PHASEALIGN: STAGE-DEPENDENT PHASE MODELING FOR LONG-TERM TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term time series forecasting relies on recurrent structure, but identifying periodic patterns alone does not indicate where they will reappear in the future. Phase provides this temporal reference, motivating different explicit uses for historical modeling and future generation. Historical modeling relies on relative phase relations among positions in the observed window, whereas future generation preserves the phase reference of the observed window to determine where recurrent structure will appear next. Motivated by this distinction, we propose PhaseAlign, which extracts a shared complex-spectrum representation from the observed window and uses its phase information differently for historical modeling and future generation. For historical modeling, dominant frequencies are used to construct relative harmonic relations that guide attention over the observed sequence. For future generation, the corresponding window-referenced phase states are propagated to future positions along the forecast horizon, providing phase-aware conditioning for future queries. Experiments on seven standard long-term forecasting benchmarks demonstrate strong forecasting performance, while ablation results show consistent advantages of stage-dependent over unified phase modeling. Code is available at this repository:https://github.com/Time358/PhaseAlign-main

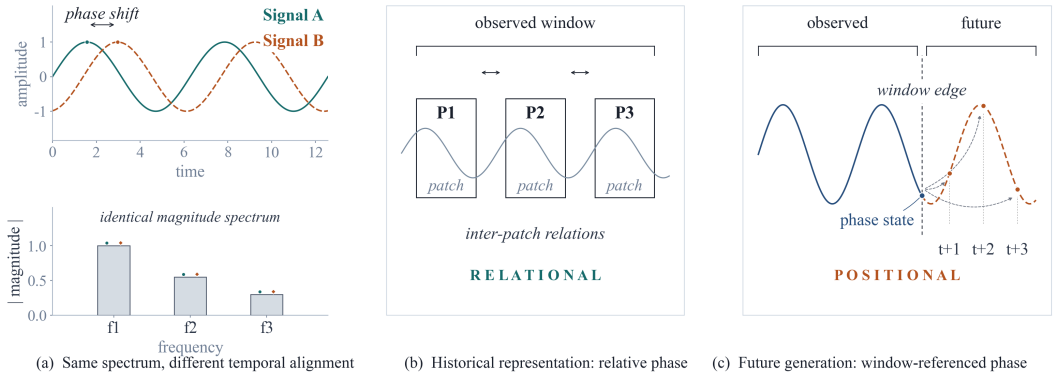
1 INTRODUCTION

Time series forecasting underpins applications in energy systems, traffic management, weather analysis, and industrial monitoring. However, for long-horizon prediction, a model must answer two related but distinct questions: what recurrent components are present in the observed window, and where should they reappear over the forecast horizon? Recent methods have become increasingly effective at representing long histories and extracting recurring structure (Nie et al., 2022; Liu et al., 2024b; Wang et al., 2024a; Zhou et al., 2022; Lin et al., 2024a). These advances primarily address the first question; forecasting also requires determining where that structure should reappear in the future. Knowing that a component is periodic does not determine where its future peaks, troughs, or turning points should occur. Long-term forecasting therefore requires both identifying recurrent structure and placing that structure correctly in the future.

The distinction between identifying recurrent structure and determining its future temporal placement naturally brings phase into focus. Two signals can share nearly identical dominant frequencies and spectral magnitudes while placing their peaks, troughs, and turning points at substantially different times. Frequency and magnitude describe which oscillations are present and how strong they are, whereas phase determines where those oscillations lie on the time axis. Frequency-domain forecasting methods have extensively exploited spectral information for long-horizon prediction (Xu et al., 2024; Piao et al., 2024; Yi et al., 2024; Lin et al., 2024a; Li et al., 2025; Liu et al., 2025), and more recent phase-aware approaches further show that phase itself provides useful forecasting information (Wang et al., 2026; Niu et al., 2026; Zhou et al., 2026; Li et al., 2026). What remains less explored is a more specific question: *should historical modeling and future generation use phase in the same explicit way?*

As illustrated in Figure 1, we treat historical modeling and future generation as two distinct forecasting stages and argue that they require different explicit uses of phase. Historical modeling is primarily relational: the model must characterize how observations at different temporal positions relate

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Same spectrum, different temporal alignment (b) Historical representation: relative phase (c) Future generation: window-referenced phase

Figure 1: Stage-dependent roles of phase in historical modeling and future generation.

within recurring oscillations. For a given harmonic component, the common phase origin cancels when two observed positions are compared, leaving a relation determined only by their temporal separation and frequency. This motivates phase-origin-invariant harmonic relations for modeling historical interactions. In contrast, future generation is positional: each forecast step lies beyond the observed window, so the model must retain the current phase reference and propagate it along the horizon to determine where recurrent structure should appear next. The distinction therefore lies not in what spectral information is available, but in how that information is used. Both stages can rely on the same spectral information while using its phase component differently: relative harmonic relations for historical modeling and window-referenced phase propagation for future generation.

Motivated by this distinction, we propose **PhaseAlign**. PhaseAlign extracts a compact complex-spectrum summary from each input window and shares it across the forecasting pipeline. The key design is how this shared spectral information is used by stage-specific phase operators: dominant frequencies induce relative harmonic relations for historical attention, whereas observed phase states are propagated along the forecast horizon to condition future generation. Time–frequency interaction and a local temporal pathway provide complementary spectral and short-range context without altering this stage-specific design. We evaluate both forecasting accuracy and, more importantly, whether matching the explicit phase operation to the role of each stage matters empirically. Our contributions are summarized as follows:

- We identify stage-dependent phase roles: historical modeling benefits from phase-origin-invariant relative harmonic relations, whereas future generation requires preserving the observed-window phase reference for temporal placement.
- We introduce PhaseAlign, which shares a complex-spectrum representation across both stages, using dominant frequencies to guide historical attention and propagated phase states to condition future queries.
- We validate the stage-dependent design through matched phase controls and targeted alignment analyses, while achieving competitive performance on seven long-term forecasting benchmarks.

2 Related Work

General Time-Series Forecasting. General time-series forecasting has advanced through efficient attention and decomposition (Zhou et al., 2021; Wu et al., 2021), patch-based and variate-centric Transformers (Nie et al., 2022; Liu et al., 2024b), and convolutional or multiscale architectures (Liu et al., 2022; Luo & Wang, 2024; Wang et al., 2024a). Complementary directions study cross-variable interaction, MLP mixing, adaptive routing, channel alignment, sparse or graph-based structure, and non-stationarity (Zeng et al., 2023; Zhang & Yan, 2023; Ekambaram et al., 2023; Chen et al., 2024; Zhou et al., 2024; Lin et al., 2024b; Wang et al., 2025b; Hu et al., 2025; Qiu et al., 2025; Wang et al., 2025c; Liu et al., 2024a). These developments provide increasingly strong temporal backbones for long-horizon prediction.

Frequency- and Periodicity-Aware Forecasting. Frequency-aware forecasting introduces periodic inductive bias through spectral decomposition and transformation, period discovery, frequency filtering, frequency-specific representation learning, and explicit cycle modeling (Zhou et al., 2022; Yi et al., 2023a;b; Wu et al., 2022; Xu et al., 2024; Piao et al., 2024; Yi et al., 2024; Dai et al., 2024; Wang et al., 2024b; Huang et al., 2025; Wang et al., 2025a; Lin et al., 2024a). Recent methods further explore multi-horizon frequency attention, adaptive energy decomposition, super-token frequency modeling, and frequency-aware fusion (Li et al., 2025; Liu et al., 2025; Lan et al., 2025; Chen et al., 2026). Collectively, these approaches improve how recurrent structure is extracted, filtered, and represented.

Phase-Aware Time-Series Forecasting. Recent forecasting methods make phase explicit in different ways. AlignTime constructs phase-aligned subsequences and models dynamic shifts across periods (Wang et al., 2026). PhaseFormer treats phase as a prediction unit through compact embeddings and cross-phase interaction (Niu et al., 2026). PAMNet separates phase positioning from amplitude modulation (Zhou et al., 2026), while PARCNet derives local Hilbert amplitude–phase descriptors for adaptive residual correction (Li et al., 2026). Together, these approaches establish phase as an alignment cue, representation unit, modulation variable, or correction signal.

PhaseAlign studies a complementary design axis: whether the explicit use of phase should depend on the forecasting stage. Specifically, the same spectral information supports phase-origin-invariant relations for historical modeling and window-referenced phase propagation for future generation.

3 METHOD

3.1 PROBLEM FORMULATION AND OVERVIEW

Given a multivariate time series $\mathbf{X} \in \mathbb{R}^{L\times C}$ observed over L time steps, the task is to predict $\mathbf{Y} \in \mathbb{R}^{H\times C}$. PhaseAlign follows the channel-independent paradigm (Nie et al., 2022): the same forecasting function is applied to each variable. We therefore describe a normalized univariate input $\mathbf{x} \in \mathbb{R}^{L}$ and omit batch and channel indices.

For a component with frequency f_k and window-referenced phase ϕ_k, define $\theta_k(t) = \phi_k + 2\pi f_k t$. The phase information required by the two forecasting stages follows directly:

$$\Delta\theta_{k,ij} = \theta_k(\tau_i) - \theta_k(\tau_j) = 2\pi f_k(\tau_i - \tau_j), \qquad \theta_{k,h} = \phi_k + 2\pi f_k(L+h), \quad h = 0, \dots, H-1. \tag{1}$$

The two expressions motivate an operator-level distinction in how phase enters historical modeling and future generation. Under a common phase shift $\phi_k \mapsto \phi_k + \delta$, the pairwise historical relation remains unchanged, $\Delta\theta'_{k,ij} = \Delta\theta_{k,ij}$, whereas the propagated future phase state changes by the same offset, $\theta'_{k,h} = \theta_{k,h} + \delta$. This motivates the central design of PhaseAlign: historical modeling uses phase-origin-invariant harmonic relations, while future generation explicitly retains and propagates the phase reference observed in the input window. Both stages may still access the full complex-spectrum context, including frequency, magnitude, and phase information. Accordingly, this phase-shift distinction is imposed on the dedicated phase operators rather than as a global invariance or equivariance constraint on the entire encoder–decoder network.

Figure 2 summarizes the architecture. PhaseAlign first builds a compact spectrum from the current window. It then uses the dominant frequencies to shape historical token interactions and propagates the associated phase states into horizon-wise future queries. A local evolution stream is fused with the spectral pathway throughout the encoder.

3.2 SHARED COMPLEX-SPECTRUM CONTEXT

We subtract the temporal mean and compute the one-sided real discrete Fourier transform,

$$X_m = \sum_{t=0}^{L-1} \widetilde{x}_t \exp\left(-j\frac{2\pi m t}{L}\right) = A_m e^{j\phi_m}, \qquad m = 0, \dots, \left\lfloor \frac{L}{2} \right\rfloor, \tag{2}$$

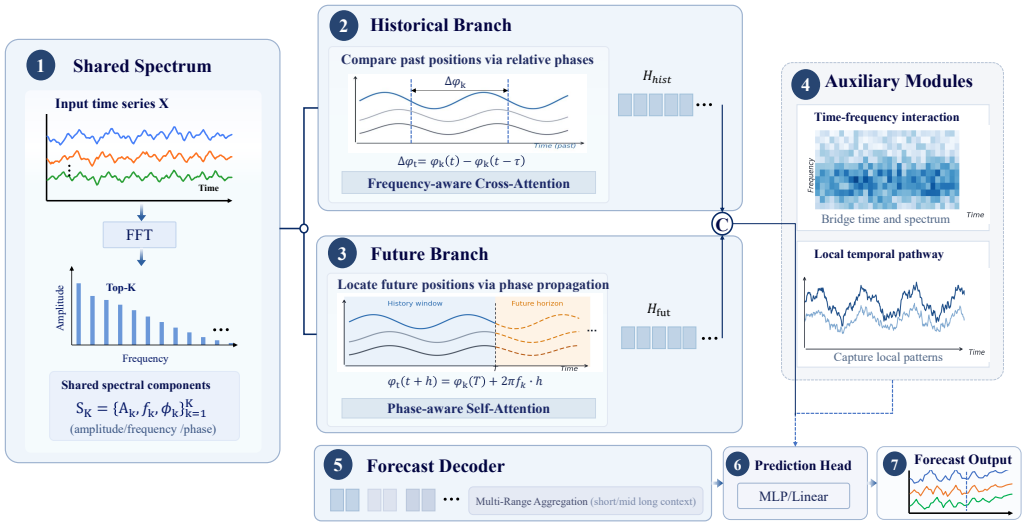


Figure 2: Overview of the PhaseAlign architecture.

where $\widetilde{x}_t = x_t - \frac{1}{L}\sum_{u=0}^{L-1} x_u$. After removing the DC term, we retain the K components with the largest magnitudes. For the k-th selected component,

$$\mathbf{s}_k = \left[\bar{A}_k,\ f_k,\ \sin\phi_k,\ \cos\phi_k\right], \qquad f_k = \frac{m_k}{L}, \qquad \bar{A}_k = \frac{A_{m_k}}{\sum_{r=1}^{K} A_{m_r}}. \tag{3}$$

The circular phase coordinates avoid discontinuity at $\pm\pi$. We also form a window-level summary $\mathbf{g}$ by concatenating the retained component statistics with the total selected magnitude and the mean and standard deviation of the selected frequencies.

Each component descriptor modulates a learnable component basis,

$$\mathbf{F}_k = \mathbf{b}_k + \tanh(\eta)\,\mathcal{P}_{\text{peak}}(\mathbf{s}_k), \qquad \mathbf{z}_{\text{spec}} = \mathcal{P}_{\text{global}}(\mathbf{g}) + \mathcal{P}_{\text{token}}\left(\sum_{k=1}^{K} \bar{A}_k \mathbf{F}_k\right), \tag{4}$$

where $\mathbf{b}_k$ is a learnable embedding for the k-th magnitude-ranked component slot. Magnitude rank provides only a deterministic slot ordering across samples: $\mathbf{b}_k$ is not assigned a fixed physical frequency, and the actual frequency always enters through $\mathbf{s}_k$. The input is partitioned into overlapping patches, and $\mathbf{z}_{\text{spec}}$ conditions every patch on the current window spectrum.

