TIMESELECT: VARIABLE-WISE HISTORICAL CONTEXT SELECTION FOR TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Most time series forecasting methods operate on a fixed-length look-back window and implicitly treat all historical observations as equally informative. In practice, however, the informative context varies across forecasting instances and variables, and distant observations may introduce outdated or irrelevant patterns. To address this issue, we propose **TIMESELECT**, a forecasting framework that adaptively selects a contiguous, variable-specific historical context within a predefined maximum look-back window. Prior to training, **TIMESELECT** derives temporal bases from lagged correlations in the training data to capture stable dependencies across temporal positions. Given an input window, a lightweight selector estimates an input-conditioned context length for each variable and constructs a differentiable mask that suppresses observations outside the selected context. The retained history is then projected onto the correlation-derived bases and fed into a learnable backbone, which models inter-variable dependencies and generates the forecast. Because global modeling tends to smooth out recent periodic variations, a period-aware refinement module further corrects the initial forecast using gated residuals derived from the most recent observed cycles. Extensive experiments on diverse real-world benchmarks demonstrate that **TIMESELECT** consistently outperforms state-of-the-art baselines. The code is available in an anonymous repository.

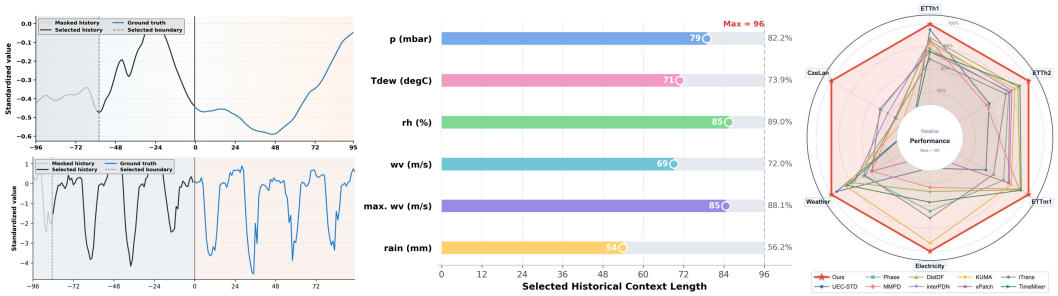
1 INTRODUCTION

Time-series forecasting plays an important role in energy management, traffic analysis, weather prediction, finance, and industrial monitoring (Liu et al., 2025; 2022; Liang et al., 2024). Recent studies have substantially improved forecasting accuracy through Transformer architectures, patch-based representations, multiscale modeling, and lightweight linear predictors (Zhou et al., 2021; Zeng et al., 2023b; Wang et al., 2026d; Jin et al., 2024; Cao et al., 2024). Despite their architectural differences, most existing methods are trained with a globally fixed look-back window and implicitly allow every observation within this window to influence the prediction.

However, the effective context required for forecasting can vary substantially across variables and input instances, even within the same predefined historical window. Such uniform treatment overlooks the fact that observations at different temporal distances do not contribute equally to forecasting. Distant observations may reflect outdated regimes, obsolete trends, or patterns inconsistent with the current forecasting state, thereby introducing irrelevant information. More importantly, as illustrated by the Weather dataset in Figure 1(a), variables with heterogeneous temporal dynamics exhibit substantially different effective historical ranges. Slowly evolving variables may require longer contexts to capture persistent trends, whereas rapidly changing variables often depend primarily on recent observations. Therefore, imposing the same historical context on all variables may obscure their variable-specific predictive patterns and ultimately limit forecasting accuracy.

Existing approaches do not explicitly resolve this issue. Multiscale temporal models (Wang et al., 2024; Stitsyuk & Choi, 2025a; Hu et al., 2025a) represent the input at different temporal resolutions, but they generally aggregate information from all scales rather than determining how much recent history is useful for each variable. Attention mechanisms (Liu et al., 2024a; Zhou et al., 2022a; Ma et al., 2026d) can assign different weights to temporal positions, but their weights are usually learned for feature aggregation and do not directly define an interpretable, contiguous context range anchored at the forecasting boundary.

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) Not all input is useful. (b) Variable-specific input selection. (c) Competitive performance.

Figure 1: Motivation and supporting results.

To address this challenge, we propose **TIMESELECT**, a forecasting framework that performs variable-wise effective-window selection within a predefined maximum historical range. For each forecasting instance, a lightweight selector estimates a personalized context length for each variable and masks out observations beyond the selected boundary, producing a contiguous recent context anchored at the forecasting boundary. Figure 1(b) visualizes the effective input contexts selected for several variables on the Weather dataset under the input-96/prediction-96 setting. However, adaptive context selection may cause the effective historical information to vary across variables and instances, making it difficult to learn a consistent mapping from history to future. To mitigate this issue, we construct correlation-based temporal bases from the training data, which organize the selected historical information into a shared and stable temporal representation. Moreover, selective context modeling and global temporal mapping may attenuate recent periodic fluctuations near the forecasting boundary. We therefore introduce a period-aware refinement module that exploits the most recently observed cycles to compensate for such local deviations. Extensive experiments on a wide range of real-world datasets demonstrate the effectiveness of **TIMESELECT**, and some of the results are shown in Figure 1(c).

Our main **contributions** are summarized as follows:

- **Problem Perspective.** We revisit the uniform-window utilization assumption in multivariate forecasting and distinguish the nominal input length from the effective historical context. We argue that useful context is both instance-dependent and variable-specific, and that not all observations within a fixed input window should contribute equally to prediction.
- **Novel Methodology.** We propose **TIMESELECT**, which predicts a variable-wise, input-conditioned context boundary and masks out observations outside the selected range. Correlation-derived temporal bases and cycle-aware refinement further capture stable and recent temporal patterns.
- **Empirical Evaluation.** Extensive experiments on 13 real-world datasets demonstrate the superior forecasting performance of **TIMESELECT**, supported by comprehensive ablation and context-selection analyses.

2 RELATED WORK

The growing demand for accurate prediction in energy, transportation, finance, and environmental monitoring has made time series forecasting an active research area (Miller et al., 2024; Guo et al., 2026; Jin et al., 2024). Recent progress has largely focused on representation learning, with different architectures designed to capture long-range dependencies, trends, periodicity, multiscale patterns, and inter-variable interactions (Qiu et al., 2024; Fang et al., 2026). Transformer-based models facilitate global temporal interaction, while decomposition, frequency-domain analysis, and patch-based modeling improve the representation of complex temporal structures (Zhang et al., 2025; Qin et al., 2026; Yang et al., 2026; Nematirad et al., 2025; Wang et al., 2026a). Meanwhile, lightweight MLP-based approaches demonstrate that carefully designed feature transformations can achieve competitive accuracy with substantially lower computational cost (Lin et al., 2025; Tang & Zhang, 2025; Murad et al., 2025; Yi et al., 2023). Beyond improving representations over fixed inputs, ALW (Tong & Yuan, 2026) learns scale-specific look-back windows through wavelet decomposition and attention. However, ALW shares each selected window across variables, whereas **TIMESELECT** pre-

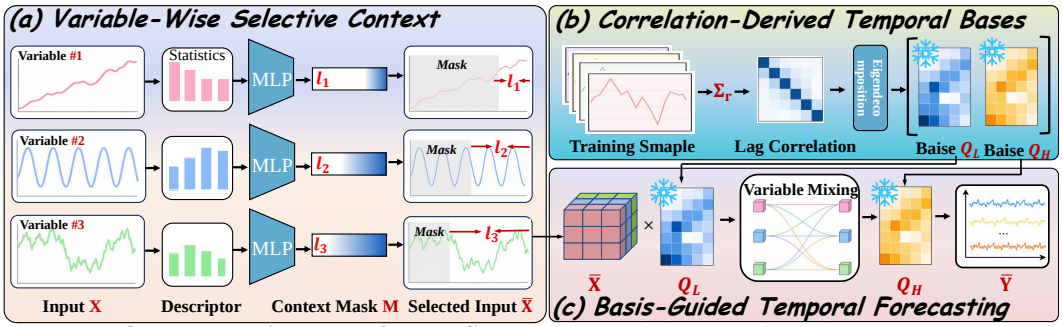


Figure 2: **Overall architecture of TIMESELECT.** It mainly comprises three components: (a) variable-wise context selection, (b) offline correlation-derived temporal basis construction, and (c) basis-guided forecasting for final prediction.

dicts an input-conditioned context length independently for each variable, thereby capturing both instance-specific states and heterogeneous variable dynamics.

Beyond these mainstream architectures, state-space models improve long-sequence efficiency (Wu et al., 2025), KAN-based networks enhance nonlinear modeling (Han et al., 2024), and pretrained language models have been adapted to numerical time series (Jin et al., 2023; Qiu et al., 2026). Multimodal learning and diffusion-based methods further broaden the representation and generation capabilities of forecasting models (Zhong et al., 2025; Mansouri et al., 2025; Tashiro et al., 2021).

3 METHOD

3.1 PROBLEM FORMULATION

Let $\mathbf{x}_t \in \mathbb{R}^N$ denote the observation of an N-variate time series at time step t. Given a look-back window of length L, the historical input ending at time t is defined as $\mathbf{X} = [\mathbf{x}_{t-L+1}, \ldots, \mathbf{x}_t]^\top \in \mathbb{R}^{L\times N}$ The forecasting target consists of the subsequent H observations denoted as $\mathbf{Y} = [\mathbf{x}_{t+1}, \ldots, \mathbf{x}_{t+H}]^\top \in \mathbb{R}^{H\times N}$, where H denotes the prediction horizon.

3.2 OVERALL ARCHITECTURE

As shown in Figure 2 and Algorithm 1, **TIMESELECT** consists of three key modules. First, two correlation-derived temporal bases, $\mathbf{Q}_L \in \mathbb{R}^{L\times L}$ and $\mathbf{Q}_H \in \mathbb{R}^{H\times H}$, are estimated from the training split and kept fixed to encode stable dataset-level temporal structures. Given an input sequence $\mathbf{X} \in \mathbb{R}^{L\times N}$, the variable-wise context selector estimates the relevance of different temporal positions and generates a soft mask $\mathbf{M} \in \mathbb{R}^{L\times N}$ to attenuate less informative observations, producing the selected history $\widetilde{\mathbf{X}} \in \mathbb{R}^{L\times N}$. The forecasting backbone then maps $\widetilde{\mathbf{X}}$ to an initial prediction $\widetilde{\mathbf{Y}}_{\text{base}} \in \mathbb{R}^{H\times N}$ using the input and output temporal bases. Finally, the cycle-aware refinement module incorporates the most recent periodic pattern to correct local deviations in the initial prediction and produces the final forecast $\widehat{\mathbf{Y}} \in \mathbb{R}^{H\times N}$.

3.3 CORRELATION-DERIVED TEMPORAL BASES

We first construct temporal bases that provide structured coordinates for representing both historical and future sequences. Rather than adopting a predefined transform, the bases are estimated from temporal correlations in the training data. For a temporal length $r \in \{L, H\}$, each variable is processed independently.

Lag-Correlation Estimation Given a training sequence $a_{1:T}^{(n)}$ for variable n, we handle invalid values, remove variables with insufficient variation, and standardize each retained series. We then collect all overlapping subsequences into a lagged matrix:

$$[\mathbf{Z}_r^{(n)}]_{i,j} = a_{i+j-1}^{(n)}, \qquad 1 \le i \le r, \quad 1 \le j \le T - r + 1. \tag{1}$$

Here, $\mathbf{Z}_r^{(n)} \in \mathbb{R}^{r\times(T-r+1)}$, where each row corresponds to a relative temporal position and each column represents an overlapping training window.

After centering each row, we calculate its covariance and convert it into a Pearson correlation matrix $\mathbf{C}_r^{(n)} \in \mathbb{R}^{r\times r}$. To reduce variable-specific noise, the correlation matrices are averaged over the valid variable set $\mathcal{V}$ and regularized through diagonal shrinkage:

$$\mathbf{\Sigma}_r = (1-\rho)\frac{1}{|\mathcal{V}|}\sum_{n\in\mathcal{V}}\mathbf{C}_r^{(n)} + \rho\mathbf{I}_r. \tag{2}$$

where $\mathbf{\Sigma}_r \in \mathbb{R}^{r\times r}$ captures the average correlation between temporal positions within a length-r window across all valid variables. $\mathbf{I}_r \in \mathbb{R}^{r\times r}$ is the identity matrix, and $\rho \in [0,1]$ controls the shrinkage strength.

