Investigating PDE Residual Attentions in Frequency Space for Diffusion Neural Operators

Medha Sawhney[1]* **Abhilash Neog**[1] **Mridul Khurana**[1] **Arka Daw**[2]

Anuj Karpatne[1]*

[1]Virginia Tech
[2]Oak Ridge National Laboratory

Abstract

Diffusion models for solving partial differential equations (PDEs) are gaining rapid attention, with many approaches using PDE residuals as loss guidance during test-time optimization. While effective under sparse/noisy observations, these frameworks face key limitations such as slow inference, optimization instabilities, and reliance on knowing the noise structure in the observations as a binary mask during inference. To overcome these limitations, we propose PRISMA (PDE Residual Informed Spectral Modulation with Attention), a conditional diffusion neural operator that informs the architecture of diffusion models with PDE residuals via gated attention mechanisms. In contrast to baselines, PRISMA does not require sensitive hyperparameter tuning of loss terms during training or inference, is mask-free, and is aware of the spatial and spectral distributions of PDE residuals. Over an extensive set of four benchmark PDEs with high (97%) noise settings, we show that PRISMA matches or exceeds baseline accuracy while using 10× to 100× fewer denoising steps (20 vs. 200/2000) and achieving 6-7× faster inference speed than state-of-the-art diffusion models.

1 Introduction

A common goal in solving parametric partial differential equations (PDEs) is to learn the *forward problem* of mapping input parameters $\mathbf{a}$ to solution fields $\mathbf{u}$, or the *inverse problem* of inferring $\mathbf{a}$ from observations of $\mathbf{u}$. There is a rich literature on using learning-based methods of solving PDEs including seminal works in *operator learning* such as the Fourier Neural Operator (FNO) [1, 2, 3], which learns mappings between $\mathbf{u}$ and $\mathbf{a}$ in the function space, thereby achieving resolution independence. There is also a growing interest in using **generative models for solving PDEs** using diffusion-based backbones [4, 5, 6, 7, 8, 9, 10, 11, 12]. These models offer two key advantages: (i) they generate full posterior distributions of $\mathbf{u}$ or $\mathbf{a}$, enabling principled ways of quantifying uncertainty (which is especially important in ill-posed inverse problems), and (ii) they naturally accommodate sparse or noisy observations that challenge standard neural operators.

For example, *DiffusionPDE* [4] learns the joint probability distribution of $(\mathbf{a}, \mathbf{u})$ using a diffusion model during training, and employs PDE-residual guidance as a loss term during inference to produce physically consistent solutions. By minimizing PDE residuals during inference, DiffusionPDE can work with arbitrary forms of sparsity in $\mathbf{a}$ and/or $\mathbf{u}$ that has not been seen during training. However, since DiffusionPDE operates directly in the native spatial domain, it remains tied to a fixed spatial resolution of $\mathbf{a}$ and $\mathbf{u}$. To address this limitation, a recent line of work on *FunDPS*

[*] Correspondence to: `medha@vt.edu`, `karpatne@vt.edu`

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

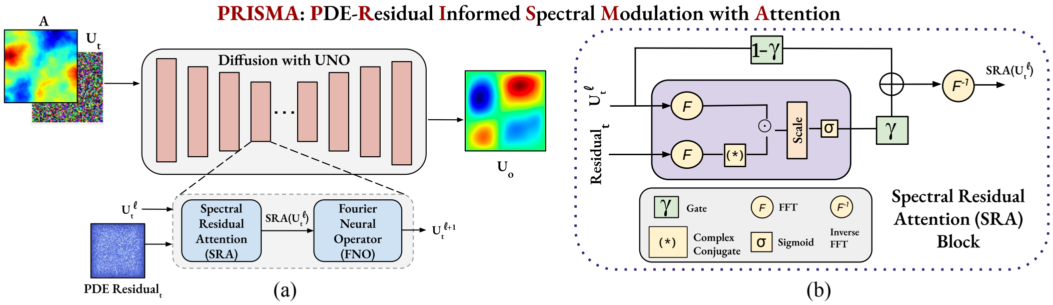

Figure 1: **(a)** Overview of our proposed framework for solving PDEs, PRISMA. At every denoising step, the model takes the current *noisy estimate* u_t (forward PDE case) together with known PDE parameters $\mathbf{a}$ and the PDE residuals, and produces u_{t-1} toward the *clean solution* u_0. **(b)** *Spectral Residual Attention (SRA)* block translates u_t and the residual to the frequency space, measures their alignment to compute frequency-modulated attentions, and applies a *learnable weighted skip connection* to gate the attention before mapping back to the native spatial domain and passing to FNO.