PhaseAlign then performs bidirectional time–frequency interaction,

$$\mathbf{P}' = \text{LN}(\mathbf{P} + \text{MHA}(\mathbf{P}, \mathbf{F}, \mathbf{F})), \qquad \mathbf{F}' = \text{LN}(\mathbf{F} + \text{MHA}(\mathbf{F}, \mathbf{P}', \mathbf{P}')), \tag{5}$$

where $\mathbf{P} \in \mathbb{R}^{N\times D}$ and $\mathbf{F} \in \mathbb{R}^{K\times D}$. Patch-to-frequency attention injects sample-specific spectral information into temporal tokens, while frequency-to-patch attention refines the component tokens using observed local patterns. This interaction provides shared spectral context to the temporal stream; it is a supporting component and does not alter the two stage-specific phase operators defined next.

3.3 Relative-Harmonic Historical Encoder

Let $\tau_i = iS$ be the anchor of patch i, where S is the patch stride, and let R denote the number of attention heads. The selected components are ordered by magnitude. The first $\min(K, R)$ heads use observed frequencies; if $K < R$, the remaining heads use deterministic, linearly spaced harmonic fillers defined in Appendix C. For head r, the relative harmonic relation is

$$B_{ij}^{(r)} = \cos\left(2\pi \hat{f}_r(\tau_i - \tau_j)\right). \tag{6}$$

Eq. (1) shows why ϕ_k does not appear in $B_{ij}^{(r)}$: a common phase origin cancels when two historical positions are compared. The resulting harmonic relation is therefore invariant to a global phase offset and measures cycle-position compatibility under a sample-specific dominant frequency. This invariance applies to the dedicated relation operator; the surrounding encoder remains free to use the shared, phase-bearing spectral context through its content and time–frequency pathways. Because the cosine kernel is even, the harmonic term intentionally represents cycle-position compatibility rather than temporal direction. Sequential order and directional information remain in the gated rotary transformation and the content-dependent attention term, separating harmonic similarity from temporal orientation.

Each encoder block combines a gated rotary transformation with the harmonic relation,

$$\mathbf{S}^{(r)} = \frac{\widetilde{\mathbf{Q}}^{(r)}\widetilde{\mathbf{K}}^{(r)\top}}{\sqrt{d_h}} + \tanh(\gamma_r)\mathbf{B}^{(r)}, \tag{7}$$

where $(\widetilde{\mathbf{Q}}^{(r)}, \widetilde{\mathbf{K}}^{(r)})$ are obtained from a learnable gated rotary positional transform (Su et al., 2024). The content term selects semantically compatible patches, and the harmonic term reshapes their interaction according to the current window's recurring structure.

A local evolution stream supplies complementary short-range context. We patch-embed first-order differences, process them with depthwise convolutions of kernel sizes 3 and 5 followed by pointwise mixing, and use a learned gate to combine the local and attention outputs. The resulting memory $\mathbf{M} \in \mathbb{R}^{N\times D}$ therefore retains local dynamics alongside the relative harmonic relations that define the historical phase operator.

3.4 Window-Referenced Phase Decoder

Future generation retains the phase reference observed in the input window. For each selected component and future step h, PhaseAlign evaluates $\theta_{k,h}$ from Eq. (1) and constructs amplitude-weighted circular coordinates,

$$\mathbf{u}_h = \mathrm{Concat}_{k=1}^{K}\left[\bar{A}_k \sin\theta_{k,h},\ \bar{A}_k \cos\theta_{k,h}\right], \qquad \boldsymbol{\psi}_h = \tanh(g_f)\,\mathcal{P}_{\mathrm{phase}}(\mathbf{u}_h). \tag{8}$$

Under a common phase shift $\phi_k \mapsto \phi_k + \delta$, each propagated angle $\theta_{k,h}$ changes by the same offset. Thus, the circular coordinates $\mathbf{u}_h$ explicitly carry the window-referenced phase state into every forecast step, and $\boldsymbol{\psi}_h$ provides learned conditioning derived from that state. This construction preserves an explicit phase reference in the decoder input without imposing an exact equivariance constraint on the full decoder.

The decoder uses one effective learned horizon-specific query term. In the implementation, a learnable query and a horizon embedding are parameterized separately but enter the decoder only through their sum,

$$\bar{\mathbf{q}}_h^{\mathrm{learn}} \equiv \mathbf{q}_h^{\mathrm{learn}} + \mathbf{e}_h^{\mathrm{hor}}, \qquad \mathbf{q}_h = \bar{\mathbf{q}}_h^{\mathrm{learn}} + \mathbf{e}_h^{\mathrm{fix}} + \boldsymbol{\psi}_h. \tag{9}$$

Only the sum enters the decoder, so we assign no separate functional interpretation to the two stored parameter tables and treat $\bar{\mathbf{q}}_h^{\mathrm{learn}}$ as one learned horizon-specific vector. The remaining query terms provide generic horizon position, whereas the sample-specific temporal reference enters through $\boldsymbol{\psi}_h$. The fixed basis is defined in Appendix C; known future covariates are added when available.

The decoder attends to the historical memory once and aggregates the resulting attention map over three supports: the most recent quarter, the most recent half, and the full history. Let $\mathcal{I}_s$, $\mathcal{I}_m$, and $\mathcal{I}_l$ denote these supports. For each horizon step,

$$\mathbf{C}_h^u = \mathbf{A}_{h,\mathcal{I}_u}\mathbf{V}_{\mathcal{I}_u}, \quad u \in \{s,m,l\}, \qquad \mathbf{d}_h = \alpha_s\mathbf{C}_h^s + \alpha_m\mathbf{C}_h^m + \alpha_l\mathbf{C}_h^l, \tag{10}$$

with

$$[\alpha_s, \alpha_m, \alpha_l] = \mathrm{Softmax}([g_s, g_m, g_l]). \tag{11}$$

Each horizon-specific query induces its own attention row $\mathbf{A}_h$, while the mixture coefficients are shared across horizons. The stage-specific future phase information enters through $\boldsymbol{\psi}_h$; multi-range aggregation only determines how each query retrieves historical context. Details are in Appendix D. The decoder state is then projected back to the forecasting space.

3.5 Learning Objective

The stage-specific architecture is trained with point forecasting together with two auxiliary regularizers: amplitude-weighted spectral phase consistency and cross-stage representation alignment,

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_p \mathcal{L}_{\text{phase}} + \lambda_a \mathcal{L}_{\text{align}}. \tag{12}$$

The point objective is mean squared error,

$$\mathcal{L}_{\text{pred}} = \frac{1}{HC} \|\widehat{\mathbf{Y}} - \mathbf{Y}\|_F^2. \tag{13}$$

For the phase term, let Ω denote the non-DC positive-frequency bins and define, for channel c,

$$w_{\omega,c} = \frac{|Y_{\omega,c}|}{\sum_{\nu\in\Omega} |Y_{\nu,c}| + \epsilon}, \qquad \mathcal{L}_{\text{phase}} = \frac{1}{C} \sum_{c=1}^{C} \sum_{\omega\in\Omega} w_{\omega,c} \Big[1 - \cos\Big(\angle Y_{\omega,c} - \angle \widehat{Y}_{\omega,c}\Big)\Big]. \tag{14}$$

Amplitude weighting reduces the influence of low-energy bins whose phase estimates are unstable. For representation alignment, the historical side pools the main encoder memory independently for each variable, while the target future is mapped by a lightweight temporal stem using the same channel-independent treatment. The per-variable representations are averaged only after encoding to obtain $\mathbf{r}_{\text{hist}}$ and $\mathbf{r}_{\text{future}}$, preserving the channel-independent backbone while forming a sample-level alignment view. We optimize their cosine distance,

$$\mathcal{L}_{\text{align}} = 1 - \frac{\mathbf{r}_{\text{hist}}^\top \mathbf{r}_{\text{future}}}{\|\mathbf{r}_{\text{hist}}\|_2 \|\mathbf{r}_{\text{future}}\|_2 + \epsilon}. \tag{15}$$

Both additional objectives are training-time regularizers rather than definitions of the stage-specific phase mechanism; the target-side alignment stem is discarded at inference. All stage-specific controls in Section 4.3 use the same objective, so their comparison isolates the phase form. Appendix G.5 attributes the auxiliary gains and checks representation dispersion, while Appendix G studies weight sensitivity.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate PhaseAlign on seven standard long-term forecasting benchmarks: ETTh1, ETTh2, ETTm1, ETTm2, Exchange, Weather, and Electricity. These datasets span different sampling frequencies, dimensionalities, and temporal structures. We use the standard chronological train/validation/test splits adopted by prior work (Zeng et al., 2023; Nie et al., 2022; Liu et al., 2024b; Wang et al., 2024a) and evaluate horizons $H \in \{96, 192, 336, 720\}$.

Baselines. We compare with PhaseFormer (Niu et al., 2026), PMDFormer (Hu et al., 2026), AlignTime (Wang et al., 2026), TQNet (Lin et al., 2025), RAFT (Han et al., 2025), PatchMLP (Tang & Zhang, 2025), iTransformer (Liu et al., 2024b), and TimesNet (Wu et al., 2022). The comparison covers phase-aware forecasting, temporal alignment, patch-based modeling, retrieval-enhanced forecasting, and Transformer-style architectures. We use official implementations and recommended configurations when available.

Implementation details. PhaseAlign uses one shared channel-independent model across variables. Validation MAE selects checkpoints. PhaseAlign and all internal controls are trained independently with seeds $\{2025, 2026, 2027\}$; the main tables report three-seed means and Appendix G.1 reports uncertainty. We evaluate MSE and MAE. Experiments proceed from forecasting accuracy, to direct stage-specific controls, to targeted mechanism and robustness tests; full settings and horizon-wise results are in the appendix.

4.2 Overall Forecasting Accuracy

Table 1 shows that PhaseAlign achieves the best or tied-best average MAE on all seven datasets and the best average MSE on five. On ETTm1 and ETTm2, it reduces average MSE by about 9.0% relative to the strongest competitor. Horizon-wise results in Appendix A rank PhaseAlign first on both metrics at every horizon on ETTh2, ETTm1, ETTm2, and Electricity, while ETTh1 and Exchange remain closely contested.

Table 1: Average results on seven benchmarks over horizons $\{96, 192, 336, 720\}$. PhaseAlign entries are three-seed means; baselines follow official/reproduced protocols. Best/second-best results are bold/underlined; lower is better.

Dataset	Ours		PhaseFormer (2026)		PMDFormer (2026)		AlignTime (2026)		TQNet (2025)		RAFT (2025)		PatchMLP (2025)		iTransformer (2024)		TimesNet (2023)	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.421	**0.421**	**0.419**	0.422	0.446	0.440	0.423	0.426	0.440	0.433	0.431	0.438	0.463	0.446	0.449	0.441	0.458	0.450
ETTh2	**0.365**	**0.388**	0.387	0.413	0.379	0.400	0.378	0.403	0.385	0.413	0.385	0.413	0.397	0.414	0.386	0.407	0.414	0.427
ETTm1	**0.343**	**0.373**	0.385	0.398	0.402	0.393	0.396	0.399	0.377	0.393	0.381	0.401	0.392	0.400	0.408	0.411	0.400	0.405
ETTm2	**0.252**	**0.311**	0.291	0.336	0.282	0.327	0.279	0.326	0.277	0.322	0.281	0.330	0.283	0.324	0.292	0.334	0.291	0.332
Exchange	0.343	**0.391**	0.370	0.411	0.357	0.402	**0.342**	**0.391**	0.355	0.404	0.403	0.415	0.372	0.411	0.366	0.413	0.416	0.443
Weather	**0.238**	**0.263**	0.257	0.280	0.255	0.269	0.254	0.279	0.242	0.268	0.273	0.308	0.248	0.275	0.261	0.281	0.259	0.286
Electricity	**0.145**	**0.237**	0.160	0.250	0.147	0.240	0.195	0.280	0.164	0.258	0.160	0.258	0.192	0.288	0.198	0.286	0.192	0.303

4.3 Testing the Stage-Specific Phase Hypothesis

We first test the central hypothesis of whether the two stages should use different explicit phase forms. *Unified-Rel* applies a relative-only convention across both stages, whereas *Unified-Abs* applies a window-referenced absolute convention throughout. Both controls preserve the same spectrum, objective, capacity, and positional backbone, changing only the phase form exposed to the two stages. We also report operator-removal and time–frequency interaction ablations, with exact definitions in Appendix A.2.

Table 2 shows that the stage-matched design consistently outperforms both unified alternatives, reducing mean MSE by 3.5% and 4.3% over Unified-Rel and Unified-Abs, respectively. Removing either stage-specific phase pathway also degrades performance. These results indicate that the gain comes not merely from supplying phase information, but from matching its explicit form to historical comparison and future placement. All controls share the same auxiliary objective; its separate contribution and run-to-run uncertainty are reported in Appendices G.5 and G.1.

Table 2: Three-seed mean ablation results on ETTh1, ETTh2, and Exchange. Standard deviations are in Appendix G.1. Best results are bold; lower is better.