Temporal Basis Extraction Since $\mathbf{\Sigma}_r$ is symmetric, we extract its principal temporal directions through eigendecomposition:

$$\mathbf{\Sigma}_r = \mathbf{U}_r\mathbf{\Lambda}_r\mathbf{U}_r^\top, \qquad \mathbf{Q}_r = \mathbf{U}_r^\top. \tag{3}$$

where $\mathbf{U}_r$ contains the temporal eigenvectors and $\mathbf{\Lambda}_r$ contains their corresponding eigenvalues. The temporal basis is defined as $\mathbf{Q}_r = \mathbf{U}_r^\top$. The eigenvectors are ordered by decreasing eigenvalue, such that the leading basis directions correspond to the strongest temporal correlation patterns. SVD-based projection is additionally used to remove minor numerical deviations from orthogonality.

Following this procedure, we construct $\mathbf{Q}_L \in \mathbb{R}^{L\times L}$ for the historical window and $\mathbf{Q}_H \in \mathbb{R}^{H\times H}$ for the prediction horizon. Since the full bases are retained, the transformations reorganize temporal information without reducing its dimensionality. The following proposition shows that the basis transformation preserves information while decoupling temporal correlations.

Proposition 1 (Properties of the temporal basis)**.** *Let* $\mathbf{\Sigma}_r = \mathbf{U}_r\mathbf{\Lambda}_r\mathbf{U}_r^\top$ *and* $\mathbf{Q}_r = \mathbf{U}_r^\top$*. Then* $\mathbf{Q}_r$ *is orthogonal and, for any* $\mathbf{A} \in \mathbb{R}^{m\times r}$*,*

$$\left\|\mathbf{A}\mathbf{Q}_r^\top\right\|_F = \|\mathbf{A}\|_F, \qquad \mathbf{A} = \left(\mathbf{A}\mathbf{Q}_r^\top\right)\mathbf{Q}_r. \tag{4}$$

Moreover, the transformed temporal correlation matrix is diagonal:

$$\mathbf{Q}_r\mathbf{\Sigma}_r\mathbf{Q}_r^\top = \mathbf{\Lambda}_r. \tag{5}$$

Proof. Since $\mathbf{\Sigma}_r$ is symmetric, its eigenvectors form an orthonormal matrix, and therefore $\mathbf{Q}_r^\top\mathbf{Q}_r = \mathbf{Q}_r\mathbf{Q}_r^\top = \mathbf{I}_r$. The Frobenius norm is consequently preserved:

$$\left\|\mathbf{A}\mathbf{Q}_r^\top\right\|_F^2 = \mathrm{tr}\left(\mathbf{A}\mathbf{Q}_r^\top\mathbf{Q}_r\mathbf{A}^\top\right) = \mathrm{tr}\left(\mathbf{A}\mathbf{A}^\top\right) = \|\mathbf{A}\|_F^2 \tag{6}$$

Orthogonality also gives $(\mathbf{A}\mathbf{Q}_r^\top)\mathbf{Q}_r = \mathbf{A}$, which proves invertibility. Finally,

$$\mathbf{Q}_r\mathbf{\Sigma}_r\mathbf{Q}_r^\top = \mathbf{U}_r^\top\left(\mathbf{U}_r\mathbf{\Lambda}_r\mathbf{U}_r^\top\right)\mathbf{U}_r = \mathbf{\Lambda}_r. \tag{7}$$

Proposition 1 shows that the proposed transformation does not discard temporal information. Instead, it expresses the sequence in coordinates aligned with the dominant correlation patterns of the training data, while diagonalizing the training-derived temporal correlation matrix in the transformed coordinate system. □

3.4 Variable-Wise Selective Context

The predefined historical window may contain observations that are no longer relevant to the current forecasting state. We therefore introduce a differentiable context selector that determines a contiguous effective context for each variable The input is first normalized using reversible instance normalization (Kim et al., 2021), yielding $\overline{\mathbf{X}} = \mathrm{RevIN}(\mathbf{X}) \in \mathbb{R}^{L\times N}$. For each variable n, the resulting feature sequence is summarized using statistics that describe its level, variability, recent state, overall change, and activation magnitude. The resulting context descriptor is denoted by $\mathbf{s}_n \in \mathbb{R}^5$. Then we use an MLP to map $\mathbf{s}_n$ to a continuous effective context length:

$$\ell_n = \ell_{\min} + (L-\ell_{\min})\sigma(\mathrm{MLP}(\mathbf{s}_n)), \qquad \ell_n \in [\ell_{\min}, L]. \tag{8}$$

The scalar ℓ_n determines the effective historical range of variable n, while the lower bound $\ell_{\min}$ ensures that a minimum amount of recent information is always retained. Directly truncating the sequence according to ℓ_n would make the selection operation non-differentiable. We instead construct a temporal mask with and apply it to the normalized history:

$$M_{t,n} = \sigma\left(\frac{\ell_n - (L-t+1)}{\tau}\right), \qquad \widetilde{\mathbf{X}} = \mathbf{M} \odot \overline{\mathbf{X}}. \tag{9}$$

where $\mathbf{M} \in \mathbb{R}^{L\times N}$ is the variable-wise context mask, $\widetilde{\mathbf{X}} \in \mathbb{R}^{L\times N}$ is the selected history input, τ controls the smoothness of the context boundary. The mask assigns large weights to observations within the estimated context and gradually suppresses older positions.

3.5 Basis-Guided Temporal Forecasting

After selecting the effective historical context, the model maps it to the future horizon using the correlation-derived temporal bases. This stage combines stable dataset-level temporal coordinates with learnable historical-to-future mapping and cross-variable interaction.

Latent Lifting and Input-Basis Transformation We first lift each selected scalar observation into a D-dimensional latent space using a shared learnable vector $\mathbf{e} \in \mathbb{R}^D$:

$$E_{n,t,d} = \widetilde{X}_{t,n} e_d \in \mathbb{R}^{N\times L\times D}, \tag{10}$$

This multiplicative lifting preserves the magnitude of the original observation while introducing multiple latent dimensions for subsequent representation learning. After rearranging the latent representation as $\mathbf{E} \in \mathbb{R}^{N\times D\times L}$, we transform its temporal dimension using the input basis $\mathbf{Q}_L$:

$$\mathbf{S}_{n,d,:} = \mathbf{E}_{n,d,:}\mathbf{Q}_L^\top + \delta_{\text{in},n,d,:} \in \mathbb{R}^{N\times D\times L}. \tag{11}$$

The fixed basis reorganizes the selected history according to the data-aligned temporal coordinates estimated from the training data. The learnable residual adjustment δ_{in} provides additional flexibility to compensate for task-specific deviations from these dataset-level coordinates.

Cross-Variable Representation Learning For each variable, the transformed temporal and latent dimensions are first flattened and projected into a shared hidden space:

$$\mathbf{H}_{n,:} = \phi_{\text{in}}\left(\text{vec}\left(\mathbf{S}_{n,:,:}\right)\right), \qquad \mathbf{H} \in \mathbb{R}^{N\times D_h}, \tag{12}$$

where $\phi_{\text{in}} : \mathbb{R}^{DL} \to \mathbb{R}^{D_h}$ is a learnable linear projection and D_h is the hidden dimension. $\text{vec}(\cdot)$ flattens the temporal and latent dimensions into a single vector. Then, a stack of lightweight encoder blocks then captures dependencies among variables. Each block learns a nonnegative, row-normalized mixing matrix and uses it to aggregate information across variable representations:

$$\begin{aligned}
\mathbf{A} &= \text{RowNorm}_1\left(\text{softplus}\left(\mathbf{A}_\theta\right)\right) \in \mathbb{R}^{N\times N}, \\
\mathbf{H}' &= \text{LN}\left(\mathbf{H} + \mathbf{A}\left(\mathbf{H}\mathbf{W}_V\right)\mathbf{W}_O\right)\mathbb{R}^{N\times D_h}, \\
\mathbf{H}^+ &= \text{LN}\left(\mathbf{H}' + \text{FFN}\left(\mathbf{H}'\right)\right)\mathbb{R}^{N\times D_h}.
\end{aligned} \tag{13}$$

Here, $\mathbf{A}_\theta \in \mathbb{R}^{N\times N}$ is a learnable unnormalized variable-interaction matrix. The softplus function constrains its entries to be nonnegative, while $\text{RowNorm}_1(\cdot)$ normalizes each row to sum to one. Consequently, $\mathbf{A} \in \mathbb{R}^{N\times N}$ represents the normalized contribution of different variables. The matrices $\mathbf{W}_V \in \mathbb{R}^{D_h\times D_v}$ and $\mathbf{W}_O \in \mathbb{R}^{D_v\times D_h}$ are the learnable value and output projections, respectively. $\text{LN}(\cdot)$ denotes layer normalization along the feature dimension, and $\text{FFN}(\cdot)$ denotes a feed-forward network shared across variables.

Future-Basis Reconstruction For each variable n, the encoded representation $\mathbf{H}^+_{n,:} \in \mathbb{R}^{D_h}$ is projected into a DH-dimensional vector and reshaped into $\mathbf{G}_{n,:,:} \in \mathbb{R}^{D\times H}$:

$$\begin{aligned}
\mathbf{G}_{n,:,:} &= \text{reshape}_{D\times H}\left(\mathcal{F}_\theta\left(\mathbf{H}^+_{n,:}\right)\right) \in \mathbb{R}^{N\times D\times H}, \\
\mathbf{O}_{n,d,:} &= \mathbf{G}_{n,d,:}\mathbf{Q}_H + \boldsymbol{\delta}_{\text{out},n,d,:} \in \mathbb{R}^{N\times D\times H}, \\
\widetilde{\mathbf{Y}}_{\text{base},:,n} &= \mathcal{P}_\theta\left(\text{vec}\left(\mathbf{O}_{n,:,:}\right)\right) \in \mathbb{R}^{H\times N},
\end{aligned} \tag{14}$$

where $\mathcal{F}_\theta : \mathbb{R}^{D_h} \to \mathbb{R}^{DH}$ is a learnable future-projection layer. . The prediction head $\mathcal{P}_\theta : \mathbb{R}^{DH} \to \mathbb{R}^H$ jointly maps the reconstructed future latent representation of each variable to its complete normalized forecast. Applying this head to all variables produces the prediction $\widetilde{\mathbf{Y}}_{\text{base}} \in \mathbb{R}^{H\times N}$.

3.6 Forecaster with Cycle Refinement

The basis-guided predictor captures the dominant temporal structure but may smooth recent fluctuations or misalign with the latest periodic phase. We therefore introduce a cycle-aware refinement module that uses the most recent cycle as a local forecasting reference. Specifically, let c denote the cycle length. We extract the latest $\min(c, L)$ observations from the normalized input and repeat them until the prediction horizon is covered, producing a cycle anchor $\mathbf{R} \in \mathbb{R}^{H\times N}$. Then, **TimeSelect** combines it with the base prediction and estimates a residual correction:

$$\overline{\mathbf{Y}} = \widetilde{\mathbf{Y}}_{\text{base}} + \alpha_{\text{cyc}}\, h_\psi\left(\left[\widetilde{\mathbf{Y}}_{\text{base}}, \mathbf{R}\right]\right), \qquad \alpha_{\text{cyc}} = \sigma(a_{\text{cyc}}). \tag{15}$$

The concatenated representation lies in $\mathbb{R}^{H\times 2N}$, while both the residual correction produced by h_ψ and the refined normalized forecast $\overline{\mathbf{Y}}$ lie in $\mathbb{R}^{H\times N}$. The refined prediction is finally mapped back to the original scale through inverse RevIN: $\overline{\mathbf{Y}} = \text{RevIN}^{-1}(\overline{\mathbf{Y}})$, where $\overline{\mathbf{Y}} \in \mathbb{R}^{H\times N}$.

