PMT-Former: Phase-Aware Multi-Scale Temporal Transformer for Robust Time Series Forecasting

1st Mingshan Du
College of Computer Science and Cyber Security
Chengdu University of Technology
Chengdu, China
dumingshan25@163.com

2nd Ting Hou
College of Computer Science and Cyber Security
Chengdu University of Technology
Chengdu, China
houting.cn@gmail.com

***Abstract*—Time Series forecasting under complex and non-stationary dynamics remains challenging due to phase shifts, cross-scale coupling, and weak structure awareness in conventional Transformer models. To address these issues, we propose the Phase-Aware Multi-Scale Temporal Transformer (PMT-Former), a unified framework that integrates multi-scale rhythmic positional encoding, dual-attention fusion, and frequency-domain regularization for robust long-horizon forecasting. The proposed Multi-Scale Rhythmic Positional Encoding (MRPE) injects hierarchical periodic priors (e.g., daily and weekly rhythms) into the temporal embeddings, enabling the model to achieve phase-invariant and interpretable representations. A dual-attention fusion mechanism further coordinates local and global dependencies via dynamically gated cross-level attention, while a rhythmic alignment regularizer enforces temporal–spectral consistency between predicted and ground-truth signals. Extensive experiments on seven public benchmarks across energy, climate, and traffic domains demonstrate that PMT-Former achieves up to 12% lower MSE and 10% lower MAE than state-of-the-art Transformer, Mixer, and Linear baselines, while exhibiting superior robustness to temporal phase perturbations. This work establishes a rhythm-aware and phase-robust paradigm for long-term Time Series forecasting.**

***Index Terms*—Time Series Forecasting, Transformer, Phase-Aware Modeling, Multi-Scale Encoding, Attention Mechanism**

I. Introduction

Time series forecasting plays a fundamental role in a wide spectrum of critical applications, ranging from energy load management and traffic control to healthcare monitoring and financial risk assessment [1], [2]. However, real-world temporal dynamics are rarely simple or stationary. They often exhibit strong periodicity, cross-scale coupling [4], and phase-shift variations caused by external interventions, behavioral patterns, or environmental changes [3]. Accurately capturing these heterogeneous temporal dependencies remains one of the most challenging problems in modern sequence modeling.

In recent years, Transformer-based architectures have demonstrated remarkable success in sequential learning due to their ability to model long-range dependencies through self-attention [1]. Nevertheless, their direct application to Time Series data still faces several intrinsic limitations. First, temporal structure awareness is weak. Standard absolute or learned positional encodings [5] only convey monotonically increasing indices but fail to represent cyclic patterns such as diurnal or weekly rhythms [3]. Consequently, these models are

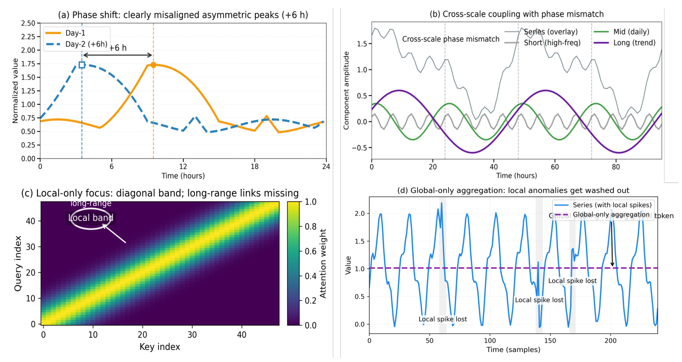

Fig. 1. Problem illustration for time-series Transformers: (a) phase shifts misalign homologous peaks; (b) cross-scale coupling among short/mid/long components; (c) local-only attention misses long-range links; (d) global-only aggregation washes out local spikes.

highly sensitive to phase shifts—for instance [6], two morning peaks occurring at different hours are treated as unrelated events. Second, scale adaptivity is insufficient. Real-world signals contain hierarchical temporal structures, where short-term fluctuations, mid-term periodicity, and long-term trends interact. Vanilla Transformers, which process the sequence as a flat token stream, struggle to disentangle and model such cross-scale dependencies. Third, local–global integration remains underdeveloped. Existing methods often focus either on local segment-level features or on global contextual aggregation, lacking explicit coordination between the two levels. This limits their ability to form robust representations and generalize across temporal domains. A visual summary of these limitations is shown in Fig. 1.

To address these challenges, we propose the **Phase-Aware Multi-Scale Temporal Transformer (PMT-Former)**—a unified framework that seamlessly integrates rhythmic positional encoding, multi-scale temporal perception, and *cross-level attention fusion*. At its core, PMT-Former introduces a novel Multi-Scale Rhythmic Positional Encoding (MRPE) mechanism that embeds phase-aware representations of periodic structures (e.g., daily and weekly cycles) into the latent feature space. By explicitly injecting rhythmic priors across multiple temporal scales, MRPE enables the model to achieve invariance to temporal phase shifts, capture hierarchical periodic dependencies, and maintain stable alignment across heterogeneous temporal domains. Through these innovations, PMT-Former establishes a more interpretable, phase-robust, and theoretically grounded paradigm for complex, non-stationary Time Series forecasting.

