DYWPE: SIGNAL-AWARE DYNAMIC WAVELET POSITIONAL ENCODING FOR TIME SERIES TRANSFORMERS

Habib Irani, Vangelis Metsis

Computer Science Department, Texas State University
San Marcos, TX 78666, USA

ABSTRACT

Existing positional encoding methods in transformers are fundamentally signal-agnostic, deriving positional information solely from sequence indices while ignoring the underlying signal characteristics. This limitation is particularly problematic for time series analysis, where signals exhibit complex, non-stationary dynamics across multiple temporal scales. We introduce Dynamic Wavelet Positional Encoding (DyWPE), a novel signal-aware framework that generates positional embeddings directly from input time series using the Discrete Wavelet Transform (DWT). Comprehensive experiments on ten diverse time series datasets demonstrate that DyWPE consistently outperforms state-of-the-art positional encoding methods, with particularly significant improvements on longer sequences and complex biomedical signals.

***Index Terms*—** position encoding, transformer, time series, wavelet transform, signal processing.

1. INTRODUCTION

The transformer architecture has revolutionized sequential data modeling across diverse domains, from natural language processing to time series analysis [1]. A fundamental component enabling transformers to process sequential data is positional encoding, which addresses the inherent permutation invariance of self-attention mechanisms by injecting positional information into input representations.

In time series analysis, the importance of positional encoding is amplified due to the intrinsic temporal dependencies and complex multi-scale patterns characteristic of temporal data [2, 3]. However, existing positional encoding methods, ranging from sinusoidal encodings [1] to sophisticated relative positioning schemes [4, 5], share a fundamental limitation: they are signal-agnostic. These methods derive positional information exclusively from abstract sequence indices (0, 1, ..., L-1) while remaining completely oblivious to the underlying signal characteristics. For instance, consider two time series segments occurring at identical absolute positions but exhibiting vastly different temporal dynamics: one representing a quiet, stable period with minimal variation, and another capturing volatile, high-frequency oscillations. Traditional positional encodings assign identical positional representations to both contexts, failing to capture the distinct temporal signatures that are crucial for effective time series modeling. This signal-agnostic approach becomes particularly problematic when dealing with non-stationary signals, where statistical properties change over time, or multi-scale phenomena where different frequency components carry distinct semantic meanings. Recent comprehensive studies have highlighted significant performance variations across different positional encoding strategies in time series applications [6], yet no existing method addresses the fundamental limitation of signal-independent positioning.

To address this, we introduce Dynamic Wavelet Positional Encoding (DyWPE), a novel signal-aware framework that provides a more powerful inductive bias for temporal data. While a sufficiently large transformer can theoretically learn the complex relationships between signal-agnostic positions and signal content, this forces the model to learn the fundamental principles of time series dynamics from scratch. DyWPE makes this learning task more efficient by directly encoding the signal's local, multi-scale characteristics *into* the positional representation itself. It achieves this by leveraging the Discrete Wavelet Transform (DWT) and learnable gating mechanisms to generate positional embeddings from the signal's content, creating a rich representation that dynamically adapts to local behavior. This approach offloads the burden of local pattern recognition, allowing the self-attention layers to focus more effectively on capturing long-range, higher-level dependencies.

Our key contributions are: (1) The first signal-aware positional encoding framework that derives positional information from signal content rather than sequence indices; (2) A computationally efficient implementation using DWT/IDWT operations with linear $\mathcal{O}(L)$ complexity; (3) Comprehensive experimental validation across ten diverse datasets demonstrating consistent superiority over eight established methods; (4) An ablation study examining the effectiveness and necessity of the different components of our algorithm, such as dynamic modulation and multi-scale wavelet decomposition.