3.7 Training Objective

Let $\rho(\cdot)$ denote the point-wise prediction loss function, such as L1 loss function. To account for the varying importance of different forecasting positions, we assign the h-th horizon a decay weight $h^{-\alpha}$, where α controls the rate at which the weight decreases with the forecasting horizon (Zhou et al., 2024). The objective for one forecasting instance is defined as

$$\mathcal{L} = \frac{1}{HN}\sum_{h=1}^{H}\sum_{n=1}^{N} h^{-\alpha}\rho\left(\widehat{\mathbf{Y}}_{h,n} - \mathbf{Y}_{h,n}\right) + \lambda_{\text{ctx}}\frac{1}{NL}\sum_{n=1}^{N}\ell_n. \tag{16}$$

The second penalizes the average normalized context length, encouraging the selector to discard less informative history instead of retaining the full input window. The coefficient λ_{ctx} balances forecasting accuracy and context compactness.

4 Experiment

4.1 Experiment Setup

Dataset We evaluate the proposed method on 13 widely used real-world datasets covering both short-term and long-term forecasting scenarios. For long-term forecasting, we consider the ETT series (four subsets) (Zhou et al., 2021), Electricity, Weather (Wu et al., 2021), Exchange, AQWan, CzeLan, and ZafNoo (Qiu et al., 2024), with H set to $\{96, 192, 336, 720\}$. For short-term forecasting, we use Air Quality, Mobile Traffic (Barlacchi et al., 2015), and ILI, with the prediction horizon H set to $\{12, 24, 48\}$. The look-back window L is fixed to 96 for all datasets. These datasets cover diverse application domains and vary substantially in temporal scale, data size, and number of variables. Detailed dataset statistics are summarized in Table 1.

Implementation All experiments were run on an NVIDIA 4090D 24G with PyTorch framework. The model name was optimized using the ADAM optimizer (Zhang, 2018). Training was performed for up to 50 epochs with early stopping, which halted training if the validation performance did not improve for 10 consecutive epochs. The temporal bases $\mathbf{Q}_L$ and $\mathbf{Q}_H$ are constructed using only the training split and remain fixed throughout model optimization. The diagonal shrinkage coefficient in the correlation estimator ρ is set to 0.05. For the variable-wise context selector, the minimum effective context length $\ell_{\min}$ is set to 24. The context-mask temperature τ is set to 12. For the cycle-aware refinement module, the initial value of the cycle-refinement gate α_{cyc} is set to 0.03.

Baselines We select **13** time series forecasting baselines, including DLinear (Zeng et al., 2023a), TimeMixer (Wang et al., 2024), PatchTST (Nie et al., 2023), iTransformer (Liu et al., 2024a), xPatch(Stitsyuk & Choi, 2025a), WPMixer (Murad et al., 2025), ALW (Tong & Yuan, 2026), interPDN (Kong & Hong, 2026b), DistDF (Wang et al., 2026c), MixLinear (Ma et al., 2026b), PhaseFormer (Niu et al., 2026), MMPD (Zhang et al., 2026b), and UEC-STD (Nguyen et al., 2026).

Table 1: Multivariate time series forecasting results. The look-back length L is fixed at 96, and "*" denotes former methods for clarity. The 1st and 2nd best results are highlighted.