(Function-space Diffusion Posterior Sampling) [6] uses U-shaped neural operators (UNO) [2] as the denoising backbone in diffusion models, leading to a new framework of *diffusion neural operators*. By marrying the strengths of neural operators and diffusion models, FunDPS can generate posterior distributions of $\mathbf{a}$ or $\mathbf{u}$ while being resolution-agnostic.

Despite these advances, current frameworks of diffusion models for solving PDEs suffer from *three major limitations*. First, diffusion models are inherently **slow during inference** because they require sequential denoising over large number of iterations for sampling outputs. To exacerbate this challenge, some methods use PDE-residual guidance during inference, adding to the cost of test-time optimization using expensive PDE loss gradients. Second, a related challenge in the minimization of PDE loss terms during training or inference is the extremely sensitive nature of PDE residuals w.r.t. model parameters, requiring highly tailored routines for tuning gradient descent hyper-parameters to avoid **optimization instabilities**. Third, even though baseline methods such as DiffusionPDE and FunDPS are able to handle arbitrary forms of sparsity in inputs, they assume that the locations of noisy/missing values is known via a binary mask during inference. However, this may not be practical in real-world settings with **unknown structure of noise** in the data due to sensor errors.

To overcome these limitations, we propose **PRISMA** (PDE Residual Informed Spectral Modulation with Attention), a *conditional* diffusion neural operator that takes an entirely different approach of informing the diffusion architecture with PDE residuals via *gated attention mechanisms*, as opposed to loss guidance during training or inference (Figure 1 a). A key innovation in PRISMA is a novel **Spectral Residual Attention (SRA)** block(Figure 1 b) that is injected at every layer of a UNO denoiser backbone, to architecturally inform the denoising process of diffusion models with PDE residuals at both training and inference stages. The SRA block is designed to capture the distribution of PDE residuals in both spectral and spatial domains, including a gating mechanism that uses frequency-modulated attention weights of PDE residuals to steer the denoising process toward physically consistent solutions.

Here are the *key contributions* of **PRISMA** differentiating it from diffusion neural-operator baselines:

- **Fast conditional inference.** By modeling the *conditional* distribution and injecting physics *architecturally* (no expensive test-time optimization), PRISMA attains comparable or better accuracy with **10–100× fewer** denoising steps (20 vs. 200/2000) and runs **6–7× faster** than state-of-the-art diffusion baselines.
- **Free from Sensitive Hyperparameter Tuning.** Instead of using problem-specific PDE-loss weights, PRISMA uses learnable, stepwise residual attention that is recomputed per layer and noise level, mitigating optimization fragility during training or inference.
- **Mask-free Generation.** PRISMA can work even in realistic settings where structure of the noise is *unknown*, since it conditions on available inputs and uses residual feedback to remain physically consistent without binary masks.
- **PDE Residual Structure awareness (Spectral & Spatial).** PRISMA takes into account the distribution of PDE residuals across spatial and spectral domains at every denoising step to impart

fine-grained steerability toward physically consistent solutions. This is fundamentally different from baselines such as DiffusionPDE and FunDPS that aggregate PDE residuals over the entire spatial domain as a scalar loss term.

