Scaling Prior-Data Fitted Networks for Physical System Learning

Kaustubh Sharma[1*], **Simardeep Singh**[2*], **and Parikshit Pareek**[1†]
Indian Institute of Technology Roorkee (IIT Roorkee), Uttarakhand, India,
{kaustubh_s;pareek}@ee.iitr.ac.in; simardeep_s@mt.iitr.ac.in

Abstract

Prior-Data Fitted Networks (PFNs) offer a path toward **foundation models for scientific and engineering systems**– by learning to emulate Bayesian and Gaussian Process (GP) inference in a single forward pass. Instead of re-optimizing kernel parameters or retraining surrogates for each configuration, a pre-trained PFN can instantly produce posterior predictions across families of physical equations—effectively automating GP learning. However, conventional Transformer-based PFNs struggle in high-dimensional regimes due to entangled input–output attention. We attempt a **Decoupled-Value Attention (DVA)**, which computes similarities purely from inputs while propagating outputs through values, mirroring the GP update rule but remaining kernel-free. This localized attention shows substantial reduction in bias and scaling bottlenecks, cutting PFN validation loss by over 50% in five- and ten-dimensional tasks and matching GP accuracy on 64-dimensional power flow learning—at over $80\times$ computational speedup. The findings highlight that *attention, not architecture*, governs PFN generalization, and position PFNs as a scalable, physics-aware foundation modeling framework for automated scientific inference. Full pre-print: https://arxiv.org/abs/2509.20950.

1 Introduction

Prior-Data Fitted Networks (PFNs) perform **amortized Bayesian inference** by learning to approximate the posterior predictive of a Gaussian Process (GP) in a single forward pass, using synthetic tasks drawn from a prior [Müller et al., 2022, Hollmann et al., 2025]. This makes them strong candidates for **foundation models of science and engineering**—capable of generalizing across families of physical systems without repeated GP training or kernel tuning. While GP inference offers principled uncertainty quantification and smooth function estimation [Williams and Rasmussen, 2006], its cubic scaling in training and need for frequent re-optimization hinder deployment in dynamic settings such as power grids [Tan et al., 2025] or stochastic design problems. PFNs bypass this cost by mapping context data directly to predictive distributions, effectively automating GP learning. However, standard Transformer-based PFNs face scaling and bias issues in high-dimensional tasks because their self-attention over concatenated $(\mathbf{x}, \mathbf{y})$ embeddings dilutes input locality and increases compute load [Müller et al., 2022, Hollmann et al., 2025, Wang et al., 2025, Nagler, 2023], while Transformer backbones remain memory-intensive and less suited to domain-specific architectures.

To address these limitations, we propose **Decoupled-Value Attention (DVA)**, a localized attention mechanism that mirrors GP inference while remaining kernel-free. The Proposed DVA computes attention affinities (queries and keys) solely from inputs and propagates labels only through values, aligning with the GP property that predictive means are weighted sums of training targets based on input similarity [Williams and Rasmussen, 2006]. This design restores locality, reduces bias, and enables PFNs to scale across architectures—including both Transformers and CNNs—without

*Equal Contribution. †Corresponding Author. [1]Department of Electrical Engineering,
[2]Department of Metallurgical and Materials Engineering

2nd Conference on Applied AI and Scientific Machine Learning (CASML 2025).

sacrificing accuracy. Summary of our main findings are: **Localized PFN inference.** DVA enforces input-space localization, cutting PFN validation loss by over 50% on 5D and 10D regression tasks compared to standard attention [Müller et al., 2022]. **Architecture independence.** With DVA, CNN-based PFNs perform on par with Transformer-based ones, confirming that attention—not the backbone—governs generalization and bias reduction [Nagler, 2023]. **Scalable physical equation learning.** On a 64D power-flow benchmark, DVA-equipped PFNs reach mean absolute errors of $\mathcal{O}(10^{-3})$, matching GP surrogates while being over $80\times$ faster [Pareek and Nguyen, 2021, Liu and Srikantha, 2022, Molzahn et al., 2019].