Model		**TimeSelect (Ours)**		UEC-STD (2026)		Phase* (2026)		MMPD (2026)		MixLinear (2026)		DistDF (2026)		interPDN (2026)		ALW (2026)		WPMixer (2025)		xPatch (2025)		iTrans* (2024)		PatchTST (2023)		TimeMixer (2023)		DLinear (2023)	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.363**	**0.384**	0.387	0.423	0.383	0.386	0.376	0.393	0.389	0.417	0.373	0.393	0.375	0.396	0.406	0.429	0.386	0.397	0.383	0.401	0.386	0.405	0.414	0.419	0.378	0.400	0.386	0.400
	192	**0.414**	**0.413**	0.438	0.442	0.444	0.418	0.425	0.424	0.434	0.446	0.428	0.425	0.429	0.416	0.469	0.470	0.436	0.431	0.431	0.427	0.441	0.436	0.460	0.445	0.432	0.429	0.437	0.432
	336	**0.456**	0.437	0.468	0.475	0.474	**0.436**	0.467	0.448	0.489	0.464	0.472	0.449	0.472	0.449	0.529	0.503	0.487	0.451	0.483	0.455	0.487	0.458	0.501	0.466	0.481	0.451	0.481	0.459
	720	0.461	0.462	0.484	0.483	0.469	**0.450**	0.472	0.461	0.495	0.487	**0.453**	0.453	0.462	0.451	0.545	0.538	0.511	0.485	0.486	0.474	0.503	0.491	0.500	0.488	0.493	0.477	0.519	0.516
	Avg	**0.424**	0.424	0.444	0.456	0.442	**0.423**	0.435	0.432	0.452	0.454	0.432	0.430	0.435	0.428	0.487	0.485	0.455	0.441	0.446	0.439	0.454	0.448	0.469	0.455	0.446	0.439	0.456	0.452
ETTh2	96	0.286	**0.330**	0.309	0.363	0.293	0.337	0.322	0.364	0.337	0.393	0.287	0.336	0.310	0.345	0.312	0.367	0.299	0.350	0.291	0.343	0.297	0.349	0.302	0.348	**0.284**	0.331	0.333	0.387
	192	**0.364**	**0.381**	0.395	0.415	0.380	0.388	0.416	0.422	0.431	0.446	0.373	0.388	0.367	0.383	0.403	0.428	0.371	0.395	0.372	0.392	0.380	0.400	0.388	0.400	0.364	0.382	0.477	0.476
	336	0.408	**0.415**	0.456	0.455	0.426	0.424	0.436	0.442	0.475	0.460	0.408	0.421	0.413	0.420	0.463	0.475	0.426	0.433	0.423	0.433	0.428	0.432	0.426	0.433	**0.406**	0.416	0.594	0.541
	720	**0.415**	**0.431**	0.536	0.519	0.419	0.437	0.440	0.454	0.467	0.487	0.416	0.435	0.417	0.434	0.479	0.499	0.433	0.449	0.430	0.451	0.427	0.445	0.431	0.446	0.423	0.436	0.831	0.657
	Avg	**0.368**	**0.391**	0.424	0.438	0.380	0.397	0.404	0.421	0.428	0.447	0.371	0.395	0.377	0.396	0.414	0.442	0.382	0.407	0.379	0.405	0.383	0.407	0.387	0.407	0.369	**0.391**	0.559	0.515
ETTm1	96	**0.303**	**0.334**	0.340	0.385	0.326	0.344	0.313	0.350	0.374	0.419	0.316	0.357	0.324	0.347	0.352	0.388	0.329	0.364	0.332	0.369	0.334	0.368	0.329	0.367	0.315	0.343	0.345	0.372
	192	**0.356**	**0.364**	0.390	0.418	0.398	0.376	0.358	0.376	0.403	0.429	0.358	0.380	0.367	0.383	0.402	0.419	0.383	0.395	0.371	0.387	0.377	0.391	0.367	0.385	0.358	0.380	0.380	0.389
	336	0.388	**0.386**	0.454	0.455	0.425	0.401	0.400	0.404	0.470	0.443	0.392	0.404	0.409	0.417	0.449	0.454	0.400	0.408	0.399	0.407	0.426	0.420	0.399	0.410	**0.385**	0.400	0.413	0.413
	720	0.449	**0.424**	0.571	0.513	**0.436**	**0.424**	0.450	0.438	0.530	0.477	0.448	0.437	0.461	0.434	0.530	0.504	0.487	0.453	0.463	0.443	0.491	0.459	0.454	0.439	0.455	0.443	0.474	0.453
	Avg	**0.374**	**0.377**	0.439	0.443	0.396	0.386	0.380	0.392	0.444	0.442	0.379	0.395	0.390	0.395	0.433	0.441	0.400	0.405	0.391	0.402	0.407	0.410	0.387	0.400	0.378	0.392	0.403	0.407
ETTm2	96	**0.169**	**0.249**	0.179	0.266	0.179	0.258	0.175	0.260	0.207	0.326	0.174	0.256	0.173	0.250	0.187	0.274	0.175	0.257	0.177	0.262	0.180	0.264	0.175	0.259	0.178	0.260	0.193	0.292
	192	**0.233**	**0.291**	0.242	0.315	**0.233**	0.296	0.242	0.305	0.311	0.386	0.239	0.298	0.237	0.293	0.262	0.327	0.238	0.297	0.244	0.306	0.250	0.309	0.241	0.302	0.246	0.306	0.284	0.362
	336	**0.292**	**0.329**	0.307	0.362	0.310	0.340	0.308	0.349	0.417	0.462	0.300	0.338	**0.292**	0.332	0.331	0.375	0.305	0.342	0.304	0.342	0.311	0.348	0.305	0.343	0.296	0.339	0.369	0.427
	720	**0.391**	**0.389**	0.419	0.432	0.392	0.398	0.403	0.403	0.583	0.596	0.397	0.394	0.394	0.391	0.449	0.447	0.402	0.400	0.392	0.392	0.404	0.399	0.412	0.407	0.402	0.400	0.396	0.554	0.522
	Avg	**0.271**	**0.314**	0.287	0.344	0.279	0.323	0.282	0.329	0.380	0.443	0.278	0.322	0.274	0.317	0.307	0.356	0.278	0.322	0.282	0.327	0.288	0.332	0.281	0.326	0.279	0.325	0.350	0.401
Electricity	96	**0.141**	**0.232**	0.166	0.262	0.154	0.246	0.161	0.457	0.214	0.314	0.167	0.265	0.845	0.761	0.150	0.252	0.165	0.259	0.191	0.277	0.148	0.240	0.181	0.270	0.156	0.248	0.197	0.282
	192	**0.157**	**0.246**	0.185	0.283	0.167	0.263	0.180	0.277	0.210	0.313	0.179	0.275	0.166	0.266	0.178	0.281	0.177	0.266	0.193	0.278	0.162	0.253	0.188	0.274	0.170	0.259	0.196	0.285
	336	0.179	**0.266**	0.214	0.310	0.183	0.283	0.194	0.295	0.237	0.320	0.203	0.287	0.183	0.279	0.195	0.301	0.196	0.286	0.206	0.291	**0.178**	0.269	0.204	0.293	0.186	0.276	0.209	0.301
	720	**0.204**	**0.289**	0.286	0.369	0.221	0.309	0.233	0.324	0.274	0.350	0.212	0.302	0.236	0.314	0.222	0.328	0.238	0.327	0.244	0.321	0.225	0.317	0.246	0.324	0.228	0.312	0.245	0.333
	Avg	**0.170**	**0.258**	0.213	0.306	0.181	0.275	0.192	0.338	0.234	0.324	0.190	0.282	0.358	0.405	0.186	0.291	0.194	0.285	0.209	0.292	0.178	0.270	0.205	0.290	0.185	0.274	0.212	0.300
Weather	96	**0.151**	**0.190**	0.163	0.219	0.172	0.204	0.183	0.222	0.213	0.286	0.164	0.209	0.163	0.197	0.170	0.218	0.169	0.215	0.178	0.221	0.174	0.214	0.177	0.210	0.163	0.210	0.196	0.255
	192	**0.200**	**0.235**	0.215	0.269	0.219	0.246	0.232	0.267	0.252	0.326	0.212	0.252	0.219	0.247	0.229	0.273	0.216	0.257	0.231	0.263	0.221	0.254	0.225	0.250	0.208	0.251	0.237	0.296
	336	**0.262**	**0.282**	0.282	0.309	0.276	**0.288**	0.294	0.313	0.320	0.361	0.270	0.295	0.275	0.288	0.302	0.330	0.272	0.298	0.283	0.300	0.278	0.296	0.278	0.290	0.268	0.295	0.283	0.335
	720	**0.341**	**0.336**	0.365	0.357	0.362	0.346	0.344	0.337	0.386	0.401	0.348	0.345	0.345	**0.336**	0.400	0.395	0.352	0.348	0.360	0.350	0.358	0.349	0.354	0.340	0.346	0.348	0.345	0.381
	Avg	**0.242**	**0.263**	0.251	0.288	0.257	0.271	0.263	0.285	0.293	0.344	0.249	0.275	0.251	0.267	0.275	0.304	0.252	0.280	0.263	0.284	0.258	0.278	0.259	0.273	0.246	0.276	0.265	0.317
Exchange	96	**0.081**	0.199	0.088	0.221	0.086	0.205	0.087	0.205	0.095	0.243	0.089	0.221	0.085	0.203	0.091	0.219	0.086	0.206	**0.081**	**0.198**	0.107	0.234	0.092	0.218	0.085	0.218	0.088	0.218
	192	**0.173**	**0.294**	0.183	0.324	0.196	0.312	0.199	0.301	0.189	0.347	0.203	0.303	0.177	0.298	0.194	0.328	0.177	0.299	0.181	0.302	0.226	0.344	0.181	0.317	0.175	0.315	0.176	0.315
	336	0.328	**0.411**	0.351	0.446	0.356	0.432	0.337	0.440	0.355	0.456	0.376	0.515	0.341	0.421	0.351	0.453	0.331	0.417	0.342	0.423	0.367	0.448	0.324	0.431	0.323	0.427	**0.321**	0.427
	720	0.834	0.686	0.881	0.740	0.935	0.735	0.865	0.714	0.939	0.730	0.883	0.726	0.870	0.703	0.956	0.781	0.847	0.691	**0.823**	**0.684**	0.964	0.746	0.856	0.693	0.851	0.695	0.839	0.695
	Avg	**0.354**	**0.397**	0.376	0.433	0.393	0.421	0.372	0.415	0.395	0.444	0.388	0.441	0.368	0.406	0.398	0.445	0.360	0.403	0.357	0.402	0.416	0.443	0.363	0.415	0.359	0.414	0.356	0.414
AQWan	96	**0.783**	**0.464**	0.791	0.500	0.803	0.479	0.831	0.486	0.855	0.553	0.912	0.550	0.797	0.470	0.955	0.608	0.787	0.481	0.827	0.489	0.828	0.496	0.803	0.489	0.809	0.490	0.784	0.510
	192	**0.842**	**0.492**	0.885	0.506	0.878	0.506	0.875	0.516	0.899	0.602	0.963	0.563	0.863	0.495	1.215	0.604	0.858	0.505	0.871	0.517	0.885	0.514	0.865	0.513	0.877	0.515	0.843	0.535
	336	0.872	**0.504**	0.899	0.520	0.908	0.524	0.889	0.524	0.967	0.586	0.967	0.570	0.885	0.507	1.102	0.662	0.880	0.517	0.893	0.511	0.912	0.532	0.905	0.526	0.894	0.524	**0.865**	0.547
	720	**0.932**	**0.528**	0.965	0.536	0.951	0.546	0.957	0.539	1.065	0.592	0.982	0.736	0.947	0.531	1.204	0.750	0.937	0.539	0.948	0.542	0.957	0.546	0.950	0.546	0.951	0.547	**0.932**	0.563
	Avg	**0.856**	**0.497**	0.885	0.516	0.885	0.514	0.888	0.516	0.947	0.583	0.956	0.605	0.873	0.501	1.119	0.656	0.866	0.511	0.885	0.515	0.896	0.522	0.881	0.519	0.883	0.519	**0.856**	0.539
CzeLan	96	**0.189**	**0.228**	0.233	0.276	0.211	0.243	0.211	0.258	0.219	0.270	0.215	0.294	0.203	0.254	0.233	0.287	0.193	0.242	0.214	0.263	0.217	0.262	0.203	0.252	0.229	0.276	0.352	0.422
	192	**0.222**	**0.248**	0.257	0.290	0.237	0.258	0.252	0.289	0.257	0.304	0.253	0.296	0.235	0.283	0.288	0.319	0.224	0.264	0.241	0.286	0.246	0.287	0.235	0.276	0.262	0.297	0.471	0.475
	336	0.254	**0.271**	0.297	0.329	0.266	0.282	0.277	0.308	0.298	0.337	0.280	0.317	0.275	0.314	0.327	0.342	0.260	0.297	**0.233**	0.275	0.282	0.313	0.265	0.299	0.297	0.328	0.597	0.527
	720	**0.303**	**0.311**	0.359	0.377	0.328	0.323	0.346	0.358	0.376	0.391	0.343	0.351	0.352	0.359	0.391	0.412	0.306	0.327	0.351	0.343	0.341	0.354	0.324	0.340	0.358	0.371	0.803	0.604
	Avg	**0.242**	**0.265**	0.287	0.318	0.261	0.277	0.272	0.303	0.288	0.326	0.273	0.315	0.266	0.303	0.310	0.340	0.246	0.283	0.260	0.292	0.272	0.304	0.257	0.292	0.287	0.318	0.556	0.507
ZafNoo	96	**0.450**	**0.377**	0.468	0.427	0.465	0.385	0.476	0.419	0.492	0.482	0.786	0.650	0.458	0.387	0.543	0.459	0.453	0.402	0.456	0.413	0.473	0.416	0.454	0.409	0.484	0.425	0.454	0.433
	192	**0.520**	**0.423**	0.563	0.464	0.557	0.437	0.558	0.462	0.556	0.524	0.845	0.680	0.536	0.425	0.651	0.502	0.532	0.442	0.526	0.447	0.558	0.461	0.525	0.448	0.567	0.470	0.521	0.465
	336	**0.559**	**0.458**	0.608	0.504	0.617	0.471	0.626	0.501	0.628	0.517	0.608	0.488	0.599	0.464	0.710	0.534	0.578	0.472	0.561	0.491	0.598	0.483	0.577	0.474	0.626	0.504	**0.561**	0.483
	720	**0.631**	**0.492**	0.735	0.550	0.709	0.517	0.721	0.544	0.731	0.532	0.719	0.545	0.703	0.539	0.831	0.596	0.651	0.505	0.660	0.512	0.684	0.524	0.661	0.515	0.728	0.552	0.640	0.506
	Avg	**0.540**	**0.437**	0.594	0.486	0.587	0.453	0.595	0.482	0.602	0.514	0.740	0.591	0.574	0.454	0.684	0.523	0.554	0.455	0.551	0.466	0.578	0.471	0.554	0.462	0.601	0.488	0.544	0.472
Air	12	**0.663**	**0.613**	0.844	0.692	0.773	0.659	0.742	0.654	0.969	0.794	1.022	0.799	0.737	0.647	0.811	0.720	0.735	0.663	1.018	0.915	0.740	0.657	0.712	0.640	0.830	0.712	0.832	0.715
	24	**0.838**	**0.702**	0.886	0.826	0.906	0.726	0.888	0.729	1.077	0.870	1.042	0.848	0.876	0.718	0.982	0.806	0.901	0.744	1.327	1.047	0.881	0.721	**0.838**	0.713	0.841	0.733	0.850	0.730
	48	**0.945**	**0.753**	1.058	0.842	1.000	0.783	0.979	0.863	1.908	0.941	0.955	0.783	0.956	0.790	1.123	0.912	0.956	0.773	1.323	1.092	0.967	0.785	0.957	0.771	0.951	0.791	0.951	0.772
	Avg	**0.815**	**0.689**	0.929	0.787	0.893	0.723	0.870	0.749	1.318	0.868	1.006	0.810	0.856	0.718	0.972	0.813	0.864	0.727	1.270	1.045	0.863	0.721	0.836	0.708	0.874	0.745	0.878	0.739
MTraffic	12	**0.227**	**0.291**	0.398	0.432	0.280	0.336	0.315	0.402	0.845	0.714	0.904	0.693	0.277	0.335	0.313	0.363	0.389	0.451	0.371	0.399	0.251	0.298	0.248	0.312	0.414	0.437	0.761	0.674
	24	**0.245**	**0.335**	0.431	0.519	0.380	0.401	0.392	0.514	1.418	0.856	1.232	0.824	0.369	0.397	0.390	0.477	0.361	0.392	0.429	0.511	0.290	0.341	0.307	0.349	0.434	0.523	0.820	0.700
	48	**0.284**	**0.354**	0.549	0.572	0.470	0.467	0.438	0.590	1.900	0.950	1.079	0.757	0.424	0.429	0.447	0.481	0.517	0.583	0.493	0.521	0.338	0.371	0.374	0.398	0.550	0.568	0.826	0.701
	Avg	**0.252**	**0.327**	0.459	0.508	0.377	0.401	0.382	0.502	1.388	0.840	1.072	0.758	0.357	0.387	0.383	0.440	0.422	0.475	0.431	0.477	0.293	0.337	0.310	0.353	0.466	0.509	0.802	0.692
ILI	12	**1.109**	**0.652**	3.435	1.039	2.168	0.966	3.367	1.332	5.442	1.612	4.051	1.494	1.719	0.783	1.968	0.948	1.967	0.967	2.152	1.025	1.515	0.728	2.523	1.131	3.460	1.035	4.061	1.463
	24	**1.609**	**0.767**	3.593	1.189	2.536	1.001	3.819	1.423	5.661	1.827	4.324	1.563	1.729	0.798	3.181	1.532	2.037	1.048	2.853	1.334	2.170	0.982	3.281	1.277	3.562	1.225	3.851	1.416
	48	**1.534**	0.779	3.679	1.269	2.645	1.135	4.083	1.482	6.359	1.944	4.908	1.623	1.565	**0.762**	2.713	1.204	2.315	1.291	2.727	1.196	1.816	0.885	5.807	1.737	3.601	1.304	3.901	1.409
	Avg	**1.417**	**0.733**	3.569	1.166	2.450	1.034	3.756	1.412	5.821	1.794	4.428	1.560	1.671	0.781	2.621	1.228	2.106	1.102	2.577	1.185	1.834	0.865	3.870	1.382	3.541	1.188	3.938	1.429
1st Count		**105**		0		7		1		0		1		4		0		0		5		1		1		5		5	

4.2 Experimental Performance Analysis

As shown in Table 1, the proposed **TimeSelect** achieves the best or second-best performance on the majority of datasets and forecasting horizons. DLinear performs relatively poorly because it relies only on trend–seasonal decomposition followed by linear mappings. In contrast, iTransformer explicitly models inter-variable dependencies through attention, leading to clear improvements over DLinear. More recent methods, including PhaseFormer, ALW, UEC-STD, and interPDN, further improve forecasting accuracy through phase alignment, Koopman-based dynamics separation, error correction, and predictive distribution modeling, respectively. Among the competing baselines, in-

terPDN exhibits the most stable overall performance, mainly due to its step-wise distribution modeling and interleaved dual-branch architecture. However, these methods generally still rely on a fixed-length historical window and cannot adaptively identify a compact, input-conditioned historical context for different variables and input instances. By contrast, **TIMESELECT** suppresses redundant historical observations through variable-wise selective context modeling and combines basis-guided representations with cycle-aware refinement to preserve both stable temporal structures and recent periodic variations. Consequently, it achieves consistently competitive performance across datasets with diverse temporal characteristics. Compared with interPDN, the strongest overall baseline, **TIMESELECT** reduces the average MSE and MAE by 7.68% and 4.04%, respectively. The largest improvement is observed on the ILI dataset at $H = 12$, where the MSE decreases from 1.515 for the best baseline to 1.109, corresponding to a relative improvement of **26.80%**.

4.3 ABLATION EXPERIMENT ANALYSIS

We created the following variants to evaluate the effectiveness of each component: (1) *w/o Context Selector* (CS) that we remove the context selection component; (2) *Static Variable-wise Context (SVC)* where each variable have the fixed input length; (3) *w/o Basis* that we remove the temporal basis $\mathbf{Q}_L$ and $\mathbf{Q}_H$; (4) *w/o Cycle Refinement (CR)* that we remove the cycle refinement module.

As shown in Figure 3, the complete model consistently achieves the lowest MAE and MSE, confirming the effectiveness of all proposed components. Removing the temporal bases, denoted as *w/o Basis*, results in the most substantial performance degradation, demonstrating that the orthogonal bases $\mathbf{Q}_L$ and $\mathbf{Q}_H$ are crucial for extracting informative temporal structures and enabling accurate forecasting. We further evaluate the ablation performance of different temporal bases in **Appendix Section B.3**. Removing the context selector, denoted as *w/o CS*, also leads to a clear decline in performance. Although SVC retains variable-wise independent processing, its fixed context length remains inferior to the adaptive selector, indicating that different variables require distinct historical contexts. Finally, removing Cycle Refinement consistently worsens performance, suggesting that explicitly refining periodic structures can further improve forecasting accuracy.

