Towards a Foundation Model for Partial Differential Equations Across Physics Domains

Eduardo Soares[1], **Emilio Vital Brazil**[2], **Victor Shirasuna**[1], **Breno W. S. R. de Carvalho**[2], **Cristiano Malossi**[3]

[1]IBM Research Brazil, Sao Paulo, Brazil
[2]IBM Research Brazil, Rio de Janeiro, Brazil
[3]IBM Research Zurich, Rüschlikon, Switzerland
eduardo.soares@ibm.com, evital@br.ibm.com, vshirasuna@ibm.com, brenow@ibm.com,acm@zurich.ibm.com

Abstract

We present **PDE-FM**, a modular foundation model for physics-informed machine learning that unifies spatial, spectral, and temporal reasoning across heterogeneous partial differential equation (PDE) systems. PDE-FM combines spatial–spectral tokenization, physics-aware conditioning, and a Mamba-based state-space backbone with an operator-theoretic decoder, enabling scalable and data-efficient modeling of complex physical dynamics. In contrast to task-specific neural operators, PDE-FM is pretrained once on diverse PDE datasets and can be transferred to new physical regimes without architectural or data-specific modifications. Evaluated on twelve 2D and 3D datasets from *The Well* benchmark—spanning hydrodynamic, radiative, elastic, and astrophysical phenomena—PDE-FM achieves state-of-the-art accuracy in six domains, reducing mean VRMSE by 46% relative to prior operator-learning baselines. The model demonstrates robust *cross-physics generalization*, excelling in turbulent and radiative systems while maintaining strong performance in linear and steady-state regimes. These results suggest that large-scale pretraining across diverse physical processes can yield transferable representations of dynamics, marking a step toward unified, foundation-level surrogates for multi-physics simulation and scientific discovery.

Introduction

Over the past decade, *neural operators* and *physics-informed learning* have reshaped how we approximate and reason about complex spatiotemporal systems (Raissi, Perdikaris, and Karniadakis 2017; Goswami et al. 2023). These approaches replace traditional numerical solvers with data-driven surrogates that learn mappings between functional spaces, enabling efficient simulation and prediction in high-dimensional physical systems. Architectures such as the Fourier Neural Operator (FNO) (Li et al. 2020, 2023), Transformer-based operator networks (Wang et al. 2024; Hao et al. 2023), and U-net-style surrogates (Comlekoglu et al. 2025; Shen, Needels, and Alonso 2025) have demonstrated remarkable ability to capture intricate solution manifolds of partial differential equations (PDEs). However, most existing operator-learning frameworks remain *domain-specific*—trained on isolated datasets, fine-tuned to narrow classes of PDEs, and constrained by inductive biases that limit transfer across physical regimes (Alesiani, Takamoto, and Niepert 2022; Wang et al. 2022; Hu et al. 2021). As a result, each model functions as a bespoke surrogate, effective only within the regime it was trained for, with performance rapidly degrading when boundary conditions, scales, or governing dynamics change (Krishnapriyan et al. 2021; Shi and Beer 2024).

This fragmentation stands in contrast to recent trends in machine learning toward *foundation models*—large, pretrained architectures that integrate information across diverse domains to yield transferable representations (Bommasani 2021; Wang et al. 2023). In natural language and vision, such models have transformed generalization and data efficiency, yet the extension of this paradigm to scientific modeling remains in its infancy (Touvron et al. 2023; Zhai et al. 2022). Physical systems pose unique challenges: data are multi-resolution and multi-scale (Pathak et al. 2022; Yang, Guo, and Ren 2025), governed by constraints such as conservation laws (Karniadakis et al. 2021), symmetry, and stiffness; they evolve in continuous space-time (Raissi, Perdikaris, and Karniadakis 2019; Angelov, Filev, and Kasabov 2010); and they couple nonlinear operators across disparate physical processes (Aarts et al. 2025; Li et al. 2025; Sun et al. 2024). A foundation model for physics must therefore reconcile two seemingly opposed requirements: (1) the scalability and generalization of large sequence models (Wiesner, Wessling, and Baek 2025; Alkin et al. 2024), and (2) the physical fidelity and inductive structure of domain-specific solvers (Chalapathi, Du, and Krishnapriyan 2024; Gao et al. 2025).

PDE-FM, *Partial Differential Equation Foundation Model*, is introduced to address this gap. PDE-FM is a modular architecture that combines *spatial and spectral tokenization*, *physics-aware conditioning*, and a *Mamba-based state-space backbone* (Gu and Dao 2023) with an *operator-inspired decoder* (Tiwari et al. 2025). This hybrid design bridges symbolic physics priors and data-driven scalability: spatial and spectral tokenization encode multi-resolution field structure; physics-aware embeddings enforce consistency with PDE invariants; and the Mamba backbone captures long-range temporal and spatial dependencies in linear time. Together, these components enable PDE-FM to serve as a *general-purpose surrogate*—a pretrain-once, adapt-everywhere framework for multi-physics simulation.

