FreqMoE: Dynamic Frequency Enhancement for Neural PDE Solvers

Tianyu Chen[1], **Haoyi Zhou**[2], **Ying Li**[3], **Hao Wang**[3], **Zhenzhe Zhang**[3], **Tianchen Zhu**[4], **Shanghang Zhang**[3], **Jianxin Li**[1]
[1]SKLCCSE, School of Computer Science and Engineering, Beihang University, China
[2]School of Software, Beihang University, China
[3]SKLMIP, School of Computer Science, Peking University, China
[4]School of Reliability and Systems Engineering, Beihang University, China
{tianyuc, haoyi, lijx}@buaa.edu.cn

Abstract

Fourier Neural Operators (FNO) have emerged as promising solutions for efficiently solving partial differential equations (PDEs) by learning infinite-dimensional function mappings through frequency domain transformations. However, the sparsity of high-frequency signals limits computational efficiency for high-dimensional inputs, and fixed-pattern truncation often causes high-frequency signal loss, reducing performance in scenarios such as high-resolution inputs or long-term predictions. To address these challenges, we propose FreqMoE, an efficient and progressive training framework that exploits the dependency of high-frequency signals on low-frequency components. The model first learns low-frequency weights and then applies a sparse upward-cycling strategy to construct a mixture of experts (MoE) in the frequency domain, effectively extending the learned weights to high-frequency regions. Experiments on both regular and irregular grid PDEs demonstrate that FreqMoE achieves up to 16.6 percent accuracy improvement while using merely 2.1 percent parameters (47.32x reduction) compared to dense FNO. Furthermore, the approach demonstrates remarkable stability in long-term predictions and generalizes seamlessly to various FNO variants and grid structures, establishing a new Low frequency Pretraining, High frequency Fine-tuning" paradigm for solving PDEs.

1 Introduction

Efficient solutions to large-scale partial differential equations (PDEs) play a crucial role in numerous scientific computing applications, ranging from weather forecasting through Navier-Stokes equations to quantum simulations in physics [Bi *et al.*, 2022; Pathak *et al.*, 2022; Li *et al.*, 2023b; Childs *et al.*, 2021]. As spatial resolution and temporal steps increase, traditional numerical solvers face prohibitive computational costs, spurring the development of neural approaches that promise to balance accuracy with efficiency.

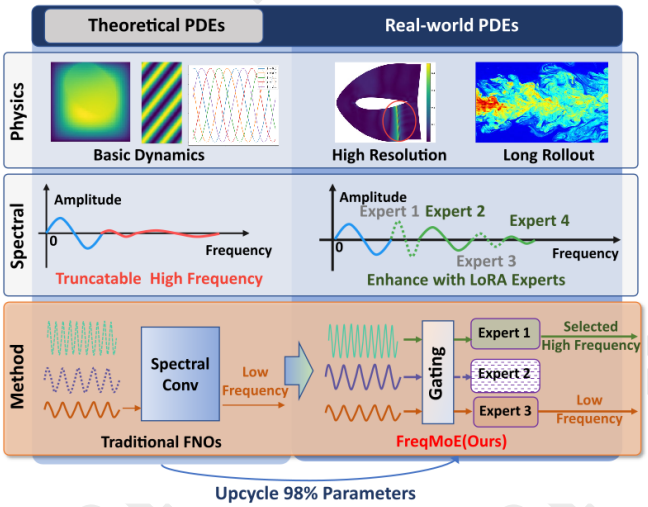

Figure 1: **Motivation of FreqMoE.** Traditional FNO directly truncates high-frequency components (left), while FreqMoE(ours) efficiently preserves them through sparse dynamic experts (right). This design enables high-frequency modeling with negligible computational overhead.

Among these approaches, Fourier neural operators (FNO [Li *et al.*, 2021] and its variants (Geo-FNO [Li *et al.*, 2023a], FFNO [Tran *et al.*, 2023], TFNO [Kossaifi *et al.*, 2023]) have emerged as particularly promising, leveraging the inherent sparsity of physical fields in frequency domain. By operating on a compact window of low-frequency signals while truncating higher frequencies, these methods achieve scale-free processing across arbitrary resolutions with reduced computational complexity. However, this frequency truncation presents a fundamental trade-off: while enabling computational efficiency, the loss of high-frequency information can significantly degrade performance in high-resolution scenarios and, as a result, accumulate errors in long-term predictions [Lippe *et al.*, 2023; Cao *et al.*, 2024].

Recent efforts to overcome high-frequency limitations have explored post-training refinement strategies. These methods aim to recover truncated frequency information through various approaches, such as diffusion-based iterative refinement [Lippe *et al.*, 2023] and numerical solver guid-

[1]The corresponding author is Haoyi Zhou (haoyi@buaa.edu.cn).

ance [Cao *et al.*, 2024]. However, existing solutions often incur substantial computational overhead or are restricted to specific scenarios, highlighting the need for a more computationally efficient approach.