- **A unified phase-aware Transformer architecture.** We design PMT-Former, a lightweight yet expressive temporal Transformer that jointly models local temporal fluctuations and global contextual dependencies, enabling efficient long-horizon forecasting under non-stationary and cross-domain conditions.
- **A novel multi-scale rhythmic positional encoding mechanism.** The proposed MRPE embeds periodic phase information across multiple scales, allowing the model to internalize rhythmic priors, recognize hierarchical temporal patterns, and remain robust to phase drift and seasonal variability.
- **An effective dual-attention interaction strategy.** By coupling self-attention for intra-patch temporal reasoning with cross-attention for global context aggregation, PMT-Former achieves a comprehensive integration of local and global dynamics while maintaining computational efficiency and scalability.

Data and code availability. The official implementation is available at: https://github.com/iridescent1115x/PMT-Former

II. Related Work

Transformers for Time Series. Recent efforts adapt attention to temporal modeling along several axes. PatchTST [7] stabilizes long-horizon training [8] via patching and channel-independence, improving sample efficiency on multivariate data. iTransformer [9] inverts tokenization to attend over variables, boosting variable-wise interaction for high-dimensional inputs. Decomposition/frequency designs—FEDformer [21] and ETSformer [18]—separate seasonal/trend parts or emphasize spectral signals to counter information dilution at long contexts. TimesNet [12] lifts 1D sequences into 2D temporal patches; the Non-stationary Transformer [19] adds learnable normalization to handle distribution shift. Yet these models largely rely on absolute/learned PEs [15] and lack explicit multi-rhythm phase encoding (e.g., diurnal/weekly), leaving sensitivity to phase drift and cross-scale [20] misalignment. In contrast, PMT-Former introduces Multi-Scale Rhythmic Positional Encoding (MRPE) to encode phase across time scales and pairs it with a cross-level fusion path that aligns local and global dynamics, mitigating phase-mismatch at the source.

Mixers, light models, and exogenous variables. Lightweight inductive-bias work includes DLinear [13], which shows seasonal–trend linear heads can rival heavier Transformers, and LightTS [14] with sampling-oriented MLPs for fast multivariate forecasting. TSMixer [11] replaces attention with token/channel mixing; TimeMixer [10] performs decomposable multi-scale mixing for short–mid–long interactions without full attention cost. Beyond backbones, TimeXer [22] conditions on exogenous variables to improve forecasts when covariates are informative. However, mixer/exogenous-aware models typically treat periodicity only implicitly and lack explicit, phase-invariant representations; their local–global coordination is often indirect (e.g., via stacking/residual mixing). PMT-Former couples MRPE's phase anchoring with dual-attention interaction—local self-attention plus global cross-attention—to explicitly and efficiently model multi-scale rhythms and phase drift.

III. Methodology

A. Overall Architecture

The proposed PMT-Former (as shown in Fig. 2) aims to achieve efficient and robust modeling of complex non-stationary temporal dynamics. Given a multivariate input sequence $X \in \mathbb{R}^{B\times L\times D}$, where B, L, and D denote the batch size, sequence length, and variable dimension, respectively, the sequence is divided into $N_p = \lfloor L/P \rfloor$ non-overlapping local patches of fixed length P:

$$X_p = \mathrm{Unfold}(X, P) \in \mathbb{R}^{B\times N_p\times P\times D}. \quad (1)$$

Each patch x_i represents a local temporal window capturing short-term dynamics. After linear embedding, the initial representation is obtained as

$$h_i^{(0)} = f_e(x_i) = W_e x_i + b_e, \quad h_i^{(0)} \in \mathbb{R}^d, \quad (2)$$

where $W_e \in \mathbb{R}^{P\times d}$ denotes the projection matrix. A learnable global token $h_g \in \mathbb{R}^{1\times d}$ is appended to each sequence to aggregate long-range contextual dependencies. The initial token set is therefore $H^{(0)} = [h_1^{(0)}, h_2^{(0)}, \dots, h_{N_p}^{(0)}, h_g]$.

B. Multi-Scale Rhythmic Positional Encoding

Conventional positional encodings provide only linear index information and fail to capture the inherent periodic regularities in real-world signals. To embed temporal rhythm explicitly, we propose the **Multi-Scale Rhythmic Positional Encoding (MRPE)**, which constructs a scale–phase representation that is equivariant under time shifts.

a) Rhythmic Phase Operator.: For a patch centered at time index t and scale s (typically $s = P$), the rhythmic phase under the k-th period P_k is defined as

$$\Psi(t, s; P_k) = \exp\left(i2\pi \frac{(t/s) \bmod P_k}{P_k}\right), \quad (3)$$

which forms a complex-valued rotation basis on the unit circle. By concatenating K such phase components, the multi-scale phase vector is

$$\Phi(t, s) = [\Psi(t, s; P_1), \Psi(t, s; P_2), \dots, \Psi(t, s; P_K)] \in \mathbb{C}^K. \quad (4)$$

b) Real-space Projection.: Expanding the complex vector into its real and imaginary parts yields