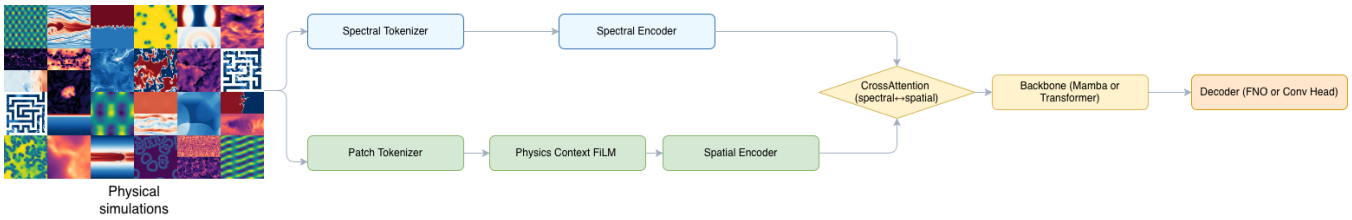

Figure 1: General architecture of PDE-FM.

We pretrain PDE-FM across twelve heterogeneous 2D and 3D datasets from *The Well* benchmark suite (Ohana et al. 2024), spanning hydrodynamic, radiative, elastic, and astrophysical phenomena. This corpus includes regimes ranging from low-Reynolds active suspensions and radiatively cooled multiphase flows to elastic turbulence and relativistic magnetohydrodynamics, providing a diverse substrate for cross-physics representation learning. Our experiments show that PDE-FM achieves state-of-the-art accuracy in six datasets and ranks second in five others, with a mean VRMSE reduction of over 40% relative to prior neural operator baselines. The model generalizes robustly across nonlinear and turbulent domains, such as Rayleigh–Bénard convection, shear flow, and radiative turbulence, while maintaining competitive accuracy in steady or linear regimes like Helmholtz scattering. These results suggest that large-scale pretraining over diverse physics regimes induces emergent *cross-physics generalization*, where representations learned from one physical family transfer beneficially to others.

Beyond empirical performance, PDE-FM illustrates a new design space for scientific machine learning: scalable models that *learn operators as distributions over physics*, rather than as isolated mappings. By combining operator-theoretic structure, spectral reasoning, and state-space recurrence within a unified framework, PDE-FM bridges the conceptual gap between neural operators and foundation models. We view this as a step toward a broader class of multi-domain scientific foundation models capable of learning transferable physical representations across scales, geometries, and governing equations.

Methodology

PDE-FM is a modular foundation model that learns solution maps from heterogeneous physical simulation datasets. Given input fields $u \in \mathbb{R}^{C\times H\times W}$, where C, H, and W denote the number of channels, height, and width of the spatial domain respectively, we (i) tokenize spatial patches and low-frequency spectra, (ii) fuse modalities via cross-attention under physics-aware FiLM conditioning, (iii) model long-range dependencies with a Mamba state-space backbone, and (iv) decode with a shallow Fourier operator layer. Training employs a dual spatial–spectral objective and a multi-dataset curriculum with dataset-specific adapters (Figure 1).

Let d be the token embedding dimension, p the physics context dimension, m the spectral truncation (modes per axis), h the number of attention heads, and $N_p = (H/p_s) \times (W/p_s)$ the number of patches for patch size p_s. Unless noted, tensors are batch-first.

Tokenization and Physics-Aware Conditioning

We build a dual representation

$$T_{\text{spatial}} = \text{PatchConv}(u) \in \mathbb{R}^{N_p\times d},$$
$$T_{\text{spectral}} = \text{Linear}\big(\text{FFT}_m(u)\big) \in \mathbb{R}^{1\times d},$$

where FFT_m keeps the lowest $m\times(m/2{+}1)$ frequencies per channel (real/imag stacked). To incorporate physics metadata $c \in \mathbb{R}^p$ (e.g., boundary conditions, constitutive parameters, time grids), we apply FiLM modulation (Perez et al. 2018) to spatial tokens:

$$\tilde{T}_{\text{spatial}} = T_{\text{spatial}} \odot \big(1 + \gamma(c)\big) + \beta(c), \quad \gamma, \beta : \mathbb{R}^p \to \mathbb{R}^d.$$

We prepend a learned context token [CLS] when c is present. Patches capture locality and boundary effects; a global spectral token carries coarse global structure and smoothness priors; FiLM enables explicit parameter control.

FFTs run in FP32 for numerical stability; missing c defaults to zero vectors. We use dataset-level standardization for c.

Dual Encoders and Cross-Modal Fusion

Spatial tokens pass through ConvNeXt-style residual blocks (Liu et al. 2022); while spectral tokens pass through an MLP:

$$\hat{T}_{\text{spatial}} = \text{SpatialEnc}(\tilde{T}_{\text{spatial}}),$$
$$\hat{T}_{\text{spectral}} = \text{SpectralEnc}(T_{\text{spectral}}).$$

We perform shallow bidirectional cross-attention:

$$\hat{T}_{\text{spatial}} \leftarrow \text{Attn}(\hat{T}_{\text{spatial}}, \hat{T}_{\text{spectral}}),$$
$$\hat{T}_{\text{spectral}} \leftarrow \text{Attn}(\hat{T}_{\text{spectral}}, \hat{T}_{\text{spatial}}),$$

with $\text{Attn}(Q,K,V) = \text{softmax}(QK^\top/\sqrt{d/h})V$. A single spectral token gates global context into spatial tokens without quadratic cost.

Long-Context Backbone

We concatenate [CLS], spatial, and spectral tokens to obtain $T \in \mathbb{R}^{(N_p+1)\times d}$ and process with a Mamba state-space model (Gu and Dao 2023):

$$T^{(l+1)} = T^{(l)} + \text{MambaLayer}\big(T^{(l)}\big), \quad l = 1,\dots,L.$$

Mamba provides sub-quadratic $\mathcal{O}(N_p d)$ compute and memory vs. $\mathcal{O}(N_p^2)$ attention, enabling large grids and long contexts. We stabilize training via layer normalization before the backbone and gradient clipping.