To address these limitations, we propose FreqMoE, a lightweight post-training framework inspired by upcycled MoE Models [He *et al.*, 2024; Zhang *et al.*, 2024]. This framework enables FNO trained on low-frequency domains to adapt to high-frequency signals dynamically. Our method leverages a pre-trained FNO as a base expert for low-frequency components while initializing specialized high-frequency experts(Fig.1). To handle the inherent sparsity of high-frequency components, we incorporate a gating mechanism to selectively activate the most relevant high-frequency Experts during prediction. Our approach is motivated by a fundamental observation in physical systems: high-frequency signals typically exhibit strong dependencies on low-frequency components, as exemplified by the energy cascade phenomenon in fluid mechanics [McKeown *et al.*, 2023]. Capitalizing on this physical insight, we initialize high-frequency experts re-using the base expert's weights through a LoRA-like strategy [Hu *et al.*, 2022].

Specifically, we decompose the high-frequency expert's weights into two components: a shared base weight R_{base} and a low-rank delta weight ΔR. For the i-th high-frequency Expert, its weights are constructed as $R_i = R + \Delta R_i$, where the shared base weights R could be intialized by the pre-trained dense FNO. This architecture offers significant computational efficiency improvement through two key design choices: (1) during post-training, low-rank delta weights are parameter-efficient, and (2) during inference, only the Top-K selected experts participate in prediction. The extremely low-rank nature of ΔR ensures minimal computational overhead in both stages, making our method particularly practical for real-world applications.

Through extensive evaluation on both regular and irregular grid PDEs, we demonstrate the effectiveness of FreqMoE in high-resolution and long-term prediction scenarios. Our experiments reveal compelling advantages: in high-resolution tasks (512×512), FreqMoE achieves up to 16.6% accuracy improvement while using merely 2.1% parameters (47.32× reduction) compared to conventional FNO. This efficiency extends to unstructured meshes, where FreqMoE maintains superior performance with 27.37× parameter reduction. Furthermore, long-term rollout experiments showcase FreqMoE's stability in mitigating error accumulation, particularly in challenging high-resolution scenarios.

The key contributions of this work are threefold:

1. We propose FreqMoE, a lightweight post-training framework that dynamically enhances high-frequency processing capabilities in neural PDE solvers. Our approach generalizes seamlessly across the FNO family on both structured and unstructured grids, establishing an efficient "low-frequency pretraining, high-frequency fine-tuning" paradigm.
2. Inspired by physical principles of frequency dependencies in PDEs, we develop a LoRA-based expert initialization scheme that efficiently reuses low-frequency weights. This design achieves remarkable parameter efficiency (47.32× reduction) while maintaining competitive performance through sparse dynamic computation.
3. Through comprehensive evaluation on diverse PDE systems, we demonstrate that FreqMoE significantly outperforms conventional FNO variants, achieving up to 16.6% accuracy improvement in high-resolution tasks (512×512) and superior stability in long-term predictions, all while maintaining minimal computational overhead.

2 Related Works

2.1 Fourier Neural Operators

Fourier Neural Operators (FNO) [Li *et al.*, 2021] have revolutionized PDE solving by introducing FFT-based spectral convolution layers to learn mappings between infinite-dimensional function spaces. This foundational work has sparked numerous architectural innovations: Geo-FNO [Li *et al.*, 2023a] and SFNO [Bonev *et al.*, 2023] extended the framework to handle irregular grids and spherical geometries, while F-FNO [Tran *et al.*, 2023] enhanced scalability through separable spectral convolutions and advanced training strategies. T-FNO [Kossaifi *et al.*, 2023] further improved parameter efficiency and generalization by implementing global tensor decomposition. Despite these advancements, the issue of high-frequency truncation—a critical limitation in FNO—remains largely unaddressed. Our work directly tackles this gap by enhancing high-frequency signal processing capabilities, offering a complementary approach that integrates seamlessly with existing FNO architectures to further improve performance.

2.2 Sparse Upcycling Techniques

Sparse upcycling has emerged as a powerful paradigm for efficient model enhancement, leveraging sparsely activated Mixture-of-Experts (MoE) initialized from pre-trained dense models. This approach has demonstrated remarkable success across diverse domains, from language models (T5) [Komatsuzaki *et al.*, 2023] to vision-language systems (MoE-LLAVA) [Lin *et al.*, 2024] and medical applications (MoE-Med) [Jiang *et al.*, 2024], consistently outperforming sparse models trained from scratch while significantly reducing computational costs. Building on these principles, our work pioneers the application of sparse upcycling to frequency-domain learning. We introduce a novel framework that utilizes pre-trained FNO to efficiently enhance high-frequency components, achieving improved performance with minimal additional training overhead.

3 Method

FreqMoE extends FNO with dynamic frequency processing through a lightweight expert system. As shown in Fig. 2, our framework splits the frequency spectrum into chunks and processes high-frequency components via specialized experts derived from pre-trained FNO. We first introduce the frequency-domain MoE design (Sec. 3.2), then present our efficient expert initialization scheme (Sec. 3.3), followed by the training