Dataset	Ours		w/o RelPhase		w/o FutPhase		w/o BothPhase		w/o TF-Interaction		Unified-Rel		Unified-Abs	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	**0.421**	**0.421**	0.438	0.430	0.441	0.432	0.450	0.437	0.445	0.434	0.431	0.425	0.434	0.427
ETTh2	**0.365**	**0.388**	0.389	0.403	0.392	0.405	0.401	0.412	0.397	0.408	0.379	0.396	0.381	0.398
Exchange	**0.343**	**0.391**	0.374	0.410	0.379	0.412	0.394	0.420	0.387	0.416	0.360	0.401	0.365	0.404

4.4 Targeted Tests of the Phase Mechanism

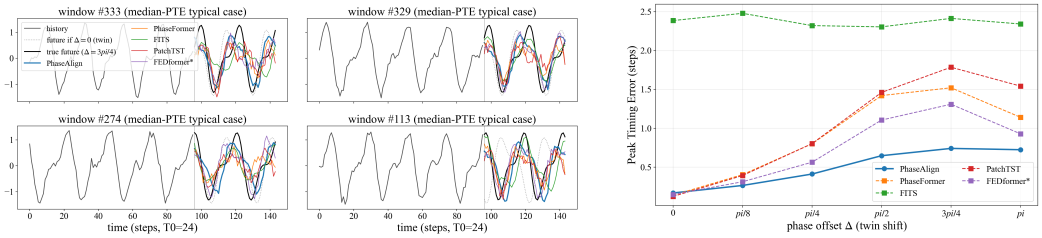
We next ask two focused questions: does temporal placement remain stable when the phase relationship shifts beyond training, and does phase corruption differ from magnitude corruption under matched history-domain distortion? Spectral twins and matched phase–magnitude interventions address these questions directly. Additional drift and DFT/window-misalignment tests are grouped in Section 4.5; full protocols are in Appendix F.

4.4.1 Spectral Twins

We construct a two-component harmonic process with fixed component amplitudes $[1.0, 0.5]$ and continuously sampled base frequency, while the relative phase Δ of the second harmonic is varied over $\{0, \pi/8, \pi/4, \pi/2, 3\pi/4, \pi\}$. Models are trained only at the reference relationship $\Delta = 0$ and evaluated without adaptation under the unseen offsets. This Δ-OOD protocol preserves the generative frequency/amplitude structure while changing the relative-phase relationship that determines future temporal placement. We use Peak Timing Error (PTE), defined in Appendix F.1, as the displacement of the dominant future oscillatory extremum.

Figure 3 shows that several models are competitive near the training condition, but their behavior separates as the unseen phase shift grows. At $\Delta = \pi$, PhaseAlign reaches a PTE of 0.7150 ± 0.0500,

compared with 0.8068 ± 0.0515 for Unified-Abs, 1.2157 ± 0.1982 for Unified-Rel, and 1.4832 ± 0.1676 for PhaseFormer. Its endpoint degradation from $\Delta = 0$ to π is also the smallest among the compared models. The advantage therefore emerges primarily when the relative-phase relationship shifts away from training. Together with the unified controls, this result directly supports the central claim that the relational phase operator for historical modeling and the reference-preserving phase operation for future generation should not be collapsed into one phase convention.



(a) Representative forecasts at $\Delta = 3\pi/4$.

(b) PTE under unseen relative-phase offsets.

Figure 3: Relative-phase OOD robustness. Models are trained only at $\Delta = 0$ and evaluated without adaptation as the second-harmonic phase relationship shifts.

4.4.2 Matched Phase–Magnitude Intervention

We next examine phase-specific sensitivity on real ETTh1/OT data by perturbing only the observed history while keeping the future target fixed. Phase and magnitude interventions use the same frequency bins and random signs, and are matched window by window by the history-domain relative ℓ_2 distortion $\rho(\mathbf{x}, \widetilde{\mathbf{x}})$. All intervention parameters are estimated solely from the observed history.

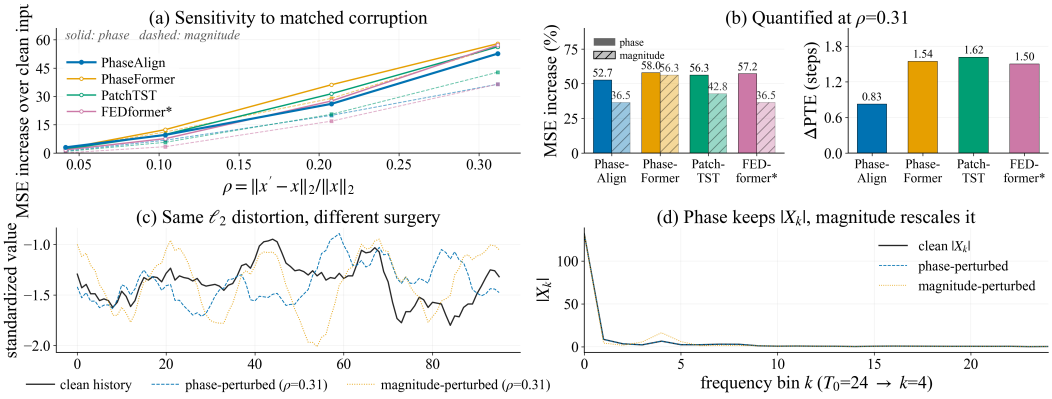


Figure 4: Matched phase–magnitude perturbations on ETTh1/OT at equal relative-ℓ_2 strength.

Figure 4 shows that phase perturbations become more damaging than matched magnitude perturbations as ρ increases. At $\rho \approx 0.31$, PhaseAlign's MSE increases by 52.7% under phase perturbation versus 36.5% under magnitude perturbation. It also shows the smallest peak-timing degradation ($\Delta\text{PTE} = 0.83$) among the plotted models. Thus, equal history-domain distortion can have different forecasting consequences depending on whether phase or magnitude is altered. This complements the stage-specific controls by showing that phase is not merely another spectral scalar: corruption of temporal alignment has a distinct effect, while PhaseAlign better preserves forecast timing under strong phase perturbation.

4.5 Additional Robustness Stress Tests

These secondary stress tests examine whether the phase mechanism remains useful when the spectral estimate itself becomes imperfect.

4.5.1 Progressive Frequency Drift

We evaluate stationary-trained models under progressive frequency drift, where small frequency deviations accumulate into increasingly large future phase mismatch. Figure 5 shows that the models remain comparable under weak drift, while PhaseAlign becomes increasingly robust as the accumulated mismatch grows. From π onward, it achieves positive gains at every tested horizon; at the strongest 4π condition, these gains range from 4.16% to 26.13%. This pattern suggests that the benefit of the phase-aware design becomes more apparent when small frequency errors accumulate into substantial future temporal displacement. Absolute-error and mismatch curves in Appendix F.3 further distinguish robustness from exact tracking of the changing future frequency.

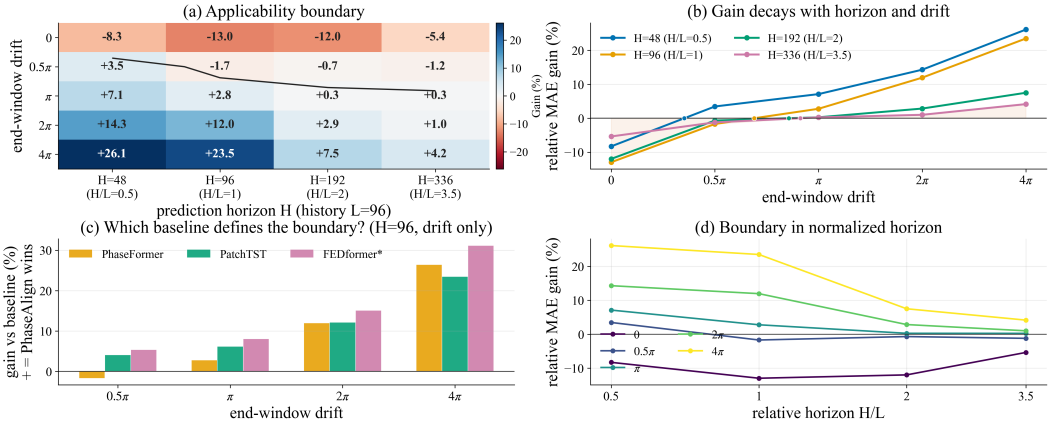


Figure 5: Progressive frequency-drift stress test as frequency deviations accumulate into larger future phase mismatch.

4.5.2 Off-Grid Frequency and Window Shift

Appendix F.4 further examines robustness to DFT-bin and observation-window misalignment. Performance remains stable for $\delta \leq 0.375$, while differences become more pronounced at the hardest half-bin offset ($\delta = 0.5$), where PhaseAlign achieves an MSE of 0.0711 compared with 0.1067 for the next-best model. Across 24 window shifts, it also yields the lowest mean error and the smallest variation. Together with the frequency-drift results, these tests indicate that the proposed phase mechanism remains effective when the dominant-frequency estimate is imperfect.

5 Conclusion

PhaseAlign is built around one simple distinction: historical modeling is relational, whereas future generation is positional. A shared complex spectrum therefore serves two explicit phase roles—phase-origin-invariant harmonic relations organize historical interactions, while propagated phase states retain the observed window reference for future placement. Across seven forecasting benchmarks, PhaseAlign shows strong and consistent performance. More importantly, matched relative-only and absolute-only controls and targeted phase interventions support the central claim that the phase operation should be matched to the role of each forecasting stage. Additional drift and off-grid tests further indicate that this design remains useful when the spectral estimate is imperfect.

Ethics Statement

This work focuses on general-purpose time-series forecasting using benchmark datasets and synthetic signals. It does not involve human subjects or personally identifiable information. No specific high-stakes deployment is studied in this work.

Reproducibility Statement

Method is in Section 3. Results and experimental details are in Appendix A and Appendices B–G.

AI Use Statement

Generative AI tools assisted with LaTeX, retrieval and discovery. All AI-assisted material was reviewed by the authors, who take full responsibility for the final scientific content and artifacts.

References

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 10527–10545, 2024.

Shijun Chen, Hua Wang, and Fan Zhang. Firformer: Transformer with frequency-aware fusion and recalibrated attention for time series forecasting. *Expert Systems*, 43(9):e70378, 2026.

Tao Dai, Beiliang Wu, Peiyuan Liu, Naiqi Li, Jigang Bao, Yong Jiang, and Shu-Tao Xia. Periodicity decoupling framework for long-term series forecasting. In *International conference on learning representations*, volume 2024, pp. 48067–48078, 2024.

Vijay Ekambaram, Arindam Jati, Nam Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. Tsmixer: Lightweight mlp-mixer model for multivariate time series forecasting. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 459–469, 2023.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. *arXiv preprint arXiv:2505.04163*, 2025.

Ao Hu, Liangjian Wen, Jiang Duan, Yong Dai, He Yan, Dongkai Wang, Jun Wang, Yukun Zhang, Ruoxi Jiang, and Zenglin Xu. Pmdformer: Patch-mean decoupling information transformer for long-term forecasting. *arXiv preprint arXiv:2606.26549*, 2026.

Yifan Hu, Guibin Zhang, Peiyuan Liu, Disen Lan, Naiqi Li, Dawei Cheng, Tao Dai, Shu-Tao Xia, and Shirui Pan. Timefilter: Patch-specific spatial-temporal graph filtration for time series forecasting. *arXiv preprint arXiv:2501.13041*, 2025.

Songtao Huang, Zhen Zhao, Can Li, and Lei Bai. Timekan: Kan-based frequency decomposition learning architecture for long-term time series forecasting. In *International Conference on Learning Representations*, volume 2025, pp. 28514–28529, 2025.

Chengdong Lan, Yifeng Xu, and Yaofeng Zeng. Sftformer: A super token and frequency-aware transformer for long-term multivariate forecasting. *Digital Signal Processing*, pp. 105726, 2025.

Ruiqi Li, Maowei Jiang, Quangao Liu, Kai Wang, Kaiduo Feng, Yue Sun, and Xiufang Zhou. Faith: Frequency-domain attention in two horizons for time series forecasting. *Knowledge-Based Systems*, 309:112790, 2025.

Ziqiong Li, Heyu Chai, Zhangyao Song, Shengjun Liu, and Xinru Liu. Parcnet: Phase-aware residual correction network for efficient multivariate time series forecasting. *Knowledge-Based Systems*, 348:116370, 2026.

Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. Cyclenet: Enhancing time series forecasting through modeling periodic patterns. *Advances in Neural Information Processing Systems*, 37:106315–106345, 2024a.

Shengsheng Lin, Weiwei Lin, Wentai Wu, Haojun Chen, and Junjie Yang. Sparsetsf: Modeling long-term time series forecasting with 1k parameters. *arXiv preprint arXiv:2405.00946*, 2024b.

Shengsheng Lin, Haojun Chen, Haijie Wu, Chunyun Qiu, and Weiwei Lin. Temporal query network for efficient multivariate time series forecasting. *arXiv preprint arXiv:2505.12917*, 2025.

Minhao Liu, Ailing Zeng, Muxi Chen, Zhijian Xu, Qiuxia Lai, Lingna Ma, and Qiang Xu. Scinet: Time series modeling and forecasting with sample convolution and interaction. *Advances in neural information processing systems*, 35:5816–5828, 2022.

Peiyuan Liu, Beiliang Wu, Yifan Hu, Naiqi Li, Tao Dai, Jigang Bao, and Shu-tao Xia. Timebridge: Non-stationarity matters for long-term time series forecasting. *arXiv preprint arXiv:2410.04442*, 2024a.

Xinru Liu, Heyu Chai, and Shengjun Liu. Faformer: A frequency-aware transformer with adaptive energy decomposition for multivariate time series forecasting. *Knowledge-Based Systems*, pp. 114952, 2025.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024b.

Donghao Luo and Xue Wang. Moderntcn: A modern pure convolution structure for general time series analysis. In *International conference on learning representations*, volume 2024, pp. 31728–31770, 2024.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Yiming Niu, Jinliang Deng, and Yongxin Tong. Phaseformer: From patches to phases for efficient and effective time series forecasting. In *International Conference on Learning Representations*, volume 2026, pp. 8262–8285, 2026.

Xihao Piao, Zheng Chen, Taichi Murayama, Yasuko Matsubara, and Yasushi Sakurai. Fredformer: Frequency debiased transformer for time series forecasting. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 2400–2410, 2024.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. Duet: Dual clustering enhanced multivariate time series forecasting. In *Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining v. 1*, pp. 1185–1196, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Peiwang Tang and Weitai Zhang. Unlocking the power of patch: Patch-based mlp for long-term time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 12640–12648, 2025.