2 PRISMA: PDE Residual Informed Spectral Modulation with Attention

In the following, we describe PRISMA in the context of solving the forward problem using a conditional diffusion operator (a separate reciprocal model is learned for the inverse case). In particular, given input parameters $\mathbf{a}$, we want to learn a diffusion model that generates $\mathbf{u}$ such that PDE residuals, $\mathcal{R}(\mathbf{a}, \mathbf{u}) = 0$. Let us denote $\mathbf{u}_\sigma = \mathbf{u} + \sigma\boldsymbol{\epsilon}$ as a noisy sample in the forward diffusion process at noise level σ with $\boldsymbol{\epsilon}$ sampled from a Gaussian random field [6]. In the reverse process, a denoiser $D_\theta(\mathbf{u}_\sigma, \sigma, \mathbf{a}, \mathbf{r})$ is then trained to predict $\hat{\mathbf{u}}$ given $\mathbf{u}_\sigma$, where $\mathbf{r} = \mathcal{R}(\hat{\mathbf{u}}, \mathbf{a})$ is the PDE residual. We instantiate D_θ with a custom U-shaped Neural Operator (UNO) [2] backbone that is architecturally informed by PDE residuals as described in Section 2.1. We train D_θ using the Elucidated Diffusion Model (EDM) [13] loss defined as $\mathbb{E}_{u,\sigma}\left[\lambda(\sigma)\left\|D_\theta(\mathbf{u}_\sigma, \sigma, \mathbf{a}, \mathbf{r}) - \mathbf{u}\right\|_2^2\right]$, where λ denotes noise level-specific weighting function. During inference, we solve the reverse diffusion using a decreasing noise schedule, computing the PDE residual $\mathbf{r}$ at each step.

2.1 UNO Backbone of PRISMA

At every layer l of our proposed UNO backbone, we compute $\mathbf{u}^{l+1}$ given $\mathbf{u}^l$ as follows:

$$\mathbf{u}^{l+1} \;=\; \mathcal{F}^{-1}\Big(W^l \odot \mathcal{F}\left(SRA(\mathbf{u}^l)\right)\Big) \;+\; \psi^{\,l}(\mathbf{u}^l), \tag{1}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fast Fourier Transform (FFT) and inverse FFT, respectively, W^l is a learnable spectral weight matrix, $\psi^{\,l}(\mathbf{u}^l)$ is a spatial residual block, and SRA denotes the novel Spatial Residual Attention (SRA) block (described in Section 2.2). Note that Equation 1 follows the structure of a standard FNO except that $\mathbf{u}^l$ is first passed through an SRA block before applying FFT.

2.2 Spectral Residual Attention (SRA) Block

The goal of SRA is to perform cross-attention between PDE residual $\mathbf{r}$ and current solution $\mathbf{u}$ in the *spectral space*, and use these attentions in a gating mechanism to guide the denoising process toward physically consistent solutions. Let us denote $\widetilde{\mathbf{u}}^{\,l} = \mathcal{F}(\mathbf{u}^l)$ and $\widetilde{\mathbf{r}}^{\,l} = \mathcal{F}(\mathbf{r}^l)$ as the FFT outputs of $\mathbf{u}^l$ and $\mathbf{r}^l$, respectively. To compute cross-attentions, we treat $\widetilde{\mathbf{r}}^{\,l}$ as the key and $\widetilde{\mathbf{u}}^{\,l}$ as both the query and value. In particular, we first compute the complex inner product of $\widetilde{\mathbf{u}}^{\,l}$ and $\widetilde{\mathbf{r}}^{\,l}$ across all channels c to obtain compatibility scores S^l, which are then scaled using learnable frequency-specific weights $\boldsymbol{\alpha}$ and passed through a sigmoid activation to obtain soft attentions A^l as follows:

$$S^l \;=\; \frac{1}{\sqrt{C}}\left|\sum\nolimits_{c=1}^{C}\, \widetilde{\mathbf{u}}_c^{\,l}\,\overline{\widetilde{\mathbf{r}}_c^{\,l}}\right|, \qquad A^l(k) \;=\; \sigma\big(\boldsymbol{\alpha}^{\,l}(k)\, S^l(k)\big)\,. \tag{2}$$

where C is the number of channels, $\overline{(\cdot)}$ denotes complex conjugate (to capture phase alignments), and k represents frequency bins. To further make SRA adaptive to the noise level σ and the PDE residuals in the native spatial domain, $\mathbf{r}$, we learn a scalar gate $\gamma^l \in [0, 1]$ using a multi-layer perceptron (MLP) with 2 hidden layers followed by a sigmoid activation, which is then applied to A^l using a weighted skip connection (involving $1 - \gamma$) to obtain the final output $SRA(\mathbf{u}^l)$ as follows:

$$\gamma^l \;=\; \sigma\left(\mathrm{MLP}^{\,l}([\mathbf{r}_{avg}^{\,l},\, c_\sigma])\right), \qquad SRA(u^l) \;=\; \mathcal{F}^{-1}\left(((1-\gamma^l) + \gamma^l A^l)\cdot \widetilde{\mathbf{u}}^{\,l}\right), \tag{3}$$

where $\mathbf{r}_{avg}^l$ captures the average PDE residual across the spatial domain and c_σ is the diffusion noise embedding of σ. When the noise level is high, $\mathbf{r}_{avg}^l$ is typically unreliable and the scalar gate γ^l down-weights PDE residual correction. As the denoising trajectory progresses and the solution becomes cleaner, γ^l increases, allowing stronger PDE-driven modulation through $A^l(k)$.