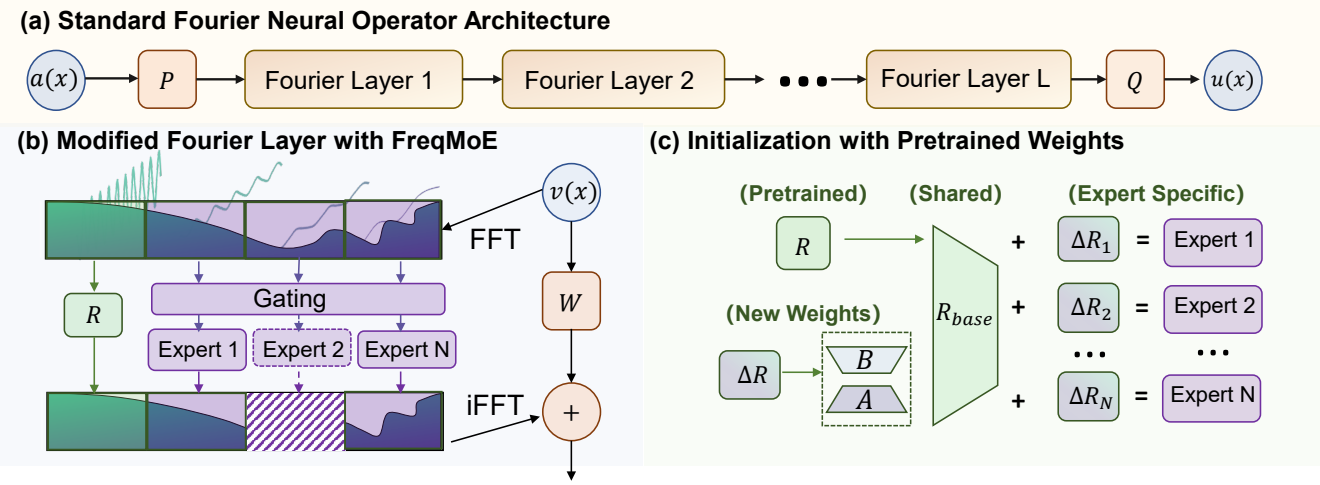

Figure 2: **Methods overview of FreqMoE. (a) The standard Fourier Neural Operator (FNO) architecture** consisting of input lifting (P), a sequence of Fourier layers, and output projection (Q). **(b) Our modified Fourier layer design with a mixture-of-experts mechanism**, where the gating networker dynamically assigns frequency components to specialized experts after FFT decomposition. High-frequency components (lighter shades) are processed by high-frequency experts, while low-frequency components are handled by the base expert. **(c) Our expert initialization strategy**, where pre-trained weights R are used as a shared base component R_{base} and expert-specific delta weights ΔR are initialized with LoRA trick, enabling efficient parameter sharing and specialized frequency processing.

strategy that enables sparse computation. We begin by reviewing the basics of FNO and MoE systems.

3.1 Preliminary

Neural PDE Solvers with FNO. Fourier Neural Operator (FNO) learns a parameterized operator $\mathcal{G}_\theta$ that maps input functions to output solutions in infinite-dimensional spaces. The core of FNO is its Fourier layer(Fig.2(a)), which performs spectral convolution through: $\mathcal{K}^{(l)}(z^{(l)}) = \text{IFFT}(R^{(l)} \cdot \text{FFT}(z))$, where $R^{(l)} \in \mathbb{C}^{H\times H\times M_1\dots M_d}$ are learnable weights operating on truncated frequency modes $\{M_{(i)} | i \in \{1, 2, \dots d\}\}$. This frequency truncation, while computationally efficient, leads to information loss in high-frequency components.

Mixture-of-Experts (MoE). A standard MoE layer consists of a gating network $\mathcal{P}_\theta$ and N expert networks E_{θ_j}, computing outputs as: $\text{MoE}(x) = \sum_{j=1}^{N} \text{TopK}(\text{Softmax}(\mathcal{P}_\theta(x)_j)) \cdot E_\theta^{(j)}(x)$. In our frequency-domain adaptation, experts specialize in different frequency chunks, with the gating network determining the activation of high-frequency computations during inference.

3.2 MoE in Frequency Domain

Traditional FNO truncates high frequencies for efficiency, but this fixed cutoff limits model capacity. Our FreqMoE design addresses this limitation by adaptively processing the frequency spectrum based on two observations: high-frequency signals in PDEs are naturally sparse, and their patterns are often localized. These properties make the frequency domain particularly suitable for expert-based processing.

Frequency Domain Partitioning. In the standard spectral convolution, for an input feature map $z \in \mathbb{R}^S$, its Fourier transform $\hat{z} = \text{FFT}(z) \in \mathbb{C}^S$ is truncated to retain only the lowest frequency bands for processing: $o_P = R_\theta \cdot \hat{z}_P$, where R_θ represents the learnable weights. We generalize this fixed truncation scheme by partitioning the frequency spectrum into $J = S/P$ bands: $\{\hat{z}_P^{(i)} \in \mathbb{C}^P \mid i = 0, 1, \dots, J-1\}$ where bands are ordered by increasing frequency, with $\hat{z}_P^{(0)}$ containing the lowest frequency components.

Expert Specialization. We assign N specialized experts $\{E_{\theta_i} \mid i = 1, \dots, N\}$ ($N \leq J-1$) to process different high-frequency bands, while keeping the original FNO weights R_θ as the base expert for low-frequency components $\hat{z}_P^{(0)}$. This design stems from a key insight in PDE solutions: low frequencies capture global patterns that require careful processing, while high frequencies reflect local details that can benefit from specialized, targeted handling.