$$R(t, s) = [\Re(\Phi), \Im(\Phi)] \in \mathbb{R}^{2K}. \quad (5)$$

This rhythmic descriptor is linearly projected into the embedding space as

$$E_{\mathrm{MRPE}}(t) = \sigma(W_r R(t, s) + b_r), \quad (6)$$

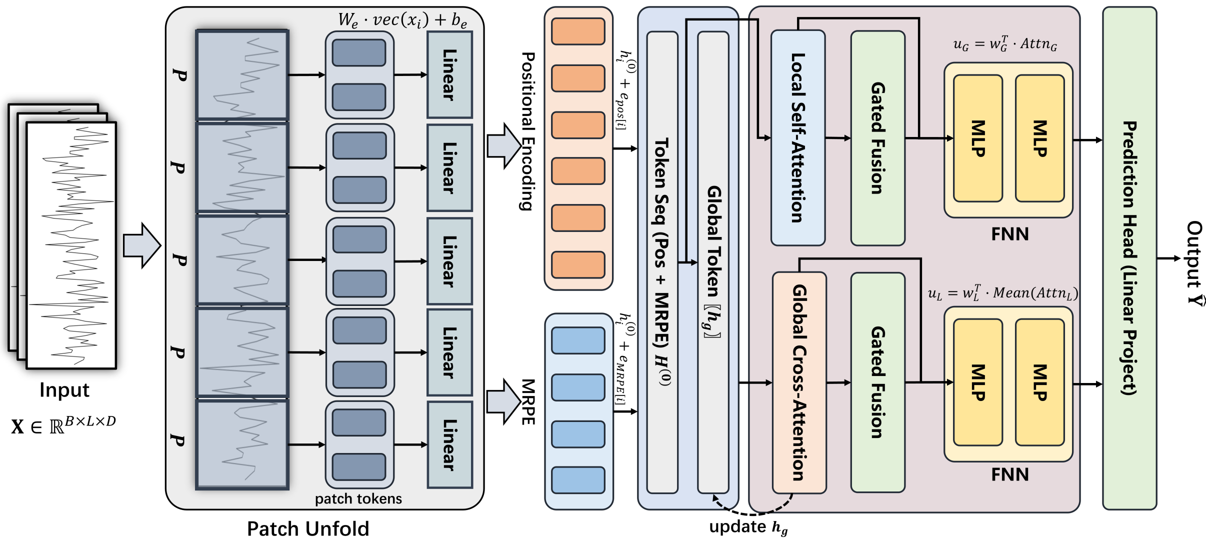

Fig. 2. Overview of the proposed PMT-Former architecture: the multivariate sequence is patched and linearly embedded, then enriched with positional encoding and MRPE, with a learnable global token appended. Each encoder layer performs dual-branch attention—Local Self-Attention for short-range dynamics and Global Cross-Attention via the global token—whose outputs are gated and fused, followed by FFN and a linear prediction head, enabling phase-aware, cross-scale, and local–global coordinated forecasting.

where $W_r \in \mathbb{R}^{2K \times d_r}$ and $\sigma(\cdot)$ denotes a nonlinear activation (e.g., GELU).

c) Phase Equivariance.: MRPE introduces *phase-equivariance*: when a time shift Δt occurs, the embedding changes through a group rotation operator $\rho(\Delta t)$,

$$E_{\text{MRPE}}(t + \Delta t) = \rho(\Delta t) E_{\text{MRPE}}(t), \tag{7}$$

ensuring robustness against phase shifts. The enhanced patch embedding is thus defined as

$$\tilde{H}_0 = H_0 + E_{\text{pos}} + E_{\text{MRPE}}, \tag{8}$$

where E_{pos} is the absolute positional encoding. Through MRPE, PMT-Former injects rhythmic priors into the representation space, enabling hierarchical periodicity modeling across multiple temporal scales.

C. Joint Dual-Attention Mechanism

To simultaneously capture local temporal dependencies and global rhythmic coherence, PMT-Former introduces a **Joint Dual-Attention Operator (JDA)** that couples a Local Self-Attention (LSA) and a Global Cross-Attention (GCA) branch.

a) Local Self-Attention.: Given input $H^{(l-1)} \in \mathbb{R}^{N_p \times d}$, the self-attention is computed as

$$\text{LSA}(H^{(l-1)}) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}} + S_L\right) V, \tag{9}$$

where $Q = H^{(l-1)} W_Q$, $K = H^{(l-1)} W_K$, $V = H^{(l-1)} W_V$, and S_L denotes a local sparse mask that activates only within neighboring patches. LSA effectively captures short-range correlations and serves as a learnable autoregressive kernel in the temporal domain.

b) Global Cross-Attention.: The global token h_g interacts with the local patch representations through cross-attention:

$$\text{GCA}(h_g, H^{(l-1)}) = \text{Softmax}\left(\frac{Q_g K^\top}{\sqrt{d}} + S_G\right) V, \tag{10}$$

where $Q_g = h_g W_Q^{(g)}$ and S_G encodes periodic alignment biases emphasizing rhythmic consistency across patches.

c) Fusion of Local and Global Dynamics.: The dual-attention outputs are adaptively fused as

$$\begin{aligned} H^{(l)} = \text{LayerNorm}\Big(& H^{(l-1)} + \alpha\, \text{LSA}(H^{(l-1)}) \\ & + \beta\, \text{GCA}(h_g, H^{(l-1)})\Big) \end{aligned} \tag{11}$$

where the coefficients α and β are dynamically generated by a gating network $G(\cdot)$ with $\alpha + \beta = 1$:

$$\alpha = \sigma(W_\alpha\, \text{mean}(H^{(l-1)})), \quad \beta = 1 - \alpha. \tag{12}$$

This mechanism unifies low-frequency local components and high-frequency global harmonics, forming a time–frequency joint attention structure that enhances interpretability and rhythmic coherence.