Hao Wang, Lichen Pan, Yuan Shen, Zhichao Chen, Degui Yang, Yifei Yang, Sen Zhang, Xinggao Liu, Haoxuan Li, and Dacheng Tao. Fredf: Learning to forecast in the frequency domain. In *International Conference on Learning Representations*, volume 2025, pp. 6893–6922, 2025a.

Min Wang, Hua Wang, and Fan Zhang. Aligntime: Interperiodic phase alignment sampling for time-series forecasting. *Information Processing & Management*, 63(1):104296, 2026.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 38626–38652, 2024a.

Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. Timemixer++: A general time series pattern machine for universal predictive analysis. In *International Conference on Learning Representations*, volume 2025, pp. 16980–17016, 2025b.

Yihang Wang, Yuying Qiu, Peng Chen, Kai Zhao, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Rose: Register-assisted general time series forecasting with decomposed frequency learning. 2024b.

Yihang Wang, Yuying Qiu, Peng Chen, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Lightgts: A lightweight general time series forecasting model. *arXiv preprint arXiv:2506.06005*, 2025c.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

Zhijian Xu, Ailing Zeng, and Qiang Xu. Fits: Modeling time series with $10k$ parameters. In *International Conference on Learning Representations*, volume 2024, pp. 26295–26318, 2024.

Kun Yi, Qi Zhang, Wei Fan, Hui He, Liang Hu, Pengyang Wang, Ning An, Longbing Cao, and Zhendong Niu. Fouriergnn: Rethinking multivariate time series forecasting from a pure graph perspective. *Advances in neural information processing systems*, 36:69638–69660, 2023a.

Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. Frequency-domain mlps are more effective learners in time series forecasting. *Advances in neural information processing systems*, 36:76656–76679, 2023b.

Kun Yi, Jingru Fei, Qi Zhang, Hui He, Shufeng Hao, Defu Lian, and Wei Fan. Filternet: Harnessing frequency filters for time series forecasting. *Advances in neural information processing systems*, 37:55115–55140, 2024.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2023.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pp. 27268–27286. PMLR, 2022.

Tian Zhou, Qingsong Wen, Jinyang Gao, Bolin Ding, Rong Jin, et al. Card: Channel aligned robust blend transformer for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11245–11283, 2024.

Yingbo Zhou, Yutong Ye, Zhiwei Ling, Shuhao Li, Rui Qian, Jian Xiong, Li Sun, and Dejing Dou. Pamnet: Cycle-aware phase-amplitude modulation network for multivariate time series forecasting. *arXiv preprint arXiv:2605.02938*, 2026.

A Additional Results

A.1 Horizon-Wise Forecasting Results

Table 3 expands the averaged results in the main paper by reporting MSE and MAE separately for each forecast horizon $H \in \{96, 192, 336, 720\}$. The horizon-wise breakdown makes explicit where the average gains are consistent across horizons and where competing methods remain close.

Table 3: Horizon-wise forecasting results on seven benchmarks. PhaseAlign reports three-seed means; best/second-best are bold/underlined.

Dataset	Horizon	Ours		PhaseFormer (2026)		PMDFormer (2026)		AlignTime (2026)		TQNet (2025)		RAFT (2025)		PatchMLP (2025)		iTransformer (2024)		TimesNet (2023)	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.358**	**0.380**	0.379	0.396	0.391	0.403	0.366	0.389	0.371	0.393	0.386	0.414	0.392	0.403	0.382	0.400	0.384	0.402
	192	**0.411**	**0.412**	0.419	0.419	0.442	0.432	0.412	0.420	0.428	0.426	0.422	0.431	0.444	0.432	0.443	0.433	0.436	0.429
	336	0.454	**0.434**	**0.449**	0.436	0.475	0.451	**0.449**	0.439	0.476	0.446	0.456	0.445	0.491	0.456	0.480	0.452	0.491	0.469
	720	0.462	0.459	**0.429**	**0.438**	0.478	0.474	0.465	0.458	0.487	0.470	0.463	0.464	0.528	0.496	0.493	0.481	0.521	0.500
	Avg	0.421	**0.421**	**0.419**	0.422	0.446	0.440	0.423	0.426	0.440	0.433	0.431	0.438	0.463	0.446	0.449	0.441	0.458	0.450
ETTh2	96	**0.285**	**0.330**	0.302	0.351	0.296	0.343	0.294	0.345	0.295	0.343	0.296	0.349	0.305	0.353	0.305	0.352	0.340	0.374
	192	**0.362**	**0.382**	0.374	0.402	0.376	0.392	0.375	0.395	0.367	0.393	0.384	0.402	0.402	0.410	0.378	0.398	0.402	0.414
	336	**0.408**	**0.414**	0.427	0.437	0.420	0.426	0.416	0.429	0.417	0.427	0.426	0.442	0.445	0.445	0.420	0.427	0.452	0.452
	720	**0.408**	**0.429**	0.447	0.462	0.425	0.439	0.430	0.445	0.433	0.446	0.434	0.459	0.437	0.449	0.441	0.451	0.462	0.468
	Avg	**0.365**	**0.388**	0.387	0.413	0.379	0.400	0.378	0.403	0.385	0.413	0.385	0.413	0.397	0.414	0.386	0.407	0.414	0.427
ETTm1	96	**0.301**	**0.338**	0.329	0.363	0.328	0.352	0.338	0.369	0.311	0.353	0.329	0.371	0.323	0.362	0.330	0.365	0.338	0.375
	192	**0.314**	**0.356**	0.366	0.385	0.377	0.376	0.374	0.383	0.356	0.378	0.363	0.388	0.369	0.385	0.388	0.399	0.374	0.387
	336	**0.348**	**0.381**	0.396	0.407	0.422	0.406	0.406	0.406	0.390	0.401	0.391	0.407	0.403	0.407	0.417	0.416	0.410	0.411
	720	**0.409**	**0.417**	0.451	0.440	0.483	0.440	0.467	0.438	0.452	0.440	0.444	0.438	0.474	0.447	0.499	0.465	0.478	0.450
	Avg	**0.343**	**0.373**	0.385	0.398	0.402	0.393	0.396	0.399	0.377	0.393	0.381	0.401	0.392	0.400	0.408	0.411	0.400	0.405
ETTm2	96	**0.161**	**0.249**	0.183	0.271	0.184	0.261	0.176	0.261	0.173	0.256	0.177	0.265	0.180	0.263	0.182	0.263	0.187	0.267
	192	**0.226**	**0.291**	0.252	0.311	0.243	0.297	0.242	0.304	0.238	0.298	0.243	0.307	0.245	0.306	0.257	0.315	0.249	0.309
	336	**0.277**	**0.329**	0.319	0.355	0.302	0.346	0.302	0.342	0.301	0.340	0.302	0.346	0.304	0.336	0.314	0.350	0.321	0.351
	720	**0.347**	**0.378**	0.410	0.408	0.402	0.404	0.398	0.400	0.397	0.396	0.402	0.404	0.406	0.394	0.418	0.409	0.408	0.403
	Avg	**0.252**	**0.311**	0.291	0.336	0.282	0.327	0.279	0.326	0.277	0.322	0.281	0.330	0.283	0.324	0.292	0.334	0.291	0.332
Exchange	96	**0.078**	**0.195**	0.089	0.207	0.084	0.200	0.082	0.198	0.084	0.203	0.084	0.200	0.094	0.215	0.105	0.228	0.107	0.234
	192	**0.167**	**0.290**	0.191	0.314	0.185	0.307	**0.167**	**0.290**	0.177	0.297	0.178	0.297	0.183	0.306	0.181	0.304	0.226	0.344
	336	0.313	0.404	0.335	0.422	0.336	0.419	**0.308**	**0.400**	0.329	0.418	0.350	0.422	0.339	0.423	0.361	0.437	0.367	0.448
	720	0.814	**0.677**	0.867	0.704	0.826	0.683	**0.813**	**0.677**	0.831	0.698	1.001	0.743	0.873	0.702	0.817	0.686	0.964	0.746
	Avg	0.343	**0.391**	0.370	0.411	0.357	0.402	**0.342**	**0.391**	0.355	0.404	0.403	0.415	0.372	0.411	0.366	0.413	0.416	0.443
Weather	96	**0.152**	**0.195**	0.177	0.219	0.171	0.203	0.172	0.220	0.157	0.200	0.188	0.246	0.165	0.210	0.175	0.214	0.172	0.220
	192	**0.200**	**0.239**	0.221	0.257	0.220	0.248	0.221	0.259	0.206	0.245	0.241	0.284	0.212	0.251	0.227	0.261	0.219	0.261
	336	0.272	**0.287**	0.278	0.298	0.275	0.288	0.275	0.296	**0.262**	**0.287**	0.296	0.326	0.269	0.294	0.285	0.301	0.280	0.306
	720	**0.330**	**0.333**	0.355	0.347	0.354	0.340	0.348	0.343	0.344	0.342	0.368	0.379	0.349	0.345	0.360	0.350	0.365	0.359
	Avg	**0.238**	**0.263**	0.257	0.280	0.255	0.269	0.254	0.279	0.242	0.268	0.273	0.308	0.248	0.275	0.261	0.281	0.259	0.286
Electricity	96	**0.120**	**0.211**	0.129	0.221	0.122	0.214	0.170	0.258	0.134	0.229	0.133	0.232	0.156	0.257	0.158	0.248	0.168	0.272
	192	**0.138**	**0.228**	0.148	0.238	0.140	0.231	0.179	0.266	0.154	0.247	0.149	0.247	0.174	0.270	0.170	0.259	0.184	0.322
	336	**0.149**	**0.242**	0.165	0.257	0.152	0.245	0.195	0.282	0.169	0.264	0.161	0.259	0.199	0.297	0.187	0.276	0.198	0.300
	720	**0.174**	**0.269**	0.201	0.285	0.177	0.272	0.237	0.316	0.201	0.294	0.197	0.297	0.242	0.331	0.280	0.363	0.220	0.320
	Avg	**0.145**	**0.237**	0.160	0.250	0.147	0.240	0.195	0.280	0.164	0.258	0.160	0.258	0.192	0.288	0.198	0.286	0.192	0.303

A.2 Horizon-Wise Ablation and Unified-Phase Controls

The removal variants follow Section 4.3 and are implemented as gating-only interventions so that registered model capacity remains matched. The gated rotary positional transform is treated as a shared positional backbone and is retained across the phase-form controls. Accordingly, w/o RelPhase removes only the sample-conditioned harmonic bias, w/o FutPhase removes only the adaptive future phase trajectory, and w/o BothPhase removes both proposed phase-specific operators. Unified-Rel removes the observed phase origin from future conditioning, whereas Unified-Abs replaces the historical harmonic bias with patch-wise absolute-phase conditioning. Differences in forecasting accuracy therefore reflect the active phase pathway rather than changes in structural parameter count.

Exact control definitions. The unified-phase controls preserve the spectral descriptor in Eq. (3) and alter only the phase form exposed to the two forecasting stages. For *Unified-Rel*, historical encoding is unchanged, while future conditioning removes the window-referenced phase origin and uses only horizon-relative harmonic advance,

$$\mathbf{u}_h^{\mathrm{UR}} = \mathrm{Concat}_{k=1}^{K}\left[\bar{A}_k \sin(2\pi f_k h),\ \bar{A}_k \cos(2\pi f_k h)\right], \qquad \boldsymbol{\psi}_h^{\mathrm{UR}} = \tanh(g_f)\mathcal{P}_{\mathrm{phase}}(\mathbf{u}_h^{\mathrm{UR}}). \tag{16}$$

Thus, future generation receives relative harmonic offsets but not the observed phase origin ϕ_k.

For *Unified-Abs*, future decoding remains identical to Eq. (8). The sample-conditioned harmonic bias is disabled, while the generic gated rotary positional transform and standard content attention are retained. Each patch instead receives a window-referenced absolute-phase feature,

$$\theta_{k,i}^{\mathrm{hist}} = \phi_k + 2\pi f_k \tau_i, \qquad \mathbf{u}_i^{\mathrm{UA}} = \mathrm{Concat}_{k=1}^{K}\left[\bar{A}_k \sin\theta_{k,i}^{\mathrm{hist}},\ \bar{A}_k \cos\theta_{k,i}^{\mathrm{hist}}\right], \tag{17}$$

$$\mathbf{P}_i \leftarrow \mathbf{P}_i + \mathcal{P}_{\mathrm{phase}}\left(\mathbf{u}_i^{\mathrm{UA}}\right), \tag{18}$$

where Unified-Abs reuses the same $\mathcal{P}_{\text{phase}}$ projection as the future phase pathway; no additional projection or trainable parameter is introduced. These controls therefore compare the proposed stage-matched pairing against using a relative-only or absolute-only phase convention across the full forecasting pipeline.

Parameter accounting. Capacity matching is audited under the same real-world configuration used in the main experiments: $L = 96$, $H = 96$, $D = 384$, $R = 8$, $E = 4$, $D_{\text{ff}} = 768$, patch length $P = 32$, stride $S = 16$, and $C = 7$. All removal variants are implemented by gating existing paths, and the unified-phase controls reuse existing projections rather than introducing new modules. In particular, Unified-Abs reuses $\mathcal{P}_{\text{phase}}$ in Eq. (18), while Unified-Rel uses the same future-phase projection after replacing the absolute phase state by a relative harmonic state. Consequently, all seven controls have the same registered trainable capacity; only the active phase form or information path changes.

Table 4: Capacity matching for the stage-specific controls under the main $H = 96$ configuration. "0" denotes no additional trainable parameters relative to PhaseAlign.

Variant	Additional trainable parameters
PhaseAlign	0
w/o RelPhase	0
w/o FutPhase	0
w/o BothPhase	0
w/o TF-Interaction	0
Unified-Rel	0
Unified-Abs	0

Table 5: Horizon-wise ablation and unified-phase controls reported as three-seed means. Lower is better.