Adaptive Frequency Gating. To exploit the natural sparsity in high-frequency signals, we design a gating mechanism g_θ that selectively activates experts based on frequency content: $g_\theta(\hat{z}_P^{(i)}) = \sigma(\frac{w_\theta \cdot \hat{z}_P^{(i)}}{\tau})$ where τ is a temperature parameter and σ is the sigmoid function. The forward computation follows:

$$o_P^{(i)} = \begin{cases} R_\theta(\hat{z}_P^{(i)}), & i = 0 \\ g_\theta(\hat{z}_P^{(i)}) \cdot E_{\theta_i}(\hat{z}_P^{(i)}), & i = 1, \dots, N \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

To encourage sparse expert utilization during training, we add a sparsity loss on gate values:

$$\mathcal{L}_{\text{sparse}} = \mathbb{E}[\sum_{i=1}^{N} g_\theta(\hat{z}_P^{(i)})]. \tag{2}$$

This regularization pushes the model to activate only the

most relevant experts for each frequency band, leading to more efficient inference.

Inference-Time Sparsity. During inference, we leverage the sparse nature of expert utilization by activating only the top-K experts ($K \leq N$) based on their gating values: active_experts $= \text{TopK}(g_\theta(\hat{z}_P^{(i)})i = 1^N, K)$.

The inference computation becomes:

$$o_P^{(i)} = \begin{cases} R_\theta(\hat{z}_P^{(i)}), & i = 0 \\ g_\theta(\hat{z}_P^{(i)}) \cdot E_{\theta_i}(\hat{z}_P^{(i)}), & i \in \text{active\_experts} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

This sparse activation strategy significantly reduces computational overhead while maintaining model performance, as high-frequency components typically require selective rather than comprehensive processing. The final output in the spatial domain is obtained through the inverse Fourier transform: $o = \text{IFFT}(o_{P\ i=0}^{(i)\ J-1})$, where the unprocessed high-frequency components are naturally zero-padded.

3.3 Sparsely Upcycle the Low-frequency Weight

After establishing the expert structure, a key challenge is how to efficiently initialize these experts. Instead of training from scratch, we propose a sparse upcycling strategy (Algorithm 1) that leverages pre-trained FNO weights while keeping the parameter count low. This approach allows experts to inherit low-frequency patterns while developing specialized high-frequency processing capabilities.

Parameter-Efficient Weight Adaptation. For each expert E_{θ_i}, we decompose its adapted weights R_{θ_i} into a shared base component and an expert-specific delta:

$$R_{\theta_i} = R_\theta + \Delta R_{\theta_i}, \quad (4)$$

where $R_\theta \in \mathbb{C}^{H \times H \times M_1 \times \ldots \times M_d}$ represents the pre-trained weights in the low-frequency domain. The expert-specific adaptation ΔR_{θ_i} is computed through low-rank decomposition:

$$\Delta R_{\theta_i} = \alpha \cdot A_{\theta^i} B_{\theta^i} \quad (5)$$

with $A^{(i)} \in \mathbb{C}^{r \times H}$ and $B_{\theta^i} \in \mathbb{C}^{H \times r \times M_1 \times \ldots \times M_d}$ being low-rank adaptation matrices with rank $r \ll H$. This formulation reduces the adaptation parameters from $O(H^2 \prod_i^d M_{(i)})$ to $O(rH(1 + \prod_i^d M_{(i)}))$ per expert.

3.4 Bridging Low and High Frequency Learning

FNO effectively addresses the challenge of learning in infinite function spaces, yet it primarily captures low-frequency patterns as resolution increases. This aligns well with PDE characteristics where dominant features reside in low frequencies, making it efficient to learn fundamental patterns. FreqMoE builds upon this insight by establishing a bridge between low and high frequencies through expert upcycling, enabling pattern transfer across frequency bands.

This design naturally leads to an efficient learning paradigm: *Low-Frequency Pretraining, High-Frequency Fine-tuning(LPHF)*. Since inference over high-resolution PDE solutions is computationally expensive, LPHF allows us to learn core patterns from abundant low-resolution data, then adapt to high frequencies with much fewer parameters. As demonstrated in our experiments on both regular (CFD) and irregular (AirFoil) grids, this paradigm significantly accelerates neural operator inference while maintaining high accuracy across resolutions.

Algorithm 1 Sparsely Upcycling of FNO

Input: Pretrained FNO Model F, number of experts N, rank r, scaling factor α
Output: Upcycled FreqMoE Model

```
// Initialize expert parameters
for each Fourier layer l do
    R_θ^(l) ← F.get_pretrained_weights(l)
    Initialize gating network g_θ^(l)
    for i in 1 to N do
        Initialize expert R_θ^i(l) ← R_θ^(l) + α · A_θ^i(l) B_θ^i(l)
    end for
end for
return New FreqMoE Model F
```

4 Experiments

We conduct systematic evaluations of FreqMoE across two critical scenarios demanding effective high-frequency modeling: high-resolution inputs and long-term prediction rollouts. Our experimental framework follows a progressive approach: (1) Training base FNO models on low-frequency regimes, then (2) transforming them into sparse FreqMoE architectures through our parameter-efficient upcycling strategy (Section 3.3). We measure model effectiveness through activated parameter counts(# Params) and prediction accuracy (L2 relative error), with comprehensive ablation studies on frequency adaptation mechanisms.

4.1 Datasets

To demonstrate FreqMoE's versatility across different PDE domains and discretization schemes, we select benchmark problems from both regular and irregular grid settings.