Positioning: We do not propose a general-purpose attention, but a specialized design enabling localized PFNs to emulate GP inference efficiently. Unlike kernel-based attentions such as linearized [Katharopoulos et al., 2020], Nyström [Xiong et al., 2021], random-feature [Choromanski et al., 2021], or cross-kernel [Wang and Others, 2025] methods aimed at language and vision scaling, DVA makes PFNs scalable, bias-reduced, and domain-adaptive for physical equation learning [Hollmann et al., 2022]. By learning data-driven similarity instead of fixed kernels, DVA-equipped PFNs remain robust across functions and operating regimes, as shown for AC power flow [Tan et al., 2025].

1.1 PFN Background and Limitations of Joint Attention

The PFNs approximate the *posterior predictive distribution* (PPD) of a Bayesian model in a single forward pass. Trained on many *synthetic datasets* $\mathcal{D}^k \sim p(\mathcal{D})$, PFNs minimize the negative log-likelihood Müller et al. [2022]

$$\text{Negative Log-Likelihood (NLL)} \quad \ell_\theta = \sum_{k=1}^{K} \big[-\log q_\theta(\mathbf{y}^k \mid \mathbf{x}^k, \mathcal{D}^k) \big]. \tag{1}$$

where q_θ outputs over discretized target bins. After training, PFNs perform *amortized Bayesian inference*: for a new dataset $\mathcal{D}_{\text{train}}$ and query x_{test}, they produce $q_{\theta^\star}(y_{\text{test}} \mid x_{\text{test}}, \mathcal{D}_{\text{train}}) \approx p(y_{\text{test}} \mid x_{\text{test}}, \mathcal{D}_{\text{train}})$ in one forward pass. PFNs usually use a Transformer with self-attention over joint embeddings $\mathbf{z}_i = \text{enc}(\mathbf{x}_i) + \text{enc}(\mathbf{y}_i)$. This design faces two key issues in high-dimensional regression [Hollmann et al., 2025, Wang et al., 2025]: **Curse of dimensionality:** Similarity computed in joint (x, y) space becomes dominated by spurious output variation, degrading performance beyond $\sim$10 dimensions. **Bias from broken localization:** Mixing inputs and outputs in attention conflicts with theory [Nagler, 2023] that only local inputs should influence posteriors, adding bias that worsens with dimension. These limitations motivate *decoupled value attention* (DVA), where queries and keys depend only on inputs, while values carry outputs, preserving localization and GP-like behavior.

2 Proposed Decoupled-Value Attention

We propose **Decoupled-Value Attention (DVA)**, an input-localized attention mechanism for training PFNs. The proposed DVA is structurally aligned with GP inference by treating input $\mathbf{x}$ and output $\mathbf{y}$ separately at the attention stage. We enforce a strict separation of roles: attention affinities (queries and keys) are computed solely from the inputs, while the aggregated information (values) comes from the corresponding outputs– during both PFN training and prediction. Below, we explain DVA mathematically along with comparative assessment against Vanilla Attention (VA) Müller et al. [2022] and a kernel-based attention Wang and Others [2025].

Consider a PFN training dataset $\mathcal{D} = \{X, \mathbf{y}\}$ where $X \in \mathbb{R}^{N\times d}$ and $\mathbf{y} \in \mathbb{R}^{N\times 1}$ with N input samples of dimension d. In DVA we calculate query Q, key K and value V as

$$Q = W_q\, \varphi_x(X), \quad K = W_k\, \varphi_x(X), \quad V = W_v\, \varphi_y(\mathbf{y}), \tag{2}$$

with encoders φ_x, φ_y and trainable linear maps $W_q \in \mathbb{R}^{d\times d_k}, W_k \in \mathbb{R}^{d\times d_k}, W_v \in \mathbb{R}^{d\times 1}$. Then, attention is then computed as $\text{Att}(Q, K, V) = \text{softmax}\big(QK^T/\sqrt{d_k}\big)\, V$. Now, via (2), proposed DVA enforces that similarity is calculated purely in input space, while labels flow only through values. This is unlike VA used in PFNs, which mixes inputs and outputs in a joint embedding.

