RELATIVE POSITION BIASES FOR TRANSFORMER PINNS

Fedor Buzaev[1], **Andrei Ermakov**[1,2], **Mariia Ivanova**[1], **Fedor Ratnikov**[1], **Denis Derkach**[1] **and Ilya Makarov**[2,3,4]
[1]HSE University,
[2]AXXX,
[3]Trusted AI, Center, RAS,
[4]Research Center of the Artificial Intelligence Institute, Innopois University

ABSTRACT

Transformer-based physics-informed neural networks (PINNs) have recently improved PDE solving by modeling spatiotemporal interactions with self-attention, yet most variants still rely on absolute coordinate embeddings. We ask whether *relative* positional structure – a key inductive bias in modern Transformers can be made coordinate-aware and yield an accuracy boost with minimal overhead for PDE PINNs. Building on a PINNsFormer-style decoder-only baseline (S-Pformer), we introduce two drop-in modifications: (i) a spatial-distance ALiBi attention bias and (ii) spatial RoPE applied to attention queries/keys using normalized physical coordinates. Across multiple PDE benchmarks and random seeds, both relative schemes consistently improve over the baseline, with spatial RoPE providing the strongest gains in our experiments. Our results suggest that injecting coordinate-relative structure into attention is a simple and effective upgrade for Transformer PINNs.

1 INTRODUCTION

Numerically solving partial differential equations (PDEs) is a core problem in science and engineering. Classical discretization-based solvers, including finite element and pseudo-spectral methods, can be highly accurate but may become computationally expensive in multiscale regimes, for long time horizons, or when repeated solves are required Bathe (2008); Fornberg (1996). Physics-informed neural networks (PINNs) provide a mesh-free alternative by approximating the solution field $u(\mathbf{x}, t)$ with a neural surrogate $u_\theta(\mathbf{x}, t)$ trained to minimize a physics loss function that enforces PDE residuals and initial/boundary constraints via automatic differentiation Lagaris et al. (1997); Raissi et al. (2019); Buzaev et al. (2023); Chuprov et al. (2024). Despite their flexibility, standard MLP-based PINNs can fail on oscillatory, multiscale, or transport-dominated dynamics, often exhibiting over-smoothing and poor propagation of boundary/initial information across the domain Krishnapriyan et al. (2021); Wang et al. (2022).

A complementary line of work strengthens architectural inductive bias by leveraging sequence modeling in physics-informed learning. Transformer-based PINNs (e.g., PINNsFormer) use self-attention to improve long-range temporal coupling and have shown better accuracy than MLP-based PINNs on several PDE benchmarks Zhao et al. (2024). Related approaches replace attention with state-space models to further stabilize long-range dependency modeling in physics-informed training Xu et al. (2025). However, most physics-informed Transformers retain a design choice inherited from language modeling: they encode position through *absolute* coordinate embeddings (e.g., learned embeddings and/or Fourier features). For PDEs, where the governing operators are naturally expressed in terms of *relative* distances and local interactions, absolute encodings may be a mismatched inductive bias – the model must rediscover locality and translation structure implicitly from residual-based supervision.

In this paper, we investigate whether *relative position mechanisms* from modern Transformers can serve as a low-overhead, physics-aligned upgrade for Transformer PINNs. Specifically, we take a state-of-the-art PINNsFormer-style decoder-only baseline (S-Pformer; a streamlined attention PINN

with strong empirical performance) Arni & Blanco (2025) and introduce two drop-in modifications *inside* self-attention: (i) **distance-based ALiBi** Press et al. (2022), an additive attention bias constructed from pairwise physical distances (e.g., $|x_i - x_j|$, adapting ALiBi from sequence modeling to continuous coordinates), and (ii) **spatial RoPE** Su et al. (2023), where rotary embeddings are driven by normalized continuous coordinates rather than token indices (adapting RoPE to PDE coordinates). Both changes preserve the overall training setup, parameter scale, and physics-informed objective, but explicitly encourage attention to respect coordinate-relative structure.