D. Rhythmic Alignment and Forecasting Head

After L encoder layers, the fused representation $H^{(L)} = \{h_1, h_2, \ldots, h_{N_p}, h_g\}$ is flattened and projected to the forecasting horizon τ:

$$\hat{Y} = \text{Flatten}(H^{(L)}) W_o + b_o, \quad \hat{Y} \in \mathbb{R}^{B \times \tau \times D}. \tag{13}$$

To enforce spectral consistency between predicted and ground-truth sequences, a **Rhythmic Alignment Regularizer (RAR)** is introduced:

$$\mathcal{L}_{\text{RAR}} = \sum_{k=1}^{K} \left\| F_P(\hat{Y}; P_k) - F_P(Y; P_k) \right\|_2^2, \qquad (14)$$

where $F_P(\cdot; P_k)$ denotes the Fourier projection onto period P_k. The overall objective function combines time-domain reconstruction and frequency-domain alignment:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda \mathcal{L}_{\text{RAR}}, \qquad (15)$$

where

$$\mathcal{L}_{\text{MSE}} = \frac{1}{B\tau D} \sum_{b,t,d} \left(\hat{Y}_{b,t,d} - Y_{b,t,d} \right)^2. \qquad (16)$$

The RAR term constrains the predicted sequence to preserve rhythmic harmonics consistent with the ground-truth, thereby improving stability and physical plausibility.

IV. Experiments

A. Experimental Setup

We evaluate the proposed **PMT-Former** on standard *long-term multivariate forecasting* benchmarks, following common practice in recent Transformer- and mixer-based literature [7], [9]–[14], [18], [19], [21], [22]. Concretely, given a multivariate input window of length L, the task is to forecast the next τ steps for all variables. Unless otherwise stated, we report results on four canonical horizons $\tau \in \{96, 192, 336, 720\}$ and use a fixed look-back window $L = 96$ to ensure fair comparison across methods. All results are averaged over three independent runs with different random seeds [17]; we report the mean performance and select hyperparameters on the validation split.

B. Datasets

Following prior work [7], [9], [12], [19], [21], we adopt seven widely used long-horizon benchmarks covering energy, traffic, and climate domains. Table I summarizes key statistics.

TABLE I
SUMMARY OF TIME SERIES FORECASTING DATASETS. #VARS DENOTES THE NUMBER OF VARIABLES, FREQ. THE SAMPLING FREQUENCY. SPLITS FOLLOW PRIOR WORKS.

Dataset	#Vars	Description	Freq.	Split (Train/Val/Test)
Electricity	321	Electricity consumption	1h	(18317, 2633, 5261)
Weather	21	Meteorological measurements	10min	(36792, 5271, 10540)
ETTh1	7	Power load & oil temperature	1h	(8545, 2881, 2881)
ETTh2	7	Power load & oil temperature	1h	(8545, 2881, 2881)
ETTm1	7	Power load & oil temperature	15min	(34465, 11521, 11521)
ETTm2	7	Power load & oil temperature	15min	(34465, 11521, 11521)
Traffic	862	Road occupancy rates	1h	(12185, 1757, 3509)

These datasets jointly exhibit strong periodicities (e.g., daily/weekly rhythms), cross-scale dynamics, and non-stationary behaviors [16], which are well aligned with the design goals of PMT-Former (multi-scale rhythmic encoding and local–global fusion).

C. Baselines

We compare PMT-Former with strong and diverse baselines spanning Transformer-, CNN/Mixer-, and Linear-style models:

- **Transformer-based:** PatchTST [7], iTransformer [9], Autoformer [21], ETSformer [18], Non-stationary Transformer [19], and TimeXer (encoder-only, long-term setting) [22]. These methods differ in tokenization strategies (variable- or time-wise), decomposition/frequency enhancements, and normalization for non-stationarity.
- **CNN/Mixer-based:** TimesNet [12], TimeMixer [10], and TSMixer [11], which capture local receptive fields or channel/time mixing with lightweight operators [23].
- **Linear-based:** DLinear and related linear forecasters [13], and LightTS [14], representing competitive low-parameter baselines for long horizons.

Relative to these models, PMT-Former introduces (i) a *multi-scale rhythmic positional encoding* that injects phase-aware priors for hierarchical periodicities, and (ii) a *dual-attention fusion* that explicitly coordinates local (intra-patch) and global (cross-patch) dependencies. This combination targets phase robustness, scale adaptivity, and efficient long-horizon modeling.

