Physics-Guided Transformer (PGT): Physics-Aware Attention Mechanism for PINNs

Ehsan Zeraatkar*, Rodion Podorozhny, Jelena Tešić

Computer Science, Texas State University, San Marcos, Texas, US

Abstract

Reconstructing continuous physical fields from sparse, irregular observations is a fundamental challenge in scientific machine learning, particularly for nonlinear systems governed by partial differential equations (PDEs). Dominant physics-informed approaches enforce governing equations as soft penalty terms during optimization, a strategy that often leads to gradient imbalance, instability, and degraded physical consistency when measurements are scarce. Here we introduce the **Physics-Guided Transformer (PGT)**, a neural architecture that moves beyond residual regularization by embedding physical structure directly into the self-attention mechanism. Specifically, PGT incorporates a heat-kernel–derived additive bias into attention logits, endowing the encoder with an inductive bias consistent with diffusion physics and temporal causality. Query coordinates attend to these physics-conditioned context tokens, and the resulting features drive a FiLM-modulated sinusoidal implicit decoder that adaptively controls spectral response based on the inferred global context. We evaluate PGT on two canonical benchmark systems spanning diffusion-dominated and convection-dominated regimes: the one-dimensional heat equation and the two-dimensional incompressible Navier–Stokes equations. In 1D sparse reconstruction with as few as 100 observations, PGT attains a relative L^2 error of 5.9×10^{-3}, representing a 38-fold reduction over physics-informed neural networks and more than 90-fold reduction over sinusoidal implicit representations. In the 2D cylinder-wake problem reconstructed from 1500 scattered spatiotemporal samples, PGT uniquely achieves strong performance on both axes of evaluation: a governing-equation residual of 8.3×10^{-4} — on par with the best residual-based methods — alongside a competitive overall relative L^2 error of 0.034, substantially below all methods that achieve comparable physical consistency. No individual baseline simultaneously satisfies these dual criteria. Convergence analysis further reveals sustained, monotonic error reduction in PGT, in contrast to the early optimization plateaus observed in residual-based approaches. These findings demonstrate that structural incorporation of physical priors at the representational level, rather than solely as an external loss penalty, substantially improves both optimization stability and physical coherence under data-scarce conditions. Physics-guided attention provides a principled and extensible mechanism for reliable reconstruction of nonlinear dynamical systems governed by partial differential equations.

Keywords: physics-informed neural networks, implicit neural representations, transformer architecture, physics-guided attention, sparse reconstruction, Navier–Stokes equations, scientific machine learning, partial differential equations

1. Introduction

Scientific systems governed by partial differential equations (PDEs) describe a wide range of natural and engineered phenomena, from heat diffusion and fluid transport to climate dynamics and material deformation. Accurately solving these equations is central to advancing predictive science and engineering. However, traditional numerical solvers — finite difference, finite volume, or spectral methods [17, 6] — often require finely discretized spatiotemporal grids to maintain stability and accuracy, leading to prohibitive computational costs for modeling high-dimensional or multiscale systems. This motivates the growing field

Email address: ehsanzeraatkar@txstate.edu (Ehsan Zeraatkar*)

of *Scientific Machine Learning* (SciML) [14, 4], which seeks to learn surrogate models that embed physical knowledge into data-driven neural architectures, enabling efficient yet physically consistent approximations of PDE-governed processes.

Among the early and influential developments in SciML are *Physics-Informed Neural Networks* (PINNs) [25], which enforce PDE constraints as soft penalties within the loss function. Although conceptually elegant, PINNs exhibit well-known challenges: gradient pathologies in stiff or multiscale regimes, multiscale convergence in high dimensions, and limited ability to represent oscillatory or high-frequency components. To overcome these issues, several operator-learning frameworks, such as *Fourier Neural Operators* (FNOs) [19], *DeepONets*, and *Galerkin Transformers* [7], have been proposed to learn mappings between function spaces rather than on discrete fields. Despite their success, these models often rely on purely spectral priors or dense Fourier convolutions, lacking explicit awareness of underlying physical causality or PDE structure.

In parallel, the *Transformer* architecture [29] has revolutionized sequence and vision modeling by capturing long-range dependencies through self-attention. Recent works have adapted Transformers to physical systems, demonstrating their potential for spatiotemporal forecasting and operator learning. Yet, conventional Transformers are *data-driven but not physics-driven* — their attention weights are learned solely from data, without constraints that enforce physical consistency, such as causality in time, locality in diffusion, or conservation laws. As a result, these models may achieve high predictive accuracy while violating fundamental PDE dynamics, particularly when training data are sparse or partially observed.

To bridge this gap, we introduce the **Physics-Guided Transformer (PGT)**, a unified architecture that couples Transformer-based context modeling with explicit physical priors derived from PDE theory. PGT embeds the heat-kernel Green's function [11] as an additive bias within the self-attention mechanism, enabling the model to respect the causal and diffusive structure of parabolic PDEs. Contextual patches extracted from low-resolution data are encoded through this physics-guided attention, producing a latent representation that captures both local interactions and physically meaningful dependencies. Query coordinates (x, t) then attend to these encoded context tokens to generate *physics-conditioned features*, which are decoded by a *FiLM-modulated SIREN*—an implicit neural representation that adaptively adjusts its frequency response based on the learned context.

Unlike purely data-driven Transformers or physics-agnostic INRs, PGT integrates physical reasoning directly into the attention kernel while retaining the flexibility of neural implicit representations. The resulting model can infer continuous spatiotemporal fields, satisfy governing PDEs via autodifferentiation, and integrate multiple sources of supervision, including high-resolution data, low-resolution averages, and boundary or initial constraints.

The key contributions of this work are summarized as follows:

- **Physics-guided attention:** We formulate an additive attention bias derived from the heat-kernel Green's function, introducing an inductive bias consistent with diffusion physics and temporal causality.
- **Context-conditioned implicit decoding:** We design a FiLM-modulated SIREN decoder that adaptively controls spectral bias via frequency gating, enabling accurate reconstruction of high-frequency details.
- **Unified physics–data training framework:** We propose a composite uncertainty-weighted loss combining PDE residuals, boundary/initial conditions, and data-fidelity terms.
- **Demonstration on canonical benchmarks:** Through extensive experiments on 1D heat diffusion and 2D incompressible Navier–Stokes problems, PGT achieves competitive reconstruction accuracy alongside markedly reduced governing-equation residuals compared to PINNs, FNOs, and Transformer-based baselines.

By integrating physical inductive biases into the Transformer attention mechanism, PGT moves beyond black-box learning toward interpretable, generalizable, and computationally efficient scientific models. The framework is broadly extensible to other parabolic and hyperbolic PDEs, laying the foundation for scalable physics-aware neural operators across scientific domains.

2. Related Work

Scientific machine learning has emerged as a framework for integrating data-driven models with governing physical laws [14, 4, 24]. A central objective is to reconstruct or predict solutions of partial differential equations (PDEs) from limited observations while preserving physical consistency. Early efforts focused on embedding differential constraints directly into neural network training, leading to the development of Physics-Informed Neural Networks (PINNs) [25, 26]. PINNs incorporate PDE residuals as soft penalties in the loss function and have been successfully applied to diffusion, fluid dynamics, elasticity, and multiphysics systems [14, 21]. However, optimization instability, gradient imbalance, and spectral bias often limit their performance under sparse or noisy supervision [16, 30].

To address representational limitations, implicit neural representations (INRs) have been explored for modeling continuous physical fields [27, 28, 22]. Sinusoidal activation functions [27] and Fourier feature embeddings [28] mitigate spectral bias and enable improved reconstruction of high-frequency components. INR-based approaches have been extended to scientific computing tasks, including PDE solution approximation and super-resolution of physical fields [18]. Despite their expressiveness, pure INRs typically lack explicit physical constraints unless combined with residual regularization.

Parallel developments in operator learning aim to learn mappings between function spaces rather than discrete solutions. DeepONet [20] and its physics-informed variants [13] approximate nonlinear operators from data, while Fourier Neural Operators (FNO) [19] leverage spectral convolution to model global interactions efficiently. These operator-based models scale favorably and have demonstrated strong performance in parametric PDE settings [15]. However, they often require extensive training data and may exhibit reduced robustness in sparse-measurement regimes.

Transformer architectures have recently been introduced into scientific machine learning to capture long-range dependencies in spatiotemporal systems [29, 1]. PINNsFormer [32] extends classical PINNs by incorporating self-attention mechanisms to enhance global feature modeling. Transformer-based PDE solvers and neural operators have also been proposed for fluid simulation and spatiotemporal forecasting [32]. While these approaches improve long-range interaction modeling, they typically enforce physics through residual penalties rather than embedding physical structure directly into the attention mechanism.

