

Long Input Sequence Network for Long Time Series Forecasting

Chao Ma
HRBUST

Yikai Hou
HRBUST

Xiang Li
HIT

Yinggang Sun
HIT

Haining Yu
HIT

Abstract

Short fixed-length inputs are the main bottleneck of deep learning methods in long time-series forecasting tasks. Prolonging input length causes overfitting, rapidly deteriorating accuracy. Our research indicates that the overfitting is a combination reaction of the multi-scale pattern coupling in time series and the fixed focusing scale of current models. First, we find that the patterns exhibited by a time series across various scales are reflective of its multi-periodic nature, where each scale corresponds to specific period length. Second, We find that the token size predominantly dictates model behavior, as it determines the scale at which the model focuses and the context size it can accommodate. Our idea is to decouple the multi-scale temporal patterns of time series and to model each pattern with its corresponding period length as token size. We introduced a novel series-decomposition module(MPSD), and a Multi-Token Pattern Recognition neural network(MTPR), enabling the model to handle *inputs up to $10\times$ longer*. Sufficient context enhances performance(*38% maximum precision improvement*), and the decoupling approach offers *Low complexity*($0.22\times$ cost) and *high interpretability*. Code: <https://github.com/Houyikai/MTE>

1 Introduction

Accurate long-term time series forecasting is essential for facilitating effective long-term planning and optimizing returns across diverse domains, such as finance[17], energy[7, 25], transportation[20]. Recent advancements in deep neural network methods show promise, particularly in leveraging time series properties to design more suitable neural network architectures. These properties include Multi-Periodicity[23], Seasonality[24, 29], Multi-Scale[21, 1], Segmented[27, 13] and non-Stationarity[14, 12, 8].

However, this advancement has encountered a bottleneck: the best-performing context window size(input sequence length) of these methods is often significantly shorter compared to the prediction length[22]. We tested several current mainstream methods for prolonging input sequences on the ETTm2 dataset, as illustrated in Figure 1(left). These methods fail to benefit from long context window and exhibit performance degradation. According to observations during the experiments, the primary reason for the performance decline is overfitting caused by long context window. This results in insufficient context for the model to effectively learn and generalize.

Our research indicates that the overfitting are the combined result of *multi-scale pattern coupling* within sequences and the *fixed focusing scale* of the model. In sequence perspective, distinct patterns manifest at varying scales. Two relevant sequence properties are multi-scale variations and multi-periodicity. Previous work[21, 23] have leveraged these properties by incorporating down-sampling and multi-period information, thereby enhanced representation capability. However, they failed to decouple the overlapping multi-scale patterns, which led to the small-scale and short-period sequence in the enhanced representation constitutes a short board for model to handle larger context. In model perspective, Transformer-based approaches operate on each segments of inputs, i.e. tokens[2, 3, 26].

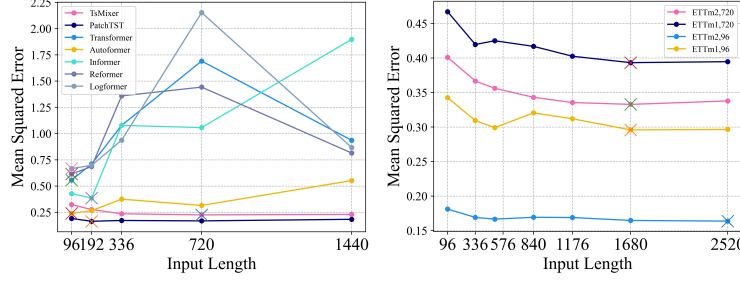


Figure 1: The MSE comparisons of forecasting 96 steps for ETTm2 dataset with prolonging input length. Current method(left) benefits primarily from a fixed short context windows, while our MTE(right) is adaptable to longer context window, thereby yielding benefits in long-term time series forecasting. The cross indicates the best-performing length.

The attention mechanism performs contextual embedding to distinguish different tokens, followed by regressive mapping via a same feed-forward layers. In contrast, MLP-based approaches map the entire input sequence(whole context as a token) directly. Small token size narrow the focusing scale of model, due to limited context provided to the feed-forward layer. Limited context are also insufficient to distinguish tokens, resulting less adaptation to long inputs. Conversely, MLPs exhibit adaptability to longer inputs, but their inherent disregard for local features leads to poorer performance.

