ASNO: An Interpretable Attention-Based Spatio-Temporal Neural Operator for Robust Scientific Machine Learning

Anonymous Author [*1]

Abstract

Scientific machine learning (SciML) aims to model complex physical processes from data, but a key challenge lies in capturing long-range spatio-temporal dependencies while generalizing across varying environmental conditions. This is especially critical in real-world applications such as additive manufacturing, where machine parameters are controlled but environmental factors fluctuate unpredictably. Traditional models often learn temporal dynamics but lack physical interpretability and robustness to distributional shifts. To address these limitations, we introduce the Attention-based Spatio-Temporal Neural Operator (ASNO), a novel architecture that decouples temporal and spatial modeling using separable attention mechanisms. ASNO is inspired by the implicit–explicit decomposition of the backward differentiation formula (BDF): it employs a Transformer encoder to forecast homogeneous temporal dynamics and a nonlocal attention-based neural operator (NAO) to integrate spatial interactions and external forcing. This design enhances interpretability by isolating contributions of historical states and external fields, while enabling zero-shot generalization to new physical regimes. Experiments on standard SciML benchmarks—including chaotic ODEs, PDEs, and additive manufacturing—demonstrate that ASNO consistently outperforms baseline models in accuracy, stability, and generalizability, making it a promising framework for interpretable and adaptive physics-informed learning.

1. Introduction

Foundation models (FMs) have rapidly advanced scientific machine learning (SciML), enabling breakthroughs in modeling complex spatio-temporal phenomena across fluid dynamics, climate science, and materials engineering (Zou et al., 2024). These models promise scalable, general-purpose reasoning capabilities, but their deployment in high-stakes, physically grounded domains demands more than raw accuracy (Ellington et al., 2024; Luo et al., 2023). To be truly impactful and trustworthy, scientific FMs must exhibit robustness to distributional shifts, interpretability of internal mechanisms, and generalizability to evolving or unseen physical regimes (Huang et al., 2025; Zhang et al., 2024). As FMs begin to influence decision-making pipelines in safety-critical settings—such as additive manufacturing or autonomous experimentation—there is growing concern over their opaque behavior and limited adaptability, echoing broader societal concerns around responsible AI (Rudin et al., 2022; Molnar, 2020).

While neural operators like the Fourier Neural Operator (FNO) (Li et al., 2020) and DeepONet (Lu et al., 2021) have shown promise in learning solution maps between function spaces, they often fail to generalize in the presence of variable environmental parameters, boundary conditions, or non-stationary forcing terms. Similarly, spatio-temporal transformers (Vaswani et al., 2017; Zhou et al., 2021) capture long-range dependencies but entangle physical structure with data-driven correlations, making their outputs difficult to interpret or trust. These challenges motivate a shift in perspective: instead of optimizing only for predictive performance, we need models that are interpretable by design, modular in architecture, and reflective of physical principles—especially when deployed in real-world scientific systems with partially observable dynamics.

To this end, we propose the *Attention-based Spatio-Temporal Neural Operator* (ASNO), a novel architecture grounded in classical numerical methods—specifically, the backward differentiation formula (BDF Fredebeul (1998)). ASNO decouples temporal extrapolation and spatial correction via a Transformer encoder (Zhou et al., 2021) and a Nonlocal Attention Operator (NAO) (Yu et al., 2024), respectively. This decomposition mirrors the implicit-explicit (IMEX) structure of traditional solvers, offering an interpretable mechanism for separating historical dynamics from external influences. By aligning with physical simulation

[*]Equal contribution [1]Anonymous University. Correspondence to: Anonymous Author <>.

Published at ICML 2025 Workshop on Reliable and Responsible Foundation Models.

structure and enabling zero-shot generalization to unseen system states, ASNO directly addresses key issues of robustness, diagnosis, and responsible generalization. Our results on benchmark systems—including additive manufacturing, PDEs, and chaotic ODEs—demonstrate that ASNO delivers state-of-the-art performance while offering transparency and adaptability, supporting its potential as a reliable foundation model for scientific applications.

2. Background and Related Work

Predicting evolving physical systems across space and time is a central challenge in scientific machine learning. A key limitation of existing models is their inability to simultaneously capture long-range dependencies and adapt to unseen environmental parameters. In this section, we provide an overview of two major methodological pillars that influence our proposed approach: transformer-based temporal modeling and operator-based methods for learning function-to-function mappings.

2.1. Temporal Modeling with Attention Mechanisms

Transformers have shown promise in modeling long-range temporal dependencies due to their self-attention mechanism, which can dynamically focus on relevant parts of a sequence (Zhou et al., 2023). Originally introduced for natural language processing, attention-based models like the Transformer (Vaswani et al., 2017) have since been adapted for time-series forecasting in various domains including energy systems, finance, and weather prediction. Architectures such as Informer (Zhou et al., 2021), Reformer (Kitaev et al., 2020), and Temporal Fusion Transformers (TFT) (Lim et al., 2021) have addressed the computational bottlenecks of vanilla transformers by introducing sparse attention, memory mechanisms, and gating modules.

Despite these innovations, most of these models process time as a univariate or multivariate sequence and fail to account for the spatial structure inherent in physical systems. This becomes problematic in settings governed by partial differential equations (PDEs), where temporal evolution depends intricately on local and nonlocal spatial interactions. Moreover, while attention improves model interpretability by providing visibility into influential timesteps, it typically lacks a principled connection to physical laws, making the insights learned from attention scores less physically grounded.

2.2. Neural Operators for Learning Physical Dynamics

Neural operators have emerged as a promising framework for learning mappings between function spaces, with applications in solving forward and inverse problems governed by PDEs. The Fourier Neural Operator (FNO) (Li et al., 2020) introduced the idea of learning a resolution-invariant operator in the Fourier domain, enabling fast predictions over continuous spatial domains. DeepONet (Lu et al., 2021) provided an alternative formulation by learning the operator through a decomposition into branch and trunk networks. These approaches have inspired a range of operator learning methods aimed at improving scalability, flexibility, and generalization.

However, most neural operator models assume static system dynamics during inference and struggle to generalize across PDEs with varying coefficients, initial conditions, or boundary inputs. Recent extensions such as GNOT (Hao et al., 2023) and Transolver (Wu et al., 2024) have incorporated attention-based mechanisms and geometric priors to improve performance on heterogeneous or unstructured domains. Yet, they often treat spatial and temporal dependencies jointly, without explicitly isolating their individual effects. This fusion can obscure causal relationships in the dynamics and hinder zero-shot generalization to new physical regimes. Furthermore, many of these models are optimized for forward prediction but are not designed to uncover underlying physical structure or handle inverse inference tasks in a unified framework.

In contrast, the proposed ASNO architecture draws on both temporal and operator-based learning paradigms, but introduces a critical separation between temporal extrapolation and spatial correction. Inspired by the implicit-explicit decomposition in backward differentiation formula (BDF) schemes, ASNO utilizes a transformer encoder to extrapolate the system's evolution based on prior temporal history, while a nonlocal attention operator captures the spatial interplay of variables and external forcings. This design enables interpretable and generalizable learning in dynamic systems, as explored in the following sections.

3. Model Architecture

This section presents ASNO, a spatio-temporal neural operator integrating the Transformer Encoder for temporal dependencies and the nonlocal attention mechanism for spatial interactions. We outline the Transformer Encoder, followed by the nonlocal attention operator model, and explain their combined role in the form of a backward differentiation formula (BDF) for complex spatio-temporal prediction.