Recent works have explored incorporating Transformer architectures into physics-informed learning, including Transformer-based PINNs and attention-based neural operators. In most of these approaches, the attention mechanism itself remains purely data-driven: positional information is typically encoded via learned relative position biases or sinusoidal encodings, while physical laws are enforced only through additional loss terms, such as PDE residual penalties. In contrast, the proposed Physics-Guided Transformer (PGT) embeds physics directly into the attention computation through a kernel-based bias term Γ. Rather than representing a generic distance-dependent bias as in standard relative position encodings, Γ is derived from PDE theory, specifically the heat kernel (Green's function) of the diffusion operator. This formulation encodes both temporal causality and the spatial diffusion structure of the underlying physical process. Consequently, PGT shifts the role of physics from an external regularization term to the attention logits themselves, thereby shaping how contextual information propagates between tokens before the softmax operation. This design fundamentally differs from prior Transformer-based PINN formulations by allowing the attention mechanism to follow physically meaningful interaction patterns rather than learning them solely from data.

Recent studies highlight the importance of architectural inductive biases in scientific learning [2, 23]. Graph neural networks and message-passing frameworks encode conservation laws and locality constraints into model structure [2]. Similarly, symmetry-preserving and equivariant networks incorporate physical priors at the representational level [9, 12]. These works collectively suggest that embedding physics into architecture, rather than solely into the objective function, may improve generalization and optimization stability.

Sparse reconstruction of fluid flows presents additional challenges due to nonlinear convection and pressure–velocity coupling [3, 10]. Classical compressed sensing and reduced-order modeling methods have been widely studied [5, 31], but they often rely on linear subspace assumptions. Data-driven neural approaches provide greater flexibility but must reconcile data fidelity with physical constraints.

Existing approaches largely fall into two categories: residual-based physics enforcement (e.g., PINNs and PINNsFormer) [32] and operator-learning frameworks (e.g., DeepONet and FNO) [12, 20]. While these methods improve either physical regularization or global modeling capacity, they typically treat governing

equations as external constraints added to the loss function. In contrast, Physics-Guided Transformers (PGT) integrate physical structure directly into the attention mechanism itself through a physics-guided bias term. This architectural integration shifts physics from a penalty-based regularizer to an intrinsic component of representational interactions. By coupling physics-guided attention with adaptive implicit decoding, PGT unifies insights from PINNs, neural operators, INRs, and Transformer architectures while explicitly addressing optimization imbalance and physical inconsistency under sparse supervision.

3. Methodology

3.1. Problem Formulation

We consider a time-dependent physical system governed by a partial differential equation (PDE)

$$\mathcal{F}\big(u(x,t), \nabla_x u(x,t), \partial_t u(x,t); \boldsymbol{\theta}_p\big) = f(x,t), \quad (x,t) \in \Omega \times [0,T], \tag{1}$$

where $u(x,t)$ denotes the physical state variable of interest, $\mathcal{F}$ is a differential operator parameterized by physical coefficients $\boldsymbol{\theta}_p$, and $f(x,t)$ is a known source or forcing term. The objective of the Physics-Guided Transformer (PGT) is to learn a continuous mapping

$$u_\Theta : (x,t) \mapsto u(x,t), \tag{2}$$

parameterized by Θ, such that the predicted solution satisfies the governing PDE while remaining consistent with available observations. By modeling the solution as an implicit function of continuous coordinates, PGT enables prediction at arbitrary spatial and temporal locations.

3.2. Overview of the PGT Architecture

PGT combines a physics-guided Transformer encoder with an implicit neural representation decoder. Given a set of sparse or coarse observations, the encoder constructs a latent token representation that captures both local measurements and global system structure. For any query coordinate (x,t), the model retrieves relevant contextual information through a cross-attention mechanism and conditions an implicit decoder to produce the solution value at that coordinate. Figure 1 illustrates the overall architecture.

3.3. Physics-Guided Transformer Encoder

Let $\{(u_i, x_i, t_i)\}_{i=1}^{P}$ denote the available spatiotemporal observations. Each observation is embedded into a latent context token by linearly projecting both the observed value and its coordinate,

$$\mathbf{c}_i = \mathbf{W}_u u_i + \mathbf{W}_p [x_i, t_i] + \mathbf{b}, \tag{3}$$

where $\mathbf{W}_u$ and $\mathbf{W}_p$ are learnable projection matrices. In addition to these context tokens, a learnable global token $\mathbf{c}_{\text{glob}}^{(0)}$ is prepended to the sequence. The resulting token matrix is

$$\mathbf{C}^{(0)} = [\mathbf{c}_{\text{glob}}^{(0)}, \mathbf{c}_1, \ldots, \mathbf{c}_P] \in \mathbb{R}^{(P+1)\times d_{\text{model}}}. \tag{4}$$

A stack of L physics-guided Transformer blocks processes the token matrix. In each block, self-attention is modified by an additive physics-based bias $\boldsymbol{\Gamma}$,

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \boldsymbol{\Gamma}\right)\mathbf{V}, \tag{5}$$

where the bias matrix $\boldsymbol{\Gamma}$ encodes known physical relations between spatiotemporal locations. For diffusion-type systems, $\boldsymbol{\Gamma}$ is derived from the logarithm of the heat kernel and enforces forward temporal influence. The output of each block is updated using residual connections and a feed-forward network,

$$\mathbf{C}^{(l+1)} = \mathbf{C}^{(l)} + \text{Attn}(\text{LN}(\mathbf{C}^{(l)})) + \text{MLP}(\text{LN}(\mathbf{C}^{(l)})). \tag{6}$$

After L layers, the final token matrix $\mathbf{C}^{(L)}$ is split by index into a global token $\mathbf{c}_{\text{glob}}$ and context tokens $\mathbf{C}_{\text{ctx}}$.

3.4. Physics Guided Bias Γ

The physics-guided bias $\mathbf{\Gamma}$ is constructed from the Green's function (fundamental solution) of the governing PDE in Eq. (1). For a pair of context tokens located at spatiotemporal coordinates $(\mathbf{x}_i, t_i)$ and $(\mathbf{x}_j, t_j)$, the bias entry is defined as

$$\Gamma_{ij} \;=\; \log G(\mathbf{x}_i - \mathbf{x}_j,\; t_i - t_j;\; \boldsymbol{\theta}_p)\,, \tag{7}$$

where G is the Green's function of the differential operator $\mathcal{F}$ and $\boldsymbol{\theta}_p$ are the physical parameters introduced in Eq. (1). Taking the logarithm maps the multiplicative kernel structure of G to an additive logit bias, consistent with the pre-softmax additive form in Eq. (5). Entries for which $G = 0$ — such as future tokens in causal problems or off-characteristic tokens in advection-dominated systems — are set to $\Gamma_{ij} = -\infty$, so they receive exactly zero attention weight after the softmax operation.

This formulation is general across PDE families. For *parabolic* systems such as the heat equation ($\partial_t u = \alpha \nabla^2 u$), the Green's function is the Gaussian heat kernel, giving

$$\Gamma_{ij} \;=\; -\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{4\alpha\,\Delta t_{ij}} \;-\; \frac{d}{2}\log(4\pi\alpha\,\Delta t_{ij})\,, \qquad \Delta t_{ij} = t_i - t_j > 0, \tag{8}$$

where d is the spatial dimension and $\alpha > 0$ is the physical diffusivity (e.g. thermal conductivity for the heat equation, kinematic viscosity ν for linearized viscous flow). The bias decays quadratically with spatial distance and logarithmically with elapsed time, encoding both diffusive locality and strict temporal causality. The effective spatial influence radius scales as $\sigma = \sqrt{2\alpha\,\Delta t_{ij}}$, matching the diffusion length of the underlying PDE. For *hyperbolic* systems (e.g. the wave equation $\partial_{tt} u = c^2 \nabla^2 u$), G has compact support on the light cone, so Γ_{ij} is finite only within the causal wavefront $\|\mathbf{x}_i - \mathbf{x}_j\| \le c\,\Delta t_{ij}$, imposing finite-speed propagation directly in the attention computation. For *elliptic* problems (e.g. Poisson or Laplace equations), G depends only on spatial separation and no temporal causal mask is applied. In each case, the physical parameters $\boldsymbol{\theta}_p$ — diffusivity, viscosity, wave speed, or boundary geometry — enter $\mathbf{\Gamma}$ directly from the problem specification, introducing no additional architectural hyperparameters beyond those already present in the governing equation.