3 Results and Discussions

We evaluate on four PDE families (Darcy, Poisson, Helmholtz, Navier Stokes) for forward & inverse directions under noisy and full observation regimes. The primary metric is relative ℓ_2 error (%).

	Steps (N)	Darcy Flow		Poisson		Helmholtz		Navier-Stokes	
		Forward	Inverse	Forward	Inverse	Forward	Inverse	Forward	Inverse
FunDPS	200	25.6%	50.7%	72.22%	1474.96%	59.83%	632.39%	32.62%	42.33%
DiffusionPDE	2000	32.6%	47.8%	60.1%	259.23%	88.3%	**152.52%**	88.02%	80.16%
PRISMA (ours)	20	**21.7%**	**42.49%**	**14.17%**	**66.4%**	**23.3%**	195%	**9.9%**	**20.55%**

Table 1: Comparison of different models on four PDE problems with 97% Gaussian noise corruption, simulating real-world measurement noise (in L_2 relative error).

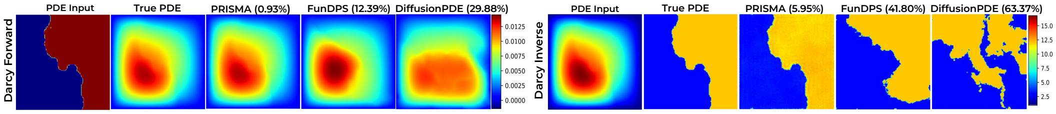

Figure 2: Visualizing Darcy flow predictions for both forward & inverse. Titles report relative ℓ_2 error (%).

Noisy regime (Table 1). Under 97% Gaussian corruption (mask-free), **PRISMA** achieves the lowest relative ℓ_2 error on the majority of forward/inverse benchmarks, reflecting strong robustness to sensor noise. We attribute this to our conditional modeling approach coupled with the SRA block, which prioritizes physics-informative frequency bands while suppressing noise-dominated ones.

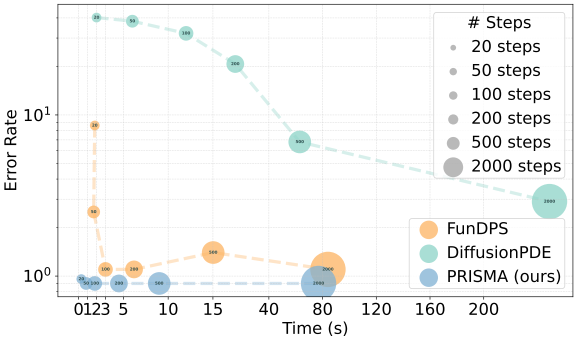

Figure 3: Comparison of PRISMA with baselines, showing comparable/better accuracy with faster sampling

Full observation (Table ??): With clean inputs, PRISMA matches or exceeds diffusion baselines while using fewer sampling steps, and is competitive with strong operator learners on several tasks.

Qualitative (Fig. 2). On *Darcy (full observation)*, PRISMA produces smooth fields with sharp interfaces and markedly cleaner error maps, yielding the lowest relative ℓ_2 error across all baselines.

Efficiency: Fig. 3 shows that PRISMA achieves comparable or better accuracy of SOTA models while requiring **10–100× fewer denoising steps**. In terms of efficiency, it delivers consistently faster inference, running **6–7× faster** than DiffusionPDE & up to **6× faster** than FunDPS across varying denoising steps.