3.1. Backward Differentiation Formula (BDF)

Backward differentiation formula (BDF) are a family of popular numerical schemes for stiff differential equations, thanks to its high-order accuracy and large region of absolute stability (Fredebeul, 1998). For a initial-valued differential equation:

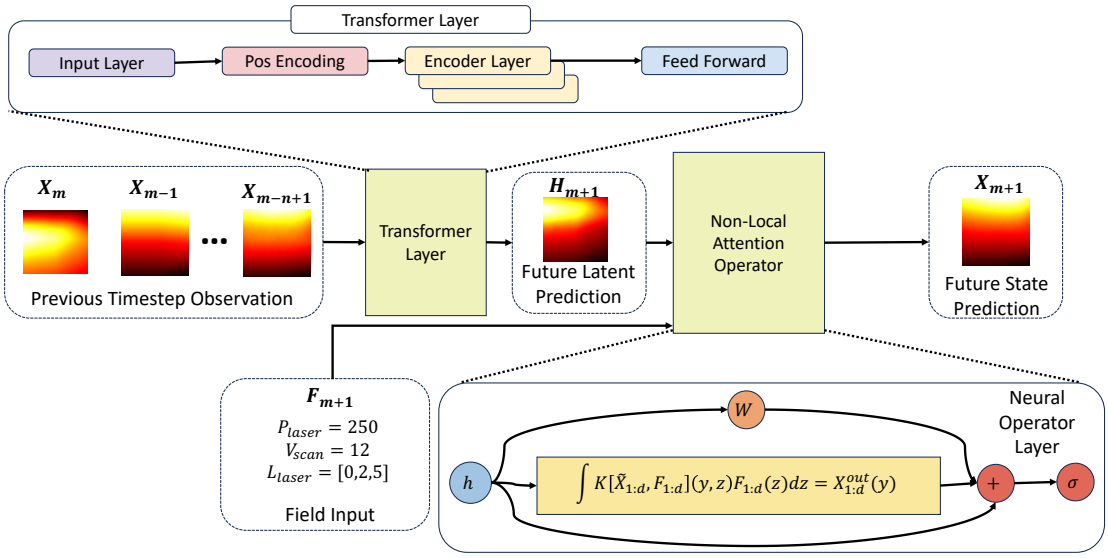

Figure 1. ASNO architecture: a Transformer Encoder captures temporal extrapolation $\tilde{X}_{m+1}$, while the NAO performs spatial correction via a learned attention-based operator.

$$\dot{X}(t) = F(t, X(t)), \quad X(t_0) = X_0, \tag{1}$$

The general n-step BDF scheme approximates the solution at discrete timesteps $t_m = t_0 + m\Delta t$ as:

$$\sum_{k=0}^{n} \alpha_k X_{m-k} = \Delta t \cdot \beta F(t_{m+1}, X_{m+1}), \tag{2}$$

where $\alpha_0 = 1$ and the coefficients α_k, β depend on the order. The left-hand side uses historical states, while the right-hand side evaluates the dynamics at the future state X_{m+1}, allowing stable integration of stiff systems.

This formulation can be decomposed into an implicit-explicit (IMEX) scheme (Ascher et al., 1995):

$$\tilde{X}_{m+1} = -\sum_{k=1}^{n} \alpha_k X_{m-k+1}, \tag{3}$$

$$X_{m+1} = \tilde{X}_{m+1} + \Delta t \cdot \beta F(t_{m+1}, X_{m+1}). \tag{4}$$

The first step extrapolates the next state assuming zero forcing ($F \equiv 0$), capturing temporal trends. The second step corrects it using spatial dynamics and external effects, often requiring nonlinear solution.

This motivates the ASNO architecture: a Transformer Encoder models the temporal forecast $\tilde{X}_{m+1}$, and a Nonlocal Attention Operator (NAO) refines it to produce X_{m+1}, reflecting the IMEX decomposition.

We consider S dynamical systems indexed by $\eta = 1, \ldots, S$, each defined by:

$$\dot{X}^{(\eta)}(t) = F(S^{(\eta)}(t), X^{(\eta)}(t)), \quad X^{(\eta)}(t_0) = X_0^{(\eta)}, \tag{5}$$

where $S^{(\eta)}(t)$ represents hidden environmental states. The dataset for each system is:

$$\mathcal{D}^{(\eta)} = \{X^{(\eta)}(m\Delta t), F^{(\eta)}(m\Delta t)\}_{m=0}^{T/\Delta t}. \tag{6}$$

ASNO is trained to: (1) learn a stable temporal extrapolation rule from past states, (2) infer and adapt to hidden system conditions via the spatial operator, and (3) generalize to new systems in a zero-shot setting, using only initial state and forcing observations. This structure enables ASNO to model diverse physical systems with strong generalization and interpretability.

3.2. Temporal Module: Transformer Encoder

The temporal module in ASNO models long-range temporal dependencies using a Transformer Encoder, which corresponds to the explicit step in the BDF formulation (Yang et al., 2024; Vaswani et al., 2017). Given n past solution fields $\{X_{m-n+1}, \ldots, X_m\}$, each $X_i \in \mathbb{R}^{N \times d}$ is projected into a latent embedding space via:

$$E_i = X_i W_E + P_i, \tag{7}$$

where $W_E \in \mathbb{R}^{d \times d_{\text{embed}}}$ is a trainable embedding matrix, and $P_i \in \mathbb{R}^{d_{\text{embed}}}$ encodes temporal position to retain ordering across timesteps. The embedded sequence $\{E_{m-n+1}, \ldots, E_m\}$ is passed through L layers of a Transformer Encoder, each composed of multi-head self-attention followed by a feed-forward network and residual connections. This allows the model to extract temporal interactions over distant steps while preserving gradient flow.

In each attention layer, the query, key, and value matrices are computed as:

$$Q = EW_{Tq}, \quad K = EW_{Tk}, \quad V = EW_{Tv}, \tag{8}$$

with $W_{Tq}, W_{Tk}, W_{Tv} \in \mathbb{R}^{d_{\text{embed}} \times d_t}$ being learnable parameters for attention. Attention scores between time steps i and j are computed via scaled dot-product attention:

$$\alpha_{ij} = \frac{\exp\left(Q_i K_j^\top / \sqrt{d_t}\right)}{\sum_k \exp\left(Q_i K_k^\top / \sqrt{d_t}\right)}. \tag{9}$$

These scores determine how much each previous timestep contributes to the current representation. The latent forecast vector H_{m+1} is then computed as the attention-weighted sum:

$$H_{m+1} = \sum_{j=1}^{n} \alpha_{m,j} V_j. \tag{10}$$

This latent representation $H_{m+1} \in \mathbb{R}^{d_{\text{embed}}}$ captures a temporally integrated forecast, aligning with the homogeneous extrapolated state $\tilde{X}_{m+1}$ in the BDF scheme. It is subsequently refined by the Nonlocal Attention Operator (NAO) in the implicit step to yield the final prediction X_{m+1}. This design allows ASNO to isolate the temporal component of the dynamics while deferring external and spatial corrections to the spatial module.

3.3. Spatial Module: Nonlocal Attention Operator (NAO)

To complete the prediction of the next system state X_{m+1}, we employ the Nonlocal Attention Operator (NAO) to process the latent extrapolation H_{m+1} (from the temporal encoder) and the external forcing field F_{m+1} (Yu et al., 2024). This step approximates the spatial operator governing physical interactions across the domain. Specifically, we model the target prediction as:

$$X_{m+1} = \text{NAO}(H_{m+1}, F_{m+1}). \tag{11}$$

This formulation reflects a general solution operator for implicit updates in dynamic PDE systems, approximating a nonlocal integral operator of the form:

$$X_{m+1}(x) = \int_\Omega K(x, x')\, F_{m+1}(x')\, dx', \tag{12}$$

where $K(x, x')$ is a learned, data-dependent kernel mapping spatial locations x and x'. We discretize the latent and forcing fields across N spatial points, forming:

$$\begin{aligned} H_{1:d} &= (H_j(y_k))_{1 \le j \le d,\, 1 \le k \le N}, \\ F_{1:d} &= (F_j(y_k))_{1 \le j \le d,\, 1 \le k \le N}, \end{aligned} \tag{13}$$

with d representing the feature dimension. The initial state is given by:

$$J_0 = (H_{1:d}, F_{1:d}). \tag{14}$$

This state is iteratively refined using attention layers with residual connections for T steps:

$$J_t = \text{Attn}(J_{t-1}; \theta_t)\, J_{t-1} + J_{t-1} := (J_t, F_t), \quad 1 \le t \le T, \tag{15}$$

where the attention mechanism is defined as:

$$\text{Attn}(J; \theta_t) = \sigma\left(\frac{1}{\sqrt{d_k}}\, J\, W_{Q_t}\, (W_{K_t})^\top\, J^\top\right), \tag{16}$$

with trainable projection matrices $W_{Q_t}, W_{K_t} \in \mathbb{R}^{2d \times d_k}$, and σ denoting a linear or ReLU activation function.