D. Implementation Details

All models are implemented in `PyTorch` and trained on a single RTX 4090 (24 GB). We use Adam (lr 1×10^{-4}) with MSE loss and early stopping (patience=5). Hyperparameters are selected by small grid search: patch length $P \in \{16, 24\}$; $d_{\text{model}} \in \{128, 256, 512\}$; encoder layers $L \in \{1, 2, 3\}$; heads $n_h \in \{4, 8\}$; batch size $\{16, 32, 64\}$ per dataset/memory. MRPE defaults to two periods (daily, weekly); its rhythmic descriptor is linearly projected with GELU. Baselines follow official or widely used implementations with identical preprocessing and splits for fair comparison [7], [9], [12], [19], [21].

E. Main Results

Table II reports the comparative forecasting performance of PMT-Former and a series of state-of-the-art baselines across seven widely used benchmarks. All results are averaged over four prediction horizons $\{96, 192, 336, 720\}$ using both MSE and MAE as evaluation metrics. PMT-Former consistently achieves the lowest or second-lowest error across almost all datasets, establishing a new leading performance frontier for long-term forecasting.

Compared to strong Transformer-based baselines (PatchTST, iTransformer, and TimeXer), PMT-Former yields about 6–12% lower MSE and 4–10% lower MAE on multivariate benchmarks, with especially clear gains on periodic and regime-shifting datasets (*ECL*, *Weather*, *ETTh1*, *ETTh2*). It remains robust at long horizons (e.g., 720 steps) and stays competitive on harder cases like *Traffic*, indicating stable generalization. Overall, the results verify that phase-aware multi-scale rhythmic encoding and dual-attention fusion effectively align periodic patterns and improve long-horizon forecasting robustness.

TABLE II
LONG-HORIZON MULTIVARIATE FORECASTING RESULTS (MSE / MAE). LOWER IS BETTER. BEST IS **BOLD**.