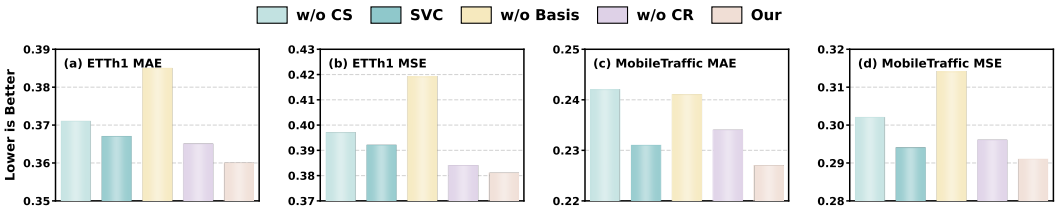


Figure 3: Ablation study analysis.

4.4 SELECT CONTEXT LENGTH ANALYSIS

We select five datasets to analyze the mean selected context lengths under different prediction horizons, as shown in Figure 4. With the input length fixed at 96, the selected context lengths exhibit clear dataset- and horizon-dependent variations. Electricity consistently requires relatively long historical contexts, whereas Exchange favors substantially shorter contexts. ETTh2, ETTm2, and Weather show moderate but varying selections across different horizons. Notably, the selected length does not monotonically increase with the prediction horizon, indicating that **TIMESELECT** adaptively determines the effective context according to the underlying temporal dynamics rather than simply retaining more history for longer-term forecasting.

We further investigate representative visualization cases. Taking Weather under the 96-input/96-prediction setting as an example, **TIMESELECT** effectively identifies the most informative historical context. In one representative case, the overall input sequence exhibits a downward trend, whereas the future sequence reverses and increases. Using the entire history may introduce outdated trend information and lead to inaccurate forecasts, as illustrated by the green curve. In contrast, **TIMESELECT** focuses on the recent context that is more relevant to the upcoming change, resulting in a more accurate prediction. As shown in Figure 6, extending the input length on Weather further

demonstrates that **TIMESELECT** can retain informative observations while filtering out redundant or potentially misleading historical information.

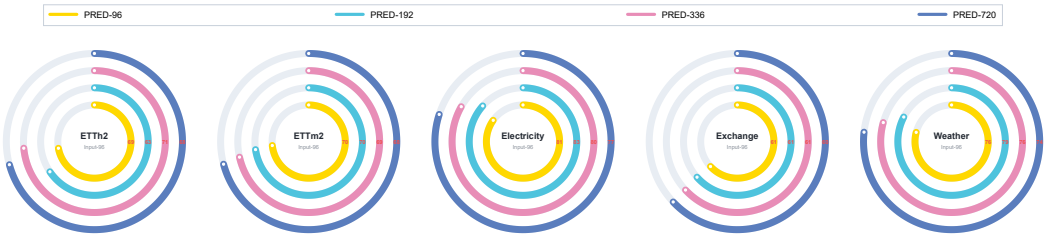


Figure 4: Statistics of valid context length.

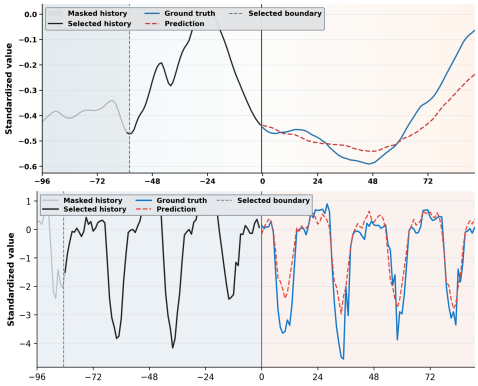


Figure 5: Predictive Visualization.

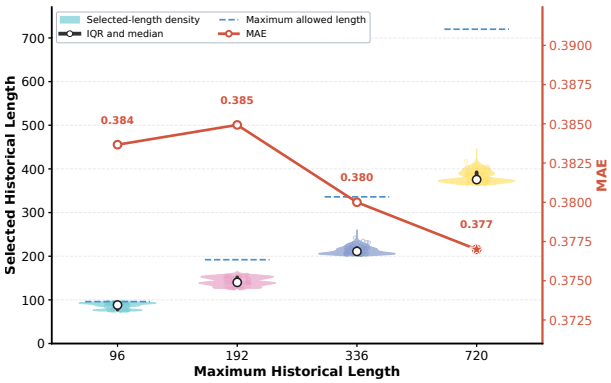


Figure 6: Selected-length distributions and forecasting performance.

4.5 EFFICIENCY ANALYSIS EXPERIMENT

Table 2 shows distinct efficiency costs across the compared models. MMPD is particularly slow at inference because probabilistic forecasting requires repeated diffusion sampling and mixture-model post-processing. DistDF has the longest training epoch: its distributional objective adds computation, and the small batch size required by its memory footprint entails more optimization steps. UEC-STD's complex backbone and error-correction stage increase inference overhead, whereas Phaseformer and iTransformer remain comparatively fast. Our method combines low training memory with fast inference, while the full-training results show strong forecasting accuracy.

Table 2: Efficiency on Mobile-Traffic (900 variables, $L = 96$, $H = 48$). Efficiency uses one epoch and 64 inference windows. Memory is peak training allocation in decimal MB. Percentages use unrounded values; red ↑ and blue ↓ mean higher and lower than ours.

Model	MSE	# MACs	# FLOPs	Epoch Time (s)	GPU Memory (MB)	Inference Time (s)
UEC-STD	0.549↑93.31%	5.77G↑28.22%	11.53G↑28.22%	9.22↑6.98%	6728↑48.87%	0.326↑545.01%
Phaseformer	0.470↑65.49%	4.53G↑0.80%	9.06G↑0.80%	9.94↑15.29%	4723↑4.50%	0.068↑35.05%
TimeMixer	0.550↑93.66%	4.65G↑3.38%	9.30G↑3.38%	31.65↑267.11%	6250↑38.29%	0.293↑479.76%
MMPD	0.438↑54.23%	4.34T↑96329%	8.67T↑96329%	40.48↑369.61%	14980↑231.45%	509.145↑1006652%
iTransformer	0.338↑19.01%	5.76G↑28.18%	11.53G↑28.18%	7.95↓7.78%	11396↑152.14%	0.056↑11.52%
DistDF	1.079↑279.93%	15.23G↑238.62%	30.45G↑238.62%	108.83↑1162.40%	13688↑202.86%	0.718↑1319.33%
Our	**0.284**	**4.50G**	**8.99G**	**8.62**	**4520**	**0.051**

5 CONCLUSION

In this work, we proposed **TIMESELECT**, a time-series forecasting framework that adaptively identifies variable-specific historical contexts instead of relying on a uniformly fixed input window. By combining differentiable context selection, correlation-derived temporal bases, and cycle-aware refinement, **TIMESELECT** effectively suppresses irrelevant historical observations while preserving stable temporal structures and recent periodic dynamics. Extensive experiments on 13 real-world benchmarks demonstrate its consistent superiority, achieving the best performance in most settings while maintaining favorable computational efficiency.

AI USE STATEMENT

An LLM-based coding assistant was used to assist with experiment orchestration, log aggregation, figure generation, language organization, and LaTeX drafting. The authors are responsible for verifying the method description, citations, experimental settings, and numerical results against the code, saved artifacts, and execution logs.

REFERENCES

Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. A multi-source dataset of urban life in the city of milan and the province of trentino. *Scientific data*, 2(1): 150055, 2015.

Defu Cao, Furong Jia, Sercan Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. In *International Conference on Learning Representations*, volume 2024, pp. 18546–18578, 2024.

Hui Chen, Viet Luong, Lopamudra Mukherjee, and Vikas Singh. Simpletm: A simple baseline for multivariate time series forecasting. In *The Thirteenth International Conference on Learning Representations*, 2025.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

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 *The twelfth international conference on learning representations*, 2024.

Jiayu Fang, Xuande Liu, Sangsha Fang, Ernie Tian, Hongwei Ma, Zhiqi Shao, and Junbin Gao. Information geometry loss for time series forecasting. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=2NbXZDl4nB.

Qi Guo, Baolin Zhao, Mingchen Song, and Guoqiang Zhong. A survey of deep learning for time series forecasting: Taxonomy, analysis and future directions. *IEEE Transactions on Knowledge and Data Engineering*, 2026.

Xiao Han, Xinfeng Zhang, Yiling Wu, Zhenduo Zhang, and Zhe Wu. Are kans effective for multivariate time series forecasting? *arXiv preprint arXiv:2408.11306*, 2024.

Yifan Hu, Peiyuan Liu, Peng Zhu, Dawei Cheng, and Tao Dai. Adaptive multi-scale decomposition framework for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17359–17367, 2025a.

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*, 2025b.

Qihe Huang, Lei Shen, Ruixin Zhang, Shouhong Ding, Binwu Wang, Zhengyang Zhou, and Yang Wang. Crossgnn: Confronting noisy multivariate time series via cross interaction refinement. *Advances in Neural Information Processing Systems*, 36:46885–46902, 2023.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. In *International conference on learning representations*, volume 2024, pp. 23857–23880, 2024.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International conference on learning representations*, 2021.

Linghao Kong and Xiaopeng Hong. Time series forecasting via direct per-step probability distribution modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22653–22661, 2026a.

Linghao Kong and Xiaopeng Hong. Time series forecasting via direct per-step probability distribution modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22653–22661, 2026b.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 6555–6565, 2024.

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*, 2024.

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.

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*, 2024a.

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*, 2024b.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-xl: Long-context transformers for unified time series forecasting. In *International Conference on Learning Representations*, volume 2025, pp. 83982–84006, 2025.

Aitian Ma, Dongsheng Luo, and Mo Sha. Mixlinear: Extreme low resource multivariate time series forecasting with $0.1k$ parameters. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=QUj0KuCumD.

Aitian Ma, Dongsheng Luo, and Mo Sha. Mixlinear: Extreme low resource multivariate time series forecasting with $0.1k$ parameters. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=QUj0KuCumD.

Jiaming Ma, Binwu Wang, Qihe Huang, Guanjun Wang, Pengkun Wang, Zhengyang Zhou, and Yang Wang. Mofo: Empowering long-term time series forecasting with periodic pattern modeling. *Advances in Neural Information Processing Systems*, 2025.

Jiaming Ma, Qihe Huang, Haofeng Ma, Guanjun Wang, Sheng Huang, Zhengyang Zhou, Pengkun Wang, Binwu Wang, and Yang Wang. Phat: Modeling period heterogeneity for multivariate time series forecasting. *arXiv preprint arXiv:2602.00654*, 2026c.

Jiaming Ma, Binwu Wang, Qihe Huang, Guanjun Wang, Pengkun Wang, Zhengyang Zhou, and Yang Wang. Mofo: Empowering long-term time series forecasting with periodic pattern modeling. *Advances in Neural Information Processing Systems*, 38:84871–84902, 2026d.

Majdi Mansouri, Khadija Attouri, and Shady S Refaat. Multimodal learning techniques for time series forecasting in renewable energy systems: A comprehensive survey. *IEEE Access*, 2025.

John A Miller, Mohammed Aldosari, Farah Saeed, Nasid Habib Barna, Subas Rana, I Budak Arpinar, and Ninghao Liu. A survey of deep learning and foundation models for time series forecasting. *arXiv preprint arXiv:2401.13912*, 2024.

Md Mahmuddun Nabi Murad, Mehmet Aktukmak, and Yasin Yilmaz. Wpmixer: Efficient multi-resolution mixing for long-term time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 19581–19588, 2025.

Reza Nematirad, Anil Pahwa, and Balasubramaniam Natarajan. Times2d: Multi-period decomposition and derivative mapping for general time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19651–19658, 2025.

Minh Hoang Nguyen, Van Dai Do, Huu Hiep Nguyen, Dung Nguyen, Kien Do, and Hung Le. Reviving error correction in modern deep time-series forecasting. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=ZlDndRDMHm.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Yiming Niu, Jinliang Deng, and Yongxin Tong. Phaseformer: From patches to phases for efficient and effective time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Lk9SqMQzhX.