In the limiting case where $\boldsymbol{\theta}_p$ drives the kernel toward a uniform distribution (e.g. $\alpha \to \infty$ for diffusion), $\mathbf{\Gamma}$ approaches a constant matrix and its effect on the softmax vanishes, recovering a standard data-driven Transformer. Conversely, as $\alpha \to 0$, the Gaussian narrows toward a Dirac delta, restricting each token to attend only to its immediate spatial neighbor. Because the vanilla Transformer is recovered exactly in the former limit, PGT strictly contains standard attention as a special case: physics-guided attention is a continuously tunable inductive bias that reduces to purely data-driven attention when physical information is absent, and progressively imposes PDE-consistent structure as the governing parameters move toward the diffusion-dominated regime.

3.5. Query Conditioning via Cross-Attention

To evaluate the solution at a query coordinate $\boldsymbol{q} = (x, t)$, the coordinate is first mapped to a latent query embedding using a small multilayer perceptron,

$$\boldsymbol{\phi}(\boldsymbol{q}) = \mathrm{MLP}_q(x, t). \tag{9}$$

The query embedding attends to the encoded context tokens through cross-attention,

$$\boldsymbol{g}(\boldsymbol{q}) = \mathrm{softmax}\left(\frac{(\mathbf{W}_q\boldsymbol{\phi})(\mathbf{W}_k\mathbf{C}_{\mathrm{ctx}})^\top}{\sqrt{d_k}}\right)(\mathbf{W}_v\mathbf{C}_{\mathrm{ctx}}), \tag{10}$$

producing a query-specific context vector that summarizes the most relevant observations for the given location.

3.6. FiLM-Modulated Implicit Decoder

The continuous solution is reconstructed using an implicit neural representation implemented as a sinusoidal representation network (SIREN). The decoder takes the raw query coordinates (x, t) as input and computes

$$\boldsymbol{h}_1 = \sin(\omega_0(\mathbf{W}_0[x, t] + \mathbf{b}_0)). \tag{11}$$

To adapt the decoder to local and global physical context, Feature-wise Linear Modulation (FiLM) is applied at each layer. A hypernetwork conditioned on the query-specific context $\boldsymbol{g}(\boldsymbol{q})$ and the global token $\mathbf{c}_{\text{glob}}$ generates modulation parameters,

$$(\boldsymbol{\alpha}_l, \boldsymbol{\beta}_l, \boldsymbol{\omega}_l) = \mathcal{H}([\boldsymbol{g}(\boldsymbol{q}), \mathbf{c}_{\text{glob}}]), \tag{12}$$

which control amplitude, bias, and frequency at layer l. Each hidden layer is computed as

$$\boldsymbol{h}_{l+1} = \sin\left(\boldsymbol{\omega}_l \odot (\boldsymbol{\alpha}_l \odot (\mathbf{W}_l \boldsymbol{h}_l) + \boldsymbol{\beta}_l)\right). \tag{13}$$

The final prediction is obtained through a linear readout,

$$u_\Theta(x, t) = \mathbf{W}_{\text{out}} \boldsymbol{h}_L + b_{\text{out}}. \tag{14}$$

The overall PGT architecture is illustrated in Figure 1.

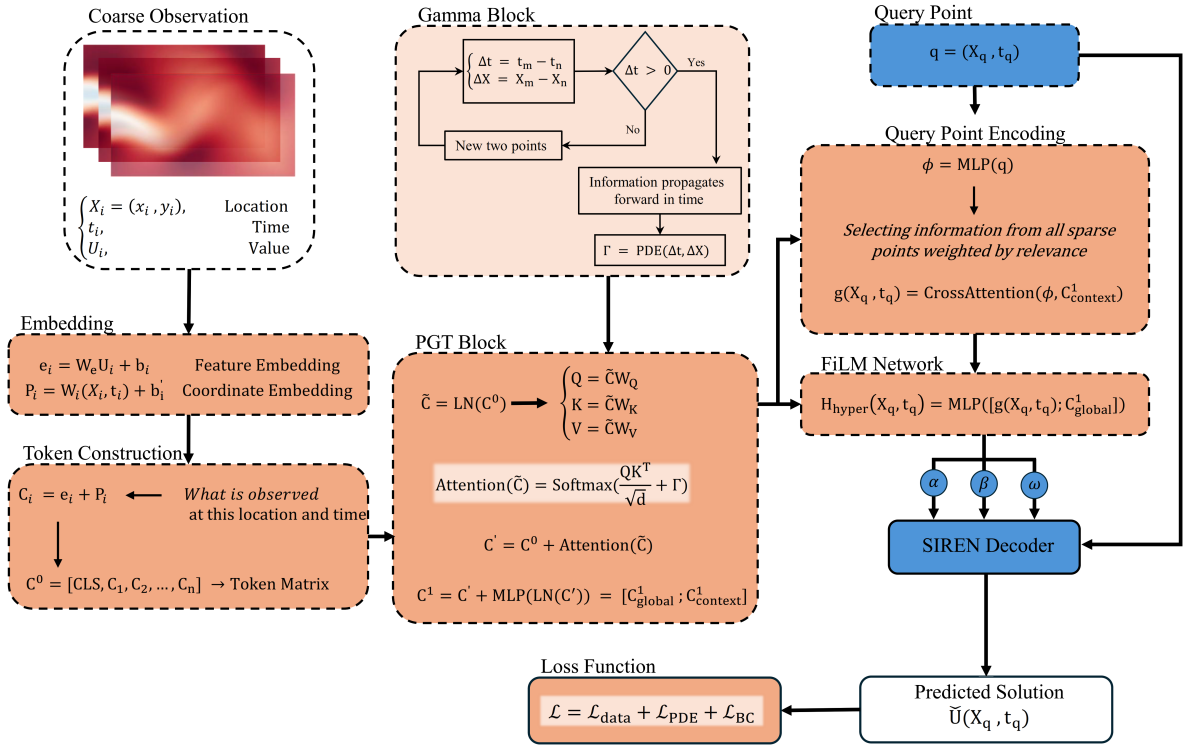


Figure 1: Overview of the PGT architecture. The physics-guided Transformer encoder processes sparse observations into latent context tokens, which are then used to condition the FiLM-modulated SIREN decoder for continuous field reconstruction.

3.7. Training Objective

PGT is trained by minimizing a composite loss that enforces both data fidelity and physical consistency across sources of supervision: observed data, PDE residuals, boundary conditions, and initial conditions.

The **data loss** penalizes discrepancies between predictions and available observations at the sampled spatiotemporal locations,

$$\mathcal{L}_{\text{data}} = \frac{1}{N_d} \sum_{i=1}^{N_d} \|u_\Theta(x_i, t_i) - u_i^{\text{obs}}\|_2^2. \tag{15}$$

Physical consistency is enforced through the **PDE residual loss**, evaluated at N_r randomly sampled collocation points,

$$\mathcal{L}_{\text{PDE}} = \frac{1}{N_r} \sum_{j=1}^{N_r} \|\mathcal{F}(u_\Theta)(x_j, t_j) - f(x_j, t_j)\|_2^2, \tag{16}$$

where $\mathcal{F}$ is the differential operator defined in Eq. (1) and f is the known forcing term. The **boundary condition loss** and **initial condition loss** enforce prescribed constraints on $\partial\Omega$ and at $t = 0$, respectively,

$$\mathcal{L}_{\mathrm{BC}} = \frac{1}{N_b}\sum_{k=1}^{N_b} \|u_\Theta(x_k, t_k) - u_k^{\mathrm{bc}}\|_2^2, \qquad \mathcal{L}_{\mathrm{IC}} = \frac{1}{N_0}\sum_{k=1}^{N_0} \|u_\Theta(x_k, 0) - u_k^{\mathrm{ic}}\|_2^2. \tag{17}$$

These four terms are combined into the total training objective using uncertainty-based weighting [8],

$$\mathcal{L} = \frac{1}{2\sigma_{\mathrm{data}}^2}\mathcal{L}_{\mathrm{data}} + \frac{1}{2\sigma_{\mathrm{PDE}}^2}\mathcal{L}_{\mathrm{PDE}} + \frac{1}{2\sigma_{\mathrm{BC}}^2}\mathcal{L}_{\mathrm{BC}} + \frac{1}{2\sigma_{\mathrm{IC}}^2}\mathcal{L}_{\mathrm{IC}}. \tag{18}$$

Here $\sigma_{\mathrm{data}}, \sigma_{\mathrm{PDE}}, \sigma_{\mathrm{BC}}, \sigma_{\mathrm{IC}} > 0$ are learnable scalar uncertainty parameters, one per loss term. Each weight $1/2\sigma_k^2$) is inversely proportional to the task-specific noise variance σ_k^2: if a supervision signal is noisy or inconsistent, the model learns a larger σ_k, automatically down-weighting that term. All four σ_k are initialized to 1 and optimized jointly with the network parameters Θ via the same gradient-based update step. The formulation eliminates the need for manual loss-weight tuning and allows PGT to adapt its supervision balance automatically as training progresses.