Regular-grid PDEs. From PDEbench [Takamoto *et al.*, 2022], we choose vortex-dominated flows under Random and Turbulent initializations. Evaluations at 128×128(CFD-Rand 128) and 512×512(CFD-Rand 512, CFD-Turb 512) resolutions test progressive frequency handling capabilities, where higher resolutions reveal finer turbulent structures.

Irregular-grid PDEs. Using Geo-FNO's [Li *et al.*, 2023a] challenging scenarios: (1)*Airfoil*, transonic flows over parameterized NACA-0012 airfoils (Mach 0.8) with shock-induced high frequencies on adapted C-grids (200×50). (2)*Elasticity*, nonlinear material deformations with central voids (radius 0.2-0.4), modeled via 1000 FEM nodes capturing stress concentrations.

4.2 Baseline and Implementation

We evaluate FreqMoE against two strong FNO variants, with implementation details summarized below.

Models	Modes	# Params ↓	Relative L2 Error(L2RE)↓ CFD-Rand 128	CFD-Rand 512	CFD-Turb 512
FNO (Dense)	(4,4)	142.69K	0.0481 ± 0.0061	0.3856 ± 0.0434	0.2445 ± 0.0259
	(16,16)	2.11M	0.0434 ± 0.0052	0.3981 ± 0.0434	0.2164 ± 0.0316
	(32,32)	8.40M	0.0410± 0.0045	0.3742 ± 0.0427	0.2436 ± 0.0267
FreqMoE (Sparse)	(32,32)*	177.53K	0.0404± 0.0047	0.3720 ± 0.0469	0.2320 ± 0.0264
	(4,4)→(32,32)†	177.53K	**0.0370 ± 0.0038**	**0.3122 ± 0.0257**	**0.1934 ± 0.0226**
Params Reduction			↓ **47.32**×		

Table 1: **Performance on Regular-Grid PDEs.** Comparison of models with varying frequency modes, where # Params indicates the number of parameters activated during inference. Underlined values represent the best performance achieved by FNO baselines. Results with blue background show our FreqMoE, where superscript * and † denote models trained from scratch and upcycled from dense FNO, respectively. The **bold** values highlight our best performance.

FNO Baselines. For regular grids, we implement vanilla FNO [Li *et al.*, 2021] with four Fourier layers (width=32) under three spectral configurations: (4,4), (16,16), and (32,32) modes. Inputs include velocity components (Vx, Vy), pressure, and density fields. Training adopts single-step prediction with Adam optimizer (initial lr=0.001).

GeoFNO Baselines. For irregular grids, we extend GeoFNO [Li *et al.*, 2023a] with task-specific designs:(1)*Airfoil*, asymmetric modes (2,4) to (16,32) capture shock waves, using 4 input channels (coordinates + physical fields). (2)*Elasticity*, symmetric modes (2,2) to (16,16) model stress concentrations, enhanced with polar coordinate encoding. Both variants employ the IPHI module for coordinate transformation, trained with 50% learning rate decay every 50 epochs.

FreqMoE Configuration. Our architecture introduces two key innovations: (1)*Dynamic Expert Selection*, expands spectral capacity from (4,4)→(32,32) for regular grids and (2,4)→(16,32)/(2,2)→(16,16) for irregular grids, activating only Top-2 experts during inference. (2) *Upcycling Strategy*, initializes weights from pre-trained base models via low-rank factorization (rank=4), contrasted with scratch training. Training stabilizes via expert sparsity loss (factor $\alpha = 0.1$) with identical hyperparameters to baselines for fair comparison. This design achieves parameter efficiency while preserving high-frequency resolution – critical for our later analyses of activation patterns and long-term stability.

4.3 Comprehensive Evaluation and Insights

Our experiments systematically validate FreqMoE's capabilities through two analytical lenses: (1) The post-training performance improvement in relative L2 error. (2) The inference efficiency improvement via activated parameters reduction. Key findings reveal that FreqMoE achieves superior high-frequency modeling with 6-28× parameter reduction compared to dense counterparts, while maintaining robust performance in all scenarios.

High-Resolution Regular Grid Analysis. Our experiments reveal fundamental limitations in conventional FNO's frequency scaling approach. As shown in Table 1, naively expanding FNO from (4,4) to (32,32) modes yields diminishing returns - while the 32×32 model achieves marginal gains on 128×128 resolution (4.81%→4.10% L2RE), it degrades performance on high-resolution CFD-Turb 512