Zining Qin, Huiling Qin, Chenhao Wang, Jianxiong Guo, Tian Wang, and Weijia Jia. Not all frequencies are equal: Energy-adaptive diffusion for time series forecasting. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=4kdkm56U5b.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. Tfb: Towards comprehensive and fair benchmarking of time series forecasting methods. *Proc. VLDB Endow.*, 17(9):2363–2377, 2024.

Xin Qiu, Junlong Tong, Yirong Sun, Yunpu Ma, and Xiaoyu Shen. Revealing scaling paradox in large-scale time series models: Implications for more efficient and accurate forecasting. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=YmFkYtD6wO.

Zongjiang Shang, Ling Chen, Binqing Wu, and Dongliang Cui. Ada-mshyper: adaptive multi-scale hypergraph transformer for time series forecasting. *Advances in Neural Information Processing Systems*, 37:33310–33337, 2024.

Robert H Shumway and David S Stoffer. Arima models. In *Time series analysis and its applications: with R examples*, pp. 85–175. Springer, 2024.

Artyom Stitsyuk and Jaesik Choi. xpatch: Dual-stream time series forecasting with exponential seasonal-trend decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20601–20609, 2025a.

Artyom Stitsyuk and Jaesik Choi. xpatch: Dual-stream time series forecasting with exponential seasonal-trend decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20601–20609, 2025b.

Tian Sun, Yuqi Chen, and Weiwei Sun. Penguin: Enhancing transformer with periodic-nested group attention for long-term time series forecasting. *arXiv preprint arXiv:2508.13773*, 2025.

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*, pp. 12640–12648, 2025.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems*, 34:24804–24816, 2021.

Suxin Tong and Jingling Yuan. Efficiently enhancing long-term series forecasting via adaptive lookback with wavelets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25966–25974, 2026.

Binwu Wang, Gaoyun Lin, Jiaming Ma, Qihe Huang, Zhengyang Zhou, Xu Wang, Pengkun Wang, and Yang Wang. Crisp: A spectral-based interaction strategy for multivariate time series forecasting. In *Forty-third International Conference on Machine Learning*, 2026a.

Eric Wang, Licheng Pan, Yuan Lu, Zhixuan Chu, Xiaoxi Li, Shuting He, Zi Ciu Chan, Haoxuan Li, Qingsong Wen, and Zhouchen Lin. DistDF: Time-series forecasting needs joint-distribution wasserstein alignment. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=VrdLwUmzBy.

Eric Wang, Licheng Pan, Yuan Lu, Zhixuan Chu, Xiaoxi Li, Shuting He, Zi Ciu Chan, Haoxuan Li, Qingsong Wen, and Zhouchen Lin. DistDF: Time-series forecasting needs joint-distribution wasserstein alignment. In *The Fourteenth International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=VrdLwUmzBy.

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, 2024.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Chen Wang, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026d.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22419–22430, 2021.

Qiang Wu, Gechang Yao, Zhixi Feng, and Yang Shuyuan. Peri-midformer: Periodic pyramid transformer for time series analysis. *Advances in Neural Information Processing Systems*, 37:13035–13073, 2024.

Yuhan Wu, Xiyu Meng, Huajin Hu, Junru Zhang, Yabo Dong, and Dongming Lu. Affirm: Interactive mamba with adaptive fourier filters for long-term time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 21599–21607, 2025.

Linxiao Yang, Xue Jiang, Gezheng Xu, Tian Zhou, Min Yang, Zhaoyang Zhu, Linyuan Geng, Zhipeng Zeng, Qiming Chen, Xinyue Gu, et al. Baguan-ts: dual in-context learning model for time series forecasting with covariates. In *Forty-third International Conference on Machine Learning*, 2026.

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, 2023.

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, 2023a.

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, 2023b.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023.

Yunhao Zhang, Wenyao Hu, Jiale Zheng, Lujia Pan, and Junchi Yan. MMPD: Diverse time series forecasting via multi-mode patch diffusion loss. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=NEUgHT8dvH.

Yunhao Zhang, Wenyao Hu, Jiale Zheng, Lujia Pan, and Junchi Yan. MMPD: Diverse time series forecasting via multi-mode patch diffusion loss. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=NEUgHT8dvH.

Zhiwei Zhang, Xinyi Du, Xuanchi Guo, Weihao Wang, and Wenjuan Han. Emaformer: Enhancing transformer through embedding armor for time series forecasting. *arXiv preprint arXiv:2511.08396*, 2025.

Zijun Zhang. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, pp. 1–2. Ieee, 2018.

Siru Zhong, Yuxuan Liang, Weilin Ruan, Ming Jin, Huan Li, and Qingsong Wen. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. In *AI for Time Series*, pp. 173–203. CRC Press, 2025.

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, 2022a.

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, 2022b.

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.

A Related Work

Time Series Forecasting Time series forecasting (TSF) aims to infer future dynamics from historical observations. Existing approaches can be broadly grouped into statistical learning, machine learning, and deep learning methods (Huang et al., 2023). Classical statistical models, such as ARIMA (Shumway & Stoffer, 2024), are effective for relatively simple processes but often struggle with long sequences, high-dimensional variables, and nonlinear interactions. Machine learning models, including XGBoost (Chen & Guestrin, 2016) and LightGBM (Ke et al., 2017), offer stronger nonlinear modeling capabilities and can perform well with limited training data. Their effectiveness, however, commonly relies on extensive feature engineering and domain knowledge. As temporal dependencies (TD) and variable correlations (VC) become increasingly complex in large-scale multivariate data, deep learning has emerged as a more flexible forecasting paradigm.

Transformer-Based Models Transformer architectures have played a central role in recent advances in time series forecasting (Miller et al., 2024). Most existing studies improve forecasting performance from two complementary perspectives: temporal pattern extraction within each variable and dependency modeling across variables.

One group of methods is designed to strengthen the representation of individual time series. These approaches explore temporal structures from different views, including spectral and joint time–frequency analysis (Zhou et al., 2022b; Ma et al., 2026c), explicit periodic pattern extraction (Wu et al., 2024; Sun et al., 2025), and hierarchical modeling over multiple temporal resolutions (Shang et al., 2024; Ma et al., 2025). A representative example is PDF (Dai et al., 2024), which identifies multiple underlying periods and constructs period-specific patches, together with frequency-domain

Algorithm 1 Overall Procedure of **TIMESELECT**

Input: Training split $\mathcal{D}_{\text{tr}}$ and historical sequence $\mathbf{X} \in \mathbb{R}^{L \times N}$
Output: Forecasted sequence $\widehat{\mathbf{Y}} \in \mathbb{R}^{H \times N}$

1. Variable-Wise Context Selection

$\overline{\mathbf{X}} \leftarrow \text{RevIN}(\mathbf{X})$
for $n = 1$ to N **do**
  $\mathbf{s}_n \leftarrow \text{ContextSummary}\left(\overline{\mathbf{X}}_{:,n}\right)$
  $\ell_n \leftarrow \ell_{\min} + (L - \ell_{\min})\sigma(\text{MLP}(\mathbf{s}_n))$
  for $t = 1$ to L **do**
    $M_{t,n} \leftarrow \sigma\left(\dfrac{\ell_n - (L - t + 1)}{\tau}\right)$
  end for
end for
$\widetilde{\mathbf{X}} \leftarrow \mathbf{M} \odot \overline{\mathbf{X}}$

2. Offline Correlation-Derived Temporal Bases

for $r \in \{L, H\}$ **do**
  for each valid variable $n \in \mathcal{V}$ **do**
    Construct the lagged matrix $\mathbf{Z}_r^{(n)} \in \mathbb{R}^{r \times (T-r+1)}$
    $\mathbf{C}_r^{(n)} \leftarrow \text{Corr}\left(\mathbf{Z}_r^{(n)}\right)$
  end for
  $\mathbf{\Sigma}_r \leftarrow (1 - \rho)\dfrac{1}{|\mathcal{V}|}\sum_{n \in \mathcal{V}} \mathbf{C}_r^{(n)} + \rho \mathbf{I}_r$
  $\mathbf{\Sigma}_r = \mathbf{U}_r \mathbf{\Lambda}_r \mathbf{U}_r^\top$
  $\mathbf{Q}_r \leftarrow \mathbf{U}_r^\top$
end for
Store the fixed temporal bases $\mathbf{Q}_L$ and $\mathbf{Q}_H$

3. Basis-Guided Forecasting and Refinement

$E_{n,t,d} \leftarrow \widetilde{X}_{t,n} e_d, \quad \mathbf{E} \in \mathbb{R}^{N \times L \times D}$
Rearrange $\mathbf{E} \in \mathbb{R}^{N \times D \times L}$
$\mathbf{S}_{n,d,:} \leftarrow \mathbf{E}_{n,d,:}\mathbf{Q}_L^\top + \boldsymbol{\delta}_{\text{in},n,d,:}$
$\mathbf{H}_{n,:} \leftarrow \phi_{\text{in}}(\text{vec}(\mathbf{S}_{n,:,:}))$
$\mathbf{A} \leftarrow \text{RowNorm}_1(\text{softplus}(\mathbf{A}_\theta))$
$\mathbf{H}' \leftarrow \text{LN}(\mathbf{H} + \mathbf{A}(\mathbf{H}\mathbf{W}_V)\mathbf{W}_O)$
$\mathbf{H}^+ \leftarrow \text{LN}(\mathbf{H}' + \text{FFN}(\mathbf{H}'))$
for $n = 1$ to N **do**
  $\mathbf{G}_{n,:,:} \leftarrow \text{reshape}_{D \times H}\left(\mathcal{F}_\theta(\mathbf{H}^+_{n,:})\right)$
  $\mathbf{O}_{n,d,:} \leftarrow \mathbf{G}_{n,d,:}\mathbf{Q}_H + \boldsymbol{\delta}_{\text{out},n,d,:}$
  $\widetilde{\mathbf{Y}}_{\text{base},:,n} \leftarrow \mathcal{P}_\theta(\text{vec}(\mathbf{O}_{n,:,:}))$
end for

4. Cycle-aware final refinement

$c' \leftarrow \min(c, L)$
$\mathbf{R} \leftarrow \text{Repeat}\left(\overline{\mathbf{X}}_{L-c'+1:L,:}, H\right)$
$\alpha_{\text{cyc}} \leftarrow \sigma(a_{\text{cyc}})$
$\overline{\mathbf{Y}} \leftarrow \widetilde{\mathbf{Y}}_{\text{base}} + \alpha_{\text{cyc}}\, h_\psi\left(\left[\widetilde{\mathbf{Y}}_{\text{base}}, \mathbf{R}\right]\right)$
$\overline{\mathbf{Y}} \leftarrow \text{RevIN}^{-1}\left(\overline{\mathbf{Y}}\right)$
return $\widehat{\mathbf{Y}}$

decomposition, to separate heterogeneous periodic behaviors. This design enables the model to characterize temporal patterns that cannot be adequately represented at a single scale or frequency.

Another group of studies concentrates on the relationships among different variables. Rather than processing each sequence independently, these methods introduce dedicated interaction mechanisms to propagate information across variables. Crossformer (Zhang & Yan, 2023) models cross-dimensional dependencies through attention over segmented representations, while iTransformer (Liu et al., 2024b) treats variables as tokens to directly learn their correlations. TimeFilter (Hu et al., 2025b) further incorporates structured interaction modules to capture heterogeneous variable relationships. By allowing information to be shared across correlated series, these models provide richer representations for multivariate forecasting.

Linear and Multilevel Models Linear forecasting models employ simple projection or decomposition operations to provide efficient alternatives to attention-based architectures. Methods such as NLinear, DLinear (Zeng et al., 2023b), SparseTSF (Lin et al., 2024), xPatch (Stitsyuk & Choi, 2025a), and SimpleTM (Chen et al., 2025) demonstrate that compact structures can be highly competitive, particularly on datasets with relatively few variables and regular temporal patterns. Nevertheless, the limited representational capacity of shallow linear modules becomes increasingly restrictive as the number of variables and interaction patterns grows. In such cases, insufficient token diversity can hinder the modeling of complex dependencies. Multilevel architectures, including WPMixer (Murad et al., 2025) and AMD (Hu et al., 2025a), exhibit smaller performance degradation under challenging settings, suggesting that hierarchical transformations may offer a practical balance between the token redundancy of Transformers and the token scarcity of shallow linear models.