2. BACKGROUND AND RELATED WORK

The application of attention in time series analysis has evolved from augmenting recurrent models to forming the core of modern transformers. Early work in sequence-to-sequence learning demonstrated the power of attention in encoder-decoder frameworks [7]. For time series, this led to methods like dual-stage attention-based RNNs, which could selectively focus on relevant features and time steps for improved forecasting [8]. The success of these hybrids motivated the development of pure self-attention models, such as Gated Transformer Networks, which use learnable gates to capture discriminative temporal patterns without recurrence [9].

The transformer architecture [1], with its self-attention mechanism, offered a powerful new paradigm. Adaptations for time series were swift, with initial frameworks focusing on stable training for multivariate data [2] and specialized multi-head attention mechanisms, as seen in the Temporal Fusion Transformer (TFT) [3]. To better handle the unique structure of temporal data, recent models have also incorporated patch-based tokenization, which segments the time series to capture both local and global features [10].

A core challenge in these adaptations is the transformer's inherent permutation invariance, which necessitates an explicit Positional Encoding (PE) to inform the model of the temporal order. The original sinusoidal PE [1] provides a fixed, deterministic mapping based on sequence indices. To improve flexibility, subsequent research introduced learnable absolute embeddings and relative position representations [4], which encode the distance between tokens. More advanced methods like Rotary Position Embedding (RoPE) [11] and Transformer with Untied Positional Encoding (TUPE) [5] further refined these concepts by integrating positional information directly into the attention computation. This evolution reflects a broader trend towards more expressive and context-dependent positional representations, often at the cost of increased computational complexity.

Despite these advances, a recent survey of PE methods in time series transformers [12] highlights a shared, fundamental limitation: they are all **signal-agnostic**. These methods operate on integer indices and remain oblivious to the signal's content, treating a volatile period the same as a stable one. While specialized designs for time series have been proposed [6, 13], they still anchor their encodings to the position index. The most proximate work, by Oka et al. [14], proposes a wavelet-based PE for language models to improve length extrapolation. However, their method is also signal-agnostic, using wavelets to create a multi-scale representation of the *relative distance between indices*, not the signal's content. To our knowledge, DyWPE is the first framework to break from this paradigm by constructing a positional encoding directly from the time series signal itself.

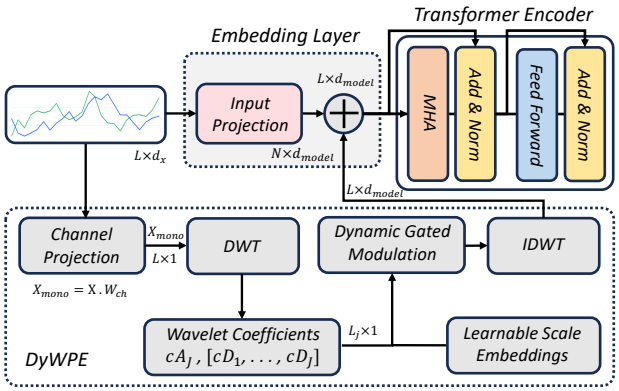

Fig. 1. Transformer with DyWPE architecture overview showing the five-step process from multivariate input to final positional encoding through wavelet-based signal analysis.

3. DYNAMIC WAVELET POSITIONAL ENCODING

3.1. Problem Formulation and Overview

Given a time series dataset $X = \{x_1, x_2, ..., x_n\}$ with n samples, where each sample $x_i \in R^{L \times d_x}$ represents a d_x-dimensional time series of length L, and corresponding labels $Y = \{y_1, y_2, ..., y_n\}$ where $y_i \in \{1, 2, ..., c\}$, our objective is to learn a positional encoding that captures signal-specific temporal characteristics.

Traditional positional encodings follow $P = f(\text{indices})$ where f is signal-independent. DyWPE introduces the paradigm $P = f(X, \theta)$, making positional encoding a learnable function of actual signal content through parameters θ. This enables the model to distinguish between different temporal contexts (e.g., high-frequency transients vs. low-frequency trends) even at identical sequence positions.