Across multiple PDE benchmarks and random seeds, we find that relative position biases consistently improve accuracy over the baseline, with spatial RoPE providing the strongest gains in our experiments. Our results suggest that *physics-aware relative attention* is a simple and effective architectural prior for Transformer PINNs, and a practical bridge between advances in foundation-model positional modeling and PDE learning Meng et al. (2025).

2 BACKGROUND

Physics-informed neural networks (PINNs). Let $u : \Omega \times [0, T] \to \mathbb{R}^{d_{\text{out}}}$ satisfy a PDE with differential operator $\mathcal{N}$, boundary operator $\mathcal{B}$, and initial condition u_0:

$$\mathcal{N}[u](\boldsymbol{x}, t) = \mathbf{0}, \quad (\boldsymbol{x}, t) \in \Omega \times (0, T], \tag{1}$$

$$\mathcal{B}[u](\boldsymbol{x}, t) = \mathbf{0}, \quad (\boldsymbol{x}, t) \in \partial\Omega \times [0, T], \tag{2}$$

$$u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega. \tag{3}$$

PINNs parameterize u by a neural surrogate $u_\theta(\boldsymbol{x}, t)$ and optimize θ by minimizing a physics loss that penalizes PDE residuals and boundary/initial conditions, using automatic differentiation for derivatives Lagaris et al. (1997); Raissi et al. (2019):

$$\mathcal{L}(\theta) = \lambda_r \, \mathbb{E}\big[\|\mathcal{N}[u_\theta](\boldsymbol{x}, t)\|_2^2\big] + \lambda_b \, \mathbb{E}\big[\|\mathcal{B}[u_\theta](\boldsymbol{x}, t)\|_2^2\big] + \lambda_0 \, \mathbb{E}\big[\|u_\theta(\boldsymbol{x}, 0) - u_0(\boldsymbol{x})\|_2^2\big]. \tag{4}$$

Transformer PINNs and pseudo-sequences (PINNsFormer). Transformer PINNs introduce local temporal coupling forming a short window of k tokens anchored at $(\boldsymbol{x}, t)$

$$\mathcal{S}_k(\boldsymbol{x}, t; \Delta t) = \{[\boldsymbol{x},\, t], [\boldsymbol{x},\, t + \Delta t], \ldots, [\boldsymbol{x},\, t + (k-1)\Delta t]\},$$

then mixing tokens with self-attention Zhao et al. (2024). The physics loss can be applied token-wise across the window (averaging residual/boundary terms and enforcing the initial term at the window start):

$$\mathcal{L}_{\text{seq}} = \lambda_r \frac{1}{k} \sum_{j=0}^{k-1} \mathbb{E}\Big[\big\|\mathcal{N}[u_\theta](\boldsymbol{x},\, t + j\Delta t)\big\|_2^2\Big] + \lambda_b \frac{1}{k} \sum_{j=0}^{k-1} \mathbb{E}\Big[\big\|\mathcal{B}[u_\theta](\boldsymbol{x},\, t + j\Delta t)\big\|_2^2\Big]$$
$$+ \lambda_0 \, \mathbb{E}\Big[\big\|u_\theta(\boldsymbol{x},\, t) - u_0(\boldsymbol{x})\big\|_2^2\Big]. \tag{5}$$

Decoder-only S-Pformer. A representative recent baseline is S-Pformer Arni & Blanco (2025), a decoder-only PINNsFormer variant. Here, each token corresponds to a spatiotemporal coordinate, and the model predicts the solution value u at the same coordinate; since there is no separate source/target sequence, an encoder–decoder split is unnecessary. S-Pformer applies stacked self-attention and feed-forward blocks over the pseudo-sequence and uses standard coordinate embeddings, typically augmented with Fourier features and WaveAct-style nonlinearities. In this work, we keep the S-Pformer backbone intact and modify only how positional structure enters self-attention (Sec. 3).