We found that **SCALE** serves as a bridge between the characteristics of the sequence and the model. We first propose the **Multi-Periodic series-Decomposition module (MPSD)** to address the *multi-scale pattern coupling* problem. Multi-Periodicity and Multi-Scale of time series are closely related, and downsampling information is actually a long-period pattern. MPSD combines two properties, iteratively extracting periodic patterns (a separate sequence) from the input, from short to long period, while performing downsampling with increasing intensity. Secondly, we propose the **Multi-Token Pattern Recognition neural network (MTPR)** to address the *fixed focusing scale* problem. The MTPR processes each periodic pattern in parallel with corresponding period length as token size. Utilizing multiple token sizes enables the model to analyze the sequence from various scales. It first folds each period sequence based on the period length and utilizes a two-stage attention layer to model intra-period(short-term) and inter-period(long-term) dependencies, then blending them together. Note that in the intra-period stage of MTPR, the token is a *phase*, while in the inter-period stage, the token is a *period*. Combined with MPSD, our tokens encompass phases and periods of varying scales. We name the model **MTE(Multi Token Encoder)**.

Experiment results affirm the state-of-the-art (SOTA) performance of our MTE across various benchmarks. Specifically, our MTE achieves up to a 22% maximum reduction in Mean Squared Error(MSE) over TimeMixer[21] and a 38% maximum reduction over PatchTST[13], the SOTA Transformer baselines, and offers lower memory and computational costs compared to the latter. Our decoupling technique enhances interpretability and uncovers new challenges in large-context scenarios. Our contributions can be summarized as follows:

- We found a positive correlation between the model’ input length and its token size. The token size also guides the model to focus on temporal patterns at a certain scale.
- We introduce the *Multi-Periodic series-Decomposition module(MPSD)* for disentangling complex temporal patterns and the *Multi-Token Pattern Recognition neural network (MTPR)* for observing sequences at different scales.
- our MTE have *long context windows(up to $10\times$)*, *SoTA performance(up to 38% precision upgrade)*, *low cost($0.22\times$ the Transformer baseline)*, and *interpret-ability*.

2 Related Work

Current mainstream methods, mostly based on MLP and Transformer, have addressed the issue of error accumulation inherent in autoregressive prediction approaches, particularly prevalent in recurrent neural networks[18, 16]. Early Transformer variants primarily focused on reducing attention complexity, exemplified by approaches like Autoformer[24] and FEDformer[29], Informer[28] which

introduced Fourier analysis and sequence decomposition methods into time series forecasting. The latest Transformer method, PatchTST[13], adopts strategies such as channel-wise independence and Patch embedding methods, while employing a simple linear head as a predictor or decoder, effectively enhancing prediction accuracy. Analogous to the Vision Transformer (ViT[3]) paradigm in computer vision, PatchTST delineates the Transformer paradigm in the domain of time series prediction. MLP-based techniques often leverage inductive biases derived from time series analysis, as seen in methods like N-BEATS[15] and N-HITS[1], which employ ensemble methods and hierarchical pooling techniques. Recent advancements such as Non-stationary[12], RevIN[8], TimesNet[23], TSMixer[4] and TimeMixer[21] utilize the periodicity, multivariate and multiscale characteristics of time series, respectively, to increase expressiveness. However, the best-performing context window size of these methods is often significantly shorter compared to the prediction, which has become a major bottleneck for long-term time series forecasting[22].

This prompts a question: **What controls the input size of neural networks for time series forecasting?** Through a comparative analysis of the design architectures and basic components of current mainstream methods, we find a positive correlation between the size of token(the minimum processing units of the model) and the input length. Models with longer tokens can handle longer input sequences. For instance, Autoformer[24] and FEDformer[29] treat each individual time point as a token, and their optimal context window size is 96. PatchTST[13] and Crossformer[27] utilize fragments ranging in size from 8 to 16 as tokens, extending their optimal context window to 336. When the entire sequence is processed using an MLP, such as TSMixer[4], Nbeat[15], and Nhits[1], the context window can expand to 512. Another important finding is that *token size also affects, dominant at some extent, the behavior of the model*. Models with smaller tokens[10, 9, 28, 29, 13] fit local variations more accurately, such as subtle periodic fluctuations, while models with larger tokens ignore subtle fluctuations and fit better to overall distribution features, such as mean and variance[15, 1, 4]. Based on the above findings, we arrive at the answer to the initial question: *Token size controls the input size that the model can accommodate*. However, long tokens may lead the model to overlook local variations. A model that simultaneously processes tokens of multiple lengths to attend to variations of different scales might be the optimal solution.