After T steps, the NAO computes the final nonlocal kernel via dual kernel maps based on the final hidden states:

$$\begin{aligned} K[H_{1:d}, F_{1:d}; \theta] = W_{P,h}\, \sigma\left(\frac{1}{\sqrt{d_k}} J_T^\top W_{Q_{T+1}} (W_{K_{T+1}})^\top J_T\right) \\ + W_{P,f}\, \sigma\left(\frac{1}{\sqrt{d_k}} F_T^\top W_{Q_{T+1}} (W_{K_{T+1}})^\top J_T\right), \end{aligned} \tag{17}$$

where $W_{P,h}, W_{P,f} \in \mathbb{R}^{d_k \times d}$ are output projections controlling the contribution of the latent and forcing channels.

The predicted next state is then computed by applying the learned kernel to the forcing input:

$$X_{m+1}(y) = \int_\Omega K[H_{1:d}, F_{1:d}](y, z)\, F_{1:d}(z)\, dz. \tag{18}$$

The NAO thus implements a data-driven spatial operator capable of capturing long-range, nonlocal dependencies across the spatial domain. Its ability to dynamically adapt its

kernel based on system state and external loadings enables ASNO to generalize across PDEs with varying dynamics, geometry, and boundary conditions—offering a robust and interpretable component for reliable scientific prediction.

3.4. Overall Formulation

ASNO integrates temporal extrapolation and spatial correction by combining the Transformer Encoder and the Nonlocal Attention Operator (NAO). Given the past n timesteps $\{X_{m-n+1}, \ldots, X_m\}$, the Transformer Encoder extracts a latent representation:

$$H_{m+1} = \text{Transformer}(X_m, X_{m-1}, \ldots, X_{m-n+1}), \tag{19}$$

which encodes the system's homogeneous temporal dynamics, analogous to the BDF approximation $\tilde{X}_{m+1}$. This latent state is then corrected using the external forcing field F_{m+1} via NAO:

$$X_{m+1} = \text{NAO}(H_{m+1}, F_{m+1}). \tag{20}$$

Together, this yields the full update rule:

$$\begin{aligned} X_{m+1} &= \text{ASNO}(X_m, \ldots, X_{m-n+1}, F_{m+1}) \\ &= \text{NAO}(\text{Transformer}(\{X\}), F_{m+1}). \end{aligned} \tag{21}$$

This formulation mirrors the IMEX BDF scheme—first performing temporal extrapolation, then spatial correction—and enables ASNO to capture rich spatio-temporal dependencies. The Transformer compresses historical observations into a robust latent code, while NAO incorporates system-specific corrections, allowing the model to generalize across systems and time horizons. Both modules are jointly trained end-to-end to minimize prediction error, ensuring stable, interpretable, and adaptive learning.

4. Experiments

In this section, we evaluate ASNO's performance on several benchmark tasks that involve evolving physics governed by partial differential equations (PDEs) and chaotic dynamics. Our experiments are designed to demonstrate ASNO's ability to model complex spatio-temporal systems, generalize across different environments, and maintain long-term predictive accuracy.

4.1. Darcy Flow Equation

We begin by evaluating ASNO on the time-dependent Darcy flow equation, a classical second-order parabolic PDE that models pressure-driven fluid movement through porous media. This equation describes the evolution of the pressure field $p(t, \mathbf{x})$ in response to a time-dependent source term $g(t, \mathbf{x})$ and a heterogeneous permeability field $b(\mathbf{x})$, defined over a unit square domain:

$$\begin{aligned} \frac{\partial p(t,\mathbf{x})}{\partial t} - \nabla \cdot (b(\mathbf{x}) \nabla p(t,\mathbf{x})) = g(t,\mathbf{x}), \quad & \mathbf{x} \in \Omega = [0,1]^2, \\ p(t,\mathbf{x}) = 0, \quad & \mathbf{x} \in \partial\Omega. \end{aligned} \tag{22}$$

Our goal is to learn the spatio-temporal solution operator of this PDE: to predict the next pressure field p_{t+1} from previous field states and the current source field. The dataset consists of 100 independently simulated trajectories, each defined by a unique sample of the permeability field $b(\mathbf{x})$ and the source term $g(t, \mathbf{x})$, resolved on a 21×21 grid over 100 time steps.

Training samples are constructed using a sliding window of five consecutive time steps, with a stride of one. For each trajectory, this yields 96 sequential training pairs. To increase data diversity and robustness, we apply 20 permutations per trajectory, leading to a total of 153,600 training samples and 38,400 held-out test samples. Each model receives as input the past five pressure fields $\{p_{t-4}, \ldots, p_t\}$ and the source field g_t to predict the future state p_{t+1}.

ASNO outperforms baseline models such as FNO, U-Net, DeepONet, Transolver, and GNOT across both in-distribution and out-of-distribution (OOD) settings (Madras & Zemel, 2021; Benitez et al., 2024; Montasser et al., 2024; Thaker et al., 2024). In OOD tests, we perturb the input distributions by altering either the permeability field (OOD-b) or the source term (OOD-f) (Hase et al., 2021). These shifts simulate scenarios where the system departs from training-time assumptions. Table 1 summarizes the test loss, OOD loss, model size, and GPU memory usage. ASNO achieves the lowest loss across all conditions, showcasing superior robustness and generalization. Additional details on the construction and distributional properties of the OOD-f and OOD-b datasets are provided in Appendix B.

ASNO not only achieves strong predictive performance but also aligns well with the underlying numerical structure of the PDE. Its explicit module—built on a Transformer Encoder—approximates the temporal behavior of a BDF5 scheme. The BDF5 coefficients used to extrapolate the latent state are:

$$\begin{aligned} \tilde{X}_m = \frac{12}{137} X_{m-5} - \frac{75}{137} X_{m-4} + \frac{200}{137} X_{m-3} \\ - \frac{300}{137} X_{m-2} + \frac{300}{137} X_{m-1}. \end{aligned} \tag{23}$$

Table 1. Performance on Darcy flow and OOD test sets.

Model	Trainable Params	GPU (MB)	Best Test Loss	Best OOD-f	Best OOD-b
ASNO	760,234	181	**0.0368**	**0.0673**	**0.0982**
FNO	900,224	214	0.0768	0.1129	0.1892
U-Net	820,994	**123**	0.1150	0.1523	0.2224
Transolver	810,573	422	0.0428	0.0721	0.1535
GNOT	760,349	208	0.0516	0.0811	0.1729
DeepONet	6,230,000	2146	0.0537	0.0826	0.1826
Transformer Enc.	1,620,394	173	0.0559	0.0927	0.1736
Linear Enc. + NAO	**720,398**	165	0.0547	0.1245	0.1394

The learned latent representation H_m closely matches $\tilde{X}_m$, showing that ASNO effectively captures high-order time-stepping schemes. For the implicit correction step, we consider the discrete approximation:

$$\tilde{X}_m = X_m + \frac{60}{137}\,\Delta t\,\big(A\,X_m + F_m\big), \tag{24}$$

$$X_m = \Big(I + \tfrac{60}{137}\,\Delta t\,A\Big)^{-1}\Big(\tilde{X}_m - \tfrac{60}{137}\,\Delta t\,F_m\Big), \tag{25}$$

which leads to a theoretical kernel:

$$K_{\text{true}} = -\frac{60}{137}\,\Delta t\,\Big(I + \tfrac{60}{137}\,\Delta t\,A\Big)^{-1}. \tag{26}$$

Figure 2 (right) shows a comparison between this ground-truth kernel and the one learned by NAO, demonstrating close alignment. The left panel reports cumulative rollout errors for all models, showing that ASNO maintains long-term stability:

$$E_T = \sum_{t=1}^{T} \|X_t^{\text{true}} - X_t^{\text{pred}}\|_{L^2}. \tag{27}$$

In summary, ASNO not only excels in predictive accuracy and out-of-distribution robustness but also captures the underlying mathematical structure of time-dependent PDEs. Its decomposable and interpretable design provides both practical reliability and theoretical alignment with physical dynamics.