Inference: At test time, the mechanism is identical except that *training dataset* forms the *context set* and the "queries" are now the real unseen inputs i.e. we do not know the true output $\mathbf{y}$ for test inputs. Given a training dataset $\mathcal{D}_{\text{train}} \equiv \mathcal{D}_{\text{context}}$ for unseen function learning via GP, we obtain the

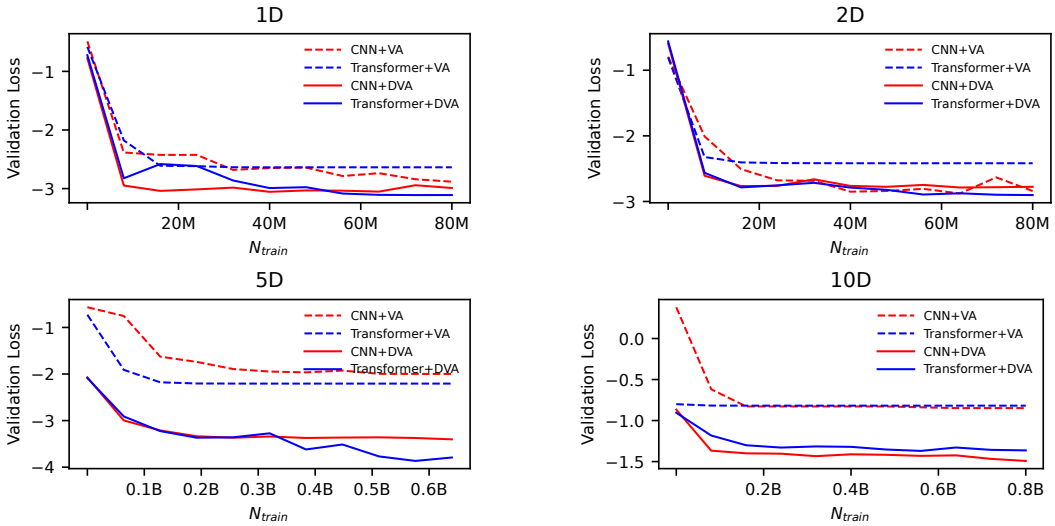

Figure 1: **Bias Reduction in PFN Training:** Validation loss (NLL) behavior with number of training points for various PFNs (Number of training points = epochs $\times$ steps per epoch $\times$ batch-size $\times$ dataset size. Dataset size is 100 for 1D/2D, 400 for 5D and 500 for 10D PFN). Validation loss was calculated on 64 out-of-sample datasets and Transformer + VA is taken from Müller et al. [2022].

predicted output with $Q_\star = W_q \varphi_x(X_\star)$ for test input $X_\star$ as

$$\hat{y}_{\text{test}} = g\Big(\text{softmax}\left(Q_\star K_{\text{tr}}^T / \sqrt{d_k}\right) V_{\text{tr}} \Big) \tag{3}$$

This ensures that the weight assigned to each context point's value $v(y_i)$ depends only on the similarity between the query input $\mathbf{x}_\star \in X_\star$ and the context input $\mathbf{x}_i \in \mathcal{D}_{train}$, mirroring the GP's use of an input-space kernel function, as discussed in the following subsection.

2.1 Localization Effect of DVA and Alignment with GP Inference

In DVA, attention weights for a test point $\mathbf{x}_\star$ are

$$\alpha_i(\mathbf{x}_\star) = \frac{\exp\big(\langle W_q \varphi_x(\mathbf{x}_\star), W_k \varphi_x(X_i)\rangle / \sqrt{d_k}\big)}{\sum_{j=1}^n \exp\big(\langle W_q \varphi_x(\mathbf{x}_\star), W_k \varphi_x(X_j)\rangle / \sqrt{d_k}\big)}. \tag{4}$$