	Steps (N)	**Darcy Flow**		**Poisson**		**Helmholtz**		**Navier-Stokes**	
		Forward	**Inverse**	**Forward**	**Inverse**	**Forward**	**Inverse**	**Forward**	**Inverse**
PINO	–	4%	**2.1%**	3.7%	**10.2%**	4.9%	**4.9%**	**1.1%**	6.8%
DeepONet	–	12.3%	8.4%	14.3%	29%	17.8%	28.1%	25.6%	19.6%
PINNs	–	15.4%	10.1%	16.1%	28.5%	18.1%	29.2%	27.3%	27.8%
FNO	–	5.3%	5.6%	8.2%	13.6%	11.1%	5.0%	2.3%	6.8%
DiffusionPDE*	2000	2.9%	13%	15.27%	21.21%	10.9%	18.97%	2.4%	8.4%
FunDPS	2000	**0.9%**	**2.1%**	–	–	–	–	1.6%	6%
FunDPS*	500	1.4%	3.0%	0.84%	19.84%	**1.08%**	13.88%	3.0%	7.0%
FunDPS*	200	1.1%	4.2%	**0.7%**	23.32%	**1.08%**	18.48%	4.9%	7.8%
FunDPS*	20	8.88%	17.75%	9.755%	39.15%	10.08%	39.39%	13.37%	13.52%
PRISMA (ours)	20	0.958%	4.2%	2.93%	6.4%	2.46%	5.0%	1.52%	**5.02%**

Table 2: Comparing different models on four PDE problems (in L_2 relative error) under Full Observation. * denotes results reproduced using authors' released code and checkpoints. **Best** is boldened, second-best underlined.

4 Conclusion & Future Work

We proposed PRISMA, a conditional diffusion neural operator that injects physics knowledge *inside* the denoiser via spectral & spatial PDE residual attention, yielding competitive accuracy under noisy, mask-free observations while converging in ~ 20 steps, achieving upto $100\times$ lower compute and substantial wall-clock speedups without requiring any test-time posterior optimization. Some future extensions of this work include, (i) extending training to the sparse-mask regime, (ii) exploring optional inference-time use of the PDE residual to further boost accuracy, and (iii) develop a unified backbone that solves forward *and* inverse problems.

References

[1] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

[2] Md Ashiqur Rahman, Zachary E Ross, and Kamyar Azizzadenesheli. U-no: U-shaped neural operators. *arXiv preprint arXiv:2204.11127*, 2022.

[3] Lu Lu, Pengzhan Jin, and George Em Karniadakis. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv preprint arXiv:1910.03193*, 2019.

[4] Jiahe Huang, Guandao Yang, Zichen Wang, and Jeong Joon Park. Diffusionpde: Generative pde-solving under partial observation. *Advances in Neural Information Processing Systems*, 37:130291–130323, 2024.

[5] Sifan Wang, Zehao Dou, Tong-Rui Liu, and Lu Lu. Fundiff: Diffusion models over function spaces for physics-informed generative modeling. *arXiv preprint arXiv:2506.07902*, 2025.

[6] Jiachen Yao, Abbas Mammadov, Julius Berner, Gavin Kerrigan, Jong Chul Ye, Kamyar Azizzadenesheli, and Anima Anandkumar. Guided diffusion sampling on function spaces with applications to pdes. *arXiv preprint arXiv:2505.17004*, 2025.

[7] Jan-Hendrik Bastek, WaiChing Sun, and Dennis M Kochmann. Physics-informed diffusion models. *arXiv preprint arXiv:2403.14404*, 2024.

[8] Edward Li, Zichen Wang, Jiahe Huang, and Jeong Joon Park. Videopde: Unified generative pde solving via video inpainting diffusion models. *arXiv preprint arXiv:2506.13754*, 2025.

[9] Salva Rühling Cachay, Bo Zhao, Hailey Joren, and Rose Yu. Dyffusion: A dynamics-informed diffusion model for spatiotemporal forecasting. *Advances in neural information processing systems*, 36:45259–45287, 2023.

[10] 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.

[11] Jae Hyun Lim, Nikola B Kovachki, Ricardo Baptista, Christopher Beckham, Kamyar Azizzadenesheli, Jean Kossaifi, Vikram Voleti, Jiaming Song, Karsten Kreis, Jan Kautz, et al. Score-based diffusion models in function space. *arXiv preprint arXiv:2302.07400*, 2023.

[12] Sifan Wang, Jacob H Seidman, Shyam Sankaran, Hanwen Wang, George J Pappas, and Paris Perdikaris. Bridging operator learning and conditioned neural fields: A unifying perspective. *arXiv preprint arXiv:2405.13998*, 2024.

[13] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.