42 | 90.56 | 0.61 | 3.75 | 78.25 | 0.50 |
| DPOT-S-FT | 41.5M | 3.93 | 78.42 | 0.54 | 2.09 | 53.49 | 0.24 |
| DPOT-L-FT | 674.2M | 3.78 | 77.50 | 0.53 | 2.08 | 53.31 | 0.24 |
| MLAverage | – | 3.94 | 84.08 | 0.55 | 2.56 | 61.69 | 0.32 |
| Ours | 102.9M | **3.43** | 79.88 | 0.52 | **2.12** | 61.12 | 0.27 |

### C.4 One-Dimensional Physical Data Generation

In this section, we consider two synthetic physical datasets with discontinuity. The Burgers' equation develops discontinuities (known as shock waves) during the evolution, while the linear transport equation is subject to discontinuous initial conditions.

**Burgers' equation.** We consider the one-dimensional forced viscous Burgers' equation,

$$\partial_t u + u\,\partial_x u = 0.01\,\partial_{xx} u + f(x,t), \tag{41}$$

on $x \in [0,1]$ and $t \in [0,8]$, with homogeneous Dirichlet boundary conditions. Initial conditions are sampled as

$$u_0(x) = a_1 \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) + a_2 \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right), \tag{42}$$

where $a_1 \sim \mathcal{U}(0,2)$, $a_2 \sim \mathcal{U}(-2,0)$, $\mu_1 \sim \mathcal{U}(0.2,0.4)$, $\mu_2 \sim \mathcal{U}(0.6,0.8)$, and $\sigma_1, \sigma_2 \sim \mathcal{U}(0.05,0.15)$, independently. The forcing comprises random Gaussian components in space and time. The dataset contains $40{,}000$ training and $8{,}000$ test trajectories, each with $120$ spatial points and $81$ solution snapshots.

**Linear transport equation.** We consider the one-dimensional linear transport equation,

$$\partial_t u + c\,\partial_x u = 0, \tag{43}$$

on the periodic domain $x \in [0,1)$ and $t \in [0,1]$, with $c \sim \mathcal{U}(0.5,2)$. Initial conditions are drawn from a mixture of top-hat profiles ($65\%$) and piecewise-constant profiles ($35\%$). Top-hat profiles have uniformly sampled locations, widths in $[0.08, 0.5]$, and two amplitudes in $[-1,1]$ with a minimum jump of $0.2$. Piecewise-constant profiles contain 3–7 segments, with uniformly sampled breakpoints and amplitudes in $[-1,1]$. We evaluate the exact solution

$$u(x,t) = u_0\big(x - ct\big) \tag{44}$$

at $256$ spatial points and $50$ time points, producing $10{,}000$ training and $2{,}000$ test trajectories.

Table 12 compares MoSE with vanilla DDPM on the Linear Transport and Burgers equations. On Linear Transport, MoSE reduces residual MSE, RMSE, and MAE by 13.3%, 4.2%, and 6.3%, respectively. On Burgers, it reduces residual RMSE by 50.7%, Fréchet distance by 26.5%, Wasserstein distance by 41.4%, and both solution and final-time MSE by 60%. The improvements across residual, distributional, and solution metrics show that MoSE produces more accurate one-dimensional physical samples than the DDPM baseline.

Table 12: One-dimensional physical data generation. Lower is better for all metrics.

| Metric | Linear Transport | | Burgers | |
|---|---|---|---|---|
| | DDPM | Ours | DDPM | Ours |
| Residual MSE (↓) | <u>39.94</u> | **34.62** | - | - |
| Residual RMSE (↓) | <u>5.51</u> | **5.28** | <u>2.78</u> | **1.37** |
| Residual MAE (↓) | <u>1.90</u> | **1.78** | - | - |
| Fréchet Distance (↓) | - | - | <u>646.358</u> | **475.363** |
| Wasserstein Distance (↓) | - | - | <u>0.58</u> | **0.34** |
| PDE Solution MSE (↓) | - | - | <u>0.05</u> | **0.02** |
| Final-Time MSE (↓) | - | - | <u>0.05</u> | **0.02** |

## D Illustration of the routing priors

Figure 3 illustrates the spectral scheduling priors. In practice, we define the priors by:

$$u_{\mathrm{F}}(\tau) = \sigma(10(\tau - 0.15))\sigma(10(0.85 - \tau)), \tag{45}$$

$$u_{\mathrm{W}}(\tau) = \sigma(12(\tau - 0.55)), \tag{46}$$

where $\sigma$ is the sigmoid function.

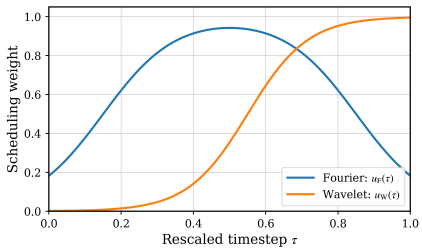



Figure 3: Illustration of the spectral scheduling priors. Accordingly, $u_{\mathrm{F}}(\tau)$ follows a smooth shape with its mass concentrated over a broad intermediate time interval, whereas $u_{\mathrm{W}}(\tau)$ is a smooth increasing function that emphasizes the later denoising process.

## E Implementation Details

Our timestep-conditioned hierarchical router uses a $37 \to 32 \to 5$ MLP with SiLU activation, adding 1,381 trainable parameters. We set the global spectral loss weight to $0.005$ and each expert's base coefficient to $0.05$. Wavelet losses use a two-level Haar decomposition with level weights $(1, 0.25)$. Each expert receives a minimum routing weight of $0.01$, while the combined energy weight is capped at $0.80$. Expert losses are normalized using detached exponential moving averages with decay $0.99$. The spectral experts and router operate only during training, leaving the inference architecture and parameter count unchanged.