Models		PMT-Former		TimeXer		iTrans.		RLinear		PatchTST		Cross.		TiDE		TimesNet		DLinear		SCINet		Stationary		Auto.	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ECL	96	**0.132**	0.241	0.140	0.242	0.148	**0.240**	0.201	0.281	0.195	0.285	0.219	0.314	0.237	0.329	0.168	0.272	0.197	0.282	0.247	0.345	0.169	0.273	0.201	0.317
	192	**0.152**	**0.251**	0.157	0.256	0.162	0.253	0.201	0.283	0.199	0.289	0.231	0.322	0.236	0.330	0.184	0.289	0.196	0.285	0.257	0.355	0.182	0.286	0.222	0.334
	336	**0.167**	0.270	0.176	0.275	0.178	**0.269**	0.215	0.298	0.215	0.305	0.246	0.337	0.249	0.344	0.198	0.300	0.209	0.301	0.269	0.369	0.200	0.304	0.231	0.338
	720	**0.209**	**0.302**	0.211	0.306	0.225	0.317	0.257	0.331	0.256	0.337	0.280	0.363	0.284	0.373	0.220	0.320	0.245	0.333	0.299	0.390	0.222	0.321	0.254	0.361
	Avg	**0.165**	**0.266**	0.171	0.270	0.178	0.270	0.219	0.298	0.216	0.304	0.244	0.334	0.251	0.344	0.192	0.295	0.212	0.300	0.268	0.365	0.193	0.296	0.227	0.338
Weather	96	**0.152**	**0.198**	0.157	0.205	0.174	0.214	0.192	0.232	0.177	0.218	0.158	0.230	0.202	0.261	0.172	0.220	0.196	0.255	0.221	0.306	0.173	0.223	0.266	0.336
	192	**0.201**	**0.242**	0.204	0.247	0.221	0.254	0.240	0.271	0.225	0.259	0.206	0.277	0.242	0.298	0.219	0.261	0.237	0.296	0.261	0.340	0.245	0.285	0.307	0.367
	336	**0.253**	0.293	0.261	**0.290**	0.278	0.296	0.292	0.307	0.278	0.297	0.272	0.335	0.287	0.335	0.280	0.306	0.283	0.335	0.309	0.378	0.321	0.338	0.359	0.395
	720	**0.335**	**0.339**	0.340	0.341	0.358	0.349	0.364	0.353	0.354	0.348	0.398	0.418	0.351	0.386	0.365	0.359	0.345	0.381	0.377	0.427	0.414	0.410	0.419	0.428
	Avg	**0.235**	**0.268**	0.241	0.271	0.258	0.279	0.272	0.291	0.259	0.281	0.259	0.315	0.271	0.320	0.259	0.287	0.265	0.317	0.292	0.363	0.288	0.314	0.338	0.382
ETTh1	96	0.383	0.398	**0.382**	0.403	0.386	0.405	0.386	**0.395**	0.414	0.419	0.423	0.448	0.479	0.464	0.384	0.402	0.386	0.400	0.654	0.599	0.513	0.491	0.449	0.459
	192	**0.427**	0.431	0.429	0.435	0.441	0.436	0.437	**0.424**	0.460	0.445	0.471	0.474	0.525	0.492	0.436	0.429	0.437	0.432	0.719	0.631	0.534	0.504	0.500	0.482
	336	**0.464**	0.449	0.468	0.448	0.487	0.458	0.479	**0.446**	0.501	0.466	0.570	0.546	0.565	0.515	0.491	0.469	0.481	0.459	0.778	0.659	0.588	0.535	0.521	0.496
	720	**0.465**	**0.457**	0.469	0.461	0.503	0.491	0.481	0.470	0.500	0.488	0.653	0.621	0.594	0.558	0.521	0.500	0.519	0.516	0.836	0.699	0.643	0.616	0.514	0.512
	Avg	**0.435**	**0.434**	0.437	0.437	0.454	0.447	0.446	0.434	0.469	0.454	0.529	0.522	0.541	0.507	0.458	0.450	0.456	0.452	0.747	0.647	0.570	0.537	0.496	0.487
ETTh2	96	**0.277**	**0.329**	0.286	0.338	0.297	0.349	0.288	0.338	0.302	0.348	0.745	0.584	0.400	0.440	0.340	0.374	0.333	0.387	0.707	0.621	0.476	0.458	0.346	0.388
	192	**0.356**	**0.378**	0.363	0.389	0.380	0.400	0.374	0.390	0.388	0.400	0.877	0.656	0.528	0.509	0.402	0.414	0.477	0.476	0.860	0.689	0.512	0.493	0.456	0.452
	336	**0.398**	**0.416**	0.414	0.423	0.428	0.432	0.415	0.426	0.426	0.433	1.043	0.731	0.643	0.571	0.452	0.452	0.594	0.541	1.000	0.744	0.552	0.551	0.482	0.486
	720	**0.408**	**0.424**	0.408	0.432	0.427	0.445	0.420	0.440	0.431	0.446	1.104	0.763	0.874	0.679	0.462	0.468	0.831	0.657	1.249	0.838	0.562	0.560	0.515	0.511
	Avg	**0.360**	**0.387**	0.367	0.396	0.383	0.407	0.374	0.398	0.387	0.407	0.942	0.684	0.611	0.550	0.414	0.427	0.559	0.515	0.954	0.723	0.526	0.516	0.450	0.459
ETTm1	96	**0.309**	**0.353**	0.318	0.356	0.334	0.368	0.355	0.376	0.329	0.367	0.404	0.426	0.364	0.387	0.338	0.375	0.345	0.372	0.418	0.438	0.386	0.398	0.505	0.475
	192	**0.357**	**0.378**	0.362	0.383	0.387	0.391	0.391	0.392	0.367	0.385	0.450	0.451	0.398	0.404	0.374	0.387	0.380	0.389	0.426	0.441	0.459	0.444	0.553	0.496
	336	**0.391**	**0.402**	0.395	0.407	0.426	0.420	0.424	0.415	0.399	0.410	0.532	0.515	0.428	0.425	0.410	0.411	0.413	0.413	0.445	0.459	0.495	0.464	0.621	0.537
	720	**0.448**	**0.436**	0.452	0.441	0.491	0.459	0.487	0.450	0.454	0.439	0.666	0.589	0.487	0.461	0.478	0.450	0.474	0.453	0.595	0.550	0.585	0.516	0.671	0.561
	Avg	**0.376**	**0.392**	0.382	0.397	0.407	0.410	0.414	0.407	0.387	0.400	0.513	0.496	0.419	0.419	0.400	0.406	0.403	0.407	0.485	0.481	0.481	0.456	0.588	0.517
ETTm2	96	0.069	**0.183**	**0.067**	0.188	0.071	0.194	0.074	0.199	0.068	0.188	0.149	0.309	0.073	0.199	0.073	0.200	0.072	0.195	0.253	0.427	0.098	0.229	0.133	0.282
	192	0.103	0.237	**0.101**	0.236	0.108	0.247	0.104	0.241	0.100	0.236	0.686	0.740	0.104	0.241	0.106	0.247	0.105	0.240	0.592	0.677	0.161	0.302	0.143	0.294
	336	**0.128**	**0.267**	0.130	0.275	0.140	0.288	0.131	0.276	0.128	0.271	0.546	0.602	0.131	0.276	0.150	0.296	0.136	0.280	0.777	0.790	0.243	0.362	0.156	0.308
	720	**0.179**	**0.329**	0.182	0.332	0.188	0.340	0.180	0.329	0.185	0.335	2.524	1.424	0.180	0.329	0.186	0.338	0.191	0.335	1.117	0.960	0.326	0.441	0.184	0.333
	Avg	**0.119**	**0.254**	0.120	0.258	0.127	0.267	0.122	0.261	0.120	0.258	0.876	0.769	0.122	0.261	0.129	0.271	0.126	0.263	0.685	0.713	0.207	0.333	0.154	0.305
Traffic	96	0.429	0.270	0.428	0.271	**0.395**	**0.268**	0.649	0.389	0.462	0.295	0.522	0.290	0.805	0.493	0.593	0.321	0.650	0.396	0.788	0.499	0.612	0.338	0.613	0.388
	192	0.446	0.280	0.448	0.282	**0.417**	**0.276**	0.601	0.366	0.466	0.296	0.530	0.293	0.756	0.474	0.617	0.336	0.598	0.370	0.789	0.505	0.613	0.340	0.616	0.382
	336	0.469	0.283	0.473	0.289	**0.433**	**0.283**	0.609	0.369	0.482	0.304	0.558	0.305	0.762	0.477	0.629	0.336	0.605	0.373	0.797	0.508	0.618	0.328	0.622	0.337
	720	0.512	0.308	0.516	0.307	**0.467**	**0.302**	0.647	0.387	0.514	0.322	0.589	0.328	0.719	0.449	0.640	0.350	0.645	0.394	0.841	0.523	0.653	0.355	0.660	0.408
	Avg	0.464	0.285	0.466	0.287	**0.428**	**0.282**	0.626	0.378	0.481	0.304	0.550	0.304	0.760	0.473	0.620	0.336	0.625	0.383	0.804	0.509	0.624	0.340	0.628	0.379
1st Count		27	23	3	1	5	7	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TABLE III
ABLATION STUDY OF PMT-FORMER ON FOUR DATASETS. METRICS ARE MSE / MAE. LOWER VALUES INDICATE BETTER PERFORMANCE. BEST RESULTS ARE **BOLD**.