4.2. Lorenz System

We next evaluate ASNO on the Lorenz system, a prototypical chaotic dynamical system characterized by extreme sensitivity to initial conditions and nonlinear feedback between variables. The governing equations, modified to include external forcing, are:

$$\frac{dx}{dt} = \sigma(y - x) + g_1(t), \tag{28}$$

$$\frac{dy}{dt} = x(\rho - z) - y + g_2(t), \tag{29}$$

$$\frac{dz}{dt} = xy - \beta z + g_3(t), \tag{30}$$

where $\sigma = 10, \rho = 28, \beta = \frac{8}{3}$ are canonical parameters and g_1, g_2, g_3 are time-varying forcing terms. The dataset contains 2,000 trajectories generated using 100 distinct parameter combinations of (σ, ρ, β), and 20 different loading profiles $g_i(t)$, each initialized from $(x_0, y_0, z_0) = (0, 1, 0)$. Each trajectory includes 1,000 time steps. We use a sliding window of five time steps with stride one to construct 995 input-output pairs per profile, yielding 1,592,000 training samples (80%) and 398,000 test samples (20%).

To preserve temporal locality and improve generalization, 5-step windows are permuted within each trajectory, and profile indices are randomized to ensure no overlap between training and test sets. All models receive as input the past five states $\{X_{m-4}, \ldots, X_m\}$ and the current forcing F_m.

We compare ASNO against multiple baselines, including DeepONet, Transolver, GNOT, and Transformer Encoder. Architectures like FNO, U-Net, and GNOT are excluded due to mismatch with the low-dimensional structure of Lorenz: U-Net and FNO assume spatial convolution or spectral filters over grids, which are not meaningful in 3D ODE settings; GNOT's geometric processing introduces unnecessary complexity without corresponding gains.

Table 2. Comparison of ASNO and baselines on the Lorenz system.

Model Type	Trainable Params	GPU (MB)	Best Test Loss
ASNO	258,808	76	**0.000794**
Transolver	395,907	95	0.000835
DeepONet	265,512	79	0.001750
Transformer Encoder	**257,647**	**71**	0.001821
GNOT	401,155	106	0.002189
Linear Enc. + NAO	305,776	87	0.005298

ASNO achieves the lowest test loss and uses fewer parameters and memory than several baselines. By explicitly mirroring the implicit-explicit decomposition of BDF schemes, ASNO separates its architecture into a Transformer Encoder for extrapolating homogeneous dynamics and a Nonlocal Attention Operator (NAO) for handling nonlinear coupling via forcing terms. The Transformer extracts dominant temporal modes over prior states, while the NAO captures fast-changing variable interactions like xy and xz, enabling the model to remain stable over long rollouts.

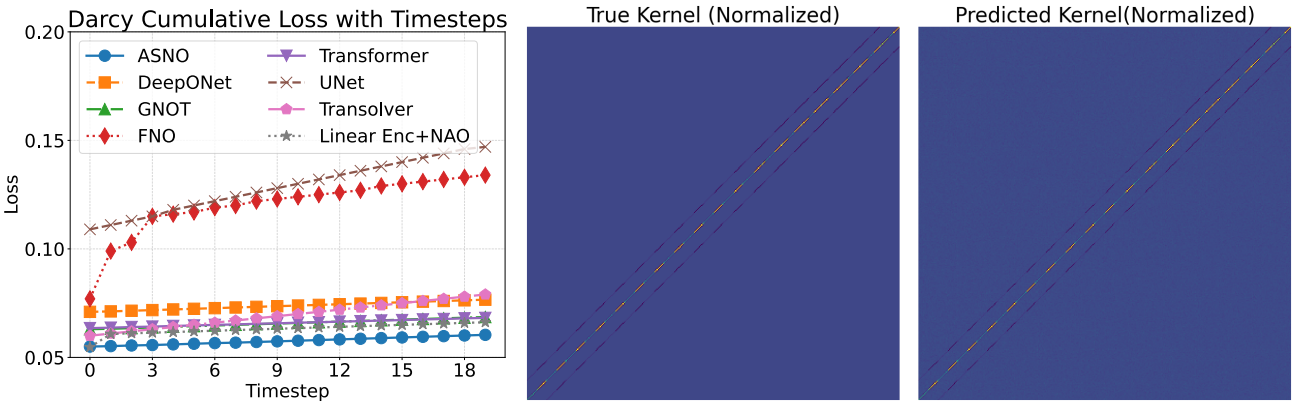

Figure 2. Left: Cumulative rollout error over time for various models on the Darcy flow benchmark. Right: Comparison between the discretized theoretical kernel and the learned NAO kernel, showing strong agreement.

This two-stage process ensures that representational burden is split: the Transformer handles regular recurrence, and NAO focuses on residual dynamics. During rollout, ASNO uses its prior output as the next input in an autoregressive fashion, improving trajectory fidelity under chaotic evolution.

Figure 3 shows that ASNO remains aligned with the ground truth much longer than other models. Its predictions track the true Lorenz attractor, especially in the sensitive y and z dimensions, where baselines diverge rapidly. The stability and structure-awareness offered by ASNO make it well-suited for complex forecasting tasks involving nonlinear, chaotic systems such as weather models, turbulence, and real-time control.

5. Uncertainty Quantification Verification

To further validate ASNO's reliability in scientific modeling tasks, we assess its ability to quantify predictive uncertainty (Zollo et al., 2024). Accurate uncertainty estimates are critical in high-stakes domains like scientific computing and engineering, where overconfident models can lead to catastrophic failures. We evaluate ASNO using two standard metrics: *Prediction Interval Coverage Probability* (PICP) and *Mean Prediction Interval Width* (MPIW), following established practices in uncertainty quantification. In our framework, uncertainty is quantified using a Linear Laplace Approximation (LA), which approximates the model's weight posterior as a Gaussian centered at the MAP estimate (Wang et al., 2018). We use a diagonal Hessian approximation to efficiently estimate epistemic uncertainty from the ASNO model parameters without prohibitive computational cost (Abdar et al., 2021; Ritter et al., 2018). Further implementation details and derivations of the LLA approach are provided in Appendix C.

These metrics are computed on the held-out Darcy flow test set consisting of 100 spatio-temporal profiles, each with 95 rollout steps and resolved on a 21×21 grid (441 spatial locations). The evaluation aggregates over all time steps and samples, yielding robust summary statistics. The results, shown in Table 3, yield a mean PICP of 0.94 ± 0.03 and a mean MPIW of 0.32 ± 0.05.

The PICP metric quantifies the proportion of true values that lie within the model's 95% confidence intervals, measuring *calibration* (Nikulchev & Chervyakov, 2023). A value close to 0.95 indicates that the model's uncertainty estimates are neither underconfident (low coverage) nor overly conservative (excessive interval width). The MPIW, on the other hand, measures *sharpness* (Xue et al., 2024), or the average width of the prediction intervals. Low MPIW is desirable, provided coverage is maintained, as it implies confident, precise predictions.