3 METHOD

3.1 PHYSICS-INFORMED RELATIVE POSITION MECHANISMS

We start from a decoder-only PINNsFormer-style baseline (S-Pformer) that processes a pseudo-sequence of spatiotemporal coordinates and predicts the solution field u_θ at each token Arni & Blanco (2025). The baseline uses standard coordinate embeddings (e.g., linear lifting and/or Fourier features) and applies multi-head self-attention (MHSA) over the token dimension. Our key change is to augment self-attention with *relative* position structure that is explicitly tied to continuous PDE coordinates.

Distance-based ALiBi (continuous coordinates). ALiBi adds a head-specific additive bias to attention logits based on relative position Press et al. (2022). For PDE tokens we replace discrete index distance with *physical* distance along a chosen coordinate s (e.g., space x in 1D). For tokens $\{\boldsymbol{x}_i\}_{i=1}^{L}$, define the (optionally normalized) distance and head bias:

$$\widetilde{d}_{ij} = \frac{\left|x_i^{(s)} - x_j^{(s)}\right|}{\alpha}, \qquad b_{ij}^{(h)} = -m_h \, \widetilde{d}_{ij}, \tag{6}$$

where $m_h > 0$ are standard ALiBi slopes and α is a distance scale (e.g., the known domain span, or an estimated span from the current batch). Attention logits become

$$A_{ij}^{(h)} = \frac{\langle q_i^{(h)}, k_j^{(h)} \rangle}{\sqrt{d_k}} + b_{ij}^{(h)}. \tag{7}$$

Spatial RoPE (normalized coordinates). RoPE injects relative position by rotating query/key channels as a function of position Su et al. (2023). We drive RoPE by *continuous* coordinates: let

$$p_i = \frac{x_i^{(s)} - x_{\min}}{(x_{\max} - x_{\min}) + \varepsilon} \in [0, 1], \tag{8}$$

(using known $[x_{\min}, x_{\max}]$ or batch min–max, with $\varepsilon \approx 10^{-6}$). For each head, we rotate $q_i^{(h)}, k_i^{(h)}$ with angles proportional to p_i:

$$\hat{q}_i^{(h)} = \mathrm{RoPE}\Big(q_i^{(h)}, p_i\Big), \quad \hat{k}_i^{(h)} = \mathrm{RoPE}\Big(k_i^{(h)}, p_i\Big), \tag{9}$$

and compute attention logits as

$$A_{ij}^{(h)} = \frac{\langle \hat{q}_i^{(h)}, \hat{k}_j^{(h)} \rangle}{\sqrt{d_k}}. \tag{10}$$

3.2 Motivation: relative geometry as an inductive bias for PDE attention

Unlike language tokens, PDE tokens are continuous coordinates, so interactions are naturally governed by *relative* geometry (e.g., distances in space/time), not just discrete token indices. With purely absolute coordinate embeddings, the model must learn locality and translation structure implicitly from the physics loss. Distance-based ALiBi encourages attention toward nearby points by adding a distance penalty to the logits (Eq. 6–7), while spatial RoPE injects coordinate-relative structure by rotating query/key channels as a smooth function of normalized coordinates (Eq. 8–10). Both are drop-in changes: they only change how attention incorporates positional information and lleave the physics-informed objective and the overall training pipeline unchanged.

3.3 Overall architecture

Our models follow the decoder-only pseudo-sequence PINNsFormer template. Given a collocation point $(\boldsymbol{x}, t)$, we form a local window of k tokens $\mathcal{S}_k(\boldsymbol{x}, t; \Delta t)$ and embed each token with shared linear layers (optionally combined with standard coordinate features as in the baseline) Zhao et al. (2024). A stack of N identical decoder blocks alternates MHSA (token mixing) and a feed-forward network (channel mixing), using WaveAct-style nonlinearities as in prior PINNsFormer variants. We instantiate two variants that differ only in the attention module:

- **S-P + ALiBi:** MHSA adds the coordinate-distance bias to attention logits (Eq. equation 7).
- **S-P + RoPE:** MHSA applies coordinate-driven RoPE to Q, K before dot-product attention (Eq. equation 9).