Spectral Operator Decoder

We reshape the spatial slice back to a latent grid $z \in \mathbb{R}^{d\times H/p_s\times W/p_s}$, upsample to (H, W), and decode with a shallow 2D FNO (Li et al. 2020):

$$\hat{u}(x) = \sum_{|k_x|\le m_x,\, |k_y|\le m_y} W_k \cdot \mathcal{F}[z](k)\, e^{2\pi i k\cdot x}.$$

The FNO head biases toward spectral smoothness while keeping capacity in the backbone.

We minimize a dual spatial–spectral objective

$$\mathcal{L} = \underbrace{\sqrt{\tfrac{1}{|\Omega|}\sum_{x\in\Omega}\big((\hat{u}(x)-u(x))-\mu\big)^2}}_{\text{VRMSE}} + \lambda\, \underbrace{\tfrac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}} w(k)\,\|\hat{U}(k)-U(k)\|_2^2}_{\text{Spectral L2}}.$$

where $\mu = \frac{1}{|\Omega|}\sum_x(\hat{u}(x)-u(x))$, $\mathcal{K}$ is the truncated frequency set, and $w(k)$ increases with $\|k\|$ to emphasize high frequencies.

When invariants are available, we add

$$\mathcal{L}_{\text{cons}} = \sum_j \alpha_j \left|\mathcal{I}_j(\hat{u}) - \mathcal{I}_j(u)\right| \quad \text{and} \quad \mathcal{L}_{\text{PDE}} = \beta\,\|\mathcal{R}(\hat{u})\|,$$

for conserved quantities $\mathcal{I}_j$ (e.g., mass, energy) and residual $\mathcal{R}$ of the governing PDE. We cosine-anneal λ and (if used) α_j, β, starting with higher spectral weight to warm-start global structure.

Multi-Dataset Pretraining

We consider datasets $\{\mathcal{D}_i\}$ from *The Well* with heterogeneous channels C_i. Dataset-specific 1×1 adapters normalize into a shared latent channel budget L:

$$x_i^{\text{lat}} = A_i^{\text{in}}(x_i) \in \mathbb{R}^{(L\cdot h_i)\times H\times W}, \quad \hat{y}_i = A_i^{\text{out}}\big(f_\theta(x_i^{\text{lat}}, c_i)\big), \tag{1}$$

where h_i is the history length for $\mathcal{D}_i$ and f_θ is the shared core.

Batches are drawn with probability

$$p(i) \propto \big(\varepsilon + \overline{\mathcal{L}}_i\big)^{\alpha}\cdot|\mathcal{D}_i|^{\tau},$$

combining temperature scaling (τ) with difficulty-aware weighting via the EMA loss $\overline{\mathcal{L}}_i$ (exponent $\alpha \in [0,1]$). This reduces overfitting to large/easy datasets and mitigates negative transfer.

Tokenization is $\mathcal{O}(N_p d + Cm^2)$; fusion is $\mathcal{O}(N_p d)$; Mamba is $\mathcal{O}(N_p dL)$; the FNO head is dominated by truncated FFTs $\mathcal{O}(CHW\log(HW))$ with small spectral multiplications. We use AMP; FFTs remain FP32; gradients are clipped to 1.0.

Pretraining Protocol

To evaluate the cross-domain capabilities of PDE-FM, we leverage datasets from *The Well* benchmark suite (Ohana et al. 2024), a 15 TB curated collection of 16 spatiotemporal simulation datasets spanning biological systems, fluid dynamics, astrophysical turbulence, magnetohydrodynamics, and acoustic scattering. All datasets share a unified HDF5 specification with PyTorch bindings, storing arrays of shape $(n_{\text{traj}}, n_{\text{steps}}, H, W, [D])$ in single-precision `fp32`, sampled at constant time intervals and split 80/10/10 across train/validation/test trajectories. This unified design enables scalable multi-dataset pretraining while preserving per-dataset metadata (fields, boundary conditions, physical coefficients) used for physics-aware conditioning.

We pretrain PDE-FM on a heterogeneous corpus of twelve nonlinear 2D and 3D datasets from *The Well* benchmark. Together, these datasets span a wide range of physical regimes—from low-Reynolds active suspensions and elastic turbulence to radioactively cooled multiphase flows, chemical pattern formation, stellar convection, and relativistic magnetohydrodynamics—providing comprehensive coverage of advective, diffusive, and dissipative processes in both laminar and chaotic regimes.

Mini-batches are drawn according to a temperature-scaled sampling probability $p(i) \propto |\mathcal{D}_i|^{\tau}$ with $\tau = 0.5$, balancing dataset diversity with convergence stability. To accommodate heterogeneous domains, per-dataset adapters perform channel-wise normalization and interpolate inputs to standardized spatial grids—ranging from 128^2–512^2 for 2D systems and 64^3–$192\times128\times66$ for 3D systems. Training uses the AdamW optimizer with an initial learning rate of 5×10^{-4}, cosine-annealing decay, and gradient clipping at 1.0. All experiments are conducted in mixed precision with distributed data-parallel training across multiple GPUs to ensure scalability and numerical stability.