B Experimental Details

B.1 Datasets

We evaluate **TimeSelect** on 13 commonly used time series forecasting benchmarks covering multiple application domains. Their statistical characteristics are reported in Table 3, while a brief introduction to each dataset is provided below.

- **Exchange-Rate (Exchange)** consists of daily exchange-rate records from eight countries, spanning the period from 1990 to 2016 (Lin et al., 2025).
- **Weather** includes 21 meteorological variables measured by the Weather Station of the Max Planck Institute for Biogeochemistry in 2020. The observations are sampled every 10 minutes (Wu et al., 2021).
- **Electricity (ECL)** records the hourly electricity consumption of 321 clients between 2012 and 2014 (Zhou et al., 2021).
- **Electricity Transformer Temperature (ETT)** comprises transformer load and oil-temperature measurements collected from July 2016 to July 2018. The dataset contains four subsets, namely ETTm1, ETTm2, ETTh1, and ETTh2. The subsets marked with "m" are sampled at 15-minute intervals, whereas those marked with "h" are recorded hourly (Zhou et al., 2021).
- **AQWan** contains hourly air-quality and meteorological observations collected at the Wanliu monitoring station. It comprises 11 variables and 35,064 time points (Qiu et al., 2024).
- **CzeLan** is a nature-domain dataset containing 11 environmental variables. It consists of 19,934 observations sampled at 30-minute intervals (Qiu et al., 2024).
- **ZafNoo** is a nature-domain dataset comprising 11 environmental variables and 19,225 observations recorded every 30 minutes (Qiu et al., 2024).
- **ILI** overs the period from 2002 to 2021 and comprises 966 weekly observations across seven variables. It records the proportion of patients with influenza-like illness relative to the total number of patient visits in the United States, as reported by the Centers for Disease Control and Prevention (CDC). Characterized by pronounced seasonality, epidemic fluctuations, and complex temporal dependencies, the dataset is widely used to evaluate forecasting models for public health time series.
- **Mobile Traffic** (Barlacchi et al., 2015) is part of the Telecom Italia Big Data Challenge 2014, which integrates multimodal data from telecommunications, weather, news, social networks, and electricity in Milan and the Trentino region of Italy, providing a rich and publicly available benchmark. We use one year of telecommunications data collected from 900 spatial regions in Milan, with an hourly temporal resolution.

B.2 Baselines

We select 13 representative models as baselines. These models have been extensively cross-tested across a wide range of prior studies and are publicly available in open-source form, ensuring both reproducibility and fair comparison.

Table 3: Description of the datasets used in our experiments. Variates denotes the number of variables, Time Steps denotes the number of time points, Granularity denotes the sampling frequency, and Domain denotes the application category. "min" and "h" represent minute(s) and hour(s), respectively.

Dataset	ETTh1	ETTh2	ETTm1	ETTm2	Electricity	Weather	Exchange	AQWan	CzeLan	ZafNoo
Variates	7	7	7	7	321	21	8	11	11	11
Time Steps	17,420	17,420	69,680	69,680	26,304	52,696	7,588	35,064	19,934	19,225
Granularity	1 h	1 h	15 min	15 min	1 h	10 min	1 day	1 h	30 min	30 min
Domain	Electricity	Electricity	Electricity	Electricity	Electricity	Environment	Economic	Environment	Nature	Nature

- **DLinear** (Zeng et al., 2023a) decomposes a time series into trend and seasonal components and applies simple linear layers to model them separately, providing an efficient yet competitive baseline for long-term forecasting.
- **iTransformer** (Liu et al., 2024a) treats individual variables as tokens and applies attention across variates, enabling the model to capture multivariate dependencies while preserving temporal information within each token.
- **PatchTST** (Nie et al., 2023) divides each time series into temporal patches and processes variables independently with a Transformer, improving long-range representation learning while reducing computational cost.
- **TimeMixer** (Wang et al., 2024) adopts a multiscale mixing architecture that decomposes and integrates temporal patterns across different resolutions, allowing efficient modeling of both short-term variations and long-term trends.
- **xPatch** (Stitsyuk & Choi, 2025b) employs exponential seasonal–trend decomposition and a dual-stream architecture, using an MLP-based linear stream for trend modeling and a CNN-based nonlinear stream for capturing seasonal patterns.
- **PhaseFormer** (Niu et al., 2026) replaces conventional temporal patches with phase-aligned representations and employs lightweight routing to model interactions across phases, enabling accurate and parameter-efficient forecasting.
- **MMPD** (Zhang et al., 2026a) introduces a multi-mode patch diffusion loss that enables patch-based forecasting models to capture complex future distributions and generate diverse predictions with corresponding probabilities.
- **MixLinear** (Ma et al., 2026a) combines segment-based temporal modeling with adaptive low-rank spectral filtering, effectively capturing complementary time- and frequency-domain patterns with an extremely small number of parameters.
- **DistDF** (Wang et al., 2026b) aligns the joint distributions of forecasts and target sequences through a Wasserstein-based objective, reducing the estimation bias caused by temporal autocorrelation in future observations.
- **interPDN** (Kong & Hong, 2026a) directly models a discrete predictive distribution at each future time step and uses an interleaved dual-branch architecture to improve forecasting accuracy and robustness.
- **ALW** (Tong & Yuan, 2026) is a wavelet-transform-driven approach for multiscale adaptive look-back window selection. It decomposes time series into distinct frequency components through wavelet transforms, quantifies the contribution of each historical time step using scale-specific attention mechanisms, and dynamically determines the optimal window length through backward information accumulation and soft truncation.
- **UEC-STD** (Nguyen et al., 2026) introduces an architecture-agnostic error corrector that decomposes forecasts into trend and seasonal components and adjusts them separately, mitigating autoregressive error accumulation without retraining the original forecasting model.

B.3 Ablation Study of Temporal Basis

To investigate the effect of temporal basis design on forecasting performance, we keep the data splits, normalization scheme, forecasting network, Adaptive-L module, training hyperparameters, and random seeds identical across all experiments, while varying only the temporal basis matrices used at the input and output ends of the model.

Table 4: Ablation study of different temporal basis designs on ETTh1, ETTm1, and Weather. Ours denotes the complete model. Lower values indicate better performance, and the best results are highlighted in bold. Rankings are determined using the unrounded values.

Dataset	Metric	Identity	Random orth.	DCT	Learnable	Input only Q_L	Output only Q_H	Ours
ETTh1	MSE	0.368	0.367	0.365	0.363	0.369	0.362	**0.363**
	MAE	0.387	0.392	0.386	0.387	0.389	0.383	**0.384**
ETTm1	MSE	0.341	0.340	0.341	0.341	0.338	0.337	**0.303**
	MAE	0.307	0.306	0.305	0.306	0.304	0.304	**0.334**
Weather	MSE	0.159	0.158	0.167	0.158	0.161	0.151	**0.151**
	MAE	0.199	0.203	0.201	0.196	0.197	0.195	**0.190**

Given an input sequence $\mathbf{X}$ of length L, the input is first projected into the temporal basis space as

$$\overline{\mathbf{X}} = \mathbf{X}\mathbf{Q}_L^\top. \tag{17}$$

The forecasting network then generates a representation $\widetilde{\mathbf{Y}}$ of length H in the transformed space, which is mapped back to the original output domain as

$$\widehat{\mathbf{Y}} = \widetilde{\mathbf{Y}}\mathbf{Q}_H. \tag{18}$$

We compare six temporal basis designs: ❶ *Identity* uses identity matrices, i.e., $\mathbf{Q}_L = \mathbf{I}_L$ and $\mathbf{Q}_H = \mathbf{I}_H$; ❷ *Random Orthogonal* first samples random square matrices from a standard Gaussian distribution and then obtains fixed orthogonal bases through QR decomposition; ❸ *DCT* adopts normalized DCT-II bases, whose rows correspond to cosine components at different frequencies; ❹ *Learnable Basis* uses the same random orthogonal initialization as *Random Orthogonal*, but treats the basis matrices as learnable parameters and reparameterizes them through QR decomposition during each forward pass to preserve orthogonality; ❺ *Input Basis Only* applies the data-driven basis $\mathbf{Q}_L$ only at the input end while setting $\mathbf{Q}_H = \mathbf{I}_H$; and ❻ *Output Basis Only* sets $\mathbf{Q}_L = \mathbf{I}_L$ and applies the data-driven basis $\mathbf{Q}_H$ only at the output end.

The data-driven bases are constructed exclusively from the training data. Specifically, each variable in the training sequence is first standardized, after which sliding windows of length L or H are extracted. Windows from all variables are treated as independent samples to estimate the covariance matrix along the temporal dimension. The corresponding eigenvectors are then sorted in descending order according to their eigenvalues to form $\mathbf{Q}_L$ and $\mathbf{Q}_H$, respectively. As shown in Table 4, experiments are conducted on ETTh1 ($96 \to 96$), ETTm1 ($96 \to 96$), and Weather ($96 \to 96$), with forecasting errors evaluated using test-set MSE and MAE. The experimental results demonstrate that the proposed data-driven temporal basis achieves the best overall forecasting performance.

B.4 Robustness to Maximum Historical Length

We examine how **TimeSelect** behaves as the maximum historical input length varies. As shown in Table 5, **TimeSelect** consistently achieves comparable or lower forecasting errors than its full-window counterpart while retaining only a portion of the available history. The selected context lengths vary across datasets and input settings, indicating that the model does not simply preserve a fixed proportion of the historical window. The advantage of adaptive selection is generally more evident when longer histories are available, where uniformly using the full window may introduce redundant or less relevant observations. Overall, these results suggest that **TimeSelect** can adjust its effective context according to the available historical range and maintain reliable forecasting performance with shorter effective contexts.

B.5 Effect of Context Assignment

Figure 7 examines whether the correspondence between each forecasting state and its selected context contributes to prediction quality. We compare the original assignments with three controlled

Table 5: Robustness to the maximum historical input length. Full Window and **TIMESELECT** use identical data and optimization settings. Retained denotes the ratio between the average selected context length and the maximum input length L.

Dataset	L	Full Window		TIMESELECT		Selected Context	
		MSE ↓	MAE ↓	MSE ↓	MAE ↓	Avg. Length	Retained
ETTh1	96	0.364	0.385	**0.363**	**0.384**	86.5	90.1%
	192	0.360	0.387	0.360	**0.385**	142.0	74.0%
	336	0.378	0.403	**0.358**	**0.388**	213.3	63.5%
	720	0.371	0.401	**0.359**	**0.395**	382.1	53.1%
ETTm1	96	0.305	0.336	**0.303**	**0.334**	74.3	77.4%
	192	**0.275**	0.327	0.278	**0.324**	150.5	78.4%
	336	0.281	0.332	**0.275**	**0.326**	236.2	70.3%
	720	0.287	0.338	**0.284**	**0.335**	419.1	58.2%
Weather	96	0.152	**0.189**	**0.151**	0.189	75.9	79.1%
	192	0.148	0.187	0.148	**0.186**	117.0	60.9%
	336	0.149	0.191	**0.147**	**0.190**	236.4	70.4%
	720	0.156	0.197	**0.150**	**0.191**	344.0	47.8%

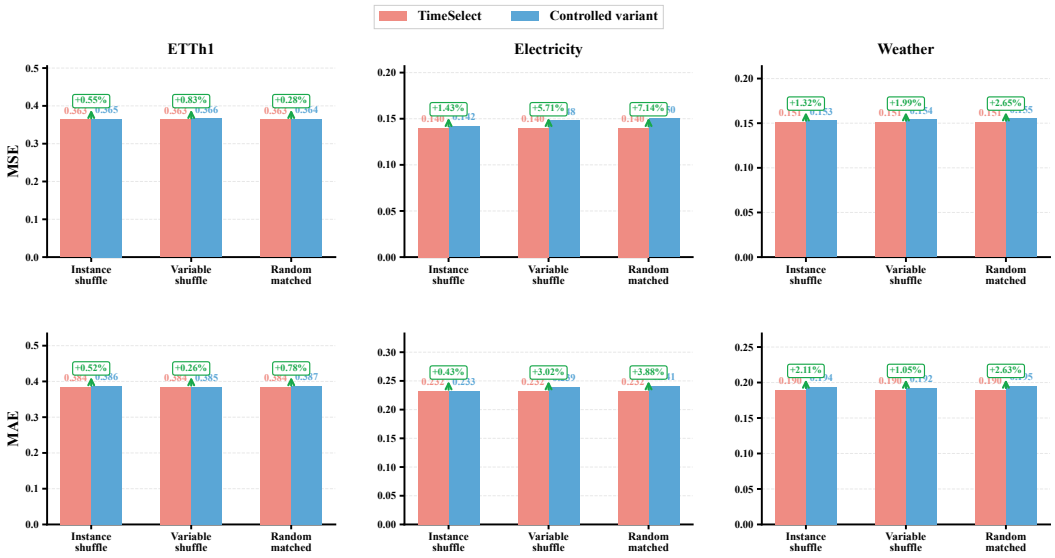


Figure 7: Context-permutation controls.

variants that shuffle context lengths across instances, shuffle them across variables, or replace them with randomly matched lengths while preserving the overall context-length distribution. All three controls increase both MSE and MAE across the evaluated datasets, although the magnitude varies by dataset. On ETTh1, the increases remain moderate, ranging from 0.28% to 0.83% in MSE and from 0.26% to 0.78% in MAE. Larger differences are observed on Electricity, where MSE and MAE increase by up to 7.14% and 3.88%, respectively, while Weather exhibits increases of up to 2.65% in MSE and 2.63% in MAE. Since these controls retain comparable context-length budgets but alter their assignments, the results suggest that the learned association between the current forecasting state and its selected historical context contributes to forecasting performance, with its effect depending on the temporal characteristics of each dataset.