The intervals are computed using a diagonal Linear Laplace approximation of the model's parameter posterior. This approach provides an efficient method for estimating epistemic uncertainty in large neural models. For N spatial locations, these metrics are defined as:

$$\text{PICP} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left\{y_i \in [\mu_i - 1.96\,\sigma_i,\ \mu_i + 1.96\,\sigma_i]\right\}, \tag{31}$$

$$\text{MPIW} = \frac{1}{N}\sum_{i=1}^{N} \left[(\mu_i + 1.96\,\sigma_i) - (\mu_i - 1.96\,\sigma_i)\right], \tag{32}$$

where y_i denotes the ground-truth value at spatial location i, and $\mu_i \pm 1.96\,\sigma_i$ represents the predicted 95% confidence interval derived from the model's posterior variance.
The observed PICP of 94.0% is well-aligned with the

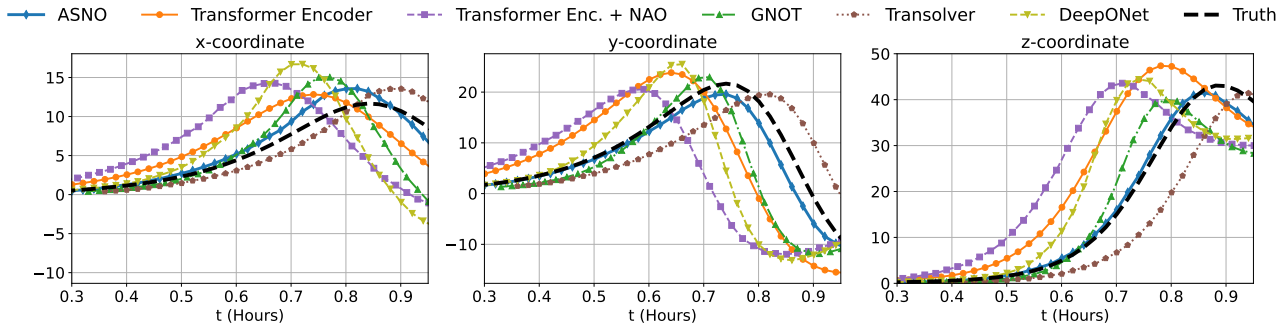

Figure 3. Long-time prediction comparison of predicted vs. true (x, y, z) trajectories for the Lorenz system using different models. ASNO shows better alignment and stability over extended timesteps.

Table 3. Uncertainty quantification metrics on Darcy flow test set (averaged over 100 test samples × 95 timesteps).

Metric	Value
PICP (Coverage Probability)	94.00 %
MPIW (Interval Width)	0.3046

nominal 95% target, indicating that ASNO produces well-calibrated uncertainty intervals. Furthermore, the mean MPIW of 0.3046 is relatively narrow considering the Darcy pressure field typically ranges from 0 to 1. Together, these findings demonstrate that ASNO offers reliable and informative uncertainty quantification—crucial for enabling robust deployment in scientific modeling, digital twins, and autonomous decision-making systems.

6. Melt Pool Temperature Field Prediction in Additive Manufacturing

We now evaluate ASNO on a real-world engineering task: predicting the spatio-temporal evolution of melt pool temperature fields in Directed Energy Deposition (DED) processes. Accurate full-field thermal prediction is vital for controlling process quality, minimizing defects, and improving part integrity in metal additive manufacturing. These temperature fields serve as global descriptors from which critical local features—such as melt pool width, depth, and solidification rate—are inferred. Therefore, learning accurate and interpretable spatio-temporal surrogates for melt pool behavior is crucial for digital twin applications in advanced manufacturing.

6.1. Full-Field Thermal Modeling and Prediction with ASNO

We evaluate ASNO on high-fidelity transient thermal simulations from GAMMA, a GPU-accelerated Finite Element Analysis (FEA) code for additive manufacturing (Liao et al., 2023). GAMMA solves the transient heat conduction equation on a structured grid:

$$\rho\, C_p(T)\, \frac{\partial T}{\partial t} + \nabla \cdot q = 0, \tag{33}$$

where T is temperature (K), ρ is density (g/mm^3), and $C_p(T)$ is temperature-dependent specific heat capacity (J/g/K). The heat flux q obeys Fourier's law:

$$q = -k(T)\, \nabla T, \tag{34}$$

with $k(T)$ denoting thermal conductivity (W/m·K). Boundary conditions include Gaussian laser heating and convective/radiative losses:

$$q{\cdot}n = -\frac{2\eta P_{\text{laser}}}{\pi r_{\text{laser}}^2} \exp\left(-\frac{2r^2}{r_{\text{laser}}^2}\right) + \sigma\epsilon(T^4 - T_0^4) + h(T - T_0), \tag{35}$$

where η is absorption, P_{laser} laser power, r_{laser} beam radius, σ Stefan–Boltzmann constant, ϵ emissivity, and h convection coefficient. The laser intensity profile follows:

$$I_{\text{laser}}(r) = \frac{2P_{\text{laser}}}{\pi r_{\text{laser}}^2} \exp\left(-\frac{2r^2}{r_{\text{laser}}^2}\right). \tag{36}$$

We generate 100 simulation trajectories, each with unique time-varying laser power and scan speed profiles. Temperature fields $T_{\text{pool}}(m) \in \mathbb{R}^{21\times 21}$ are recorded per time step. Using a 5-step sliding window, we extract 726 training samples per simulation, producing 1,161,600 total training samples. Materials include 316L stainless steel and 1018 carbon steel, with nonlinear thermal properties—for example, $k(T)$ increases from 0.01396 to 0.03439 W/m·K (300–1600 K), and $C_p(T)$ rises from 0.512 to 0.770 J/g·K (400–1800 K), introducing strong field nonlinearity.

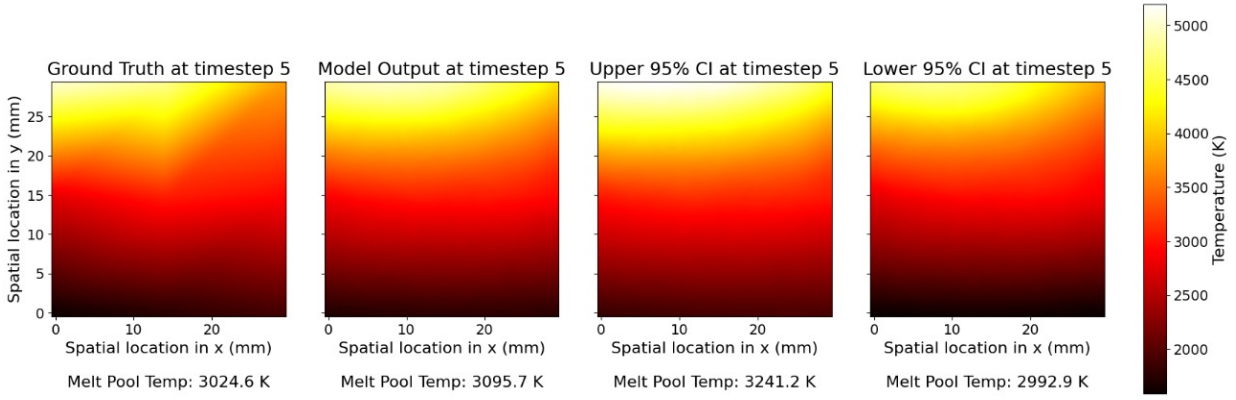

Figure 4. Comparison of ASNO-predicted full-field temperature (left) and GAMMA ground truth (right) for a representative timestep. Color indicates temperature (K).