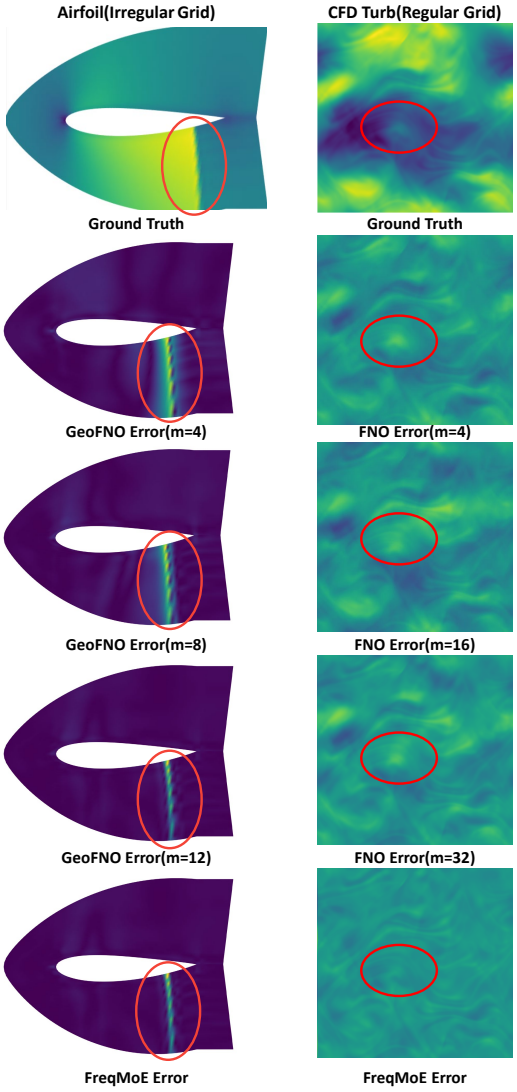

Figure 3: **Visualization of prediction errors. Left Column**: Irregular Grid Results from AirFoil. **Right Column**: Regular Grid Results from CFD-Turb 512. Red circles highlight regions with high-frequency components, where our FreqMoE demonstrates better capability in capturing fine-grained spatial details compared to FNO.

Models	AirFoil Modes	AirFoil # Params↓	AirFoil L2RE↓	Elasticity Modes	Elasticity # Params↓	Elasticity L2RE↓
Geo-FNO (Dense)	(2,4)	74.27k	0.0270 ± 0.0038	(2,2)	49.06K	0.0236 ± 0.0034
	(4,8)	270.88k	0.0161 ± 0.0020	(4,4)	171.94k	0.0386 ± 0.0057
	(8,16)	1.06M	0.0153 ± 0.0016	(8,8)	663.46k	0.0312 ± 0.0037
	(12,24)	2.37M	<u>0.0152</u> ± 0.0016	(12,12)	1.48M	<u>0.0229</u> ± 0.0025
	(16,32)	4.20M	0.0708 ± 0.0100	(16,16)	2.62M	0.0540 ± 0.0067
FreqMoE (Sparse)	(16,32)*	148.06k	0.0432 ± 0.0038	(16,16)*	94.57k	0.0397 ± 0.0046
	(2,4)→(16,32)†	148.06k	**0.0154** ± 0.0013	(2,2)→(16,16)†	94.57k	**0.0217** ± 0.0018
Params Reduction		**↓ 27.37×**			**↓ 26.70×**	

Table 2: **Performance on Irregular-Grid PDEs.** Comparison of models on two representative irregular-grid tasks: AirFoil and Elasticity, where # Params indicates the number of parameters activated during inference. <u>Underlined</u> values represent the best performance achieved by Geo-FNO baselines. Results with blue background show our FreqMoE approach, where superscript * and † denote models trained from

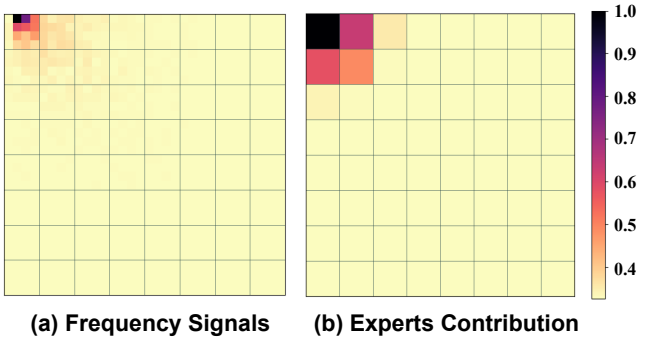

Figure 4: **Visualization of Experts.** (a) Distribution of frequency signals after FFT transformation. (b) Activation patterns of experts in FreqMoE, where each grid cell represents a frequency mode chunk. Beyond capturing low-frequency signals in the top-left corner, FreqMoE dynamically activates experts to capture surrounding high-frequency components.

(24.45%→24.36%) despite 59× parameter growth. This exposes a critical tradeoff: dense spectral models overparameterize high-frequency components that rarely activate in practice.

FreqMoE breaks this tradeoff through dynamic expert specialization. With only 177.53K active parameters (47× fewer than (32,32) FNO), our sparse model reduces errors by 9.8%-16.6% across resolutions. The upcycled variant (4→32 modes) achieves particularly striking improvements: 20.6% error reduction on CFD-Turb 512 compared to its dense counterpart, demonstrating superior turbulence modeling. Spatial error maps in Figure 3 validate this behavior - while dense FNO accumulates errors in vortex cores (red circles), FreqMoE maintains accurate predictions through adaptive frequency allocation.This resolution-aware adaptation explains FreqMoE's dual advantage: preserving low-frequency stability (4.10%→3.70% on CFD-Rand 128) while capturing high-frequency details (24.36%→19.34% on CFD-Turb 512).

Results on Irregular-Grid PDEs. The challenges of irregular grids exacerbate conventional Geo-FNO methods' inefficiency in high-frequency processing. As Table 2 demonstrates, naively expanding Geo-FNO to (16,32) modes for AirFoil catastrophically degrades performance (L2RE surges from 0.0152 to 0.0708) despite 4.2M parameters - revealing dense models' vulnerability to spectral overparameterization. FreqMoE addresses this through sparse high-frequency specialization: with merely 148K parameters (28× fewer than (16,32) modes Geo-FNO), our model achieves near-identical AirFoil accuracy (0.0154 vs 0.0152) while reducing Elasticity errors by 5.2%. This efficiency stems from dynamic frequency enhancement - preserving critical high-frequency components around geometric discontinuities (airfoil edges in Figure 3) without parameter bloat. The 26.7× parameter reduction in Elasticity tasks particularly highlights FreqMoE's advantage in handling stress concentration areas where high-frequency signals dominate.