3.8. Generality of Physics-Guided Attention

The physics-guided attention formulation in Eq. (5) is intentionally general. The bias matrix $\mathbf{\Gamma}$ may encode diffusion kernels, transport directionality, or other problem-specific relational priors. By incorporating such a structure directly into the attention logits, PGT embeds domain knowledge into the Transformer encoder while retaining the flexibility and scalability of modern attention-based architectures.

4. Experimental Results

We evaluate the proposed Physics-Guided Transformer (PGT) on two canonical PDE-governed systems: (1) sparse scattered-data reconstruction for the 1D heat diffusion equation, and (2) reconstruction of 2D velocity and pressure fields governed by the incompressible Navier–Stokes equations. Across both tasks, PGT is compared with state-of-the-art baselines, including FNO, PINN, PI-DeepONet, PINNsFormer, SIREN, and WIRE, under matched sparse-sampling budgets.

4.1. 1D Heat Equation: Sparse Reconstruction Analysis

We first evaluate PGT on the 1D heat equation:

$$\partial_t u - \nu\,\partial_{xx} u = 0, \quad x \in [0, 1],\ t \in [0, 1], \tag{19}$$

with sinusoidal initial conditions

$$u(x, 0) = \sin(n\pi x), \qquad u(x, t) = e^{-\nu(n\pi)^2 t}\sin(n\pi x). \tag{20}$$

The task is sparse reconstruction: given only M randomly sampled spatiotemporal observations, the model must reconstruct the full solution field. We vary the number of sparse samples ($M = 100, 200, 500$) to study robustness under different supervision levels.

Table 1 reports the data loss, PDE residual loss, relative L^2 error, computational cost (FLOPs), training time, and parameter count for SIREN, PINN, and PGT. PGT achieves dramatically lower data error and relative L^2 error across all different observation data points, M. For example, at $M = 100$, PGT reduces the relative L^2 error to 5.90×10^{-3}, compared to 2.26×10^{-1} for PINN and 5.40×10^{-1} for SIREN. The PGT error reduction is approximately a **38× improvement over PINN** and nearly **90× improvement over SIREN**. Even as the number of sparse points increases to $M = 500$, PGT maintains relative errors on the order of 10^{-3}, while PINN and SIREN remain two orders of magnitude higher.

Although PGT's PDE residual loss remains around 6.6×10^{-2}, its significantly smaller data loss indicates superior global field reconstruction accuracy. SIREN exhibits relatively high PDE residuals due to the absence of explicit physics enforcement. PINN reduces the PDE residual compared to SIREN, but struggles to achieve comparable reconstruction fidelity under sparse supervision.

Table 1: Comparison of SIREN, PINN, and PGT models across different values of M.

Model		Loss			FLOPs	Train time	Param
		Data	PDE	Rel L^2	(G)		
SIREN	$M = 100$	0.063	0.18	0.54	6.5	126 s	**26,419**
	$M = 200$	0.045	0.14	0.45	6.8	127 s	
	$M = 500$	0.025	0.18	0.34	7.5	135 s	
PINN	$M = 100$	0.026	0.12	0.226	**1.69**	**65 s**	66,500
	$M = 200$	0.024	0.19	0.332	1.75	65.9 s	
	$M = 500$	0.021	0.13	0.313	1.94	66.8 s	
PGT(ours)	$M = 100$	0.000076	0.066	0.0059	116	9.5 min	4.05E+08
	$M = 200$	0.000029	**0.066**	0.0028	132	13 min	
	$M = 500$	**0.000017**	0.067	**0.0026**	190	27.5 min	

As the number of sparse observations increases, all models benefit from additional supervision. However, the improvement for PGT is substantially more stable and consistent. Its relative L^2 error decreases from 5.90×10^{-3} at $M = 100$ to 2.60×10^{-3} at $M = 500$, demonstrating robustness even in low-data regimes. In contrast, PINN and SIREN show less consistent trends and remain significantly less accurate.

The improved accuracy of PGT comes at a higher computational cost. While PINN requires approximately 1.7–1.9 GFLOPs and about 65 seconds of training time, PGT requires 116–190 GFLOPs and up to 1.67×10^3 seconds of training time. The parameter count is also substantially larger for PGT. The trade-off highlights that PGT prioritizes reconstruction fidelity and physics-guided generalization over lightweight deployment.

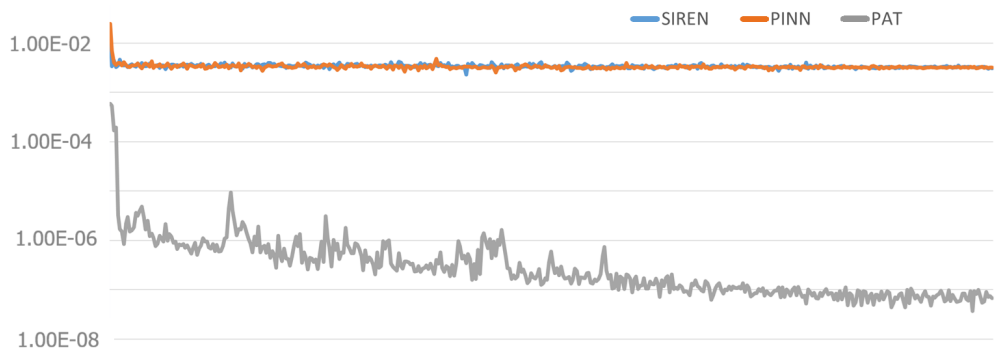
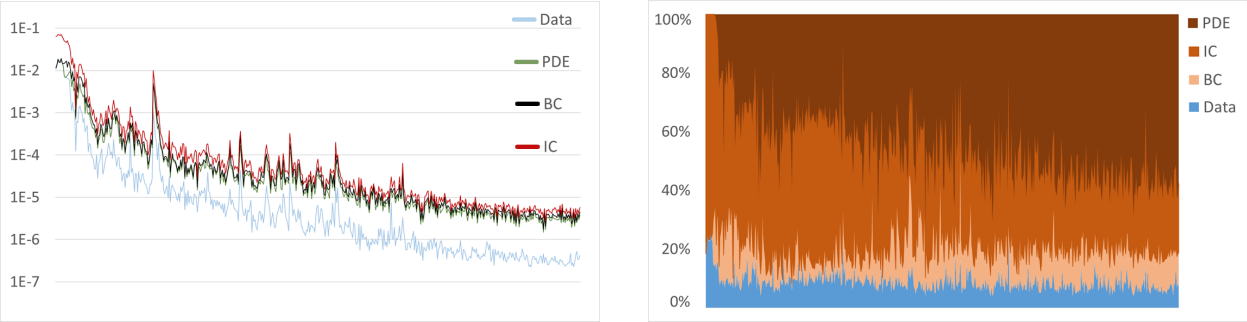


Figure 2: Training error convergence (relative L^2) for PINN, SIREN, and PGT on the 1D heat diffusion sparse reconstruction task ($M = 100$ observations). PGT exhibits sustained monotonic decay, whereas PINN and SIREN plateau at much higher error levels.

Figure 2 illustrates the evolution of the reconstruction error during training. Both SIREN and PINN show rapid initial error reduction, followed by early stagnation at relatively high error levels around 10^{-3}. In contrast, PGT exhibits continuous, sustained error decay throughout training, ultimately reaching errors on the order of 10^{-7}. Thus, the PGT avoids the optimization plateaus observed in the baseline methods and converges to a significantly more accurate solution.

To further analyze the training dynamics, Figure 3 (loss component contribution plot) shows the relative contributions of the data, PDE, boundary-condition (BC), and initial-condition (IC) losses within PGT. In the early stages of training, the IC loss contributes substantially, reflecting the model's initial effort to satisfy the prescribed initial state. As training progresses, the relative contribution of the PDE loss increases, indicating a gradual shift toward enforcing the governing physical dynamics across the domain. Notably, the contribution of the data loss remains comparatively stable throughout training. The balanced evolution suggests that PGT effectively coordinates data fitting and physics enforcement, progressively refining the

solution while maintaining consistency with both observations and governing equations.



(a) (b)

Figure 3: (a) Error components breakdown. (b) Error components contribution analysis.

Overall, the 1D experiment demonstrates that incorporating physics-guided attention and adaptive frequency modulation significantly enhances sparse reconstruction accuracy. The controlled setting validates the effectiveness of PGT before extending the comparison to more complex 2D nonlinear systems.