Dataset	Horizon	Ours		w/o RelPhase		w/o FutPhase		w/o BothPhase		w/o TF-Interaction		Unified-Rel		Unified-Abs	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.358**	**0.380**	0.368	0.387	0.370	0.389	0.375	0.392	0.372	0.390	0.364	0.383	0.366	0.385
	192	**0.411**	**0.412**	0.431	0.421	0.434	0.423	0.441	0.427	0.437	0.424	0.423	0.416	0.426	0.418
	336	**0.454**	**0.434**	0.469	0.442	0.472	0.444	0.481	0.449	0.477	0.446	0.462	0.438	0.465	0.440
	720	**0.462**	**0.459**	0.484	0.471	0.491	0.474	0.503	0.481	0.496	0.477	0.475	0.465	0.480	0.468
	Avg	**0.421**	**0.421**	0.438	0.430	0.441	0.432	0.450	0.437	0.445	0.434	0.431	0.425	0.434	0.427
ETTh2	96	**0.285**	**0.330**	0.303	0.341	0.306	0.343	0.313	0.348	0.309	0.345	0.296	0.336	0.299	0.338
	192	**0.362**	**0.382**	0.389	0.397	0.392	0.399	0.403	0.405	0.397	0.401	0.379	0.390	0.382	0.392
	336	**0.408**	**0.414**	0.435	0.432	0.438	0.434	0.450	0.441	0.444	0.437	0.423	0.424	0.426	0.427
	720	**0.408**	**0.429**	0.430	0.444	0.433	0.447	0.440	0.454	0.438	0.451	0.421	0.437	0.420	0.436
	Avg	**0.365**	**0.388**	0.389	0.403	0.392	0.405	0.401	0.412	0.397	0.408	0.379	0.396	0.381	0.398
Exchange	96	**0.078**	**0.195**	0.087	0.205	0.089	0.207	0.096	0.214	0.092	0.210	0.083	0.200	0.085	0.202
	192	**0.167**	**0.290**	0.184	0.302	0.181	0.300	0.193	0.309	0.189	0.305	0.176	0.296	0.178	0.297
	336	**0.313**	**0.404**	0.339	0.421	0.345	0.424	0.361	0.432	0.353	0.428	0.327	0.413	0.331	0.416
	720	**0.814**	**0.677**	0.889	0.713	0.901	0.717	0.929	0.728	0.916	0.723	0.856	0.696	0.868	0.701
	Avg	**0.343**	**0.391**	0.374	0.410	0.379	0.412	0.394	0.420	0.387	0.416	0.360	0.401	0.365	0.404

B Implementation Details

Unless otherwise stated, all datasets and horizons use the same architecture and optimization settings; only the number of input variables and the prediction horizon change. Inputs are normalized per sample and variable before channel-independent processing and are restored to the original scale at the output. Model selection uses the validation set.

Patch count. Let P denote patch length and S patch stride. The implementation appends one stride of replicate padding before unfolding, giving

$$N = 1 + \left\lfloor \frac{L + S - P}{S} \right\rfloor . \tag{19}$$

With $L = 96$, $P = 32$, and $S = 16$, the model uses $N = 6$ temporal patches. The patch anchor in Section 3.3 is $\tau_i = iS$, $i = 0, \ldots, N-1$.

Table 6: Default architecture and optimization settings for PhaseAlign.

Setting	Value	Setting	Value
Input length L	96	Forecast horizons H	$\{96, 192, 336, 720\}$
Dominant components K	9	Patch length / stride	32 / 16
Latent width D	384	Attention heads R	8
Encoder blocks E	4	Feed-forward width D_{ff}	768
Dropout	0.2	Time-feature embedding	timeF
Optimizer	Adam	Initial learning rate	10^{-4}
Training batch size	32	Evaluation batch size	8
Maximum epochs	10	Early-stopping patience	10
Prediction loss	MSE	Validation criterion	MAE
λ_p / λ_a	0.1 / 0.1	Random seeds	$\{2025, 2026, 2027\}$

B.1 Evaluation Metrics and Model Selection

For a test set containing N_{test} forecast windows, we compute

$$\text{MSE} = \frac{1}{N_{\text{test}}HC} \sum_{n=1}^{N_{\text{test}}} \left\| \widehat{\mathbf{Y}}^{(n)} - \mathbf{Y}^{(n)} \right\|_F^2, \qquad \text{MAE} = \frac{1}{N_{\text{test}}HC} \sum_{n=1}^{N_{\text{test}}} \left\| \widehat{\mathbf{Y}}^{(n)} - \mathbf{Y}^{(n)} \right\|_1. \quad (20)$$

All model-selection decisions are made using validation MAE. The default configuration caps training at 10 epochs and uses a patience value of 10; therefore, validation MAE primarily determines the retained checkpoint, while early termination rarely activates before the epoch cap. Test data are not used for checkpoint selection or hyperparameter tuning.

Randomness. All PhaseAlign results and all internal ablation/control results in the main paper are averages over three independent runs with seeds $\{2025, 2026, 2027\}$. The data splits, architecture, optimization budget, and validation-based model-selection protocol are held fixed across seeds. Main tables report the means, while Appendix G.1 reports the same means together with their standard deviations.

C Deterministic Frequency and Positional Bases

C.1 Fixed Harmonic Frequencies for Unassigned Heads

Let N be the number of temporal patches, S the patch stride, and R the number of attention heads. The non-learnable fallback grid linearly covers the smallest patch-scale frequency $1/(NS)$ and the Nyquist frequency $1/2$:

$$f_j^{\text{fix}} = \frac{1}{NS} + \frac{j-1}{R-1}\left(\frac{1}{2} - \frac{1}{NS}\right), \qquad j = 1, \ldots, R, \quad (21)$$

with $f_1^{\text{fix}} = 1/(NS)$ when $R = 1$. After sorting the selected input components by magnitude, the frequency assigned to head r is

$$\widehat{f}_r = \begin{cases} f_r, & r \le \min(K, R), \\ f_{r-K}^{\text{fix}}, & K < r \le R. \end{cases} \quad (22)$$

The grid is fixed and non-learnable. In the main setting $K = 9$ and $R = 8$, all heads receive observed dominant frequencies; the fallback is used only when fewer than R valid components are available.

C.2 Fixed Decoder Fourier Features

Let $M = \max(1, \lfloor D/2 \rfloor)$ and normalize each future index by

$$t_h = \frac{h}{\max(H-1, 1)}, \qquad h = 0, \dots, H-1. \tag{23}$$

The deterministic frequencies are logarithmically spaced from 1 to 100:

$$\omega_j = \exp\left[\log 1 + \frac{j-1}{\max(M-1, 1)} (\log 100 - \log 1)\right], \qquad j = 1, \dots, M. \tag{24}$$

For $M = 1$, we set $\omega_1 = 1$. The raw Fourier map is

$$\widetilde{\mathbf{e}}_h^{\text{fix}} = [\sin(2\pi t_h \omega_1), \dots, \sin(2\pi t_h \omega_M), \cos(2\pi t_h \omega_1), \dots, \cos(2\pi t_h \omega_M)]\,. \tag{25}$$

It is truncated or zero-padded to D dimensions to obtain $\mathbf{e}_h^{\text{fix}}$. This basis is deterministic, non-learnable, and independent of the input spectrum; it therefore carries no sample-specific spectral peaks or phases. It is distinct from the linear head-frequency grid in Eq. (21).

D Multi-Range Context Aggregation

The decoder first computes one full-history attention map

$$\mathbf{A} = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right) \tag{26}$$

over all N memory tokens. The short-, medium-, and long-range contexts in Eq. (10) use slices of this shared map:

$$\mathbf{C}_h^u = \mathbf{A}_{h,\mathcal{I}_u} \mathbf{V}_{\mathcal{I}_u}, \qquad u \in \{s, m, l\}. \tag{27}$$

The restricted slices are not renormalized. Their magnitude therefore reflects both the values inside a support and the probability mass assigned to that support by the full attention distribution. The coefficients $\alpha_s, \alpha_m, \alpha_l$ are global learned scalars shared across forecast steps; horizon dependence enters through the query-specific attention row $\mathbf{A}_h$. Reusing a single map avoids three separate query–memory attention computations.

E Complexity Analysis

Let B be the batch size, C the number of variables, L the input length, N the number of patches, K the retained spectral components, R the attention heads, D the latent width, D_{ff} the feed-forward width, E the encoder depth, and H the forecast horizon. Because processing is channel-independent, the effective number of univariate sequences is BC.

Table 7: Leading time and activation-memory complexity of the main PhaseAlign components.

Component	Time complexity	Attention/feature memory
RFFT and top-K selection	$\mathcal{O}(BC[L \log L + L \log K])$	$\mathcal{O}(BCL)$
Bidirectional time–frequency attention	$\mathcal{O}\big(BC[NKD + (N+K)D^2]\big)$	$\mathcal{O}(BCRNK)$
E historical encoder blocks	$\mathcal{O}\big(BCE[N^2D + ND^2 + NDD_{\text{ff}}]\big)$	$\mathcal{O}(BCERN^2)$
Future phase feature construction	$\mathcal{O}(BCHKD)$	$\mathcal{O}(BCHD)$
Query–memory decoding	$\mathcal{O}\big(BC[HND + (H+N)D^2]\big)$	$\mathcal{O}(BCRHN)$

Combining the leading terms gives

$$\mathcal{O}\Big(BC\big[L \log L + L \log K + NKD + (N+K)D^2 + E(N^2D + ND^2 + NDD_{\text{ff}}) + HND + (H+N)D^2 + HKD\big]\Big). \tag{28}$$

The stage-specific phase operations do not change the dominant quadratic order of patch self-attention. The harmonic bias adds $\mathcal{O}(BCRN^2)$ inexpensive trigonometric/bias operations, while the future phase trajectory adds a term linear in HK. Multi-range aggregation reuses the same query–memory attention map and therefore does not multiply the $\mathcal{O}(HND)$ decoder cost by three. Learned parameter growth is linear in K and H for component slots and horizon queries/embeddings, while the projection matrices scale as $\mathcal{O}(D^2)$.

F CONTROLLED-ANALYSIS PROTOCOLS

This section formalizes the controlled analyses in Section 4.4. The reported generator settings and evaluation parameters are taken directly from the corresponding experiment scripts so that each intervention can be reproduced independently of the training code.

F.1 SPECTRAL-TWIN PHASE-SHIFT ROBUSTNESS

We use a controlled spectral-twin experiment to isolate robustness to relative-phase changes while preserving the underlying component frequencies and amplitudes. The purpose of this experiment is not to replicate the real-world forecasting setting, but to test whether a model trained under one phase relationship retains the temporal placement of future oscillatory structure when that relationship changes at test time.

Signal construction. The synthetic generator contains two harmonically related components. For time index t, the clean sequence is

$$x_t^{(\Delta)} = \sin(2\pi f t + \phi) + 0.5\,\sin(2\pi(2f)t + 2\phi + \Delta + 0.7)\,, \tag{29}$$

where

$$f \sim \mathcal{U}(0.9, 1.1)/24, \qquad \phi \sim \mathcal{U}(0, 2\pi). \tag{30}$$

Thus, the experiment uses two fixed component amplitudes $[1.0, 0.5]$, while the base frequency and nuisance absolute phase are sampled independently for every sequence. The intervention variable Δ changes only the relative phase of the second harmonic; the generative component frequencies and amplitudes remain unchanged. Continuous frequency sampling is used rather than a fixed DFT-bin grid so that the task cannot be reduced to memorizing a small set of Fourier locations. Because the evaluated histories are finite, noisy, and generally off-grid, we do not require their finite-window FFT magnitudes to be numerically identical for every Δ; the controlled quantity is the underlying component amplitude/frequency structure.

Delta-OOD protocol. Each model observes $L = 96$ history points and predicts $H = 48$ future points. The nominal period is $T_0 = 24$, and Gaussian noise with standard deviation $\sigma = 0.10$ is added by the generator. Training and validation use only the reference phase relationship $\Delta = 0$, with 3,000 and 500 sequences, respectively. At test time, the trained checkpoint is evaluated without parameter updates on

$$\Delta \in \left\{0, \frac{\pi}{8}, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi\right\}. \tag{31}$$

Each offset contains 400 independently generated test sequences, giving 2,400 test examples in total. Training and validation use noisy futures, whereas test histories are noisy and evaluation is performed against the corresponding clean future. For the repeated-run analysis, the complete train-at-$\Delta = 0$ procedure is independently repeated using seeds $\{2025, 2026, 2027\}$. Each resulting checkpoint is then evaluated on all six offsets.

Peak Timing Error. The controlled analysis focuses on the temporal placement of the dominant future oscillatory extremum. For a future sequence $\mathbf{z}$, define

$$p(\mathbf{z}) = \arg\max_{h>0} |z_h|. \tag{32}$$

Peak Timing Error (PTE) is then

$$\mathrm{PTE}(\widehat{\mathbf{y}}, \mathbf{y}) = |p(\widehat{\mathbf{y}}) - p(\mathbf{y})|\,, \tag{33}$$

measured in forecast steps and averaged over the test set. No peak detector, minimum-distance threshold, or prominence threshold is used. This metric is intentionally targeted at temporal placement rather than full-spectrum phase agreement: the experiment asks whether the dominant future oscillatory structure remains correctly localized as the unseen relative-phase relationship moves away from the training condition.

Table 8: Spectral-twin protocol: train at $\Delta = 0$ and test without adaptation over six relative-phase offsets.