How should the token size be determined? In the research of self-supervised learning for natural language processing[19, 2] and computer vision[3, 6], a patch or token should contain semantic information, which naturally leads us to the periodicity of time series. We initiate our analysis by examining the *Multi-Periodicity* of the time series, a concept previously explored in TimesNet[23]. In this framework, sequences can be understood as overlapping periods with different scales, where the scale refers to the intervals at which events recur, such as hourly, daily, or weekly periods. Real-world phenomena are influenced by these scales, with distinct patterns emerging at different scales. It uses Fourier analysis to find the potential period lengths of the sequence, which can be used to determine the token size of our model.

3 Methodologies

Preliminaries. In a Deep learning paradigm for time series forecasting[5], given a multivariate historical sequence $x \in \mathbb{R}^{M \times L}$, where L is the context windows size (or input sequence length), the objective of long time series forecasting is to predict a future sequence $y \in \mathbb{R}^{M \times H}$, where H is the predict length. The multivariate time series comprises multiple dimensions, where each dimension $i \in \{1, 2, \dots, M\}$ represents a separate time series $x^{(i)} \in \mathbb{R}^L$, which was referred as a channel. our MTE handles multiple channels simultaneously, but ignores potential correlations between channels, since cross-dimension dependency[4, 27] is not the focus of this paper.

Overview. our MTE comprises four steps: (3.1)Employing Fourier analysis, the top k period lengths are identified based on the multi-periodicity of the sequence. (3.2)Recursively extracting different-scale periodic patterns from the sequence, from short to long, to form separate period sequences. (3.3)For each period sequence, its period length is used as the token size to specify an individual module for pattern recognition. (3.4)Employing separate predictors for each pattern, prediction and interpolation(corresponding pooling) are conducted, then combined to form the result. The overall framework is depicted in the Figure 2.

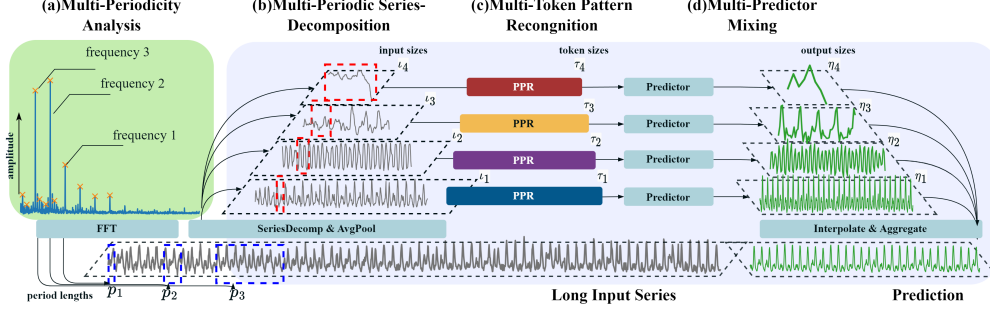


Figure 2: Overall Framework.

3.1 Multi-Periodicity Analysis

Time series exhibit multi-periodicity, as illustrated in Figure 2. The first step of our model involves identifying the different period lengths within the sequence, a task conventionally accomplished through frequency domain analysis. Initially, we apply the Fast Fourier Transform (FFT) to convert the time series from the time domain to the frequency domain, aiming to extract frequencies that exhibit significant periodicity. Due to the sparsity of the frequency domain and the potential noise introduced by high-frequency components, we select only the top k frequencies $\{f_1, f_2, \dots, f_k\}$ with the highest amplitude values \mathbf{A} . Given the conjugate nature of the frequency domain, we focus on frequencies below $\lfloor L/2 \rfloor$, where L denotes the length of the input sequence. In this context, the period is defined as the reciprocal of the frequency. Specifically, we utilize the *Period* module in TimesNet[23] to select the periods $\{p_1, \dots, p_k\}$ corresponding to the reciprocals of the top k amplitudes (The Period module is not mandatory as, in fact, it is unstable. We recommend presetting hyperparameters).

$$\mathbf{A}, \{f_1, \dots, f_k\}, \{p_1, \dots, p_k\} = \text{Period}(x). \quad (1)$$