4.2. 2D Navier–Stokes Reconstruction Under Sparse Measurements

We next evaluate PGT on the two-dimensional incompressible Navier–Stokes equations governing viscous flow:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = -\frac{\partial p}{\partial x} + \nu\nabla^2 u, \tag{21}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} = -\frac{\partial p}{\partial y} + \nu\nabla^2 v, \tag{22}$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \tag{23}$$

where (u, v) denote the velocity components, p is the pressure field, and $\nu = 1/\mathrm{Re}$ is the kinematic viscosity. We consider the canonical cylinder wake dataset and perform sparse reconstruction from $N_{\text{train}} = 1500$ randomly sampled spatiotemporal points. All models are evaluated on a full-resolution spatial snapshot at a fixed time index.

Table 2 reports the final quantitative comparison across all seven methods, including total relative L^2 error, variable-wise errors for u, v, and p, the PDE residual, and computational cost.

Table 2: Sparse reconstruction performance for the 2D Navier–Stokes cylinder wake at a fixed time snapshot ($N_{\text{train}} = 1500$). Best results are shown in **bold**. Architecture family groups methods: operator-learning (FNO, PI-DeepONet), Transformer-based (PINNsFormer), implicit representation (SIREN, WIRE), classical PINN, and the proposed PGT.

Model	Params (M)	FLOPs (G)	Train Time	Rel-L^2	Rel-$L^2(u)$	Rel-$L^2(v)$	Rel-$L^2(p)$	PDE Residual	Train Error
FNO	2.360	0.092	4.41 s	0.710	0.2200	1.4900	0.770	4.70×10^{-2}	4.20×10^{-4}
PI-DeepONet	0.215	176.12	36 s	0.095	0.0600	0.2300	0.210	4.20×10^{-1}	3.90×10^{-3}
PINNsFormer	0.545	13.200	296 min	0.080	0.0400	0.0900	0.110	8.30×10^{-2}	8.30×10^{-4}
SIREN	0.264	1.310	2.5 min	0.690	0.3100	1.0000	0.770	3.90×10^{-4}	5.50×10^{-2}
WIRE	0.528	5.200	6.3 min	0.018	0.0041	0.0350	0.014	5.10×10^{-1}	1.54×10^{-4}
PINN	0.336	1.600	2.6 min	0.110	0.0400	0.1400	0.160	8.40×10^{-4}	1.19×10^{-3}
PGT (ours)	**5.630**	**32.860**	**47 min**	**0.034**	**0.0160**	**0.0410**	**0.046**	$\mathbf{8.30 \times 10^{-4}}$	$\mathbf{6.50 \times 10^{-5}}$

The results reveal important differences across model families. Among the operator-learning methods, FNO achieves the lowest computational cost but yields the highest reconstruction error, with a relative L^2 error of 1.49 for the vertical-velocity component, indicating a near-complete failure to capture vortex-shedding dynamics under sparse sampling. PI-DeepONet substantially improves reconstruction accuracy, yet

its PDE residual remains large (4.2×10^{-1}), indicating that the predicted fields do not satisfy the governing momentum equations with adequate precision.

Among implicit representation baselines, SIREN achieves a low PDE residual (3.9×10^{-4}). Still, it exhibits very poor field reconstruction accuracy, with a relative L^2 error of 0.69 and a vertical velocity error of 1.00, reflecting the absence of explicit physics enforcement in the reconstruction process. WIRE presents an opposing behaviour: it attains the lowest overall relative L^2 error (0.018). It excels on the horizontal velocity (0.0041) and pressure (0.014) components. Yet, its PDE residual (5.1×10^{-1}) is the highest among all methods, revealing that accurate data fitting does not guarantee satisfaction of the governing equations. The classical PINN achieves a PDE residual of 8.4×10^{-4}, competitive with PGT, but its reconstruction accuracy (overall relative $L^2 = 0.11$) is substantially inferior, particularly for the vertical velocity and pressure fields.

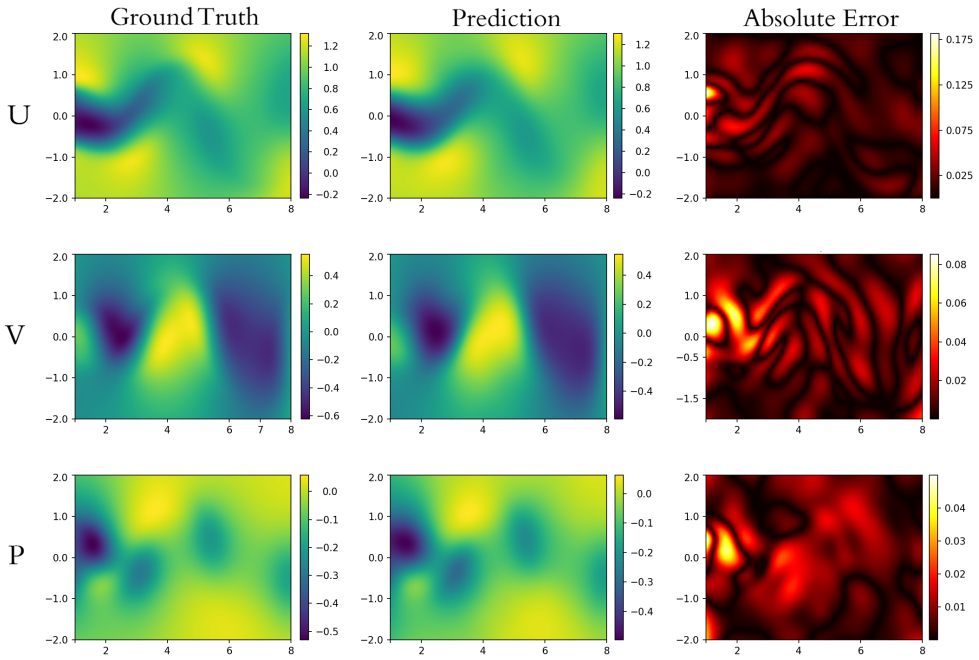


Figure 4: Qualitative comparison between ground truth and PGT reconstruction for the 2D Navier–Stokes problem. Rows correspond to u, v, and p fields. Columns show ground truth, PGT prediction, and absolute error.

PGT achieves the strongest simultaneous balance between reconstruction fidelity and physical consistency across all evaluated methods. Its overall relative L^2 error of 0.034 is surpassed only by WIRE (0.018). Yet, PGT delivers a PDE residual of 8.3×10^{-4} — comparable to PINN and SIREN in residual magnitude, but accompanied by significantly superior reconstruction accuracy. Variable-wise errors remain consistently low across u (0.016), v (0.041), and p (0.046), indicating balanced coupling between the velocity and pressure components, a property not achieved by any individual baseline. The combination — competitive data fidelity with rigorous physical consistency — distinguishes PGT from all comparators and is precisely the regime that matters for scientific reconstruction tasks.

Figure 4 illustrates the qualitative reconstruction of u, v, and p. The predicted fields faithfully reproduce the dominant vortex shedding structures and pressure gradients present in the ground truth. Absolute error maps confirm that residuals are primarily confined to regions of strong nonlinear interaction and high vorticity, while the global flow topology is accurately preserved throughout the domain.

The training convergence behavior is shown in Figure 5. PGT exhibits stable, monotonically decreasing error throughout optimization. In contrast, several baselines undergo rapid initial decay followed by early stagnation, a hallmark of optimization imbalance under sparse supervision. PGT's sustained convergence reflects the stabilizing effect of embedding physics directly within the attention mechanism, which continuously biases the model toward physically plausible solutions rather than relying on competing loss terms.

Collectively, the 2D results demonstrate that PGT uniquely addresses the fundamental tension between

data fidelity and physical consistency that afflicts all baseline approaches: pure data-fitting methods (WIRE, SIREN) achieve low reconstruction error at the expense of PDE compliance, while residual-based methods (PINN, PINNsFormer) enforce physics at the cost of reconstruction accuracy. By embedding physical priors architecturally rather than as an external penalty, PGT sidesteps this trade-off and achieves strong performance on both axes simultaneously.