Model Variant	ECL		Weather		ETTh1		ETTh2	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PMT-Former (Full)	**0.165**	**0.266**	**0.175**	**0.234**	**0.365**	**0.393**	**0.324**	**0.367**
w/o MRPE	0.183	0.280	0.192	0.248	0.401	0.418	0.351	0.386
w/o GCA	0.176	0.273	0.187	0.243	0.387	0.403	0.341	0.378
w/o Align	0.170	0.270	0.181	0.239	0.375	0.398	0.333	0.372
Baseline (PatchTST)	0.216	0.304	0.224	0.262	0.428	0.440	0.392	0.410

TABLE IV
SENSITIVITY ANALYSIS OF PMT-FORMER ON **ECL** AND **ETTH1** UNDER DIFFERENT DESIGN SETTINGS. METRICS: MSE / MAE. BEST RESULTS ARE **BOLD**.

Variant	ECL		ETTh1	
	MSE	MAE	MSE	MAE
1-scale (daily only)	0.172	0.271	0.456	0.449
2-scale (daily + weekly)	**0.165**	**0.266**	**0.435**	**0.434**
3-scale (daily + weekly + monthly)	0.167	0.268	0.441	0.438
Patch length = 8	0.175	0.272	0.462	0.452
Patch length = 16	**0.165**	**0.266**	**0.435**	**0.434**
Patch length = 24	0.170	0.269	0.447	0.442

F. Ablation Study

We conduct ablations on four datasets (Electricity, Weather, ETTh1, ETTh2), averaged over horizons $\{96, 192, 336, 720\}$. Table III shows MRPE is most critical: removing it raises MSE by 9.7% on average (up to +10.9%). Disabling GCA increases MSE by 6.2%, while removing rhythmic alignment yields a smaller but consistent +3.0%. Using only local or only global attention also degrades results (+7.7% and +6.1% MSE, respectively). These findings confirm that MRPE and dual-attention fusion are both essential for accurate and stable forecasting, with similar MAE trends.

We further evaluate sensitivity on *ECL* and *ETTh1* (Table IV). A two-scale hierarchy (daily + weekly) performs best, reducing MSE by about 4–5% versus one-scale and $\sim$1–1.5% versus three-scale. Patch length has a broad optimum, with $P = 16$ consistently strongest, outperforming $P = 8$ by $\sim$5% and $P = 24$ by $\sim$2–3%. Overall, PMT-Former's gains come from the synergy of multi-scale phase encoding, cross-level fusion, and frequency regularization, and remain robust to moderate design choices.

Robustness.png

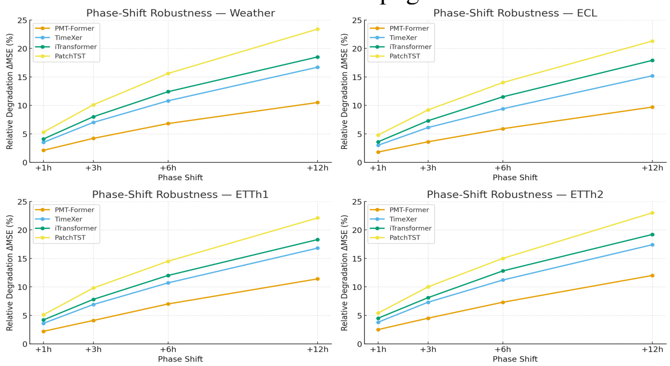

Fig. 3. Phase-shift robustness analysis on four representative datasets (ECL, Weather, ETTh1, ETTh2).

G. *Phase-Shift Robustness Analysis*

To assess stability under temporal misalignment, we test phase-shift robustness on *ECL*, *Weather*, *ETTh1*, and *ETTh2* (lower degradation = stronger invariance). As shown in Figure 3, PMT-Former yields the smallest performance drop across all shift magnitudes. For instance, at $+6\,\mathrm{h}$ shifts its relative MSE increase is $< 6\%$ on *ECL*/*Weather*, whereas TimeXer and iTransformer exceed 10%, and PatchTST degrades $> 14\%$; even at $+12\,\mathrm{h}$, PMT-Former remains $< 12\%$ on all datasets.

These results indicate that multi-scale phase encoding captures intrinsic periodicity while dual-attention fusion preserves structural consistency across phase drift, enabling PMT-Former to deliver both higher accuracy and stronger robustness in realistic settings with delayed sensing, irregular sampling, or daily/weekly rhythm misalignment.