3.2. Mathematical Formulation

Given a multivariate time series $X \in R^{B \times L \times d_x}$ where B is batch size, L is sequence length, and d_x is the number of input channels, DyWPE produces position embedding $P_{DyWPE} \in R^{B \times L \times d_{model}}$ through five sequential steps:

Step 1: Channel Projection. For multivariate signals, we create a single representative channel for wavelet analysis:

$$x_{mono} = x \cdot w_{channel}$$

where $w_{channel} \in R^{d_x}$ is a learnable projection vector capturing the most relevant temporal dynamics across input channels.

Step 2: Multi-Level Wavelet Decomposition. We apply J-level 1D Discrete Wavelet Transform to the projected signal:

$$(cA_J, [cD_J, cD_{J-1}, ..., cD_1]) = \text{DWT}(x_{mono})$$

This decomposition yields approximation coefficients cA_J capturing low-frequency, large-scale trends and detail coeffi-

cients cD_j for $j \in [1, J]$ capturing high-frequency, fine-scale patterns.

Step 3: Learnable Scale Embeddings. We introduce learnable embedding vectors serving as "prototypes" for each temporal scale:

$$E_{scales} = \{e_{A_J}, e_{D_J}, e_{D_{J-1}}, ..., e_{D_1}\}$$

where each embedding $e \in R^{d_{model}}$ corresponds to a specific scale captured by the DWT.

Step 4: Dynamic Modulation. The core innovation is the dynamic modulation mechanism, where actual wavelet coefficients modulate learnable scale embeddings through a gating function:

$$\text{gate}(e, c) = (\sigma(W_g e) \odot \tanh(W_v e)) \otimes c'$$

where $W_g, W_v \in R^{d_{model} \times d_{model}}$ are learnable weight matrices, σ is sigmoid activation, and c' is the coefficient tensor broadcasted appropriately.

This generates modulated coefficients for all scales:

$$Yl_{\text{mod}} = \text{gate}(e_{A_J}, cA_J)$$
$$Yh_{\text{mod}} = [\text{gate}(e_{D_J}, cD_J), ..., \text{gate}(e_{D_1}, cD_1)]$$

Step 5: Reconstruction. The final positional encoding is reconstructed using Inverse DWT:

$$P_{DyWPE} = \text{IDWT}(Yl_{mod}, Yh_{mod})$$

leveraging the perfect reconstruction property of wavelets to synthesize modulated multi-scale information back into a sequence of length L.

4. EXPERIMENTAL EVALUATION

4.1. Experimental Setup

We conduct comprehensive experiments across ten diverse time series datasets spanning multiple domains such as Human Activity Recognition, Audio, and EEG classification, with eight datasets from the UEA archive [15] and two additional datasets [16, 17], as shown in Table 1.

We evaluated DyWPE against eight established positional encoding methods using a patchTST. All experiments use consistent hyperparameters: 4 transformer layers, 4 attention heads, 128 hidden dimensions, dropout rate 0.2, and an Adam optimizer. The complete code implementation and benchmarks are made publicly available for reproducibility: `https://github.com/imics-lab/DyWPE`.

4.2. Comparative Results

Table 2 presents comprehensive accuracy results across all datasets and methods, and Fig. 2 visualizes the same results in the form of a box plot. DyWPE demonstrates consistent superior performance, achieving the highest accuracy on 6 out of 10 datasets and ranking in the top 2 for the remaining datasets.

Table 1. Time series dataset properties.

Dataset	Train	Test	Len	Cls	Ch	Type
Sleep (Sl)	478,785	90,315	178	5	1	EEG
ElectricDevices (ED)	8,926	7,711	96	7	1	Device
FaceDetection (FD)	5,890	3,524	62	2	144	EEG
MelbournePedestrian (MP)	1,194	2,439	24	10	1	Traffic
LSST (LS)	2,459	2,466	36	14	6	Other
SelfRegulationSCP1 (SR1)	268	293	896	2	6	EEG
SelfRegulationSCP2 (SR2)	200	180	1152	2	6	EEG
JapaneseVowels (JV)	270	370	29	9	12	AUDIO
UniMiB-SHAR (UM)	4,601	1,524	151	9	3	HAR
RoomOccupancy (RO)	8,103	2,026	30	4	18	Sensor

Table 2. Classification accuracy comparison across all PE methods and datasets. All numbers are the average of 5 runs.