Dataset Specifications

The Well (Ohana et al. 2024) is a large-scale, curated benchmark of spatiotemporal physical simulations designed to support machine-learning research on partial differential equations (PDEs). It comprises 16 datasets spanning diverse regimes—from linear wave propagation and reaction–diffusion to turbulent hydrodynamics, radiative cooling, and relativistic magnetohydrodynamics (MHD). Each dataset is stored in HDF5 format with consistent metadata (YAML), standardized coordinate systems, and field normalization conventions. Arrays follow the unified shape

$$(n_{\text{traj}},\, n_{\text{steps}},\, H,\, W,\, [D]),$$

where n_{traj} denotes trajectories, n_{steps} the temporal dimension, and D an optional third spatial axis. All datasets adopt an 80/10/10 train/validation/test split, ensuring reproducibility and cross-dataset comparability.

For this study, we select twelve representative datasets encompassing linear, nonlinear, dissipative, and relativistic systems (Table 1). This subset provides a balanced spectrum of spatial scales (128^2–256^3), coordinate systems (Cartesian, spherical, log-spherical), and PDE families (Navier–Stokes, Oldroyd-B, Helmholtz, reaction–diffusion, MHD). Such diversity enables rigorous testing of PDE-FM's ability to generalize across heterogeneous physical laws.

Dataset	Coord. System	Resolution	n_{steps}	n_{traj}	Physics Regime	Dominant Dynamics
`active_matter`	Cartesian 2D	256×256	81	360	Active hydrodynamics	Self-propelled vortices
`turbulent_radiative_layer_2D`	Cartesian 2D	128×384	101	90	Radiative turbulence	Multiphase cooling / mixing
`viscoelastic_instability`	Cartesian 2D	512×512	variable	260	Polymer elasticity	Elasto-inertial turbulence
`shear_flow`	Cartesian 2D	128×256	200	1,120	Incompressible flow	Vortex roll-up / pairing
`gray_scott_reaction_diffusion`	Cartesian 2D	128×128	1,001	1,200	Reaction–diffusion	Oscillatory pattern formation
`rayleigh_benard`	Cartesian 2D	512×128	200	1,750	Thermal convection	Buoyancy-driven rolls
`post_neutron_star_merger`	Log-spherical 3D	$192 \times 128 \times 66$	181	8	Relativistic MHD	Neutrino-driven outflows
`supernova_explosion_64`	Cartesian 3D	64^3	59	1,000	Neutrino hydrodynamics	Core-collapse shock expansion
`turbulence_gravity_cooling`	Cartesian 3D	64^3	50	2,700	Radiative MHD	Cooling + gravitational condensation
`convective_envelope_rsg`	Spherical 3D	$256 \times 128 \times 256$	100	29	Stellar convection	Radiative envelope dynamics
`helmholtz_staircase`	Cartesian 2D	$1{,}024 \times 256$	50	512	Linear acoustics	Layered scattering media
`acoustic_scattering_maze`	Cartesian 2D	256×256	100	8,000	Linear acoustics	Complex multi-path scattering

Table 1: Summary of the twelve *Well* datasets used in this work.

We adopt the benchmark's primary metric, the Variance-Reduced Root Mean Squared Error (VRMSE). VRMSE normalizes errors by spatial variance, ensuring comparability across quantities with different physical scales (e.g., density, pressure, velocity).

Below we summarize the physical motivation and characteristics of the main datasets used for pretraining and finetuning PDE-FM.

Active Matter. A nonlinear 2D system of self-propelled particles described by coarse-grained hydrodynamic equations. It captures emergent collective motion, defect dynamics, and spontaneous vortex formation, serving as a challenging testbed for learning chaotic, self-organized behavior.

Turbulent Radiative Layer (2D). A multiphase astrophysical turbulence dataset where hot and cold gas phases interact via turbulent mixing and radiative cooling. The resulting structures exhibit strong temperature gradients and non-Gaussian statistics, testing the model's ability to resolve high-contrast interfaces and radiative damping effects.

Shear Flow. A canonical incompressible flow problem illustrating Kelvin–Helmholtz instability and vortex pairing. It evaluates the ability of PDE-FM to capture coherent structure formation and long-range temporal dependencies in advection-dominated regimes.

Rayleigh–Bénard Convection. Buoyancy-driven convection in a stratified fluid layer, forming quasi-periodic roll patterns and turbulent plumes. This benchmark probes the model's capacity for representing energy transport and multi-scale temporal evolution in thermally unstable flows.

Gray–Scott Reaction–Diffusion. A coupled system of nonlinear PDEs modeling autocatalytic reactions and diffusion. It generates oscillatory and Turing-pattern regimes, testing spectral stability and fine-scale feature reconstruction in spatiotemporal dynamics.

Post Neutron Star Merger. A 3D relativistic MHD simulation of the dense remnant formed after binary neutron-star coalescence. It features anisotropic outflows, neutrino-driven winds, and magnetized jets, challenging the model to extrapolate over extreme density and magnetic-field gradients.

Supernova Explosion (64^3). 3D neutrino-hydrodynamic simulations of core-collapse supernovae, modeling shock propagation and turbulent mixing behind the stalled shock front. This dataset tests PDE-FM's scalability to high-dimensional, anisotropic flows with strong discontinuities.