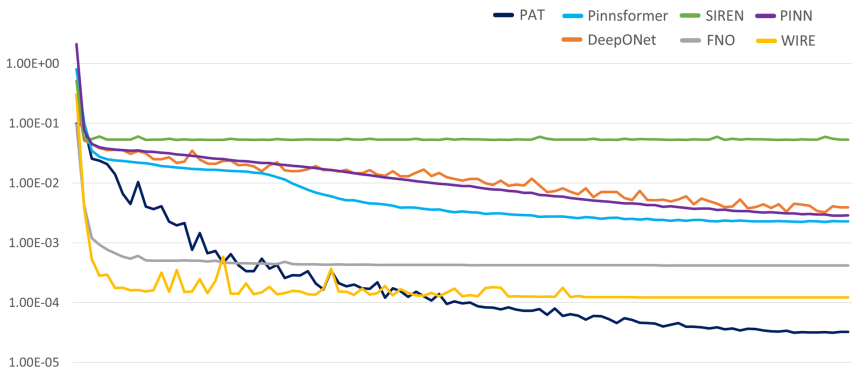


Figure 5: Training error convergence for PGT and baseline methods on the 2D Navier–Stokes problem.

4.3. Ablation Study

To rigorously attribute PGT's performance to its individual design choices, we conduct a comprehensive ablation study on the 2D Navier–Stokes cylinder wake task ($N_{\text{train}} = 1500$). The study probes three independent axes of the architecture: (i) the physics-guided attention bias $\mathbf{\Gamma}$ and the PDE residual loss $\mathcal{L}_{\text{PDE}}$, which together constitute the physics-integration mechanisms in the encoder and training objective; and (ii) the decoder design, which isolates the contribution of sinusoidal activations and FiLM-based context conditioning within the implicit reconstruction head. All seven configurations share the same encoder depth, embedding dimension, optimizer, and training budget, and all retain the data loss $\mathcal{L}_{\text{data}}$ throughout. Results are summarized in Table 3.

Table 3: Ablation study on the 2D Navier–Stokes cylinder wake ($N_{\text{train}} = 1500$). *Attn. bias* $\mathbf{\Gamma}$: heat-kernel additive bias in self-attention logits (disabled in pure-transformer variants). *PDE loss* $\mathcal{L}_{PDE}$: momentum and continuity residual terms. *Decoder*: FiLM-SIREN is the full decoder; SIREN (no FiLM) uses a plain SIREN with context concatenated at the input; MLP replaces sinusoidal activations with GELU; FiLM-MLP retains FiLM conditioning but uses GELU activations. ✓ active; ○ disabled. Data loss is active in all rows. Best results in **bold**.

Variant	Decoder	Attn. bias $\mathbf{\Gamma}$	PDE loss $\mathcal{L}_{\text{PDE}}$	Reconstruction Error ↓	PDE Residual ↓
PGT (full)	FiLM-SIREN	✓	✓	$\mathbf{6.50 \times 10^{-5}}$	$\mathbf{8.30 \times 10^{-4}}$
No PDE loss	FiLM-SIREN	✓	○	8.80×10^{-5}	2.90×10^{-3}
No attention bias	FiLM-SIREN	○	✓	3.00×10^{-4}	1.30×10^{-2}
No physics (data only)	FiLM-SIREN	○	○	3.60×10^{-4}	3.50×10^{-2}
SIREN, no FiLM	SIREN	✓	✓	1.83×10^{-4}	1.10×10^{-3}
FiLM-MLP	FiLM-MLP	✓	✓	1.21×10^{-4}	1.05×10^{-3}
Plain MLP	MLP	✓	✓	3.12×10^{-4}	1.42×10^{-3}

The results reveal three distinct and complementary insights into the architectural design of PGT.

The physics-guided attention bias $\mathbf{\Gamma}$ *is the dominant driver of reconstruction accuracy..* Comparing variants that differ only in whether $\mathbf{\Gamma}$ is active, its effect is consistent and substantial. Without the physics loss ("no PDE loss" vs. "no physics"), activating $\mathbf{\Gamma}$ reduces the reconstruction error by approximately 4× — from 3.60×10^{-4} to 8.80×10^{-5} — with no change to the training objective. This gain arises purely from the structural inductive bias introduced at the attention level: by encoding the heat-kernel Green's function into the attention logits, $\mathbf{\Gamma}$ directs the encoder to aggregate context tokens according to physically meaningful

spatiotemporal proximity, enabling more coherent field reconstruction even when no explicit PDE penalty is applied. The same pattern holds when the physics loss is active ("no attention bias" vs. "PGT full"): enabling $\mathbf{\Gamma}$ reduces the reconstruction error from 3.00×10^{-4} to 6.50×10^{-5}, a further 4.6× improvement. Across both conditions, $\mathbf{\Gamma}$ is the single most impactful design choice for reconstruction fidelity.

The PDE loss $\mathcal{L}_{PDE}$ is essential for governing-equation compliance.. When $\mathbf{\Gamma}$ is active but $\mathcal{L}_{\text{PDE}}$ is withheld ("no PDE loss"), the model achieves strong reconstruction accuracy yet yields a PDE residual of 2.90×10^{-3} — approximately 3.5× larger than the full model. Introducing $\mathcal{L}_{\text{PDE}}$ reduces the residual to 8.30×10^{-4}, confirming that explicit residual supervision at collocation points is necessary to enforce momentum conservation and incompressibility throughout the domain. This effect also appears in the absence of $\mathbf{\Gamma}$: adding the physics loss ("no attention bias" vs. "no physics") reduces the residual from 3.50×10^{-2} to 1.30×10^{-2}, a 2.7× reduction. However, the residual achieved by $\mathcal{L}_{\text{PDE}}$ alone without $\mathbf{\Gamma}$ (1.30×10^{-2}) remains more than an order of magnitude larger than that of the full model (8.30×10^{-4}), confirming that the two mechanisms are complementary rather than substitutable. Structural physics priors shape the representation. Explicit residual supervision enforces pointwise compliance with the PDE.

FiLM conditioning and sinusoidal activations both contribute to decoder quality, and their combination is necessary for optimal performance.. The bottom three rows of Table 3 isolate the decoder design while holding the encoder ($\mathbf{\Gamma}$ active) and training objective ($\mathcal{L}_{\text{PDE}}$ active) fixed.

Removing FiLM while keeping sinusoidal activations (SIREN, no FiLM) raises the reconstruction error from 6.50×10^{-5} to 1.83×10^{-4}, a degradation of 2.8×, and modestly worsens the PDE residual to 1.10×10^{-3}. Without FiLM, the decoder cannot adapt its frequency response to the local physical context inferred by the cross-attention step. Instead, it receives the entire context vector only at the first layer, forcing a single fixed set of sinusoidal frequencies to represent all query locations equally. The resulting model still benefits from the periodic activation's spectral properties but loses the per-query modulation that allows PGT to resolve fine vortex-shedding structures and pressure gradients with varying spatial complexity.

Replacing sinusoidal activations with GELU while retaining FiLM (FiLM-MLP) yields a reconstruction error of 1.21×10^{-4} and a PDE residual of 1.05×10^{-3}. This is noticeably better than the SIREN-no-FiLM variant, indicating that deep, multilayer context conditioning contributes more to reconstruction quality than the choice of activation function alone. Nevertheless, the gap with the full FiLM-SIREN model (6.50×10^{-5}) confirms that GELU activations are insufficient to represent the smooth, oscillatory solution fields typical of convection-dominated flows. The absence of periodic activations prevents the decoder from efficiently encoding the spectral content present in the cylinder wake — a limitation that FiLM conditioning alone cannot overcome.

Removing both FiLM and sinusoidal activations (plain MLP) produces the weakest decoder: reconstruction error climbs to 3.12×10^{-4} and the PDE residual reaches 1.42×10^{-3}. The reconstruction error of this variant approaches that of the no-attention-bias variant (3.00×10^{-4}), suggesting that a suboptimal decoder can negate the representational benefit of physics-guided attention in the encoder.

Together, these decoder results establish a clear performance hierarchy: FiLM-SIREN > FiLM-MLP > SIREN (no FiLM) > plain MLP. The hierarchy reveals that FiLM conditioning is the more critical of the two mechanisms: Its removal costs 2.8× in reconstruction accuracy, whereas swapping sinusoidal activations for GELU cost only 1.9×. Yet the combination uniquely achieves the lowest error, confirming that both properties — periodic spectral representation and adaptive context-driven modulation — are independently beneficial and mutually reinforcing.

The full model uniquely achieves strong performance on both evaluation axes simultaneously.. Across all seven configurations, only the full PGT variant simultaneously attains the lowest reconstruction error (6.50×10^{-5}) and the lowest PDE residual (8.30×10^{-4}). Disabling $\mathbf{\Gamma}$ primarily degrades reconstruction accuracy; disabling $\mathcal{L}_{\text{PDE}}$ primarily degrades PDE compliance; degrading the decoder hurts both metrics in proportion to the severity of the simplification. The four-way ablation traces a clear Pareto front in the accuracy–consistency space: no simplified variant achieves the Pareto frontier of the full model. These findings provide mechanistic validation of every major design choice in PGT — physics-guided attention, explicit PDE supervision, sinusoidal implicit decoding, and FiLM-based context conditioning — and confirm that all four are necessary components of the architecture.