Setting	Value
Components / amplitudes	$Q = 2$; fixed $[1.0, 0.5]$
History / forecast length	$L = 96$ / $H = 48$
Nominal period / noise	$T_0 = 24$ / $\sigma = 0.10$
Frequency sampling	$f \sim \mathcal{U}(0.9, 1.1)/24$ (continuous)
Absolute phase	$\phi \sim \mathcal{U}(0, 2\pi)$
Training phase offset	$\Delta = 0$ only
Train / validation size	$3000/500$
Test offsets	$\{0, \pi/8, \pi/4, \pi/2, 3\pi/4, \pi\}$
Test examples	400 per offset (2400 total)
Train/validation targets	noisy future
Test target	clean future
Optimization	Adam, learning rate 10^{-3}, MSE loss
Batch / epochs / patience	$256/50/10$
Repeated-run seeds	$\{2025, 2026, 2027\}$
Primary robustness metric	PTE (forecast steps)

Controlled-model configuration. For this synthetic experiment, PhaseAlign uses latent width $D = 32$, four attention heads, patch length/stride $32/16$, two encoder layers, $D_{\text{ff}} = 64$, dropout 0.1, one input channel, and no time-feature marks. The external comparison reported below uses the experiment's lightweight PhaseFormer and FEDformer implementations under the same data splits and optimization protocol. The stage-specificity study additionally trains Unified-Rel, Unified-Abs, and w/o BothPhase independently. All reported models are trained from scratch for each seed; no pretrained or shared checkpoint is used across methods.

Three-seed phase-shift robustness. Table 9 reports mean±standard deviation across the three independently trained seeds.

Table 9: Three-seed spectral-twin PTE across relative-phase offsets. Lower is better.

Model	0	$\pi/8$	$\pi/4$	$\pi/2$	$3\pi/4$	π
PhaseAlign	0.1726±0.0077	0.2788±0.0496	0.3650±0.0450	0.5750±0.0550	0.6850±0.0500	0.7150±0.0500
Unified-Abs	0.1755±0.0154	0.2469±0.0287	0.4007±0.0574	0.6389±0.0826	0.7634±0.0581	0.8068±0.0515
Unified-Rel	0.1885±0.0155	0.2812±0.0073	0.4692±0.0082	0.7909±0.0298	1.0775±0.1086	1.2157±0.1982
w/o BothPhase	0.1744±0.0072	0.2735±0.0438	0.4257±0.0876	0.7148±0.1785	0.8559±0.2165	0.9237±0.2346
FEDformer*	0.1566±0.0079	0.2800±0.0239	0.5242±0.0409	1.0098±0.0852	1.3140±0.2124	1.1091±0.1912
PhaseFormer	0.1300±0.0034	0.4271±0.0115	0.8320±0.0396	1.4124±0.0865	1.5575±0.1024	1.4832±0.1676

Near the training condition, several models achieve similar PTE and PhaseFormer obtains the lowest value at $\Delta = 0$. The behavior changes as the relative-phase shift increases. PhaseAlign grows from 0.1726 ± 0.0077 at $\Delta = 0$ to only 0.7150 ± 0.0500 at $\Delta = \pi$, whereas PhaseFormer reaches 1.4832 ± 0.1676 and FEDformer* reaches 1.1091 ± 0.1912. At $\Delta = \pi$, the PhaseAlign PTE is therefore approximately 51.8% lower than PhaseFormer and 35.5% lower than FEDformer*. The stage-specificity controls also degrade more strongly: Unified-Abs reaches 0.8068 ± 0.0515, w/o BothPhase reaches 0.9237 ± 0.2346, and Unified-Rel reaches 1.2157 ± 0.1982.

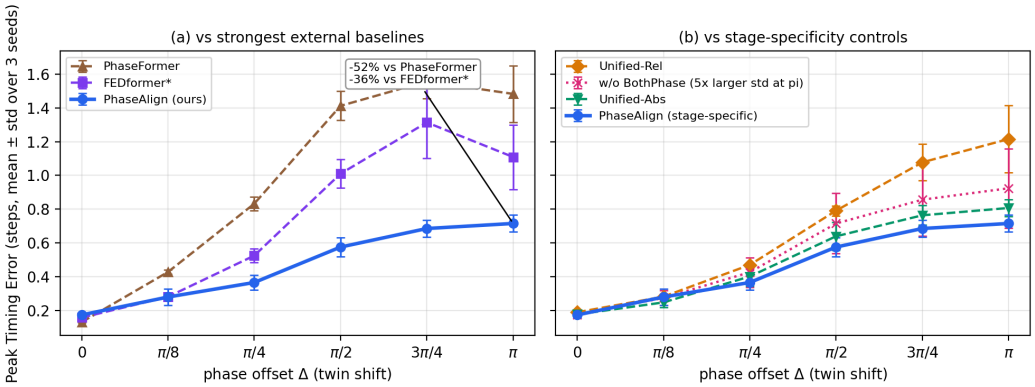


Figure 6: Three-seed spectral-twin PTE under unseen relative-phase shifts. Left: external baselines; right: stage-specificity controls.

Endpoint phase-drift summary. To summarize sensitivity over the complete phase-offset sweep, define

$$D_m = \overline{\mathrm{PTE}}_m(\pi) - \overline{\mathrm{PTE}}_m(0) \tag{34}$$

for model m. For the displayed uncertainty, we use endpoint error propagation,

$$\widetilde{\sigma}_{D_m} = \sqrt{\sigma_m^2(\pi) + \sigma_m^2(0)}. \tag{35}$$

This quantity is used only for the error bars in the endpoint summary; it is not the standard deviation of seed-wise paired drift values.

Table 10: Endpoint PTE drift from $\Delta = 0$ to $\Delta = \pi$; lower indicates stronger phase robustness.

Model	PTE(0)	PTE(π)	D_m
PhaseAlign	0.1726±0.0077	0.7150±0.0500	0.5424±0.0506
Unified-Abs	0.1755±0.0154	0.8068±0.0515	0.6313±0.0538
w/o BothPhase	0.1744±0.0072	0.9237±0.2346	0.7493±0.2347
FEDformer*	0.1566±0.0079	1.1091±0.1912	0.9525±0.1914
Unified-Rel	0.1885±0.0155	1.2157±0.1982	1.0272±0.1988
PhaseFormer	0.1300±0.0034	1.4832±0.1676	1.3532±0.1676

Overall, the spectral-twin experiment shows that the main advantage of the stage-specific phase formulation appears under distribution shift rather than at the training phase relationship itself. Although PhaseAlign is not the lowest-PTE model at $\Delta = 0$, its PTE increases substantially more slowly as the unseen relative-phase offset grows. The unified-phase and phase-removal controls exhibit larger degradation, supporting the use of different phase conventions for historical encoding and future conditioning when the phase relationship changes beyond the training regime.

F.2 Matched Phase–Magnitude Intervention

This intervention isolates the effect of changing phase from the effect of changing spectral magnitude while matching the perturbation strength window by window. The evaluation uses the ETTh1 test set on the OT channel under univariate forecasting with $L = 96$ and $H = 96$. Intervention parameters are estimated from the observed history only, and the transformed history is then fed to the forecasting model; the future target remains clean and unperturbed.

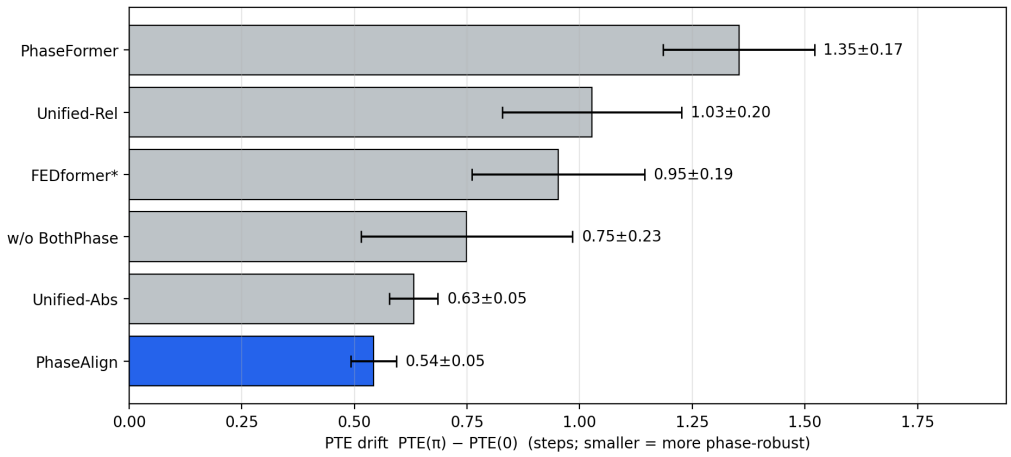


Figure 7: Endpoint PTE drift from $\Delta = 0$ to $\Delta = \pi$; lower values indicate stronger phase robustness.

History-only matched perturbation. For an observed history vector $\mathbf{x} \in \mathbb{R}^L$, let X_k denote its real-FFT coefficients. A phase intervention changes only phase on a selected set of bins $\mathcal{S}$,

$$X_k^{\text{phase}} = X_k \exp(jcs_k), \qquad k \in \mathcal{S}, \tag{36}$$

where $s_k \in \{\pm 1\}$ are fixed random signs shared with the matched magnitude intervention. A magnitude intervention changes only magnitude on the same bins,

$$X_k^{\text{mag}} = X_k(1 + cs_k), \qquad k \in \mathcal{S}. \tag{37}$$

All non-selected bins remain unchanged. The matching criterion is the relative ℓ_2 perturbation measured on the observed history alone:

$$\rho(\mathbf{x}, \widetilde{\mathbf{x}}) = \frac{\|\widetilde{\mathbf{x}} - \mathbf{x}\|_2}{\|\mathbf{x}\|_2}. \tag{38}$$

Window-by-window matching. Phase and magnitude interventions use the same top-K non-DC frequency bins and random signs, but do not share the same perturbation coefficient. For each history $\mathbf{x}_n$, we first set a common distortion target ρ_n^{tgt} as a fixed fraction of that window's phase-perturbation reach. The phase and magnitude coefficients are then determined separately to match this same target. Because magnitude distortion scales linearly with its coefficient, its value is obtained by direct scaling, whereas the phase coefficient is found by bisection on $[0, \pi]$ using the resulting time-domain distortion. Thus,

$$\rho_n^{\text{phase}} \approx \rho_n^{\text{mag}} \approx \rho_n^{\text{tgt}} \tag{39}$$

holds for each window rather than only after averaging. The reported intervention level is the mean target distortion $\bar{\rho} = \frac{1}{N}\sum_n \rho_n^{\text{tgt}}$, yielding $\{0.041570, 0.103924, 0.207848, 0.311773\}$ on ETTh1/OT.

Forecast degradation under matched perturbations. After perturbing only the history, we evaluate forecasting against the original clean future. Let MSE_{orig} denote the forecasting MSE on the unperturbed histories and MSE_{pert} the MSE on the perturbed histories. We report

$$\Delta\text{MSE} = \text{MSE}_{\text{pert}} - \text{MSE}_{\text{orig}}, \tag{40}$$

and

$$\text{RelInc}(\%) = 100 \cdot \frac{\text{MSE}_{\text{pert}} - \text{MSE}_{\text{orig}}}{\text{MSE}_{\text{orig}}}. \tag{41}$$

Table 11: Matched phase–magnitude intervention protocol on ETTh1/OT and the achieved history-domain distortion levels.

Setting	Value
Dataset / target channel	ETTh1 / OT, univariate, $L = 96$ / $H = 96$
Intervention parameters	estimated from observed history only
Selected-frequency rule	same top-K bins and random signs for both modes
Distortion fractions	$\{0.10, 0.25, 0.50, 0.75\}$
Achieved mean $\bar{\rho}$	$\{0.041570, 0.103924, 0.207848, 0.311773\}$
Matching	same per-window ρ_n^{tgt}; coefficients solved separately

For peak timing, we report $\Delta\mathrm{PTE} = \mathrm{PTE}_{\mathrm{phase}} - \mathrm{PTE}_{\mathrm{clean}}$ under phase perturbation at the strongest level. At $\rho \approx 0.31$ (frac $= 0.75$):

$$\begin{aligned}
\text{PhaseAlign:} &\quad 52.69\% \,/\, 36.45\%, \quad \Delta\mathrm{PTE} = 0.83,\\
\text{PhaseFormer:} &\quad 57.95\% \,/\, 56.27\%, \quad \Delta\mathrm{PTE} = 1.54,\\
\text{FITS:} &\quad 35.81\% \,/\, 15.76\%, \quad \Delta\mathrm{PTE} = 1.00,\\
\text{PatchTST:} &\quad 56.28\% \,/\, 42.83\%, \quad \Delta\mathrm{PTE} = 1.62,\\
\text{FEDformer}^{*}: &\quad 57.17\% \,/\, 36.51\%, \quad \Delta\mathrm{PTE} = 1.50.
\end{aligned}$$

Each percentage pair is phase/magnitude relative MSE increase. Figure 4(a) shows the full four-level curves for the plotted models, while Fig. 4(b) summarizes their strongest-level relative MSE increases and phase-induced timing degradation. FITS is included here for completeness from the same evaluation table but omitted from the main figure for visual clarity. At the strongest matched intervention, phase perturbation is more damaging than magnitude perturbation for every evaluated model, while PhaseAlign has the smallest phase-induced peak-timing degradation (0.83 steps; FITS is next at 1.00). This supports the distinction between overall forecasting sensitivity and preservation of temporal placement under phase corruption.

F.3 Progressive Frequency Drift

We further evaluate robustness when the future frequency evolves continuously beyond the stationary pattern observed during training. The controlled generator uses a linear chirp that is continuous across the history–future boundary. Its clean trajectory is

$$x_t^{\mathrm{clean}} = \sin\left[2\pi\left(ft + \frac{1}{2}\kappa t^2\right) + \phi\right], \tag{42}$$

with instantaneous frequency

$$f_{\mathrm{inst}}(t) = f + \kappa t. \tag{43}$$

For each generated sequence,

$$f \sim \mathcal{U}(0.9, 1.1)/24, \qquad \phi \sim \mathcal{U}(0, 2\pi), \tag{44}$$

and Gaussian noise with standard deviation $\sigma = 0.05$ is added by the generator. The history length is fixed at $L = 96$, while the forecast horizon varies over

$$H \in \{48, 96, 192, 336\}, \qquad H/L \in \{0.5, 1, 2, 3.5\}. \tag{45}$$

Stationary-train / drift-test protocol. For each forecast horizon, every model is trained from scratch only on stationary sequences with $\kappa = 0$. The resulting checkpoint is then evaluated without parameter updates on progressively drifting test sequences. The train/validation/test sets contain $3000/500/400$ sequences, respectively, with an independent random seed used for each horizon. To remove optimization-budget confounding, all compared methods use the same maximum budget of 80 epochs and early-stopping patience 15; validation MAE selects the retained checkpoint for every model. Forecasting performance is measured using MAE against the corresponding clean future.