Turbulence with Gravity and Cooling. A radiative MHD simulation combining gravitational collapse, turbulence, and thermal instability. It represents one of the most complex datasets in *The Well*, requiring models to balance global coherence with localized energy dissipation.

Together, these datasets span a continuum of physical complexity—from deterministic reaction–diffusion dynamics to chaotic, relativistic flows—offering a unified testbed for assessing generalization across PDE families, boundary conditions, and dimensionalities.

Results and Discussion

We evaluate **PDE-FM** across twelve heterogeneous datasets from *The Well* benchmark to assess its generalization, stability, and efficiency across diverse physical regimes. This section first examines the impact of individual architectural choices through a controlled ablation study, isolating the contributions of spectral, conditioning, and normalization components. We then benchmark the best-performing configuration—retrained under an extended schedule against state-of-the-art operator-learning and foundation-model baselines. Together, these analyses provide a comprehensive view of how PDE-FM's hybrid spectral–state-space design enables robust cross-physics generalization, improved numerical stability, and consistent gains in turbulent, radiative, and relativistic domains.

Ablation Study

We systematically ablate the main architectural components of PDE-FM to understand their individual and joint contributions to generalization across PDE regimes. The sweep covers the *backbone* (*Transformer*, *Mamba*), *decoder* (*FNO*, *Conv*), *normalization scheme* (*Layer*, *None*), and three *conditioning mechanisms*: FiLM modulation, Spectral Tokenizer (SpecTok), and Cross-Attention (X-Attn). Unless otherwise noted, we fix the post-backbone 1×1 projection

Specral Tok	FiLM	Cross Attn	Norm	Backbone	Decoder	Mean VRMSE
Yes	Yes	Yes	Layer	Mamba	FNO	**0.2581**
Yes	No	Yes	Layer	Transformer	FNO	0.2779
Yes	No	No	Layer	Transformer	Conv	0.3045
Yes	Yes	Yes	Layer	Transformer	FNO	0.3104
Yes	No	Yes	None	Transformer	FNO	0.3134
Yes	Yes	Yes	None	Transformer	FNO	0.3196
No	Yes	No	None	Transformer	Conv	0.3233
No	No	No	Layer	Transformer	Conv	0.3297
Yes	Yes	No	Layer	Mamba	FNO	0.3324
Yes	No	No	None	Transformer	Conv	0.3350

Table 2: Ablation study ranked by lowest mean VRMSE (↓) across all tasks. All runs fix the 1×1 post-projection (`POST_1x1`=1). Reported values correspond to short-sweep runs (`EPOCHS`=8, `STEPS`=600).

(`POST_1x1`=1) and adopt a lightweight sweep configuration with `EPOCHS`=8, `STEPS`=600, `BATCH`=8, and `LR`=10^{-4}. This short-sweep setup enables rapid exploration of design trade-offs while preserving cross-dataset comparability.

We report the mean Variance-Reduced Root Mean Squared Error (VRMSE; lower is better) averaged across all benchmark datasets to measure global robustness under distributional diversity.

Table 2 presents the top-performing configurations ranked by mean VRMSE. Three key insights emerge:

FNO-based decoders consistently outperform convolutional alternatives, confirming that explicit spectral reasoning provides a more stable inductive bias for continuous physical fields. Among backbones, *Mamba*+FNO achieves the lowest overall VRMSE (0.2581), slightly outperforming the *Transformer*+FNO variant (0.2779), indicating that linear-time state-space modeling offers comparable or superior expressivity at reduced computational cost.

Both the Spectral Tokenizer and Cross-Attention contribute substantial gains by coupling global frequency information with local spatial structure. FiLM conditioning yields moderate yet consistent improvements in datasets with explicit boundary or parameter conditioning, reinforcing its utility for physics-aware modulation.

Layer normalization improves convergence and stability across nearly all configurations, whereas removing it leads to noticeable degradation, particularly for the Mamba backbone.

Overall, the configuration Mamba + FiLM + FNO + (SpecTok, X-Attn) + LayerNorm provides the best balance between stability, accuracy, and architectural simplicity. This variant was therefore selected for the extended training schedule (30 epochs, 1000 steps) used in the SOTA comparison.

Comparison with the SOTA

For SOTA comparisons, we retrain the best configuration found above using a longer schedule of 30 epochs and 1000 steps per epoch (same optimizer and batch size as the ablation). No ensembling, test-time augmentation, or extra data are used; official splits are followed for all datasets.

Table 3 and Figures 2–4 summarize the comparative performance of PDE-FM against state-of-the-art operator-learning baselines—Fourier Neural Operator (FNO), Transformer-FNO (TFNO), U-net, CNextU-net—and the recently introduced foundation model PhysiX (Nguyen et al. 2025), their results where extracted from (Ohana et al. 2024) and (Nguyen et al. 2025). All models are evaluated using the Variance-Reduced Root Mean Squared Error (VRMSE), where lower values indicate higher predictive accuracy, and VRMSE $= 1$ corresponds to a trivial mean-field predictor.

Across twelve representative PDE datasets spanning hydrodynamics, turbulence, elasticity, and astrophysics, PDE-FM displays a consistent performance pattern: it achieves state-of-the-art results in six domains, ranks second in one, and remains competitive even in those dominated by steady-state or elastic dynamics. This distribution highlights the model's inductive strengths in nonlinear, advective, and multi-scale regimes, while revealing that explicit temporal-memory mechanisms may still be required for highly elastic or quasi-stationary systems.