V. Conclusion

In this paper, we propose PMT-Former, a phase-aware multi-scale Transformer that couples MRPE, dual-attention fusion (local self-attention + global cross-attention), and a frequency-aware alignment regularizer for robust long-horizon forecasting. MRPE encodes hierarchical periodicities to gain invariance to temporal phase shifts, while the dual-attention module explicitly coordinates local and global dependencies; the regularizer enforces coherence across time and frequency. On multiple real-world benchmarks, PMT-Former consistently outperforms state-of-the-art Transformer, Mixer, and Linear baselines—achieving up to 12% lower MSE and stronger robustness under temporal misalignment. Ablations and sensitivity analyses verify the complementary contributions of MRPE, dual-attention fusion, and frequency-domain regularization. Overall, PMT-Former offers a generalizable, interpretable solution for non-stationary dynamics across energy, climate, and healthcare applications.

References

[1] Y. Wang, H. Wu, J. Dong, Y. Liu, C. Wang, M. Long, and J. Wang, “Deep Time Series Models: A Comprehensive Survey and Benchmark,” arXiv:2407.13278, 2024.

[2] J. Kim, Y. Chen, S. Lee, and A. Garg, “Temporal Alignment Networks for Non-Stationary Time Series Forecasting,” in *Proc. ICLR*, 2024.

[3] H. Irani and V. Metsis, “Positional Encoding in Transformer-Based Time Series Models: A Survey,” arXiv:2502.12370, 2025.

[4] R. Gupta, H. Zhao, Y. Yuan, and B. Schölkopf, “Cross-Scale Transformers for Long-Horizon Time Series Forecasting,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.

[5] M. Li, Z. Wang, H. Zhou, and T. Zhao, “Rotary Position Embedding Revisited for Periodic Time Series Modeling,” in *Proc. ICML*, 2024.

[6] K. Sun, Y. Wang, J. Liu, and X. He, “Period-Aware Attention for Robust Sequential Forecasting,” in *Proc. AAAI*, 2023.

[7] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, “A Time Series is Worth 64 Words: Long-term Forecasting with Transformers,” in *Proc. ICLR*, 2023.

[8] H. Wang, T. He, J. Zhang, and M. Long, “Time-LLM: Time Series Forecasting with Pretrained Long-Context Models,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024.

[9] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, “iTransformer: Inverted Transformers Are Effective for Time Series Forecasting,” in *Proc. ICLR*, 2024.

[10] S. Wang, H. Wu, X. Shi, T. Hu, H. Luo, L. Ma, J. Y. Zhang, and J. Zhou, “TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting,” in *Proc. ICLR*, 2024.

[11] S.-A. Chen, C.-L. Li, N. Yoder, S. Ö. Arik, and T. Pfister, “TSMixer: An All-MLP Architecture for Time Series Forecasting,” *arXiv:2303.06053*, 2023.

[12] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, “TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis,” in *Proc. ICLR*, 2023.

[13] A. Zeng, M. Chen, L. Zhang, and Q. Xu, “Are Transformers Effective for Time Series Forecasting?” in *Proc. AAAI*, 2023.

[14] T. Zhang, Y. Zhang, W. Cao, J. Bian, X. Yi, S. Zheng, and J. Li, “Less Is More: Fast Multivariate Time Series Forecasting with Light Sampling-oriented MLP Structures,” *arXiv:2207.01186*, 2022.

[15] D. Wu, G. Cui, X. Huang, Y. Chen, G. Liu, L. Ren, and Y. Li, “An accurate and explainable ensemble learning method for carotid plaque prediction in an asymptomatic population,” *Comput. Methods Programs Biomed.*, vol. 221, p. 106842, 2022.

[16] L. Liu, R. Chen, J. Qu, C. Liu, Y. Li, and D. Wu, “Multi-scale Spatiotemporal Dynamic Graph Neural Network for Early Prediction of Mortality Risks in Heart Failure Patients,” *IEEE J. Biomed. Health Inform.*, 2025.

[17] L. Liu, G. Cui, C. Wan, D. Wu, and Y. Li, “ECG-LLM: Leveraging Large Language Models for Low-Quality ECG Signal Restoration,” in *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM)*, pp. 3537–3542, 2024.

[18] G. Woo, C. Liu, D. Sahoo, A. Kumar, and S. C. H. Hoi, “ETSformer: Exponential Smoothing Transformers for Time Series Forecasting,” *arXiv:2202.01381*, 2022.

[19] Y. Liu, H. Wu, J. Wang, and M. Long, “Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.

[20] L. Fang, T. Wu, Y. Wen, and Q. Yao, “Cross-Patch Temporal Modeling for Multivariate Time Series Forecasting,” in *Proc. KDD*, 2024.

[21] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, “FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting,” in *Proc. ICML*, Proc. Mach. Learn. Res., vol. 162, pp. 27268–27286, 2022.

[22] Y. Wang, H. Wu, J. Dong, Y. Liu, Y. Qiu, H. Zhang, J. Wang, and M. Long, “TimeXer: Empowering Transformers for Time Series Forecasting with Exogenous Variables,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024.

[23] Q. Li, C. Yang, M. Defferrard, and Y. Bengio, “SpectralGNN: Periodic Graph Neural Networks for Time Series Forecasting,” in *Proc. ICLR*, 2023.