Rather than using the same numerical κ for horizons of different lengths, the experiment uses an equal-total-phase-error grid. Let

$$T = L + H \tag{46}$$

denote the end time of the complete history-plus-future trajectory. For a dimensionless target level

$$q \in \{0, 0.5, 1, 2, 4\}, \tag{47}$$

the drift coefficient is chosen as

$$\kappa(H, q) = \frac{q}{T^2}. \tag{48}$$

Hence the total chirp contribution at the end of the complete window is

$$\pi\kappa T^2 = q\pi, \tag{49}$$

giving the common phase-target grid

$$\{0, 0.5\pi, \pi, 2\pi, 4\pi\} \tag{50}$$

for every forecast horizon.

Table 12: Drift coefficients κ for equal end-window phase targets across horizons.

H	0.5π	π	2π	4π
48	2.411×10^{-5}	4.823×10^{-5}	9.645×10^{-5}	1.929×10^{-4}
96	1.356×10^{-5}	2.713×10^{-5}	5.425×10^{-5}	1.085×10^{-4}
192	6.028×10^{-6}	1.206×10^{-5}	2.411×10^{-5}	4.823×10^{-5}
336	2.679×10^{-6}	5.358×10^{-6}	1.072×10^{-5}	2.143×10^{-5}

Accumulated future phase mismatch. The history provides only an estimate of the frequency available before forecasting. If that history-estimated frequency is extrapolated stationarily into the future, while the true process continues to chirp, the additional phase mismatch accumulated after the history boundary is

$$g_\kappa(h) = \pi\kappa\left[(L + h)^2 - L^2\right], \qquad h = 0, \ldots, H. \tag{51}$$

Thus $g_\kappa(0) = 0$, while the mismatch grows quadratically with forecast step. This quantity directly measures the error incurred when a history-estimated stationary frequency is extrapolated along the future phase trajectory.

It is important to distinguish Eq. 51 from the end-window phase target used to construct the κ grid. Combining Eqs. 48 and 51 gives

$$\frac{g_\kappa(H)}{\pi} = q\left[1 - \left(\frac{L}{L + H}\right)^2\right]. \tag{52}$$

For example, when $L = 96$ and $H = 192$, $L/(L + H) = 1/3$, so the four nonzero target levels $q = \{0.5, 1, 2, 4\}$ correspond to boundary-anchored future mismatches

$$\frac{g_\kappa(H)}{\pi} = \{0.444,\ 0.889,\ 1.778,\ 3.556\}, \tag{53}$$

which are the endpoints shown in Fig. 8(c).

Relative-gain metric. For each (H, q) condition, we compare PhaseAlign with the strongest baseline among PhaseFormer, PatchTST, and FEDformer*. Let

$$\mathrm{MAE}_{\mathrm{best}} = \min_{m \in \{\mathrm{PhaseFormer}, \mathrm{PatchTST}, \mathrm{FEDformer}^*\}} \mathrm{MAE}_m. \tag{54}$$

The reported relative gain is

$$\mathrm{Gain}(\%) = 100\frac{\mathrm{MAE}_{\mathrm{best}} - \mathrm{MAE}_{\mathrm{PA}}}{\mathrm{MAE}_{\mathrm{best}}}. \tag{55}$$

Positive values therefore indicate that PhaseAlign achieves lower MAE than the strongest baseline under the same drift condition.

Table 13: Relative MAE gain (%) of PhaseAlign over the strongest compared baseline; positive is better.

End-window phase target	$H = 48$	$H = 96$	$H = 192$	$H = 336$
0	-8.28	-12.96	-11.98	-5.36
0.5π	$+3.46$	-1.70	-0.69	-1.21
π	$+7.12$	$+2.79$	$+0.28$	$+0.26$
2π	$+14.31$	$+11.96$	$+2.85$	$+1.01$
4π	$+26.13$	$+23.50$	$+7.51$	$+4.16$

Model-resolved results at $H = 96$. Table 14 reports the absolute MAE values used in Fig. 8(a). In the stationary setting ($q = 0$), PhaseAlign is not the lowest-error model. As the future process departs from the stationary training regime, however, its degradation becomes smaller than that of the comparison models. At the strongest tested drift (4π), PhaseAlign obtains an MAE of 0.7302, compared with 0.9926 for PhaseFormer, 0.9545 for PatchTST, and 1.0607 for FEDformer*.

Table 14: Absolute MAE at $H = 96$ under progressive frequency drift. Lower is better.

Model	0	0.5π	π	2π	4π
PhaseAlign	0.0414	0.2187	0.4146	0.7193	0.7302
PhaseFormer	0.0366	0.2150	0.4264	0.8170	0.9926
PatchTST	0.0395	0.2280	0.4419	0.8186	0.9545
FEDformer*	0.0441	0.2311	0.4508	0.8472	1.0607

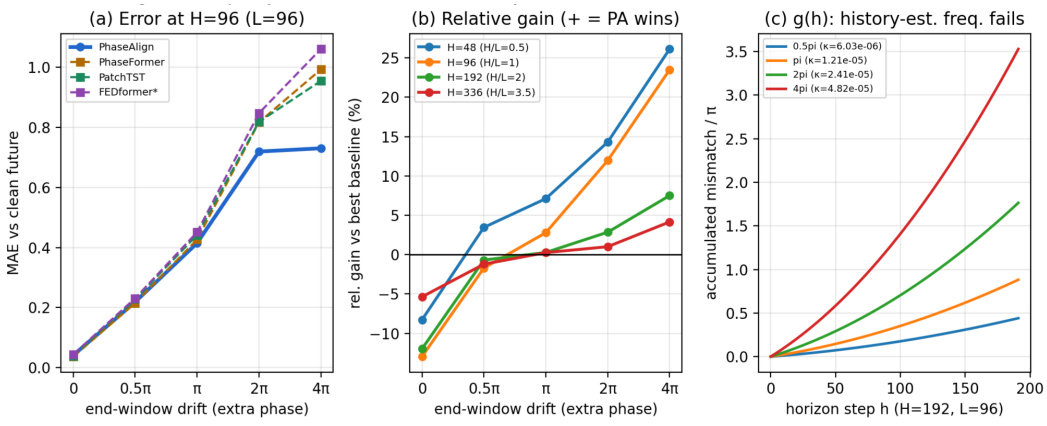


Figure 8: Progressive frequency-drift stress test. Panels show $H = 96$ MAE, relative gain across horizons, and accumulated future phase mismatch.

The stress test reveals a transition between the stationary and nonstationary regimes. At $\kappa = 0$, PhaseAlign does not outperform the strongest baseline, and at the smallest 0.5π drift its advantage is limited to $H = 48$. Once the accumulated phase error becomes larger, the relative behavior changes consistently: PhaseAlign achieves positive gain for all four horizons at π, 2π, and 4π, with gains of 26.13%, 23.50%, 7.51%, and 4.16% at the strongest 4π condition for $H = 48, 96, 192, 336$, respectively. Together with the quadratic mismatch curves in Fig. 8(c), these results characterize how the future-phase mechanism degrades when the history-estimated frequency becomes progressively less valid beyond the observation window.

Table 15: Progressive frequency-drift protocol with stationary training and drifted evaluation.

Setting	Value
Signal	$x_t = \sin\big(2\pi(ft + \frac{1}{2}\kappa t^2) + \phi\big)$
Instantaneous frequency	$f_{\text{inst}}(t) = f + \kappa t$
Base frequency	$f \sim \mathcal{U}(0.9, 1.1)/24$
Initial phase	$\phi \sim \mathcal{U}(0, 2\pi)$
Noise	$\sigma = 0.05$
History length	$L = 96$
Forecast horizons	$H \in \{48, 96, 192, 336\}$
Horizon ratios	$H/L \in \{0.5, 1, 2, 3.5\}$
Train / validation / test	$3000/500/400$ per horizon
Training drift	$\kappa = 0$ only
Drift targets	$\{0, 0.5\pi, \pi, 2\pi, 4\pi\}$
Drift coefficient	$\kappa = q/(L + H)^2$ for target $q\pi$
Training budget (all models)	80 epochs; patience 15
Evaluation target	clean future
Primary metric	MAE
Relative gain	Eq. 55; strongest compared baseline
Future mismatch diagnostic	$g_\kappa(h)$ in Eq. 51

F.4 Off-Grid Frequency and Window-Shift Robustness

We evaluate whether the frequency- and phase-aware operators remain stable when the dominant oscillation falls between discrete Fourier bins and when the same off-grid process is observed under different window origins. This analysis uses clean single-tone probes and reuses the $H = 96$ stationary checkpoints from the frequency-drift experiment; no model is retrained for the off-grid or window-shift tests.

Off-grid probe construction. The probe signal is

$$x_t = \sin(2\pi f_\delta t + \phi)\,, \qquad f_\delta = \frac{4 + \delta}{96}, \tag{56}$$

where

$$\delta \in \{0,\ 0.125,\ 0.25,\ 0.375,\ 0.5\}, \qquad \phi \sim \mathcal{U}(0, 2\pi). \tag{57}$$

The history and forecast lengths are both $L = H = 96$. No noise is added in this probe ($\sigma = 0$), and each value of δ is evaluated on $N = 400$ independently generated examples.

The models use the corresponding $H = 96$, $\kappa = 0$ checkpoints from the stationary-training protocol in Section F.3. Those checkpoints were trained using $3000/500/400$ train/validation/test synthetic sequences with $\sigma = 0.05$. The off-grid analysis therefore changes only the evaluation distribution and does not introduce additional training.

Frequency-selection diagnostics. For each observed history, let X_k denote its real-FFT coefficients. The dominant non-DC Fourier bin is selected as

$$\widehat{k} = \arg\max_{k>0} |X_k|. \tag{58}$$

Because the true tone location in bin units is

$$k_{\text{true}} = f_\delta L = 4 + \delta, \tag{59}$$

we report the bin-selection error

$$\Delta_k = \left| \widehat{k} - f_\delta L \right| \tag{60}$$

and the fraction of spectral energy captured by the selected bin,

$$\eta_{\mathrm{sel}} = \frac{|X_{\widehat{k}}|^2}{\sum_k |X_k|^2}. \tag{61}$$

These quantities depend only on the input FFT and are therefore model-independent.

Table 16: Off-grid/window-shift probe protocol using stationary $H = 96$ checkpoints.

Setting	Value
Probe signal	$x_t = \sin(2\pi f_\delta t + \phi)$
Frequency	$f_\delta = (4 + \delta)/96$
Off-grid offsets	$\delta \in \{0, 0.125, 0.25, 0.375, 0.5\}$
Absolute phase	$\phi \sim \mathcal{U}(0, 2\pi)$
History / forecast length	$L = 96$ / $H = 96$
Probe noise	none ($\sigma = 0$)
Examples per δ	$N = 400$
Model checkpoints	drift-protocol $H = 96$, $\kappa = 0$ stationary weights
Checkpoint training data	$3000/500/400$, training noise $\sigma = 0.05$
Retraining for this probe	none
Compared models	PhaseAlign, PhaseFormer, PatchTST, FEDformer*
Selected bin	$\widehat{k} = \arg\max_{k>0} \lvert X_k \rvert$
Selection error	$\Delta_k = \lvert \widehat{k} - f_\delta L \rvert$
Selected-bin energy	$\eta_{\mathrm{sel}} = \lvert X_{\widehat{k}} \rvert^2 / \sum_k \lvert X_k \rvert^2$
Window shifts	$s \in \{0, \ldots, 23\}$ at $\delta = 0.5$
Forecast metric	MSE against clean future

Off-grid forecasting error. Table 17 reports forecasting MSE as the tone moves from an exactly aligned Fourier bin ($\delta = 0$) toward the maximum half-bin offset ($\delta = 0.5$).

Table 17: MSE under off-grid frequency offsets at $L = H = 96$. Lower is better.

Model	$\delta = 0$	0.125	0.25	0.375	0.5
PhaseAlign	0.001348	0.001053	0.000939	0.004164	**0.071054**
PhaseFormer	0.000688	0.000462	0.000871	0.001582	0.108396
PatchTST	0.000547	0.000557	0.000942	0.002256	<u>0.106702</u>
FEDformer*	0.001874	0.000955	0.001024	0.003401	0.114144

For $\delta \leq 0.375$, all four models remain in a low-error regime, with MSE values on the order of 10^{-3}. The largest change occurs at the half-bin condition $\delta = 0.5$, where spectral leakage is strongest. At this point, PhaseAlign obtains an MSE of 0.071054, compared with 0.106702 for the next-lowest model, PatchTST. This corresponds to a 33.4% reduction in MSE relative to that baseline under the maximum-leakage probe.

Spectral leakage and bin-selection degradation. The FFT diagnostics make the source of the $\delta = 0.5$ stress explicit. As the tone moves away from the discrete Fourier grid, the selected-bin location becomes increasingly inaccurate and its share of the total spectral energy decreases.

Table 18: FFT bin-selection error and selected-bin energy versus off-grid offset.

δ	Δ_k (bins)	η_{sel}
0	0.000	1.0000
0.125	0.125	0.9505
0.25	0.250	0.8143
0.375	0.375	0.6174
0.5	0.500	0.4383

At $\delta = 0.5$, the dominant-bin selector necessarily chooses one of the two neighboring integer bins, producing a 0.5-bin location error. At the same time, the energy captured by that single selected bin falls from 100% at $\delta = 0$ to only 43.83%. Thus, the maximum-error forecasting condition coincides with the strongest degradation of the discrete peak representation.