Unlike joint embeddings, $\mathbf{y}_i$ does not influence affinities, appearing only in the values. This ensures that the attention concentrates on inputs near $\mathbf{x}_\star$, recovering the input-space localization property [Nagler, 2023]. DVA aligns with GP inference: in a GP, predictions are weighted sums of outputs with weights depending only on inputs, $\mu(x_\star) = \sum_i \beta_i(x_\star) y_i$ with $\beta(x_\star) = k(x_\star, X)\,[K(X,X) + \sigma^2 I]^{-1}$. Similarly, DVA weights act as a positive kernel on inputs (via the exponential inner product), producing predictions as weighted sums of outputs, preserving the input-output dependency structure of a GP. Alternative kernel-based attentions (e.g., RBF) can mimic GP weights, but they restrict flexibility and require tuning of kernel parameters γ, whereas DVA learns localization from data.

3 Numerical Results and Discussion

Figure 1 summarizes PFN validation loss across increasing input dimensions (1D, 2D, 5D, 10D) for CNN and Transformer backbones using vanilla attention (VA, dashed) and decoupled value attention (DVA, solid). **Bias Reduction:** Across all dimensions, VA-equipped PFNs saturate at higher loss, revealing persistent residual bias. DVA consistently reaches lower validation loss, with the effect more pronounced in higher dimensions (5D and 10D), where VA models quickly plateau while DVA continues improving. In 5D and 10D, DVA also shows lower initial loss, indicating faster convergence and a reduced asymptotic bias floor.

Architecture-Agnostic Behavior: The choice of attention dominates performance differences: both CNN- and Transformer-based PFNs achieve comparable loss with DVA, while VA exhibits

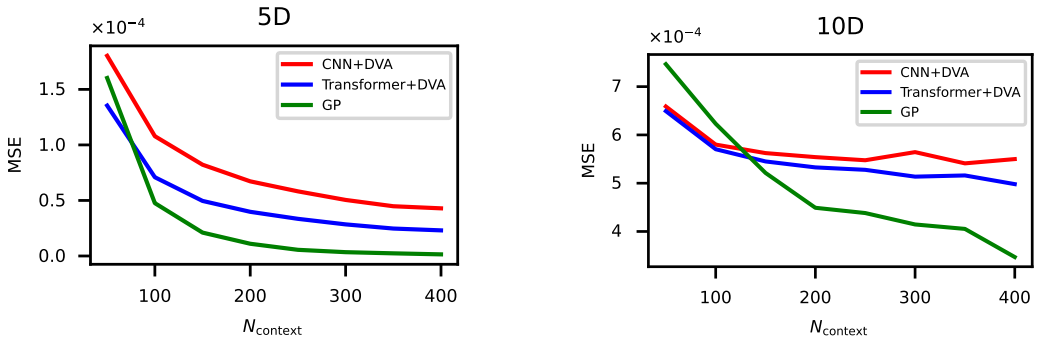

Figure 2: **Comparison with GP:** MSE for 5D and 10D PFNs as a function of context size. All models are tested using $n_{\text{test}} = 500$, for $N_{context}$. The results show that error consistently decreases with larger context sizes, and that CNN- and Transformer-based PFNs with DVA approach the performance of exact GP inference even in higher dimensions. Exact GP baselines were fit using `scikit-learn` with $N_{context}$ training samples.

Table 1: **Voltage prediction on a 64D power-flow test bed:** Trained on 500 samples; evaluated on 4,500 test samples. Time (t) is for evaluating all 32 node voltages, and MSE/MAE correspond to the maximum across buses.

	Exact GP			**CNN + DVA**			**Transformer + DVA**		
ΔLoad	**MSE**	**MAE**	t **(s)**	**MSE**	**MAE**	t **(s)**	**MSE**	**MAE**	t **(s)**
5%	2.2e-7	0.0004	10.88	4.5e-7	0.0005	0.13	1.5e-6	0.001	0.17
10%	3.5e-7	0.0004	10.94	1.7e-6	0.001	0.13	2.8e-6	0.001	0.17
30%	3.2e-7	0.0005	11.61	1.5e-5	0.003	0.14	1.6e-5	0.003	0.17
50%	2.2e-7	0.0003	11.89	4.2e-5	0.005	0.13	4.4e-5	0.005	0.17

a larger gap. DVA narrows architecture-induced variance, enabling smaller CNNs to match larger Transformers in both 5D and 10D tasks, confirming PFNs are largely architecture-agnostic once attention is specified.