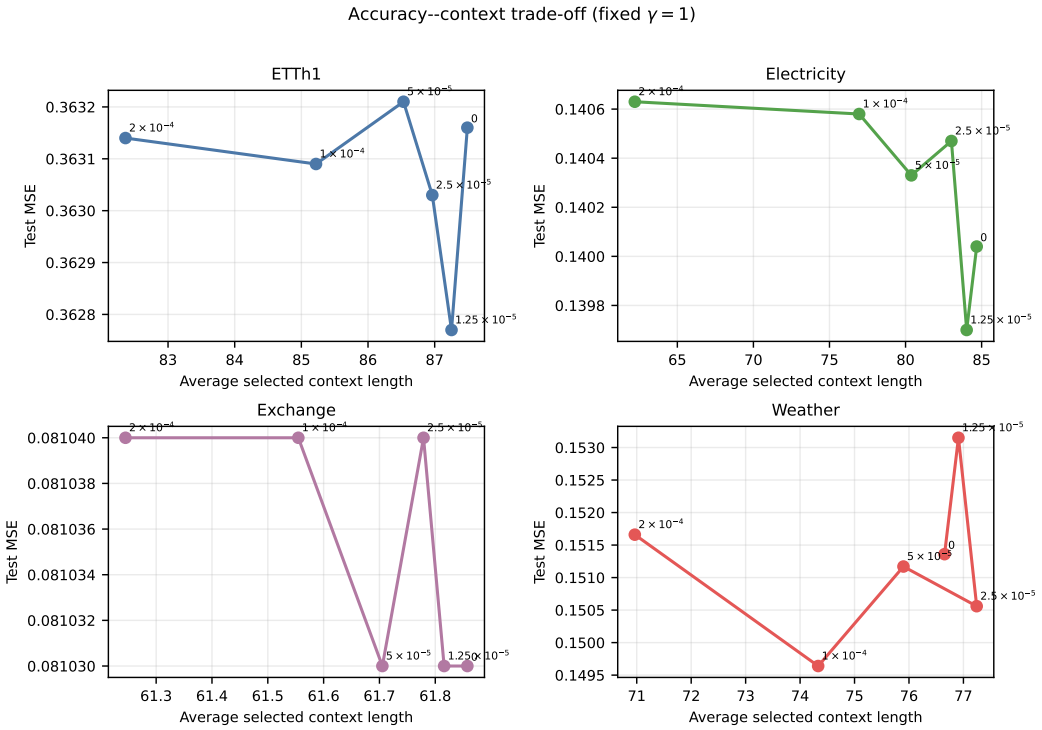


Figure 8: Accuracy–context-length trade-off for the context regularization sweep.

B.6 Effect of Context Regularization

We investigate the effect of the context regularization coefficient λ_{ctx} on forecasting accuracy and the average selected context length. As shown in Figure 8, increasing the regularization strength generally encourages the model to retain shorter historical contexts, while its influence on forecasting error varies across datasets. The relationship is not strictly monotonic: moderate regularization often maintains or improves prediction accuracy, whereas overly strong regularization may remove useful historical information and increase the error. ETTh1 and Exchange remain relatively stable across the evaluated settings, while Electricity and Weather exhibit a more noticeable accuracy–context trade-off. Overall, the results suggest that context regularization provides a controllable mechanism for balancing forecasting accuracy and historical compactness, with the preferred strength depending on the temporal characteristics of each dataset.

B.7 Comparison of Context Selection Strategies

To determine whether the gains arise from adaptive context selection rather than merely from using shorter inputs, we compare **TimeSelect** with several routing alternatives under the same backbone and training configuration. Full $L = 96$ retains the complete input history using an all-one mask. Validation-best fixed L independently selects $L^*_{d,h} \in \{24, 48, 72, 96\}$ for each dataset–horizon setting according to validation MSE, after which the selected scalar boundary is shared by all instances and variables in that setting. Static variable-wise learns one continuous boundary ℓ_n for each variable end-to-end, while keeping it fixed across instances. Instance-wise shared first predicts variable-wise candidate boundaries from each multivariate input, averages them into one instance-dependent boundary ℓ_i, and broadcasts ℓ_i to all variables.

As shown in Table 7, **TimeSelect** obtains the lowest average MSE and MAE across the evaluated settings. The validation-selected fixed-window baseline remains competitive in several short-horizon cases, but its performance varies substantially across datasets and horizons. Learning static variable-specific boundaries improves over the fixed baselines on average, yet remains behind the instance-adaptive alternatives. The stronger performance of Instance-wise shared demonstrates the

Table 6: Numerical results corresponding to the accuracy–context trade-off in Figure 8 under $L = 96$ and $H = 96$. Retained denotes the average selected context length as a percentage of L. The gray rows indicate the default setting.

$\lambda_{\mathbf{ctx}}$	MSE ↓	MAE ↓	Avg. selected L	Retained (%)
ETTh1				
0	0.363	0.384	87.5	91.1
1.25×10^{-5}	0.363	0.383	87.3	90.9
2.5×10^{-5}	0.363	0.384	87.0	90.6
5×10^{-5}	0.363	0.384	86.5	90.1
1×10^{-4}	0.363	0.384	85.2	88.8
2×10^{-4}	0.363	0.384	82.4	85.8
Electricity				
0	0.140	0.232	84.7	88.2
1.25×10^{-5}	0.140	0.232	84.0	87.5
2.5×10^{-5}	0.140	0.232	83.0	86.5
5×10^{-5}	0.140	0.232	80.4	83.7
1×10^{-4}	0.141	0.233	77.0	80.2
2×10^{-4}	0.141	0.233	62.2	64.8
Exchange				
0	0.081	0.198	61.9	64.4
1.25×10^{-5}	0.081	0.198	61.8	64.4
2.5×10^{-5}	0.081	0.198	61.8	64.4
5×10^{-5}	0.081	0.198	61.7	64.3
1×10^{-4}	0.081	0.198	61.6	64.1
2×10^{-4}	0.081	0.198	61.2	63.8
Weather				
0	0.151	0.190	76.7	79.8
1.25×10^{-5}	0.153	0.191	76.9	80.1
2.5×10^{-5}	0.151	0.189	77.2	80.5
5×10^{-5}	0.151	0.190	75.9	79.1
1×10^{-4}	0.150	0.188	74.3	77.4
2×10^{-4}	0.152	0.189	71.0	73.9

importance of conditioning context selection on the input instance. The additional improvement of **TIMESELECT**, particularly for the longer-horizon ETTh1 and Exchange settings, indicates that allowing the selected context to vary across both instances and variables can capture heterogeneous temporal dependencies that a single instance-level boundary cannot. The ALW-style router is competitive in individual settings but remains less effective on average than the two instance-adaptive routing strategies. These results therefore support modeling both instance-dependent relevance and variable-specific context requirements.

C Discussion and Limitations

Despite its strong empirical performance, **TIMESELECT** still has several limitations that warrant further investigation:

- **Limited adaptability of the fixed temporal bases.** The temporal bases $\mathbf{Q}_L$ and $\mathbf{Q}_H$ are estimated offline from the training split and remain fixed throughout model optimization and inference. This design provides stable temporal coordinates and avoids introducing substantial training overhead. However, the dominant temporal correlations learned from the training data may become less representative when the test distribution exhibits temporal drift, changing periodicity, or previously unseen operating regimes. A possible extension is

Table 7: Comparison of historical-context selection strategies on ETTm1 and Weather. Lower values are better. The best result is shown in **bold**, and the second-best result is underlined.

Dataset	Horizon	Full $L = 96$		Val.-best fixed L		Static variable-wise		Instance-wise shared		ALW		TIMESELECT	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	96	0.305	0.336	0.305	0.337	0.306	0.337	0.305	0.338	0.352	0.388	**0.303**	**0.334**
	192	**0.355**	**0.364**	0.358	0.366	0.359	0.368	0.361	0.371	0.402	0.419	0.356	**0.364**
	336	0.395	0.390	0.397	0.391	0.397	0.390	0.398	0.391	0.449	0.454	**0.388**	**0.386**
	720	0.478	0.438	0.472	0.435	0.477	0.437	0.454	0.426	0.530	0.504	**0.449**	**0.424**
Weather	96	0.152	0.189	0.152	0.189	0.155	0.193	0.154	0.191	0.170	0.218	**0.151**	**0.189**
	192	0.201	0.237	0.202	0.238	0.204	0.238	0.202	0.237	0.229	0.273	**0.200**	**0.235**
	336	**0.260**	**0.278**	0.261	0.280	0.265	0.286	0.266	0.289	0.302	0.330	0.262	0.282
	720	0.341	**0.337**	0.343	0.337	0.344	0.338	0.345	0.339	0.400	0.395	**0.341**	**0.336**

to periodically update the bases using recent observations or introduce a lightweight residual transformation that adapts the fixed bases without reconstructing them from scratch.

- **Contiguous-context assumption.** The current selector represents the effective history of each variable as a continuous interval ending at the forecasting boundary. This formulation is interpretable and computationally efficient, but it assumes that recent observations are always more relevant than earlier ones. In some cases, useful information may occur at several separated historical positions, such as observations from the same hour on previous days or the same weekday in previous weeks. Future work could extend the current soft boundary into a small number of learnable context segments or combine the contiguous mask with sparse periodic anchors, while retaining the compactness of the selected history.
- **Restricted context descriptors.** The effective context length is predicted from a compact set of statistical features describing the level, variability, recent state, overall change, and activation magnitude of each variable. Although these descriptors enable a lightweight selector, they may not fully characterize complex temporal states involving abrupt transitions, local frequency changes, or interactions with other variables. A richer context encoder could incorporate short-range temporal features or cross-variable summaries. Nevertheless, such an extension should remain lightweight so that the context-selection module does not introduce computational costs comparable to those of the forecasting backbone.
- **Shared control of context compactness.** The current training objective uses a global regularization coefficient to penalize the average selected context length. This encourages the model to suppress redundant history, but a single coefficient may not provide the optimal accuracy–compactness trade-off for all datasets, variables, or forecasting instances. Some variables may require long-term trends, whereas others can be predicted primarily from recent observations. Future extensions could introduce variable-dependent or instance-adaptive regularization, allowing the model to retain longer contexts only when they provide measurable forecasting benefits.

Overall, these limitations mainly concern the flexibility of temporal-basis construction, historical-context representation, and cycle selection. Addressing them through lightweight adaptive mechanisms may further improve the robustness of **TIMESELECT** without sacrificing its current computational efficiency.