Window-shift robustness at maximum leakage. We next fix the maximally off-grid tone at $\delta = 0.5$ and shift the observation window by

$$s \in \{0, 1, \ldots, 23\}. \tag{62}$$

This changes the phase origin under which the same off-grid oscillation is presented while preserving its frequency and leakage condition. Table 19 summarizes MSE across the 24 shifts.

Table 19: Window-shift MSE summary for the $\delta = 0.5$ off-grid tone.

Model	Mean	Min	Max	Range width
PhaseAlign	**0.07204**	0.055877	0.097907	**0.042030**
PhaseFormer	0.10325	0.055250	0.158570	0.103320
PatchTST	0.10871	0.068584	0.138225	0.069641
FEDformer*	0.11808	0.075693	0.158216	0.082523

PhaseAlign has both the lowest mean shifted-window MSE (0.07204) and the narrowest error range (0.04203). For comparison, PhaseFormer's mean is 0.10325 and its range width is 0.10332. The reduced sensitivity across window origins supports the interpretation that the phase-aware forecasting mechanism does not rely on one favorable alignment of the observation window.

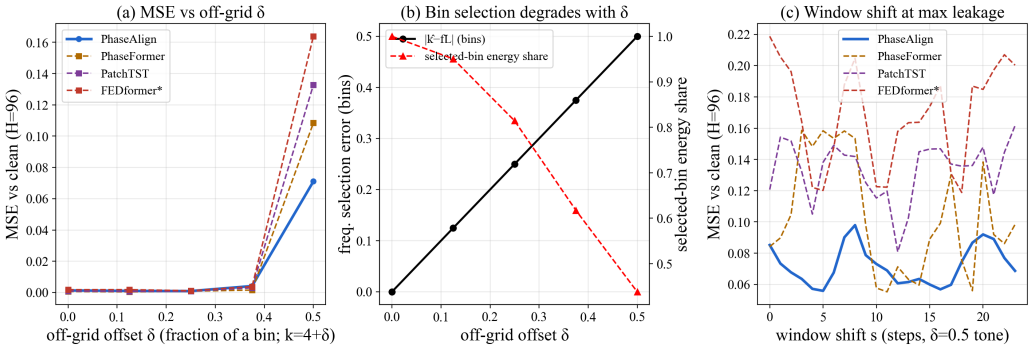


Figure 9: Off-grid frequency and window-shift robustness at $H = 96$. Panels show forecasting error, FFT selection degradation, and shift sensitivity at $\delta = 0.5$.

Figure 9 links the forecasting behavior to the underlying discrete-spectrum failure mode. Performance remains stable while the dominant Fourier bin still captures most of the tone energy, but all methods degrade when the frequency lies halfway between bins. Under this maximum-leakage condition, PhaseAlign shows lower forecasting error than the three comparison models and remains

more stable as the window origin is shifted, providing evidence that its phase handling is less sensitive to exact DFT-bin and observation-window alignment.

G Repeated-Run Robustness and Hyperparameter Sensitivity

G.1 Repeated-Run Robustness

To assess robustness to random initialization, we repeat the complete training and model-selection procedure with three independent seeds ($\{2025, 2026, 2027\}$) on ETTh1, ETTh2, and Exchange, keeping data splits, architecture, and optimization settings fixed. Table 20 reports the mean$\pm$standard deviation. All stage-specific controls are retrained independently rather than evaluated from a shared checkpoint; the equality of means with the main ablation table is therefore intentional, while exposing the run-to-run dispersion.

The mean ordering is preserved across seeds. PhaseAlign consistently achieves the lowest MSE and MAE, outperforming the closest control (Unified-Rel) by reducing MSE by roughly 2.3%, 3.7%, and 4.7%, and MAE by 0.9%, 2.0%, and 2.5% on ETTh1, ETTh2, and Exchange. Seed dispersion remains small for the full model (MSE std $\approx$ 0.006–0.007), while aggressive ablations (e.g., w/o BothPhase, w/o TF-Interaction) exhibit larger variations, confirming that the improvements are reproducible rather than driven by a single initialization.

Table 20: Three-seed robustness on ETTh1, ETTh2, and Exchange; entries are mean$\pm$standard deviation.

Dataset	Method	MSE	MAE	Seeds
ETTh1	PhaseAlign	**0.421±0.007**	**0.421±0.006**	3
	w/o RelPhase	0.438±0.010	0.430±0.008	3
	w/o FutPhase	0.441±0.011	0.432±0.009	3
	w/o BothPhase	0.450±0.014	0.437±0.012	3
	w/o TF-Interaction	0.445±0.012	0.434±0.010	3
	Unified-Rel	0.431±0.007	0.425±0.006	3
	Unified-Abs	0.434±0.006	0.427±0.005	3
ETTh2	PhaseAlign	**0.365±0.006**	**0.388±0.005**	3
	w/o RelPhase	0.389±0.009	0.403±0.008	3
	w/o FutPhase	0.392±0.010	0.405±0.009	3
	w/o BothPhase	0.401±0.013	0.412±0.012	3
	w/o TF-Interaction	0.397±0.011	0.408±0.010	3
	Unified-Rel	0.379±0.007	0.396±0.006	3
	Unified-Abs	0.381±0.005	0.398±0.005	3
Exchange	PhaseAlign	**0.343±0.006**	**0.391±0.006**	3
	w/o RelPhase	0.374±0.010	0.410±0.010	3
	w/o FutPhase	0.379±0.011	0.412±0.011	3
	w/o BothPhase	0.394±0.016	0.420±0.015	3
	w/o TF-Interaction	0.387±0.013	0.416±0.012	3
	Unified-Rel	0.360±0.008	0.401±0.008	3
	Unified-Abs	0.365±0.005	0.404±0.006	3

G.2 Sensitivity to the Number of Spectral Components

The main configuration uses $K = 9$ retained spectral components. We evaluate $K \in \{2, 4, 8, 9, 12, 16\}$ on ETTh1 with $L = 96$ and $H = 96$ while keeping the remaining sensitivity-study settings fixed. Each configuration is trained for up to 20 epochs and the checkpoint retained by the standard MAE-based model selection criterion is reported.

Table 21: Sensitivity to retained spectral components K on ETTh1 ($L = 96$, $H = 96$).

K	Best ep.	MSE	MAE
2	8	0.3723	0.3829
4	13	0.3643	0.3814
8	9	0.3595	**0.3805**
9 (default)	12	**0.3593**	0.3806
12	10	0.3597	0.3808
16	9	0.3605	0.3812

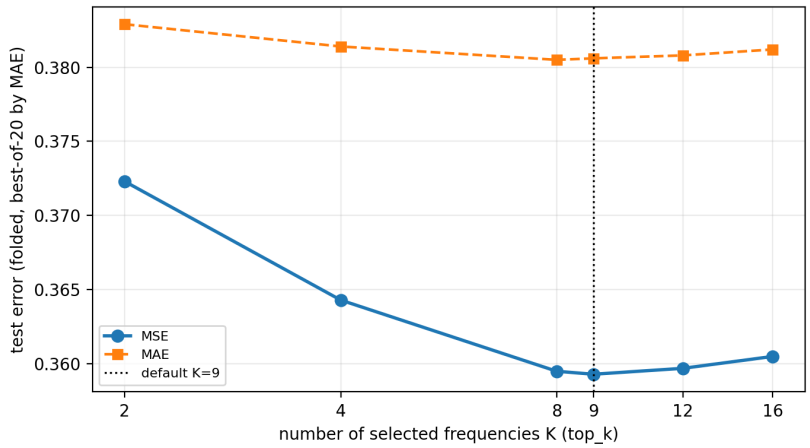


Figure 10: Sensitivity to the number of retained spectral components K on ETTh1.

Relative to the default $K = 9$, using only two components increases MSE by approximately 3.6%. In contrast, the configurations $K \in \{8, 9, 12, 16\}$ form a broad plateau: their MSE values span only 0.3593–0.3605, a range of approximately 0.33% relative to the best value. The default $K = 9$ therefore lies near the knee of the curve, providing sufficient spectral coverage without requiring a larger retained set.

G.3 Sensitivity to Auxiliary-Loss Weights

The final default uses $\lambda_p = 0.1$ and $\lambda_a = 0.1$. To examine the interaction between the phase-consistency and alignment objectives, we sweep both coefficients over

$$\lambda_p, \lambda_a \in \{0, 0.05, 0.1, 0.2\} \tag{63}$$

on ETTh1 with $L = 96$ and $H = 96$, leaving the forecasting architecture unchanged. As in the other sensitivity runs, model selection follows the MAE-based checkpoint criterion.

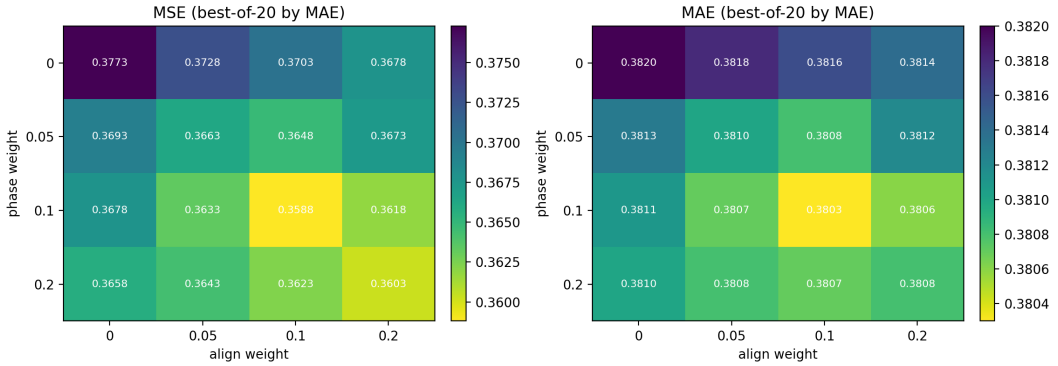


Figure 11: Sensitivity to auxiliary-loss weights on ETTh1 ($L = 96$, $H = 96$); left/right panels show MSE/MAE over (λ_p, λ_a).

The response surface exhibits a broad interior optimum rather than a monotonic preference for stronger regularization. With both auxiliary terms disabled, $(\lambda_p, \lambda_a) = (0, 0)$ yields MSE/MAE of $0.3773/0.3820$. Activating only the phase term at $(0.1, 0)$ improves MSE to 0.3678, while activating only alignment at $(0, 0.1)$ gives 0.3703. Their combination at the default $(0.1, 0.1)$ reaches $0.3588/0.3803$, reducing MSE by approximately 4.9% relative to prediction-only training. Increasing both coefficients to $(0.2, 0.2)$ slightly degrades MSE to 0.3603, indicating a stable sweet spot around the default rather than continued gains from heavier auxiliary regularization.

G.4 Sensitivity to the Fixed Positional Fourier Basis

The deterministic future positional basis is intentionally input-independent. To verify that the forecasting gains do not depend on one arbitrary frequency range, we compare removing the fixed basis entirely with logarithmic ranges $[1, 10]$, $[1, 50]$, and $[1, 100]$, while leaving the adaptive phase trajectory unchanged. Because this basis is deterministic and non-learnable, changing its frequency range does not alter the learned parameter count.

Table 22: Sensitivity to the fixed decoder Fourier basis on ETTh1 ($L = 96$, $H = 96$).

Fixed basis	Best ep.	MSE	MAE	Rel. MSE increase (%)
None	15	0.3710	0.3834	$+3.3\%$
Log-spaced $[1, 10]$	10	0.3640	0.3819	$+1.4\%$
Log-spaced $[1, 50]$	9	0.3593	0.3806	$+0.08\%$
Log-spaced $[1, 100]$ (default)	11	**0.3590**	**0.3804**	–

Removing the deterministic basis increases MSE by approximately 3.3%, whereas narrowing the range to $[1, 10]$ produces a smaller but visible 1.4% degradation. The $[1, 50]$ and $[1, 100]$ settings are effectively indistinguishable (0.3593 versus 0.3590 MSE), showing that the method is not sensitive to the precise upper frequency once the fixed basis spans a sufficient range.

G.5 Auxiliary-Objective Attribution and Representation Dispersion

To separate the contribution of the two training-only regularizers from a possible representation-collapse effect, we compare prediction-only training, phase-only regularization, alignment-only regularization, the default full objective, and a heavier $(0.2, 0.2)$ setting. In addition to forecasting error, we summarize the dispersion of the historical representations. For a batch of centered representations $\{\mathbf{r}_b\}_{b=1}^{B}$, define the mean feature variance

$$V_r = \frac{1}{D} \sum_{d=1}^{D} \mathrm{Var}_b(r_{b,d}), \tag{64}$$

and the effective rank of the covariance matrix Σ,

$$r_{\mathrm{eff}} = \exp\left(- \sum_j p_j \log(p_j + \epsilon) \right), \qquad p_j = \frac{\lambda_j(\Sigma)}{\sum_i \lambda_i(\Sigma) + \epsilon}. \tag{65}$$

Table 23: Auxiliary-objective attribution and representation-dispersion diagnostics on ETTh1 ($L = 96$, $H = 96$).

Variant	λ_p	λ_a	MSE	MAE	V_r	r_{eff}
Prediction only	0	0	0.3773	0.3820	0.9578	2.69
+ phase	0.1	0	0.3678	0.3811	0.9660	3.09
+ alignment	0	0.1	0.3703	0.3816	0.9728	3.15
Full (default)	0.1	0.1	**0.3588**	**0.3803**	0.9664	3.12
Heavy	0.2	0.2	0.3603	0.3808	0.9620	3.78

Both auxiliary terms improve forecasting when introduced individually, and their combination yields the lowest error, supporting a complementary rather than redundant role. The representation diagnostics do not show the low- variance collapse that could trivially reduce an alignment loss: V_r remains close to the prediction-only value and the effective rank increases from 2.69 to 3.12 under the default full objective. The heavier setting further raises effective rank but does not improve forecasting, reinforcing that the benefit comes from a balanced auxiliary objective rather than simply maximizing representation dispersion.