ASNO receives five prior temperature maps $T_{\text{pool}}(m-4), \ldots, T_{\text{pool}}(m)$, along with next-step process parameters: laser power $P_{\text{laser}}(m+1)$, scan speed $V_{\text{scan}}(m+1)$, and laser location $L_{\text{laser}}(m+1) \in \mathbb{R}^2$. The Transformer Encoder maps the temporal sequence to a latent forecast:

$$H_{m+1} = \text{TE}\left(T_{\text{pool}}(m-4:m)\right), \tag{37}$$

which acts as the extrapolated BDF solution $\tilde{X}_{m+1}$. The NAO then refines the prediction using process parameters:

$$\hat{T}_{\text{pool}}(m+1) = \text{NAO}\left(H_{m+1}, \{P_{\text{laser}}, V_{\text{scan}}, L_{\text{laser}}\}_{m+1}\right). \tag{38}$$

This decomposition improves robustness across dynamic regimes. A diagonal Laplace approximation estimates epistemic uncertainty, providing 95% confidence intervals at each grid point:

$$\mu_i \pm 1.96\,\sigma_i, \tag{39}$$

with μ_i and σ_i as the predicted mean and standard deviation at location i. Figure 4 compares ASNO predictions with ground truth from GAMMA, showing strong alignment in melt pool shape, peak intensity, and thermal trailing zones. ASNO achieves a mean absolute percentage error (MAPE) of 2.50%, confirming high-fidelity predictive performance. ASNO outperforms baseline models in both accuracy and computational efficiency, achieving the lowest test loss while maintaining fast inference. Its separation of temporal and spatial modules improves adaptability under varying physics, and its uncertainty quantification supports deployment in real-world AM systems. Training protocols, mesh resolution, and further material-specific parameters are provided in Appendix A.

Table 4. Comparison of models on the Additive Manufacturing dataset.

Model Type	Trainable Params	Time per Epoch (s)	Best Test Loss
ASNO	2,825,444	6.17	**0.0140**
FNO	3,654,312	7.10	0.0357
DeepONet	3,385,605	5.25	0.0456
UNet	4,564,157	6.64	0.0652
GNOT	**2,442,678**	6.60	0.0332
Transolver	2,636,9158	8.71	0.0502
Linear + NAO	2,943,435	4.75	0.0397
Transformer Encoder (TE)	2,784,920	**3.95**	0.0563

7. Conclusion

This work introduces ASNO, an attention-based spatio-temporal neural operator designed to advance the reliability, generalizability, and interpretability of foundation models applied to evolving physical systems. By emulating the implicit–explicit structure of the Backward Differentiation Formula (BDF), ASNO separates temporal dynamics—modeled via a Transformer Encoder—from spatial interactions—captured using a Nonlocal Attention Operator (NAO). This modular decomposition improves transparency by isolating historical system behavior from external forcing, while enhancing adaptability under changing conditions. Evaluated on diverse scientific benchmarks—including chaotic systems, PDE-driven phenomena, and additive manufacturing simulations—ASNO consistently achieves strong prediction accuracy and long-term stability, even under distribution shifts. Importantly, ASNO offers quantifiable uncertainty via Laplace-based epistemic estimation, addressing critical concerns around trust and responsible deployment in high-stakes environments. By combining physical structure with modern attention mechanisms, ASNO contributes to the broader effort of building reliable and responsible foundation models that are not only accurate but interpretable, robust to dynamics, and deployable in safety-critical domains.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

Ascher, U. M., Ruuth, S. J., and Wetton, B. T. Implicit-explicit methods for time-dependent partial differential equations. *SIAM Journal on Numerical Analysis*, 32(3): 797–823, 1995.

Benitez, J. A. L., Furuya, T., Faucher, F., Kratsios, A., Tricoche, X., and de Hoop, M. V. Out-of-distributional risk bounds for neural operators with applications to the helmholtz equation. *Journal of Computational Physics*, 513:113168, 2024.

Ellington, C. N., Sun, N., Ho, N., Tao, T., Mahbub, S., Li, D., Zhuang, Y., Wang, H., Song, L., and Xing, E. P. Accurate and general dna representations emerge from genome foundation models at scale. *bioRxiv*, pp. 2024–12, 2024.

Fredebeul, C. A-bdf: a generalization of the backward differentiation formulae. *SIAM journal on numerical analysis*, 35(5):1917–1938, 1998.

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

Hase, P., Xie, H., and Bansal, M. The out-of-distribution problem in explainability and search methods for feature importance explanations. *Advances in neural information processing systems*, 34:3650–3666, 2021.

Huang, Y., Gao, C., Wu, S., Wang, H., Wang, X., Zhou, Y., Wang, Y., Ye, J., Shi, J., Zhang, Q., et al. On the trustworthiness of generative foundation models: Guideline, assessment, and perspective. *arXiv preprint arXiv:2502.14296*, 2025.

Kitaev, N., Kaiser, Ł., and Levskaya, A. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

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

Liao, S., Golgoon, A., Mozaffar, M., and Cao, J. Efficient gpu-accelerated thermomechanical solver for residual stress prediction in additive manufacturing. *Computational Mechanics*, 71(5):879–893, 2023.

Lim, B., Arık, S. Ö., Loeff, N., and Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.

Lu, L., Jin, P., Pang, G., Zhang, Z., and Karniadakis, G. E. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3):218–229, 2021.

Luo, S., Ni, J., Chen, S., Yu, R., Xie, Y., Liu, L., Jin, Z., Yao, H., and Jia, X. Free: The foundational semantic recognition for modeling environmental ecosystems. *arXiv preprint arXiv:2311.10255*, 2023.

Madras, D. and Zemel, R. Identifying and benchmarking natural out-of-context prediction problems. *Advances in Neural Information Processing Systems*, 34:15344–15358, 2021.

Molnar, C. *Interpretable machine learning*. Lulu. com, 2020.

Montasser, O., Shao, H., and Abbe, E. Transformation-invariant learning and theoretical guarantees for ood generalization. *arXiv preprint arXiv:2410.23461*, 2024.

Nikulchev, E. and Chervyakov, A. Prediction intervals: A geometric view. *Symmetry*, 15(4):781, 2023.

Ritter, H., Botev, A., and Barber, D. A scalable laplace approximation for neural networks. In *6th international conference on learning representations, ICLR 2018-conference track proceedings*, volume 6. International Conference on Representation Learning, 2018.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85, 2022.

Thaker, P., Setlur, A., Wu, Z. S., and Smith, V. On the benefits of public representations for private transfer learning under distribution shift. *Advances in Neural Information Processing Systems*, 37:27088–27120, 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, J., Wells III, W. M., Golland, P., and Zhang, M. Efficient laplace approximation for bayesian registration uncertainty quantification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 880–888. Springer, 2018.

Wu, H., Luo, H., Wang, H., Wang, J., and Long, M. Transolver: A fast transformer solver for pdes on general geometries. *arXiv preprint arXiv:2402.02366*, 2024.

Xue, L., Zhou, K., and Zhang, X. Continuous optimization for construction of neural network-based prediction intervals. *Knowledge-Based Systems*, 293:111669, 2024.

Yang, X., Yao, H., and Wei, Y. One meta-tuned transformer is what you need for few-shot learning. In *Forty-first International Conference on Machine Learning*, 2024.

Yu, Y., Liu, N., Lu, F., Gao, T., Jafarzadeh, S., and Silling, S. Nonlocal attention operator: Materializing hidden knowledge towards interpretable physics discovery. *arXiv preprint arXiv:2408.07307*, 2024.

Zhang, Y., Ma, Z., Li, J., Qiao, Y., Wang, Z., Chai, J., Wu, Q., Bansal, M., and Kordjamshidi, P. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *arXiv preprint arXiv:2407.07035*, 2024.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

Zhou, X., Gupta, A., Upadhyay, S., Bansal, M., and Faruqui, M. Can sequence-to-sequence transformers naturally understand sequential instructions? In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, pp. 527–534, 2023.

Zollo, T. P., Deng, Z., Snell, J. C., Pitassi, T., and Zemel, R. Improving predictor reliability with selective recalibration. *arXiv preprint arXiv:2410.05407*, 2024.

Zou, S., Tao, T., Mahbub, S., Ellington, C. N., Algayres, R., Li, D., Zhuang, Y., Wang, H., Song, L., and Xing, E. P. A large-scale foundation model for rna function and structure prediction. *bioRxiv*, pp. 2024–11, 2024.