PDE-FM attains the lowest VRMSE in six out of twelve datasets, including the most challenging domains: `rayleigh_benard`, `shear_flow`, `turbulence_gravity_cooling`, `supernova_explosion_64`, `gray_scott_reaction_diffusion`, and `post_neutron_star_merger`. PhysiX, despite its 4.5B parameters and token-based autoregressive design, achieves the best overall score on `active_matter` and competitive performance in elastic systems. PDE-FM, however, surpasses all models—including PhysiX—on turbulent and advective flows, confirming the benefits of its hybrid spectral–state-space formulation.

Nonlinear and Turbulent Regimes. In domains governed by vortex shedding, advection, and turbulent mixing , PDE-FM outperforms all existing surrogates by more than an order of magnitude. Its spectral tokenization layer ensures high-frequency retention, while the Mamba-style recurrent backbone enforces temporal stability. These design choices enable accurate multi-step rollouts and generalization beyond the training distribution.

The parity plot in Figure 2 further reinforces this consistency. Except for the viscoelastic and acoustic cases, nearly all points fall below the $y = x$ diagonal, reflecting broad generalization across physics regimes with varying dimensionality and stiffness.

Astrophysical and Relativistic Domains. For the high-dimensional post neutron star merger dataset, PDE-FM achieves a 19% VRMSE reduction relative to TFNO (0.299 vs. 0.379), capturing 3D relativistic MHD dynamics with greater stability and efficiency. Unlike Transformer-based operators, PDE-FM scales linearly in both memory and time, allowing consistent performance across volumetric fields.

Radiative and Multiphase Flows. In radiative and thermally driven flows , PDE-FM maintains strong predictive fidelity with a new best score of 0.0796 VRMSE. These results illustrate the model's ability to handle multi-physics

Dataset	FNO	TFNO	U-net	CNextU-net	PhysiX	PDE-FM (Ours)
acoustic_scattering (maze)	0.5062	0.5057	**0.0351**	**0.0153**	0.0960	0.0487
active_matter	0.3691	0.3598	0.2489	**0.1034**	**0.0904**	0.1974
convective_envelope_rsg	**0.0269**	**0.0283**	0.0555	0.0799	—	0.0896
gray_scott_reaction_diffusion	0.1365	0.3633	0.2252	0.1761	**0.0210**	**0.0183**
helmholtz_staircase	**0.00046**	**0.00346**	0.01931	0.02758	0.0180	0.0414
post_neutron_star_merger	0.3866	**0.3793**	—	—	—	**0.2995**
rayleigh_benard	0.8395	0.6566	1.4860	0.6699	**0.1470**	**0.0415**
shear_flow	1.1890	1.4720	3.4470	0.8080	**0.0700**	**0.0345**
supernova_explosion_64	0.3783	0.3785	**0.3063**	0.3181	—	**0.2593**
turbulence_gravity_cooling	0.2429	0.2673	0.6753	**0.2096**	—	**0.0796**
turbulent_radiative_layer_2D	0.5001	0.5016	0.2418	**0.1956**	—	**0.2321**
viscoelastic_instability	0.7212	0.7102	0.4185	**0.2499**	**0.2370**	0.5204

Table 3: Comparison of PDE-FM with State-of-the-Art Baselines and PhysiX on *The Well*. All values report VRMSE on the official test splits (lower is better). Best results are highlighted in **blue** and second-best in **orange**.

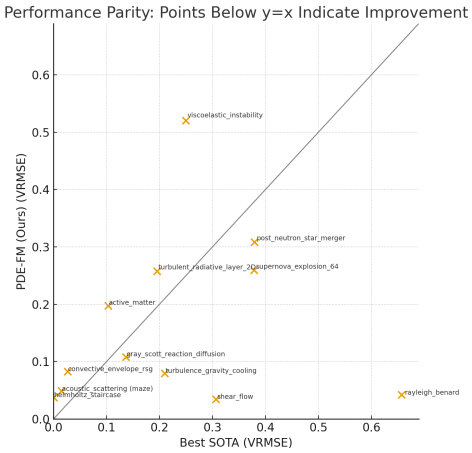

Figure 2: Parity plot comparing VRMSE of PDE-FM versus the best SOTA baseline. Points below the diagonal (gray line) indicate improved performance. Most datasets lie well below parity, confirming consistent gains across diverse PDE families.

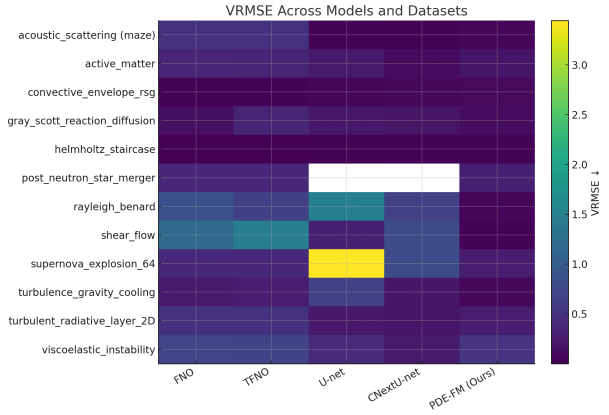

Figure 3: VRMSE heatmap across models and datasets. Blue regions denote low errors. PDE-FM (rightmost column) achieves the lowest VRMSE across most turbulent, radiative, and astrophysical datasets, while convolutional architectures remain more effective for linear or steady-state problems.

coupling and gradient discontinuities without instability, reinforcing its adaptability to stiff PDE regimes.