The prediction head maps the final token representations to u_θ at each token; during evaluation we read the token corresponding to the query time. Training uses the same physics-informed sequence loss (Eq. equation 5) for all models.

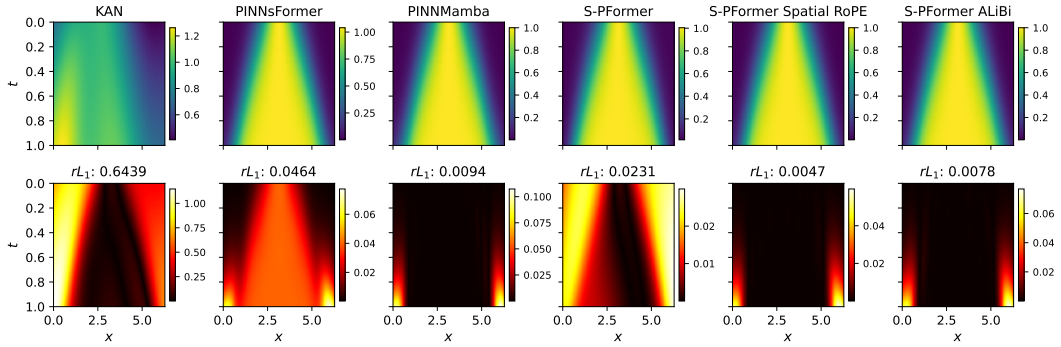

Figure 1: **Qualitative comparison on the 1D reaction equation.** Top row: predicted solution fields $u(x,t)$ over the space–time domain. Bottom row: pointwise absolute error $|u_\theta(x,t) - u(x,t)|$ (lower is better). Relative RL1 errors are reported under each method (single displayed run). Adding physics-informed *relative* position structure to S-Pformer improves accuracy: S-P + RoPE achieves the lowest error, followed by S-P + ALiBi, both outperforming the original S-Pformer and other baselines.

Table 1: **Reaction** and **2D Navier–Stokes** results (mean $\pm$ std over 5 seeds). RL1/RL2 are relative errors; Time is wall-clock training time (s).

	Reaction			Navier–Stokes (2D)		
Model	RL1	RL2	Time	RL1	RL2	Time
PINN	0.9803 ± 0.02	0.9785 ± 0.01	487.3	1.2210 ± 0.09	1.2130 ± 0.10	3843
KAN	0.6781 ± 0.07	0.6987 ± 0.07	78.4	1.1557 ± 0.28	1.1766 ± 0.32	3426
PINNsFormer	0.0346 ± 0.02	0.0438 ± 0.04	398.5	0.1223 ± 0.04	0.1531 ± 0.03	3925
PINN-Mamba	0.0105 ± 0.00	0.0248 ± 0.00	287.6	0.1013 ± 0.05	0.1172 ± 0.05	4160
S-Pformer	0.0254 ± 0.01	0.0271 ± 0.00	231.6	0.1253 ± 0.02	0.1321 ± 0.02	3960
S-P + RoPE	0.0053 ± 0.00	0.0655 ± 0.00	240.2	0.0823 ± 0.01	0.0891 ± 0.01	2980
S-P + ALiBi	0.0081 ± 0.00	0.0116 ± 0.00	243.7	0.0709 ± 0.01	0.0781 ± 0.01	2870

4 Experiments

We evaluate relative position biases as drop-in replacements in a decoder-only PINNsFormer-style baseline (S-Pformer) on four PDE benchmarks: **wave**, **reaction**, **convection**, and **2D Navier–Stokes**. All models are trained with the same physics-informed objective and sampling strategy; we report mean $\pm$ std over 5 random seeds using two relative error metrics (RL1/RL2) and wall-clock training time. Baselines include a standard PINN (MLP) Raissi et al. (2019), PINNsFormer Zhao et al. (2024), KAN Liu et al. (2025), PINN-Mamba Xu et al. (2025), and S-Pformer Arni & Blanco (2025). Our two methods are **S-Pformer + ALiBi** and **S-Pformer + spatial RoPE** (Sec. 3).