4.4. Robustness to Measurement Noise

Sensor noise is unavoidable in experimental flow measurement. To assess its impact, we corrupt the u and v context observations with additive Gaussian noise at six relative levels, $u_i^{\text{noisy}} = u_i + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $\sigma = \eta \cdot \text{std}(u_{\text{train}})$, for $\eta \in \{0, 0.01, 0.02, 0.05, 0.10, 0.20\}$ (clean to 20% of signal std). We compare standard PGT (MSE data loss) against a variant, PGT-UW, that replaces the data loss with a heteroscedastic uncertainty-weighted negative log-likelihood,

$$\mathcal{L}_{\text{data}}^{\text{UW}} = \frac{1}{N_d} \sum_{i=1}^{N_d} \left[\log \sigma_i^2 + \frac{\left\| u_\Theta(x_i, t_i) - u_i^{\text{obs}} \right\|^2}{\sigma_i^2} \right], \tag{24}$$

where σ_i^2 is a per-token aleatoric variance predicted by a small auxiliary head on the cross-attention output. All other components are identical between the two variants. Results are reported in Table 4 and Figure 6.

Table 4: Noise robustness on the 2D Navier–Stokes cylinder wake ($\sigma = \eta \cdot \text{std}(u_{\text{train}})$). PGT uses a standard MSE data loss; PGT-UW uses a heteroscedastic uncertainty-weighted loss. Best per row in **bold**.

	Avg. Rel-ℓ_2 (u, v) ↓		PDE Residual ↓	
η	PGT	PGT-UW	PGT	PGT-UW
0.00	0.0496	**0.0410**	**1.61e-03**	4.55e-03
0.01	**0.0634**	0.0861	**1.99e-03**	6.80e-03
0.02	**0.0406**	0.1054	**1.86e-03**	7.78e-03
0.05	**0.0508**	0.1102	**1.49e-03**	7.63e-03
0.10	**0.0512**	0.1551	**1.92e-03**	7.74e-03
0.20	0.0607	**0.0400**	**1.82e-03**	4.64e-03

Standard PGT is *remarkably stable*: the average Rel-ℓ_2 error for u and v remains within [0.040, 0.064] across all noise levels, and the PDE residual stays in the narrow band $[1.5, 2.0] \times 10^{-3}$ throughout. This robustness is a direct consequence of the physics-guided attention bias $\mathbf{\Gamma}$: by continuously anchoring internal representations to the heat-kernel Green's function, the architectural prior prevents noise from propagating into the reconstructed fields, acting as an implicit regularizer independent of the data loss.

PGT-UW, by contrast, exhibits a non-monotone trajectory: error rises steeply for intermediate noise levels ($\eta = 0.01$–0.10), peaking at 0.155 before partially recovering at $\eta = 0.20$. PDE residuals follow the same pattern, remaining 3–4× higher than standard PGT across all intermediate levels. The heteroscedastic head introduces optimisation complexity that the fixed training budget cannot fully resolve; at very high noise, the large predicted variances effectively suppress the data loss, allowing the physics loss to compensate. These findings indicate that when a strong architectural physics prior is present, an explicit aleatoric uncertainty mechanism offers limited benefit and may destabilise training. Standard PGT is therefore the recommended configuration for noisy measurement scenarios.

5. Discussion

This study demonstrates that embedding physical structure directly into the neural architecture, rather than solely as an external loss penalty, can substantially improve the reconstruction of nonlinear dynamical systems from sparse observations. Across both diffusion-dominated and convection-dominated regimes, PGT achieved strong physical consistency while maintaining competitive reconstruction accuracy—a combination that none of the individual baseline methods achieved simultaneously.

The expanded 2D Navier–Stokes comparison reveals a clear structural dichotomy among existing methods. Pure data-fitting approaches (WIRE, SIREN) achieve low reconstruction error but exhibit large PDE residuals, confirming that accurate interpolation of sparse observations does not guarantee satisfaction of the underlying governing equations. Conversely, residual-based methods (PINN, PINNsFormer) reduce the PDE residual at the cost of reconstruction accuracy, particularly for the vertical velocity and pressure fields where the effects of pressure–velocity coupling are most pronounced. Operator-learning methods (FNO, PI-DeepONet) occupy a middle ground but fail to excel on either axis under the sparse measurement budget considered here.

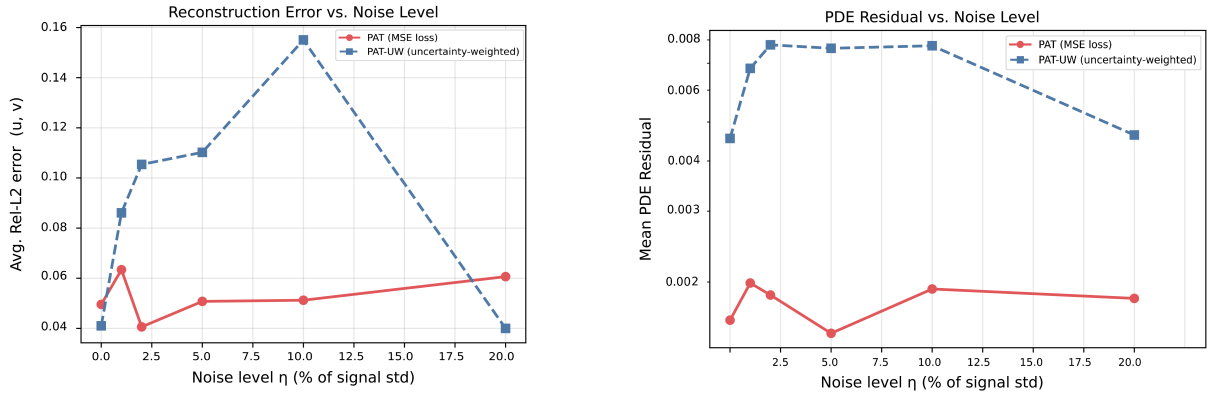


Figure 6: Reconstruction error (left) and PDE residual (right) as a function of noise level η for PGT and PGT-UW on the 2D Navier–Stokes cylinder-wake problem.

PGT sidesteps this trade-off by integrating physical priors at the representational level. The heat-kernel–derived attention bias continuously steers the model toward physically plausible solutions during optimization, without competing with data-fidelity terms as penalty-weighted losses do. This architectural constraint appears to stabilize training and promote globally coherent solutions, especially in nonlinear flow regimes where pressure–velocity coupling is critical. The result is a model that simultaneously achieves reconstruction accuracy approaching that of the best data-fitting baseline (WIRE) and a PDE residual comparable to that of the best residual-based baseline (PINN). This outcome is structurally difficult to achieve through loss reweighting alone.

The ablation study provides direct mechanistic evidence for this claim. Removing the physics-guided attention bias $\mathbf{\Gamma}$ while retaining the PDE loss degrades reconstruction error by more than an order of magnitude, confirming that the architectural bias — not the loss term — is the primary mechanism responsible for accurate field recovery. Conversely, removing the PDE loss while retaining $\mathbf{\Gamma}$ leaves reconstruction accuracy largely intact but causes the PDE residual to increase nearly sevenfold, demonstrating that explicit residual supervision remains necessary for governing-equation compliance. These two mechanisms are therefore non-redundant: $\mathbf{\Gamma}$ shapes the latent representation toward physically coherent solutions, while $\mathcal{L}_{\text{PDE}}$ enforces pointwise satisfaction of the differential constraints. Their combination uniquely achieves strong performance on both evaluation axes.

PGT is deliberately more expensive than lightweight baselines such as PINN, SIREN, and FNO, as the quadratic self-attention over context tokens and the FiLM hypernetworks are precisely the mechanisms that enable global propagation of sparse observations and adaptive spectral decoding — capabilities that cheaper architectures forgo, at the cost of the large reconstruction and PDE-residual errors documented in Tables 1 and 2. Viewed against physics-aware methods of comparable ambition, however, PGT is competitive: PINNsFormer trains substantially longer yet yields higher reconstruction error, and PI-DeepONet consumes far greater FLOPs while producing a PDE residual orders of magnitude larger. The current cost is therefore an engineering constraint of the prototype rather than a fundamental limit of physics-guided attention. The spatial locality of $\mathbf{\Gamma}$ — which decays rapidly beyond a physically determined radius — motivates sparse or hierarchical attention that would reduce complexity from $\mathcal{O}(P^2)$ toward $\mathcal{O}(P \log P)$; low-rank factorisation of the Gaussian bias matrix and mixed-precision training offer further reductions with no change to the underlying physics prior.