Comparison with GP: Figure 2 shows PFN MSE as a function of context size. DVA-equipped PFNs approach GP performance, with error decreasing steadily as more context points are available, even in high-dimensional tasks. VA-equipped models remain biased, highlighting the critical role of DVA in high-dimensional, data-efficient Bayesian inference.

Power Flow Learning: In this experiment, we model the IEEE 33-bus system, treating real and reactive power demands at 32 load buses as uncertain inputs, resulting in a 64-dimensional input space. The task is to predict steady-state bus voltage magnitudes, effectively learning the nonlinear AC power flow mapping Voltage $= f(\text{Loads})$. Table 1 benchmarks performance under 5%–50% load perturbations. Exact GPs achieve the lowest MSE and MAE but require training one model per bus, making repeated queries impractical. DVA-equipped PFNs (CNN+DVA and Transformer+DVA) slightly increase error but are over 80× faster, achieving voltage predictions accurate to 10^{-3} p.u., sufficient for practical use. Note that, vanilla-attention PFNs failed to train for this 64D problem. Training time for both DVA models is approximately 14 hours on an NVIDIA 4500ADA GPU.

4 Conclusions and Future Work

We propose Decoupled-Value Attention (DVA) to train Prior-Data Fitted Networks (PFNs) for high-dimensional GP inference, showing that DVA halves residual bias in 5D and 10D tasks and allows CNN- and Transformer-based PFNs to achieve comparable accuracy once equipped with the attention mechanism. Leveraging this, DVA enables PFNs to serve as efficient surrogates for high-dimensional power flow learning: on the IEEE 33-bus system with 64-dimensional load variations, DVA-equipped PFNs achieved voltage prediction accuracy on the order of 10^{-5} while delivering over 80× speedup versus exact GP. Future work includes scaling PFNs to larger networks and higher-dimensional uncertainties, designing architectures that better capture input-space localization, and addressing

DVA's limitation of lacking output affinities, aiming for real-time, uncertainty-aware decision making in modern power systems.

References

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations (ICLR)*, 2021.

Noah Hollmann, Samuel Müller, Eyke Hüllermeier, and Asja Fischer. Learning to learn with prior-data fitted networks. In *International Conference on Learning Representations (ICLR)*, 2022.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, jan 2025. doi: 10.1038/s41586-024-08328-6.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jingyuan Liu and Pirathayini Srikantha. Kernel structure design for data-driven probabilistic load flow studies. *IEEE Transactions on Smart Grid*, 13(4):2679–2689, 2022. doi: 10.1109/TSG.2022.3159579.

Daniel K Molzahn, Ian A Hiskens, et al. A survey of relaxations and approximations of the power flow equations. *Foundations and Trends® in Electric Energy Systems*, 4(1-2):1–221, 2019.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations (ICLR)*, 2022.

Thomas Nagler. Statistical foundations of prior-data fitted networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25660–25676. PMLR, jul 2023.

Parikshit Pareek and Hung D Nguyen. A framework for analytical power flow solution using gaussian process learning. *IEEE Trans. on Sustainable Energy*, 13(1):452–463, 2021.

Bendong Tan, Tong Su, Yu Weng, Ketian Ye, Parikshit Pareek, Petr Vorobev, Hung Nguyen, Junbo Zhao, and Deepjyoti Deka. Gaussian processes in power systems: Techniques, applications, and future works. *arXiv preprint arXiv:2505.15950*, 2025.

Author Wang and Others. Cross-kernel attention for efficient sequence modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Yuxin Wang, Botian Jiang, Yiran Guo, Quan Gan, David Wipf, Xuanjing Huang, and Xipeng Qiu. Prior-fitted networks scale to larger datasets when treated as weak learners. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 1090–1098. PMLR, may 2025.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, and et al. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.