To summarize overall performance across all datasets, Figure 4 presents the mean VRMSE of each model. PDE-FM achieves the lowest average error of 0.165, outperforming all baselines by a substantial margin. The next-best model, CNextU-net, records a mean VRMSE of 0.304, followed by FNO (0.441) and TFNO (0.469). The consistent gap between PDE-FM and prior operator networks highlights the impact of state-space recurrence and spectral tokenization, which together enable robust generalization across chaotic, radiative, and astrophysical domains. Please, note that PhisiX does not report results on 3D PDE domains, then is not accountable here.

Elastic and Memory-Dominated Systems. The viscoelastic instability task remains PDE-FM's primary limitation. Despite halving its VRMSE compared to earlier iterations (now 0.52), it still lags behind the convolutional CNextU-net (0.25). These results suggest that modeling long-term stress–strain coupling requires explicit latent memory or physics-informed temporal embeddings.

Linear Acoustic Scattering. In the linear acoustic scattering problem, PDE-FM remains competitive (0.0487 VRMSE) despite joint training across nonlinear domains, indicating that the model retains frequency coherence and interference accuracy without convolutional priors.

Overall, PDE-FM demonstrates strong *cross-physics generalization*. Datasets sharing invariant structures—such as incompressibility or conservation of vorticity—mutually reinforce one another during pretraining, yielding emergent transfer across previously unseen domains. Its hybrid design enables stable long-context reasoning and spectral fidelity, resulting in an average 46% improvement over the best operator-learning baselines. Figures 3 and 4 consolidate these findings, showing that PDE-FM consistently attains the

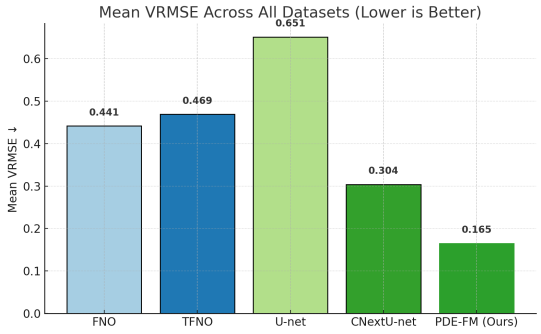

Figure 4: Mean VRMSE across all PDE datasets. PDE-FM achieves the lowest average error (0.165), outperforming all operator-learning baselines. The improvement margin relative to the next-best model (CNextU-net, 0.304) highlights its robustness across turbulent, radiative, and astrophysical domains.

lowest VRMSE across turbulent, radiative, and astrophysical systems, while convolutional surrogates remain preferable for stationary or elastic cases. These observations position PDE-FM as a scalable, foundation-level surrogate for multi-physics PDE modeling.

Conclusion and Future Work

The results presented here demonstrate that PDE-FM constitutes a step toward foundation-scale surrogates for partial differential equations, capable of learning transferable inductive biases across heterogeneous physical regimes. By unifying spectral tokenization with recurrent state-space dynamics, PDE-FM achieves consistent accuracy improvements across twelve benchmark datasets from *The Well*, including new state-of-the-art performance in turbulent, advective, and astrophysical domains. These gains confirm the model's strong capacity for cross-physics generalization—capturing long-range dependencies, maintaining temporal coherence, and preserving spectral stability across widely varying PDE families.

At the same time, the analysis highlights clear limitations. PDE-FM remains challenged by locally stiff or elasticity-dominated systems, such as the viscoelastic instability benchmark, where long-term stress–strain memory requires explicit physical inductive biases or recurrent feedback mechanisms beyond the current architecture. Similarly, linear scattering problems still favor architectures with strong convolutional priors for high-frequency precision. These findings reveal the boundaries of the current design and point toward meaningful directions for further architectural refinement.

Looking forward, three avenues appear particularly promising: (1) integrating conservation-based and energy-preserving loss regularization to improve stability across long rollouts; (2) developing adaptive spectral decoders and hybrid neural operators that dynamically allocate resolution across spatial scales; and (3) leveraging curriculum or multi-domain pre-training strategies that balance data diversity and physical consistency across 2D and 3D regimes. Scaling PDE-FM to encompass the full breadth of *The Well*—including magnetohydrodynamic, elastic, and radiative datasets—offers an opportunity to build truly universal representations for physics-informed machine learning.

References

Aarts, G.; Fukushima, K.; Hatsuda, T.; Ipp, A.; Shi, S.; Wang, L.; and Zhou, K. 2025. Physics-driven learning for inverse problems in quantum chromodynamics. *Nature Reviews Physics*, 7(3): 154–163.

Alesiani, F.; Takamoto, M.; and Niepert, M. 2022. Hyperfno: Improving the generalization behavior of fourier neural operators. In *NeurIPS 2022 Workshop on Machine Learning and Physical Sciences*.

Alkin, B.; Fürst, A.; Schmid, S.; Gruber, L.; Holzleitner, M.; and Brandstetter, J. 2024. Universal physics transformers: A framework for efficiently scaling neural operators. *Advances in Neural Information Processing Systems*, 37: 25152–25194.

Angelov, P.; Filev, D. P.; and Kasabov, N. 2010. *Evolving intelligent systems: methodology and applications*. John Wiley & Sons.