Table 1 summarize results for Reaction and 2D Navier-Stokes equations, and for Convection and Wave, Table 2 in the appendix. We show that for all the problems considered, relative position biases improve S-Pformer, with spatial RoPE achieving the best errors among attention-based PINN variants in our runs.

5 Conclusion

We studied whether *relative position* mechanisms from modern Transformers can serve as a lightweight, physics-aligned inductive bias for Transformer PINNs. Starting from a decoder-only PINNsFormer-style baseline (S-Pformer), we introduced two drop-in modifications inside self-attention: a distance-based ALiBi bias and coordinate-driven spatial RoPE. Across multiple seeds, both methods improve accuracy on the PDEs where we have complete results, with spatial RoPE providing the strongest gains in our experiments. These findings suggest that injecting coordinate-relative structure directly into attention is a simple and effective upgrade for physics-informed sequence models, and a promising ingredient for more robust Transformer-based PDE solvers. Future work will extend evaluation to additional PDE families and explore multi-dimensional relative position schemes (e.g., 2D/3D RoPE or anisotropic distance biases) for complex geometries.

ACKNOWLEDGEMENTS

The study was supported by the Ministry of Economic Development of the Russian Federation (agreement No. 139-10-2025-034 dd. 19.06.2025, IGK 000000C313925P4D0002)

REFERENCES

Rohan Arni and Carlos Blanco. Physics-informed neural networks with fourier features and attention-driven decoding. In *NeurIPS 2025 AI for Science Workshop*, 2025. URL https://openreview.net/forum?id=woq4ZAm1AH.

Klaus-Jürgen Bathe. Finite element method, June 2008. URL http://dx.doi.org/10.1002/9780470050118.ecse159.

Fedor Buzaev, Jiexing Gao, Ivan Chuprov, and Evgeniy Kazakov. Hybrid acceleration techniques for the physics-informed neural networks: a comparative analysis. *Machine Learning*, 113(6): 3675–3692, December 2023. ISSN 1573-0565. doi: 10.1007/s10994-023-06442-6. URL http://dx.doi.org/10.1007/s10994-023-06442-6.

I. A. Chuprov, J. Gao, D. S. Efremenko, F. A. Buzaev, and V. V. Zemlyakov. Solution of the multimode nonlinear schrödinger equation using physics-informed neural networks. *Doklady Mathematics*, 110(S1):S15–S24, December 2024. ISSN 1531-8362. doi: 10.1134/s1064562424602105. URL http://dx.doi.org/10.1134/S1064562424602105.

Bengt Fornberg. *A Practical Guide to Pseudospectral Methods*. Cambridge University Press, January 1996. ISBN 9780511626357. doi: 10.1017/cbo9780511626357. URL http://dx.doi.org/10.1017/CBO9780511626357.

Aditi S. Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W Mahoney. Characterizing possible failure modes in physics-informed neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Isaac E. Lagaris, Aristidis C. Likas, and Dimitrios Ioannis Fotiadis. Artificial neural networks for solving ordinary and partial differential equations. *IEEE transactions on neural networks*, 9 5:987–1000, 1997. URL https://api.semanticscholar.org/CorpusID:18698107.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov–arnold networks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ozo7qJ5vZi.

Chuizheng Meng, Sam Griesemer, Defu Cao, Sungyong Seo, and Yan Liu. When physics meets machine learning: a survey of physics-informed machine learning. *Machine Learning for Computational Science and Engineering*, 1(1), May 2025. ISSN 3005-1436. doi: 10.1007/s44379-025-00016-0. URL http://dx.doi.org/10.1007/s44379-025-00016-0.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation, 2022. URL https://arxiv.org/abs/2108.12409.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. In *Journal of Computational Physics*, 2019.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why pinns fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, January 2022. ISSN 0021-9991. doi: 10.1016/j.jcp.2021.110768. URL http://dx.doi.org/10.1016/j.jcp.2021.110768.

Chenhui Xu, Dancheng Liu, Yuting Hu, Jiajie Li, Ruiyang Qin, Qingxiao Zheng, and Jinjun Xiong. Sub-sequential physics-informed learning with state space model. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=V7VnjJxBlg.