A. Melt Pool Temperature Field Prediction in Additive Manufacturing

This section presents a detailed evaluation of melt pool predictions produced by ASNO, alongside baseline comparisons. In Directed Energy Deposition (DED), melt pool behavior is a key determinant of the resulting material properties. Accurately modeling this behavior is crucial for capturing temperature evolution, phase transitions, and thermal cycling—factors that directly influence microstructure and mechanical integrity. Moreover, insight into melt pool dynamics supports improved process control, defect avoidance, and dimensional consistency. Through this analysis, we aim to illustrate ASNO's ability to accurately reflect these critical thermal features in a manufacturing context.

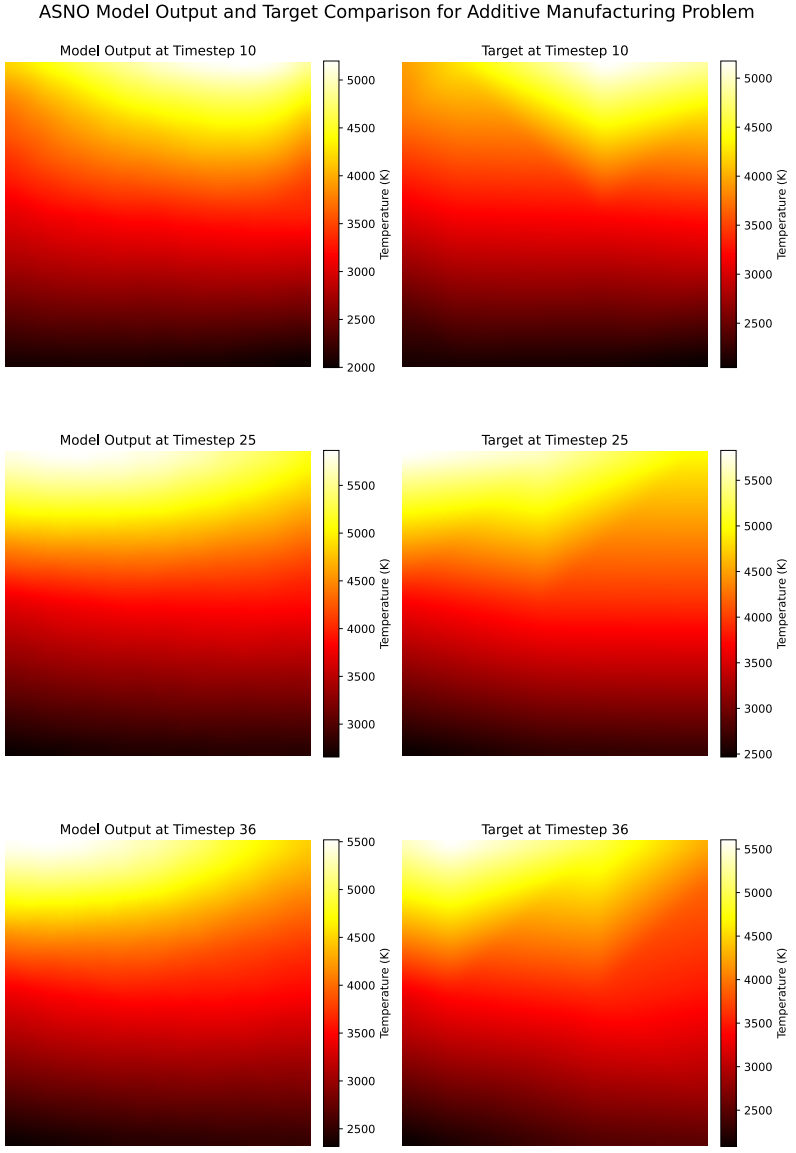

Figure 5. Comparison of ASNO model output and target temperature distributions for the melt pool in Directed Energy Deposition at Timesteps 10, 25, and 36. Each row represents a different timestep, with the model output shown on the left and the target on the right. The color scale indicates temperature in Kelvin, highlighting ASNO's accuracy in capturing the thermal profile of the melt pool.

Simulations are conducted on a 21×21 spatial grid, capturing 2D temperature fields at multiple timesteps throughout the deposition sequence. Each timestep represents the effect of the moving heat source on the substrate, with temperature gradients shaped by material response and boundary heat loss.

Figure 5 displays these comparisons at three representative timesteps (10, 25, and 36). For each timestep, ASNO's output (left column) is directly compared to the corresponding ground truth (right column). The temperature scale ranges from 2000 K to 5500 K, revealing heat concentration in the melt pool center and the cooling gradients that trail the scanning laser. Across all shown timesteps, ASNO closely matches the true temperature fields.

These results highlight ASNO's capacity to capture high-temperature regions and their spatial evolution with high fidelity. The model maintains stability over longer prediction horizons and accurately follows the transient heat conduction patterns characteristic of DED processes. The close alignment between predictions and targets supports the conclusion that ASNO is a reliable and precise surrogate for melt pool simulation—making it valuable for thermal control, quality assurance, and optimization in industrial-scale additive manufacturing workflows.

B. Additional Details of Generation of OOD-f and OOD-b Test Sets for Darcy Flow

To rigorously evaluate the robustness and generalization capabilities of ASNO under distributional shifts, as in section 4.1, we constructed two distinct out-of-distribution (OOD) test sets for the time-dependent Darcy flow benchmark: OOD-f and OOD-b. These datasets were designed to systematically alter key components of the PDE inputs—specifically, the source field $g(t, \mathbf{x})$ and the permeability field $b(\mathbf{x})$—while keeping the governing dynamics fixed. This setup enables assessment of model reliability and adaptability in scenarios where real-world conditions deviate from those observed during training.

The OOD-f dataset targets variation in the time-varying source field $g(t, \mathbf{x})$. In the training set, source fields are sampled from a Gaussian random field (GRF) with fixed hyperparameters $\alpha = 2$ and $\tau = 3$, defining the smoothness and length scale of the spatial correlation, respectively. To construct the OOD-f samples, we modulate the original GRF realizations with a sinusoidal temporal component and scale their amplitudes to introduce more dynamic variability. Specifically, each OOD-f source field evolves as $g(t, \mathbf{x}) = \tilde{g}(\mathbf{x}) \cdot \sin(t)$, where $\tilde{g}(\mathbf{x})$ is a spatial GRF sampled using the same base parameters as in training, and $t = q\Delta t$ is the timestep with $\Delta t = 5 \times 10^{-5}$. The amplitude of these signals is scaled by a factor of 200, significantly enhancing the temporal variance and introducing a richer structure in both spatial and temporal dimensions. This modification alters the distributional properties of the source field by injecting temporal oscillations that were absent during training, effectively testing the model's ability to extrapolate under novel temporal dynamics.

The OOD-b dataset focuses on spatial distribution shifts in the hidden permeability field $b(\mathbf{x})$, which controls the diffusivity of the medium and is critical to pressure evolution in Darcy flow. During training, the permeability field is represented as a binary microstructure generated from a GRF with parameters $\alpha_\chi = 4$ and $\tau_\chi = 5$. These parameters yield relatively smooth, moderate-frequency spatial variations in the binary phase structure. In contrast, the OOD-b set is generated using a GRF with increased smoothness and shorter correlation length: $\alpha_\chi = 7$ and $\tau_\chi = 6$. This leads to finer, more fragmented spatial patterns with sharper phase transitions and higher local heterogeneity. As a result, the spatial structure of the flow environment becomes significantly more complex, challenging the ability of a learned operator to generalize to unseen permeability regimes. These shifts in the spatial correlation structure directly influence the shape and propagation of pressure waves, providing a meaningful test of model generalization across geometry-induced variability.