Sparsely Activation of Experts. As described in Section 3.2, FreqMoE dynamically activates high-frequency experts based on input signals through its gating mechanism. Figure 4 illustrates both the frequency distribution and expert activation patterns. The frequency visualization (Figure 4(a)) reveals that high-frequency components in PDE solutions exhibit natural sparsity, with signal energy primarily concentrated in the low-frequency region (top-left corner). The expert activation map (Figure 4(b)) demonstrates how FreqMoE's gating mechanism responds to this spectral characteristic - while maintaining consistent engagement with low-frequency experts, it selectively activates high-frequency experts only when corresponding signal components are present. This adaptive activation pattern suggests that FreqMoE can effectively identify the sparse high-frequency patterns while preserving computational efficiency through targeted expert utilization.

Analysis of Rollout Performance. The rollout experiments demonstrate FreqMoE's effectiveness in mitigating error accumulation during long-term predictions. In low-resolution scenarios (CFD-Rand 128), all models show relatively stable performance, with FreqMoE maintaining a slight edge in accuracy. However, the advantages of FreqMoE become substantially more evident in high-resolution cases. For CFD-Rand 512, while baseline FNO models exhibit rapid er-

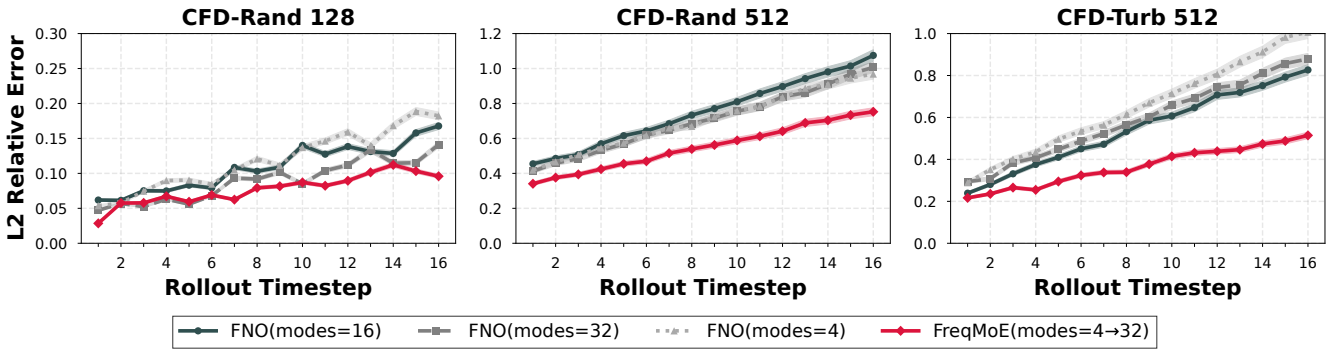

Figure 5: **Long-term Prediction Performance on Different CFD Datasets.** The plots show the L2 relative error evolution during rollout prediction across three datasets of varying complexity. FreqMoE demonstrates superior stability in long-term predictions compared to baseline FNO models with different mode configurations. This advantage becomes particularly pronounced in high-resolution scenarios (CFD-Rand 512 and CFD-Turb 512), where the error growth is significantly moderated.

ror accumulation regardless of their mode numbers, FreqMoE maintains a significantly lower error trajectory, with 31.67% reduction in final prediction error. This pattern is further amplified in the more challenging CFD-Turb 512 dataset, where turbulent flows introduce additional high-frequency components. Here, FreqMoE's adaptive frequency modeling capability proves particularly valuable, effectively containing error growth even as prediction steps extend. This performance gap suggests that FreqMoE's dynamic expert activation successfully preserves critical high-frequency information that traditional FNO models typically lose, thereby preventing the cascade of prediction errors in complex fluid simulations.

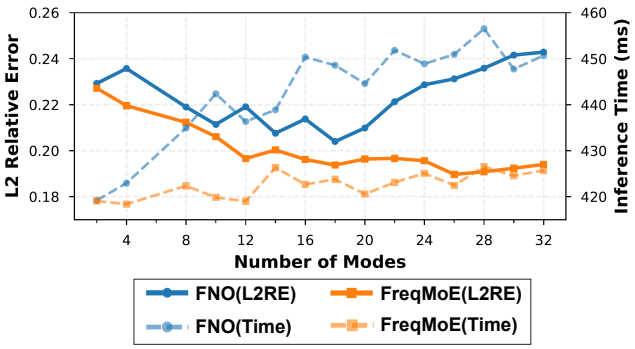

Figure 6: **Performance and Efficiency across different frequency modes.** The solid lines (left y-axis) show the L2 Relative Error (L2RE) achieved by different numbers of modes, while the dashed lines (right y-axis) represent the corresponding inference time measured on a single NVIDIA V100 (32GB) GPU. FreqMoE consistently maintains two active experts (Topk=2) across all modes.