Zhiyuan Zhao, Xueying Ding, and B. Aditya Prakash. PINNsformer: A transformer-based framework for physics-informed neural networks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DO2WFXU1Be.

Table 2: **Convection** and **Wave** results (mean ± std over 5 seeds). RL1/RL2 are relative errors; Time is wall-clock training time (s).

	Convection			Wave		
Model	RL1	RL2	Time	RL1	RL2	Time
PINN	0.9421 ± 0.15	1.0130 ± 0.18	843.7	0.4101 ± 0.05	0.4141 ± 0.05	1587.4
KAN	0.1433 ± 0.03	0.1458 ± 0.03	312.4	0.5011 ± 0.04	0.5023 ± 0.05	124.6
PINNsFormer	0.0412 ± 0.03	0.0531 ± 0.03	921.3	0.0459 ± 0.05	0.0463 ± 0.05	1456.3
PINN-Mamba	0.0588 ± 0.02	0.0601 ± 0.02	876.2	0.0207 ± 0.00	0.0213 ± 0.00	1187.5
S-Pformer	0.1759 ± 0.27	0.1878 ± 0.28	913.4	0.0093 ± 0.00	0.0098 ± 0.00	1424.0
S-P + RoPE	0.0414 ± 0.01	0.0543 ± 0.02	613.4	0.0080 ± 0.00	0.0088 ± 0.00	1400.6
S-P + ALiBi	0.0673 ± 0.02	0.0749 ± 0.02	667.4	0.0173 ± 0.00	0.0178 ± 0.00	1720.4

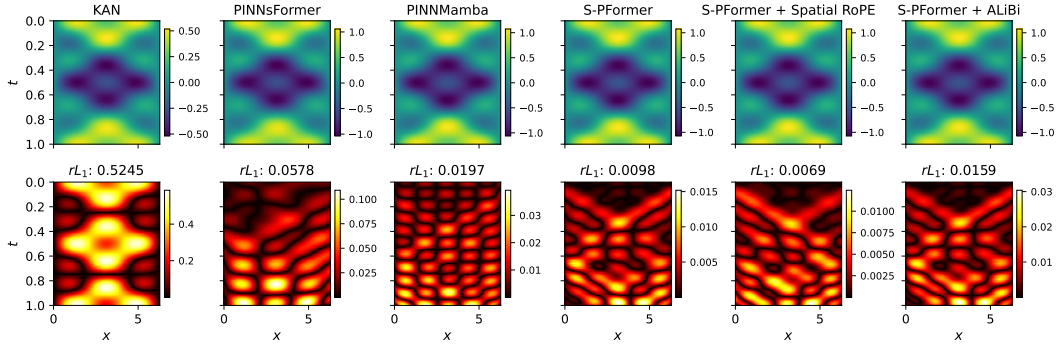

Figure 2: **1D wave equation: qualitative comparison.** Predicted solution fields $u(x, t)$ and absolute error maps $|u_\theta - u|$ for representative baselines and our methods. Relative position mechanisms reduce structured error bands and improve overall fidelity; S-P + RoPE is visually the sharpest and most accurate among the compared S-Pformer variants.

A Numerical Results

We provide additional comparison figures for **1D wave**) to complement the quantitative results in Tables 1. Each figure contrasts baseline PINN variants against our relative-position S-Pformer models (S-P + ALiBi, S-P + RoPE). We also provide the number of parameters for each model 3.

Table 3: **Model size (number of parameters) per PDE.**

Model	Wave	Reaction	Convection	Navier–Stokes (2D)
PINN	527,361	527,361	527,361	527,361
PINNsFormer	453,561	453,561	453,561	453,561
KAN	891	891	891	891
PINN-Mamba	285,763	285,763	285,763	285,763
S-Pformer	370,991	370,991	305,551	305,551
S-P + ALiBi	368,959	368,959	305,891	305,891
S-P + RoPE	366,959	366,959	305,891	305,891