Data	Absolute PE		Relative PE			Hybrid PE			Ours
	Learn.	tAPE	eRPE	ALiBi	SPE	RoPE	T-PE	TUPE	DyWPE
Sl	85.2	85.1	86.4	85.3	87.6	84.4	87.2	87.9	**88.2**
ED	69.4	69.3	76.2	68.4	**81.1**	71.3	76.3	77.9	79.1
FD	64.2	65.3	67.8	62.6	67.4	63.5	68.6	**69.5**	68.8
MP	70.2	68.2	73.3	67.2	**75.3**	69.0	74.2	74.5	74.8
LS	58.2	58.4	61.1	59.2	60.1	58.3	60.2	59.5	**62.2**
SR1	84.4	84.2	85.6	84.9	88.3	83.1	87.1	87.5	**89.3**
SR2	54.6	53.3	56.4	58.6	58.2	53.1	56.6	59.3	**61.2**
JV	95.8	95.8	96.0	98.1	98.7	96.8	98.9	97.9	**99.2**
UM	84.4	83.3	86.4	84.1	86.7	83.6	**87.1**	86.5	86.7
RO	91.1	92.2	92.9	91.5	93.4	91.7	93.1	93.7	**94.8**

4.3. Performance Analysis

Sequence Length Effects: DyWPE demonstrates strong performance across varying sequence lengths, with particularly notable advantages on longer, complex sequences. On the longest sequence (SR2, 1152 timesteps), DyWPE achieves 61.2% accuracy, substantially outperforming most methods. For other long sequences, DyWPE shows consistent improvements: Sleep (88.2%), SR1 (89.3%), and competitive performance on ElectricDevices and UniMiB-SHAR. DyWPE maintains robust performance across diverse signal types and lengths, demonstrating the value of signal-aware positioning for complex temporal patterns.

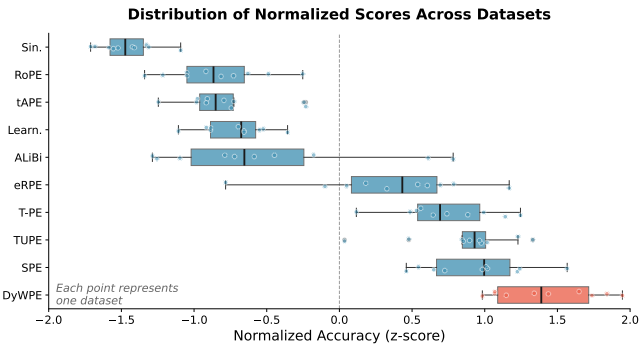

Fig. 2. Distribution of z-score normalized (mean = 0, std = 1) classification accuracy across 10 datasets for each PE method. Box plots show the median (dark line), interquartile range (box), and data range (whiskers), with outliers shown as isolated points.

Table 3. Computational Complexity Comparison of Positional Encoding Methods and Training Time Analysis for PE methods on all Dataset. Parameters: L = sequence length, d = dimension, h = attention heads, l = transformer layers, K = kernel size, R = representation dimension. Relative Overhead is average on 10 datasets based on baseline model (No PE).