Scale up frequency Modes sparsely vs densely. Figure 6 demonstrates the trade-offs between model performance and computational cost when scaling frequency modes. Dense FNO shows initial error reduction from modes 4 to 12, but experiences performance degradation with higher modes due to the inherent sparsity of frequency signals. FreqMoE maintains steady improvement through dynamic expert selection. The inference time of dense FNO grows quadratically with modes due to full spectral convolution, while FreqMoE achieves linear complexity by fixing active experts (Topk=2), where modes only affect gating computation.

5 Conclusion

We presented FreqMoE, a dynamic frequency enhancement framework that addresses high-frequency signal loss in Fourier Neural Operators through a sparse mixture-of-experts paradigm. Our "Low-Frequency Pretraining, High-Frequency Fine-tuning"(LPHF) strategy efficiently bridges frequency domains while maintaining remarkable parameter efficiency. Key innovations include: (1) a frequency-domain MoE architecture with dynamic expert activation, (2) LoRA-based weight initialization that recycles pretrained FNO weights, and (3) sparse upward-cycling training achieving 47.32× parameter reduction.

While our current approach uses predefined frequency partitioning and introduces minor routing overhead, future work will focus on adaptive partitioning strategies and extending the LPHF paradigm to broader operator learning scenarios. FreqMoE establishes a foundational framework for frequency-aware neural PDE solvers, opening new pathways for efficient high-resolution scientific computing.

Acknowlegements

This work was supported by the National Science and Technology Major Project(No.2022ZD0117800), and Young Elite Scientists Sponsorship Program by CAST(No.2023QNRC001). This work was also sponsored by CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ2023MindSpore12) and developed on openl community. Thanks for the computing infrastructure provided by Beijing Advanced Innovation Center for Big Data and Brain Computing.

References

[Bi *et al.*, 2022] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *CoRR*, abs/2211.02556, 2022.

[Bonev *et al.*, 2023] Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical fourier neural operators: Learning stable dynamics on the sphere. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2806–2823. PMLR, 2023.

[Cao *et al.*, 2024] Shuhao Cao, Francesco Brarda, Ruipeng Li, and Yuanzhe Xi. Spectral-refiner: Fine-tuning of accurate spatiotemporal neural operator for turbulent flows. *CoRR*, abs/2405.17211, 2024.

[Childs *et al.*, 2021] Andrew M. Childs, Jin-Peng Liu, and Aaron Ostrander. High-precision quantum algorithms for partial differential equations. *Quantum*, 5:574, 2021.

[He *et al.*, 2024] Ethan He, Abhinav Khattar, Ryan Prenger, Vijay Korthikanti, Zijie Yan, Tong Liu, Shiqing Fan, Ashwath Aithal, Mohammad Shoeybi, and Bryan Catanzaro. Upcycling large language models into mixture of experts. *CoRR*, abs/2410.07524, 2024.

[Hu *et al.*, 2022] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[Jiang *et al.*, 2024] Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 3843–3860. Association for Computational Linguistics, 2024.

[Komatsuzaki *et al.*, 2023] Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[Kossaifi *et al.*, 2023] Jean Kossaifi, Nikola B. Kovachki, Kamyar Azizzadenesheli, and Anima Anandkumar. Multi-grid tensorized fourier neural operator for high-resolution pdes. *CoRR*, abs/2310.00120, 2023.

[Li *et al.*, 2021] Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[Li *et al.*, 2023a] Zongyi Li, Daniel Zhengyu Huang, Burigede Liu, and Anima Anandkumar. Fourier neural operator with learned deformations for pdes on general geometries. *J. Mach. Learn. Res.*, 24:388:1–388:26, 2023.

[Li *et al.*, 2023b] Zongyi Li, Nikola B. Kovachki, Christopher B. Choy, Boyi Li, Jean Kossaifi, Shourya Prakash Otta, Mohammad Amin Nabian, Maximilian Stadler, Christian Hundt, Kamyar Azizzadenesheli, and Animashree Anandkumar. Geometry-informed neural operator for large-scale 3d pdes. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[Lin *et al.*, 2024] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *CoRR*, abs/2401.15947, 2024.

[Lippe *et al.*, 2023] Phillip Lippe, Bas Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. Pde-refiner: Achieving accurate long rollouts with neural PDE solvers. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[McKeown *et al.*, 2023] Ryan McKeown, Alain Pumir, Shmuel M Rubinstein, Michael P Brenner, and Rodolfo Ostilla-Mónico. Energy transfer and vortex structures: visualizing the incompressible turbulent energy cascade. *New Journal of Physics*, 25(10):103029, 2023.

[Pathak *et al.*, 2022] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *CoRR*, abs/2202.11214, 2022.

[Takamoto *et al.*, 2022] Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. Pdebench: An extensive benchmark for scientific machine learning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *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.

[Tran *et al.*, 2023] Alasdair Tran, Alexander Patrick Mathews, Lexing Xie, and Cheng Soon Ong. Factorized fourier

neural operators. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[Zhang *et al.*, 2024] Qizhen Zhang, Nikolas Gritsch, Dwaraknath Gnaneshwar, Simon Guo, David Cairuz, Bharat Venkitesh, Jakob N. Foerster, Phil Blunsom, Sebastian Ruder, Ahmet Üstün, and Acyr Locatelli. Bam! just like that: Simple and efficient parameter upcycling for mixture of experts. *CoRR*, abs/2408.08274, 2024.