Across both OOD-f and OOD-b, we ensure that the data generation pipeline matches the resolution and format used in training, with simulations performed on a 21×21 spatial grid over 100 time steps. For each OOD set, new trajectories were generated using the modified GRFs and their respective parameter configurations. The number of test samples (38,400) and evaluation protocol remain consistent with those used for in-distribution testing. Table 1 reports the performance of ASNO and competing baselines under these OOD conditions. The results show that ASNO achieves the lowest test losses on both OOD-f and OOD-b, confirming its superior generalizability under both temporal and spatial distribution shifts. These improvements are particularly noteworthy given ASNO's comparable parameter count and GPU footprint relative to the other models evaluated. Collectively, the OOD benchmarks validate ASNO's robustness and its ability to maintain predictive accuracy even when the data distribution deviates significantly from the training regime.

C. Additional Details of Uncertainty Quantification with Linear Laplace Approximation

This section provides implementation details for the uncertainty quantification results discussed in Section 5. In particular, we describe how a Linear Laplace Approximation (LLA) is applied to the ASNO model to quantify epistemic uncertainty in the Darcy flow benchmark.

Uncertainty quantification is essential in scientific computing and applications such as additive manufacturing, where variabil-

ity in system conditions can significantly influence predictive outcomes. In this context, epistemic uncertainty—stemming from limited training data or structural model limitations—must be captured to ensure robust decision-making, adaptive sampling, and model-based control. LLA provides a principled way to estimate such uncertainty by approximating the posterior distribution over model parameters.

The standard learning process in neural networks optimizes the regularized empirical risk using:

$$\theta_{\mathrm{MAP}} = \arg\min_{\theta} \mathcal{R}(\mathcal{D}, \theta) = \arg\min_{\theta} \left(l(\mathcal{D}, \theta) + r(\theta) \right), \tag{40}$$

where $l(\mathcal{D}, \theta)$ is the data-fitting loss and $r(\theta)$ represents a regularization term, interpreted as the negative log-prior in a Bayesian framework. The resulting optimizer θ_{MAP} serves as the maximum-a-posteriori estimate. The posterior is then written as:

$$p(\theta|\mathcal{D}) \propto \exp\Big(-\mathcal{R}(\mathcal{D}, \theta)\Big). \tag{41}$$

LLA proceeds by approximating this posterior with a Gaussian via a second-order Taylor expansion of the risk function around θ_{MAP}:

$$\mathcal{R}(\mathcal{D}, \theta) \approx \mathcal{R}(\mathcal{D}, \theta_{\mathrm{MAP}}) + \frac{1}{2}(\theta - \theta_{\mathrm{MAP}})^{\top} H \, (\theta - \theta_{\mathrm{MAP}}), \tag{42}$$

where $H = \nabla^2_{\theta\theta} \mathcal{R}(\mathcal{D}, \theta)\big|_{\theta_{\mathrm{MAP}}}$ is the Hessian of the risk at the MAP estimate. Since the gradient vanishes at the optimum, the approximate posterior becomes:

$$p(\theta|\mathcal{D}) \approx p(\theta_{\mathrm{MAP}}|\mathcal{D}) + \frac{1}{2}(\theta - \theta_{\mathrm{MAP}})^{\top} \left(\nabla^2_{\theta\theta} p(\theta|\mathcal{D}) \Big|_{\theta_{\mathrm{MAP}}} \right) (\theta - \theta_{\mathrm{MAP}}). \tag{43}$$

This posterior is thus a multivariate Gaussian with:

$$\theta \sim \mathcal{N}(\theta_{\mathrm{MAP}}, \Sigma), \quad \Sigma = \left(\nabla^2_{\theta\theta} p(\theta|\mathcal{D}) \Big|_{\theta_{\mathrm{MAP}}} \right)^{-1}. \tag{44}$$

Due to the high dimensionality of modern neural networks, computing the full Hessian is impractical. We adopt a diagonal approximation:

$$\Sigma \approx \mathrm{diag}\big(\tfrac{1}{\lambda_1}, \tfrac{1}{\lambda_2}, \ldots, \tfrac{1}{\lambda_n}\big), \tag{45}$$

where λ_i are the diagonal entries of the Hessian. This reduces memory and computational costs while retaining useful information about parameter uncertainty.

Using this Gaussian posterior, we compute the predictive variance by propagating uncertainty through the ASNO architecture. Given the ASNO update rule:

$$\mathbf{X}^{\mathrm{out}}_{m+1} = \mathrm{NAO}\Big(\mathrm{TE}(\mathbf{X}_m, \ldots, \mathbf{X}_{m-n+1}), \, \mathbf{F}_{m+1}\Big), \tag{46}$$

the predictive covariance is:

$$\mathrm{Cov}\big[\mathbf{X}^{\mathrm{out}}_{m+1}\big] \approx J_{\theta}^{(m+1)} \, \Sigma \, (J_{\theta}^{(m+1)})^{\top}, \tag{47}$$

where the Jacobian $J_{\theta}^{(m+1)}$ is given by:

$$J_{\theta}^{(m+1)} = \frac{\partial \, \mathbf{X}^{\mathrm{out}}_{m+1}}{\partial \, \theta} = \frac{\partial}{\partial \theta} \Big[\mathrm{NAO}\big(\mathrm{TE}(\mathbf{X}_m, \ldots, \mathbf{X}_{m-n+1}), \, \mathbf{F}_{m+1}\big) \Big].$$

This allows us to obtain confidence intervals around each predicted field value, supporting the evaluation metrics introduced in the main paper (see Equations 31 and 32 in Section 5). These intervals are derived from the posterior variance using the Gaussian assumption, yielding prediction bounds of the form $\mu_i \pm 1.96\,\sigma_i$, where μ_i and σ_i are the mean and standard deviation at spatial location i.

This framework enables ASNO to quantify uncertainty efficiently and reliably, providing well-calibrated intervals and supporting applications in scientific modeling, where trustworthiness and interpretability are essential. LLA offers a scalable, computationally tractable method for augmenting ASNO with epistemic uncertainty estimates, without significant changes to the training procedure or model architecture.

D. Nomenclature

Symbol	Meaning
$X,\ x$	State vector (general)
X_m	State at timestep m
$\tilde{X}_{m+1}$	Extrapolated (homogeneous) state
F	External forcing / loading field
S	Hidden system (environment) state
t	Continuous time
m	Time-index
n	History length
Δt	Time step size
$\alpha_k,\ \beta$	BDF coefficients
$\mathcal{D}^{(\eta)}$	Dataset for system η
$\mathcal{F}$	ASNO operator mapping
$\hat{y}_{t+1}$	Predicted output at time $t+1$
$y_t,\ \hat{y}_t$	True and predicted values at time t
E_T	Cumulative error over T steps
e_t	Instantaneous error at timestep t
$Q,\ K,\ V$	Query, Key, and Value matrices used in attention mechanisms; used generically, with context-specific definitions below
W_E	Embedding matrix
$W_{Tq},\ W_{Tk},\ W_{Tv}$	Time-series Transformer Encoder attention weight matrices for Query (Q), Key (K), and Value (V) used in the explicit extrapolation step
P_k	Positional encoding at position k
$d_{\text{embed}},\ d_t,\ d$	Embedding dimension, key/query dimension, and feature dimension
H_t	Penultimate-step latent features used in computing the NAO kernel weights for h and f
J_t	Intermediate NAO state at attention step t composed of latent features and forcing fields (H_t, F_t), iteratively updated across T layers
$W_{P,h},\ W_{P,f},\ W_{Q_t},\ W_{K_t}$	Weight matrices in the Nonlocal Attention Operator (NAO) used for computing kernel projections; W_{Q_t} and W_{K_t} generate attention Query and Key vectors used in the implicit interaction modeling
$K[\cdot]$	Learned nonlocal kernel operator acting over latent space in NAO
$\mathcal{H},\ \mathcal{F}$	Input and output Banach function spaces
T	Number of attention steps in NAO

E. Impact Statement

This paper contributes to advancing the field of Machine Learning by developing an innovative approach for modeling spatio-temporal dynamics in scientific and engineering applications. Our work has broad potential societal implications, including applications in manufacturing, autonomous systems, and physics-based modeling. While these implications are significant, we do not identify any immediate ethical concerns that necessitate specific emphasis at this time.