Method	Params	Memory	Time Complex.	Rel. Overh.
tAPE	Ld	$\mathcal{O}(Ld)$	$\mathcal{O}(Ld)$	1.07
RoPE	0	$\mathcal{O}(Ld)$	$\mathcal{O}(L^2d)$	1.10
Learn.	Ld	$\mathcal{O}(Ld)$	$\mathcal{O}(Ld)$	1.12
ALiBi	0	$\mathcal{O}(L^2h)$	$\mathcal{O}(L^2h)$	1.28
SPE	$3Kdh{+}dl$	$\mathcal{O}(LKR)$	$\mathcal{O}(LKR)$	1.37
TUPE	$2dl$	$\mathcal{O}(Ld{+}d^2)$	$\mathcal{O}(Ld{+}d^2)$	1.45
DyWPE	$2d^2 + \lfloor \log_2(L) \rfloor d$	$\mathcal{O}(Ld)$	$\mathcal{O}(Ld)$	1.48
eRPE	$(L^2{+}L)l$	$\mathcal{O}(L^2d)$	$\mathcal{O}(L^2d)$	1.71
T-PE	$2d^2l/h{+}(2L{+}2l)d$	$\mathcal{O}(L^2d)$	$\mathcal{O}(L^2d)$	1.95

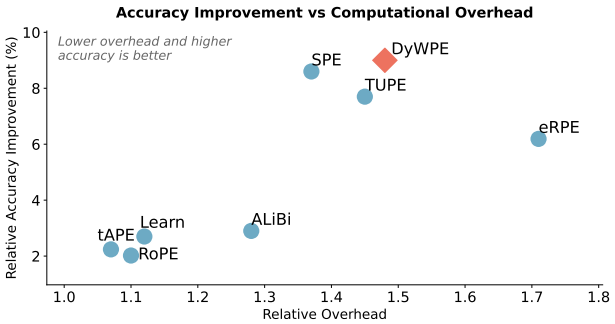

Fig. 3. Average relative accuracy improvement vs. computational overhead trade-off of different SOTA PE methods on all datasets.

Domain-Specific Performance: In biomedical signal processing (Sleep EEG, SelfRegulationSCP1, SelfRegulationSCP2), DyWPE demonstrates consistent improvements, achieving top performance on these datasets by effectively capturing multi-scale physiological dynamics. For sensor and device data, results are more variable: RoomOccupancy shows strong performance (94.8%), while ElectricDevices lags behind specialized methods like SPE.

Computational Efficiency: The asymptotic time and space complexities of our method scale linearly with the length of the sequence, unlike other SOTA PE methods, which scale quadratically. The fact that our method uses the input signal adds some practical computational overhead compared to signal-agnostic methods; however, the relative added overhead does not exceed that of other SOTA methods. Table 3 provides a detailed comparison of computational characteristics across different positional encoding methods, and Figure 3 provides a visual depiction of the tradeoff between accuracy gains for different PE methods versus practical computational overhead compared to the baseline.

4.4. Ablation Study

We conduct three critical ablation experiments to validate DyWPE's core contributions: signal-awareness, multi-scale representation, and wavelet types.

Table 4. DyWPE: complete signal-aware multi-scale approach; Static PE: multi-scale framework with fixed coefficients; Single-Scale: signal-aware with J=1 decomposition.

Method	Sl	ED	FD	MP	LS	SR1	SR2	JV	UM	RO
DyWPE	**88.2**	79.1	**68.8**	**74.8**	**62.2**	**89.3**	**59.4**	**99.2**	86.7	**94.8**
Static Wavelet	86.5	77.4	67.4	74.1	61.0	87.7	57.9	98.4	85.8	93.8
Single-Scale	86.1	75.9	66.4	74.5	59.5	**89.3**	52.1	**99.2**	**87.7**	93.3

Signal-Awareness Validation: To isolate the impact of signal-aware modulation, we compare DyWPE against Static Wavelet PE (SWPE), which retains the complete DWT/IDWT framework and learnable scale embeddings but removes signal dependency. SWPE uses fixed, learnable tensors instead of dynamic coefficient modulation, creating a signal-agnostic baseline with identical architectural complexity. Results in Table 4 demonstrate that signal-awareness provides improvements across all 10 datasets (average +1.09%), with particularly strong gains on Sleep (+1.7%), SR2 (+1.5%), and FaceDetection (+1.4%). This confirms that dynamic adaptation to signal characteristics drives performance improvements beyond the multi-scale framework alone.