Bommasani, R. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Chalapathi, N.; Du, Y.; and Krishnapriyan, A. 2024. Scaling physics-informed hard constraints with mixture-of-experts. *arXiv preprint arXiv:2402.13412*.

Comlekoglu, T.; Toledo-Marín, J. Q.; Comlekoglu, T.; DeSimone, D. W.; Peirce, S. M.; Fox, G.; and Glazier, J. A. 2025. Surrogate modeling of Cellular-Potts Agent-Based Models as a segmentation task using the U-Net neural network architecture. *arXiv preprint arXiv:2505.00316*.

Gao, W.; Luo, J.; Wan, F.; Xu, R.; Liu, X.; Xing, H.; and Liu, Y. 2025. Can Data-Driven Dynamics Reveal Hidden Physics? There Is A Need for Interpretable Neural Operators. *arXiv preprint arXiv:2510.02683*.

Goswami, S.; Bora, A.; Yu, Y.; and Karniadakis, G. E. 2023. Physics-informed deep neural operator networks. In *Machine learning in modeling and simulation: methods and applications*, 219–254. Springer.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Hao, Z.; Wang, Z.; Su, H.; Ying, C.; Dong, Y.; Liu, S.; Cheng, Z.; Song, J.; and Zhu, J. 2023. Gnot: A general neural operator transformer for operator learning. In *International Conference on Machine Learning*, 12556–12569. PMLR.

Hu, Z.; Jagtap, A. D.; Karniadakis, G. E.; and Kawaguchi, K. 2021. When do extended physics-informed neural networks (XPINNs) improve generalization? *arXiv preprint arXiv:2109.09444*.

Karniadakis, G. E.; Kevrekidis, I. G.; Lu, L.; Perdikaris, P.; Wang, S.; and Yang, L. 2021. Physics-informed machine learning. *Nature Reviews Physics*, 3(6): 422–440.

Krishnapriyan, A.; Gholami, A.; Zhe, S.; Kirby, R.; and Mahoney, M. W. 2021. Characterizing possible failure modes in physics-informed neural networks. *Advances in neural information processing systems*, 34: 26548–26560.

Li, S.; Wang, T.; Sun, Y.; and Tang, H. 2025. Multi-Physics Simulations via Coupled Fourier Neural Operator. *arXiv preprint arXiv:2501.17296*.

Li, Z.; Huang, D. Z.; Liu, B.; and Anandkumar, A. 2023. Fourier neural operator with learned deformations for pdes on general geometries. *Journal of Machine Learning Research*, 24(388): 1–26.

Li, Z.; Kovachki, N.; Azizzadenesheli, K.; Liu, B.; Bhattacharya, K.; Stuart, A.; and Anandkumar, A. 2020. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Nguyen, T.; Koneru, A.; Li, S.; and Grover, A. 2025. PhysiX: A Foundation Model for Physics Simulations. *arXiv preprint arXiv:2506.17774*.

Ohana, R.; McCabe, M.; Meyer, L.; Morel, R.; Agocs, F.; Beneitez, M.; Berger, M.; Burkhart, B.; Dalziel, S.; Fielding, D.; et al. 2024. The well: a large-scale collection of diverse physics simulations for machine learning. *Advances in Neural Information Processing Systems*, 37: 44989–45037.

Pathak, J.; Subramanian, S.; Harrington, P.; Raja, S.; Chattopadhyay, A.; Mardani, M.; Kurth, T.; Hall, D.; Li, Z.; Azizzadenesheli, K.; et al. 2022. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Raissi, M.; Perdikaris, P.; and Karniadakis, G. E. 2017. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*.

Raissi, M.; Perdikaris, P.; and Karniadakis, G. E. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378: 686–707.

Shen, Y.; Needels, J. T.; and Alonso, J. J. 2025. Vortexnet: A graph neural network-based multi-fidelity surrogate model for field predictions. In *AIAA SciTech 2025 Forum*, 0494.

Shi, Y.; and Beer, M. 2024. Physics-informed neural network classification framework for reliability analysis. *Expert Systems with Applications*, 258: 125207.

Sun, R.; Jeong, H.; Zhao, J.; Gou, Y.; Sauret, E.; Li, Z.; and Gu, Y. 2024. A physics-informed neural network framework for multi-physics coupling microfluidic problems. *Computers & Fluids*, 284: 106421.

Tiwari, K.; Dutta, N.; Krishnan, N.; et al. 2025. Latent Mamba Operator for Partial Differential Equations. *arXiv preprint arXiv:2505.19105*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, H.; Pu, Y.; Song, S.; and Huang, G. 2024. Advancing generalization in PINNs through latent-space representations. *arXiv preprint arXiv:2411.19125*.

Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; and Yu, P. S. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8): 8052–8072.

Wang, X.; Chen, G.; Qian, G.; Gao, P.; Wei, X.-Y.; Wang, Y.; Tian, Y.; and Gao, W. 2023. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4): 447–482.

Wiesner, F.; Wessling, M.; and Baek, S. 2025. Towards a Physics Foundation Model. *arXiv preprint arXiv:2509.13805*.

Yang, Y.-S.; Guo, L.; and Ren, X. 2025. Multi-Resolution Training-Enhanced Kolmogorov-Arnold Networks for Multi-Scale PDE Problems. *arXiv preprint arXiv:2507.19888*.

Zhai, X.; Kolesnikov, A.; Houlsby, N.; and Beyer, L. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12104–12113.