It is worth noting that sparse reconstruction differs fundamentally from super-resolution: whereas super-resolution operates on a dense low-resolution grid with uniform spatial coverage, sparse reconstruction must recover continuous fields from scattered, unstructured samples, leaving large portions of the domain entirely unobserved. This more ill-posed setting makes architectural inductive biases and physics-based constraints correspondingly more critical, precisely the regime that PGT targets.

More broadly, these findings suggest that future scientific machine learning models may benefit from embedding governing principles at the representational level rather than relying solely on residual regularization.

Extending this framework to higher-dimensional, multi-physics, and turbulent regimes — and quantifying its behaviour under varying Reynolds numbers and noise levels — remains an important direction for future research.

6. Conclusion

We have introduced a Physics-Guided Transformer (PGT) for reconstructing partial differential equation–governed systems from sparse observations. By embedding physical structure directly into the attention mechanism and coupling it with an adaptive implicit decoder, the proposed framework moves beyond residual-only physics enforcement. Instead, it incorporates governing principles at the architectural level.

Across diffusion and nonlinear flow problems, PGT achieved solutions that were both numerically accurate and strongly consistent with the underlying equations. A controlled ablation study further confirmed that the physics-guided attention bias $\mathbf{\Gamma}$ and the PDE residual loss $\mathcal{L}_{\mathrm{PDE}}$ operate through distinct and complementary pathways: $\mathbf{\Gamma}$ is the primary driver of reconstruction accuracy, while $\mathcal{L}_{\mathrm{PDE}}$ is essential for governing-equation compliance. Only their combination simultaneously minimizes both objectives, providing mechanistic validation of the architectural design choices.

While the approach incurs a higher computational cost than lightweight operator-learning methods, it offers a meaningful trade-off between efficiency and physical fidelity. More broadly, this work points toward a shift in scientific machine learning: from treating governing equations as external constraints to incorporating them as intrinsic components of model design. Extending such physics-guided attention mechanisms to higher-dimensional, multiscale, and multiphysics systems may provide a path to reliable and interpretable data-driven modeling in science and engineering.

References

[1] Alexey Dosovitskiy, L.B., et. al., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. URL: https://arxiv.org/abs/2010.11929, arXiv:2010.11929.

[2] Brandstetter, J., Worrall, D., Welling, M., 2023. Message passing neural pde solvers. URL: https://arxiv.org/abs/2202.03376, arXiv:2202.03376.

[3] Brunton, B.W., Brunton, S.L., Proctor, J.L., Kutz, J.N., 2016. Sparse sensor placement optimization for classification. SIAM Journal on Applied Mathematics 76, 2099–2122. doi:10.1137/15M1036713.

[4] Brunton, S.L., Noack, B.R., Koumoutsakos, P., 2020. Machine learning for fluid mechanics. Annual Review of Fluid Mechanics 52, 477–508.

[5] Candès, E.J., Romberg, J., Tao, T., 2006. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory 52, 489–509.

[6] Canuto, C., Quarteroni, A., 2017. Spectral Methods. John Wiley and Sons, Ltd. chapter 5. pp. 1–16. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119176817.ecm2003m, doi:https://doi.org/10.1002/9781119176817.ecm2003m, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119176817.ecm2003m.

[7] Cao, S., 2021. Choose a transformer: Fourier or galerkin. arXiv:2105.14995.

[8] Cipolla, R., Gal, Y., Kendall, A., 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: Proceedings of the IEEE Conference on Computer Vision and Pattern, pp. 7482–7491. doi:10.1109/CVPR.2018.00781.

[9] Cohen, T.S., Welling, M., 2016. Group equivariant convolutional networks. arXiv:1602.07576.

[10] Erichson, N.B., Mathelin, L., Yao, Z., Brunton, S.L., Mahoney, M.W., Kutz, J.N., 2020. Shallow neural networks for fluid flow reconstruction with limited sensors. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 476, 20200097. URL: https://doi.org/10.1098/rspa.2020.0097, doi:10.1098/rspa.2020.0097, arXiv:https://royalsocietypublishing.org/rspa/article-pdf/doi/10.1098/rspa.2020.0097/637169/rspa.202

[11] Evans, L.C., 2010. Partial Differential Equations. volume 19 of *Graduate Studies in Mathematics*. 2 ed., American Mathematical Society, Providence, RI.

[12] Finzi, M., Stanton, S., Izmailov, P., Wilson, A.G., 2020. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data, in: Proceedings of the 37th International Conference on Machine Learning, PMLR. pp. 3165–3176. URL: https://proceedings.mlr.press/v119/finzi20a.html.

[13] Goswami, S., Bora, A., Yu, Y., Karniadakis, G.E., 2022. Physics-informed deep neural operator networks. URL: https://arxiv.org/abs/2207.05748, arXiv:2207.05748.

[14] Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L., 2021. Physics-informed machine learning. Nature Reviews Physics 3, 422–440.

[15] Kovachki, N., Li, Z., Liu, B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., Anandkumar, A., 2023. Neural operator: learning maps between function spaces with applications to pdes. The Journal of Machine Learning Research 24.

[16] Krishnapriyan, A., Gholami, A., Zhe, S., Kirby, R., Mahoney, M.W., 2021. Characterizing possible failure modes in physics-informed neural networks, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 26548–26560.

[17] LeVeque, R.J., 2007. Finite Difference Methods for Ordinary and Partial Differential Equations. Society for Industrial and Applied Mathematics. URL: https://epubs.siam.org/doi/abs/10.1137/1.9780898717839, doi:10.1137/1.9780898717839, arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9780898717839.

[18] Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., Anandkumar, A., 2020a. Neural operator: Graph kernel network for partial differential equations. URL: https://arxiv.org/abs/2003.03485, arXiv:2003.03485.

[19] Li, Z.Y., Kovachki, N.B., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A.M., Anandkumar, A., 2020b. Fourier neural operator for parametric partial differential equations. ArXiv abs/2010.08895. URL: https://api.semanticscholar.org/CorpusID:224705257.

[20] Lu, L., Jin, P., Pang, G., Zhang, Z., Karniadakis, G.E., 2021a. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. Nature Machine Intelligence 3, 218–229.

[21] Lu, L., Meng, X., Mao, Z., Karniadakis, G.E., 2021b. Deepxde: A deep learning library for solving differential equations. SIAM Review 63, 208–228. doi:10.1137/19M1274067, arXiv:https://doi.org/10.1137/19M1274067.

[22] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R., 2020. Nerf: Representing scenes as neural radiance fields for view synthesis, in: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (Eds.), Computer Vision – ECCV 2020, Springer, Cham. pp. 405–421.

[23] Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A., Battaglia, P.W., 2021. Learning mesh-based simulation with graph networks. URL: https://arxiv.org/abs/2010.03409, arXiv:2010.03409.

[24] Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., Skinner, D., Ramadhan, A., Edelman, A., 2021. Universal differential equations for scientific machine learning. URL: https://arxiv.org/abs/2001.04385, arXiv:2001.04385.

[25] Raissi, M., Perdikaris, P., Karniadakis, G., 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. URL: https://www.sciencedirect.com/science/article/pii/S0021999118307125, doi:https://doi.org/10.1016/j.jcp.2018.10.045.

[26] Sirignano, J., Spiliopoulos, K., 2018. Dgm: A deep learning algorithm for solving partial differential equations. Journal of Computational Physics 375, 1339–1364. URL: https://www.sciencedirect.com/science/article/pii/S0021999118305527, doi:https://doi.org/10.1016/j.jcp.2018.08.029.

[27] Sitzmann, V., Martel, J.N.P., Bergman, A.W., Lindell, D.B., Wetzstein, G., 2020. Implicit neural representations with periodic activation functions, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. pp. 7462–7473.

[28] Tancik, M., Srinivasan, P.P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J.T., Ng, R., 2020. Fourier features let networks learn high frequency functions in low dimensional domains. arXiv:2006.10739.

[29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2023. Attention is all you need. URL: https://arxiv.org/abs/1706.03762, arXiv:1706.03762.

[30] Wang, S., Teng, Y., Perdikaris, P., 2021. Understanding and mitigating gradient pathologies in physics-informed neural networks. SIAM Journal on Scientific Computing 43, A3055–A3081.

[31] Willcox, K., Peraire, J., 2012. Balanced model reduction via the proper orthogonal decomposition. AIAA Journal 40, 2323–2330. doi:doi.org/10.2514/2.1570.

[32] Zhao, Z., Ding, X., Prakash, B.A., 2024. Pinnsformer: A transformer-based framework for physics-informed neural networks. URL: https://arxiv.org/abs/2307.11833, arXiv:2307.11833.