Multi-Scale Analysis Validation: We evaluate the necessity of hierarchical temporal decomposition by comparing full DyWPE against simplified single-scale variants. The multi-scale analysis shows variable effectiveness across datasets (7 out of 10 show improvements, average +1.60%), with exceptional gains on SR2 (+7.3%) and substantial improvements on Sleep (+2.1%) and FaceDetection (+2.4%). However, UniMiB-SHAR shows negative results (-1.0%), indicating that multi-scale decomposition benefits depend on signal complexity. Datasets with rich temporal hierarchies benefit most from multi-scale analysis, while simpler patterns may not require deep decomposition.

Effects of Different Wavelet Types: We evaluated 11 wavelet families (Daubechies, Coiflets, Biorthogonal, Haar). Our experiments show that while db4 remains a robust default, bior2.2 showed a slight improvement on complex signals, suggesting that biorthogonal wavelets' reconstruction properties may further benefit signal-aware encoding.

5. CONCLUSION

We introduced Dynamic Wavelet Positional Encoding (DyWPE), the first signal-aware position encoding framework for time series classification. By analyzing actual signal content through multi-scale wavelet decomposition and dynamically modulating learnable scale embeddings, DyWPE creates rich positional representations that adapt to local temporal characteristics. Comprehensive experiments demonstrate consistent superiority over SOTA methods, with particularly significant improvements on longer sequences and complex signals.

6. REFERENCES

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

[2] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 2114–2124.

[3] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[4] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.

[5] G. Ke, D. He, and T.-Y. Liu, "Rethinking positional encoding in language pre-training," in *International Conference on Learning Representations*.

[6] N. M. Foumani, C. W. Tan, G. I. Webb, and M. Salehi, "Improving position encoding of transformers for multivariate time series classification," *Data Mining and Knowledge Discovery*, vol. 38, no. 1, pp. 22–48, 2024.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[8] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," *arXiv preprint arXiv:1704.02971*, 2017.

[9] M. Liu, S. Ren, S. Ma, J. Jiao, Y. Chen, Z. Wang, and W. Song, "Gated transformer networks for multivariate time series classification," *arXiv preprint arXiv:2103.14438*, 2021.

[10] J.-B. Cordonnier, A. Mahendran, A. Dosovitskiy, D. Weissenborn, J. Uszkoreit, and T. Unterthiner, "Differentiable patch selection for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2351–2360.

[11] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

[12] H. Irani and V. Metsis, "Positional encoding in transformer-based time series models: a survey," *arXiv preprint arXiv:2502.12370*, 2025.

[13] A. Bell, M. Gray, and R. King, "Adaptive positional encoding mechanisms for dynamic time series," *Expert Systems with Applications*, vol. 220, p. 119678, 2023.

[14] Y. Oka, T. Hasegawa, K. Nishida, and K. Saito, "Wavelet-based positional representation for long context," in *The Thirteenth International Conference on Learning Representations*.

[15] A. Bagnall, H. A. Dau, J. Lines, M. Flynn, J. Large, A. Bostrom, P. Southam, and E. Keogh, "The uea multivariate time series classification archive, 2018," *arXiv preprint arXiv:1811.00075*, 2018.

[16] D. Micucci, M. Mobilio, and P. Napoletano, "Unimib shar: A dataset for human activity recognition using acceleration data from smartphones," *Applied Sciences*, vol. 7, no. 10, p. 1101, 2017.

[17] A. P. Singh and S. Chaudhari, "Room Occupancy Estimation," UCI Machine Learning Repository, 2018, DOI: https://doi.org/10.24432/C5P605.