3.2 Multi-Periodic Series-Decomposition

To enable the model focusing simultaneously on multiple periodic patterns at different scale, it is imperative to disentangle the overlapping multi-periodic patterns within the sequence. Multi-Periodic Series-Decomposition module recursive extracts periodic patterns in a sequence base on Multi-Periodicity Analysis results $\{p_1, \dots, p_k\}$. According to the nature of the sequence decomposition algorithm, the reasonable extraction order is based on the period lengths from short to long. Therefore, we first sort the periods:

$$\{p_1, \dots, p_k\} = \text{Asc}(\{p_1, \dots, p_k\}), \quad (2)$$

Here, $\text{Asc}(\cdot)$ denote the ascending ordering. Base on the period lengths $\{p_1, \dots, p_k\}$, the Series-Decomposition module[24] is used to extract larger-scale period features from the trend components decomposed from preceding round, in a recursive manner, from short to long period:

$$\mathbf{s}_{j+1}, \mathbf{t}_{j+1} = \text{SeriesDecomp}_{\omega=p_{j+1}, v=1}(\mathbf{t}_j), j \in \{0, \dots, k-1\}, \mathbf{t}_0 = x, \quad (3)$$

$\mathbf{s}_j \in \mathbb{R}^{M \times L}$ and $\mathbf{t}_j \in \mathbb{R}^{M \times L}$ are the season and trend component of round j respectively, where in the first round the trend components \mathbf{t} is just input sequence x . $\text{SeriesDecomp}_{\omega, v}(\cdot)$ denote the Series-Decomposition module and its subscripts ω, v represent the sliding window size and step size, respectively. Given that long period features typically entail less information, average pooling, $\text{AvgPool}_{\omega, v}(\cdot)$, is applied, wherein the period length of preceding round p_{j-1} are used as parameters to ω and v , thereby mitigating redundancy:

$$\begin{aligned} \tilde{\mathbf{s}}_j &= \text{AvgPool}_{\omega=p_{j-1}/2, v=p_{j-1}/2}(\mathbf{s}_j), j \in \{1, \dots, k\}, p_0 = 2, \\ \tilde{\mathbf{s}}_{k+1} &= \tilde{\mathbf{t}}_k = \text{AvgPool}_{\omega=p_k/2, v=p_k/2}(\mathbf{t}_k), \end{aligned} \quad (4)$$

In addition to the season component $\tilde{\mathbf{s}}_j \in \mathbb{R}^{M \times \ell_j}$ for each round, the trend component $\tilde{\mathbf{t}}_k$ for the last round is also retained as $\tilde{\mathbf{s}}_{k+1}$. We have a series group $\{\tilde{\mathbf{s}}_1, \dots, \tilde{\mathbf{s}}_{k+1}\}$ as the final result of this module, and each size is $\ell_j = \lceil \frac{L-\omega_j}{\delta_j} \rceil + 1$, and their corresponding period length is $\tau_j = \frac{\ell_j}{L} p_j$ and

prediction length is $\eta_j = \lceil \frac{H-\omega_j}{\delta_j} \rceil + 1$ after pooling, where ω_j, δ_j is the parameters for the j -th round of pooling.

$$\{\iota_1, \dots, \iota_{k+1}\}, \{\eta_1, \dots, \eta_{k+1}\}, \{\tau_1, \dots, \tau_{k+1}\}, \{\tilde{s}_1, \dots, \tilde{s}_{k+1}\} = \text{MPSD}(x). \quad (5)$$

3.3 Multi-Token Pattern Recognition

As shown in Figure 2, we organize multiple Periodic Pattern Recognition(PPR) modules in parallel for $k+1$ period sequences. Each module uses a set of parameters $(\iota_j, \eta_j, \tau_j)$ to define for an input sequence \tilde{s}_j , which can be formalized as:

$$\mathbf{z}_j = \text{PPR}_{(\iota_j, \eta_j, \tau_j)}(\tilde{s}_j), j \in \{1, \dots, k+1\}, \quad (6)$$

where $\text{PPR}_{(\iota_j, \eta_j, \tau_j)}(\cdot)$ use ι_j, η_j as input and output length, and τ_j as token size, containing two main modules, Period & Phase Embedding and Two-stage Attention Encoding Layer.

Period & Phase Embedding. For each sequence \tilde{s}_j , we use its period length τ_j as the token size for embedding. We first truncate \tilde{s}_j to length $\rho\tau_j$, where ρ is the preset maximum number of periods, to prevent small-scale sequences from becoming a bottleneck. To allow focusing on the changes between different periods, namely *inter-period* and the changes between different phases of each period, namely *intra-period*, we need to embed each *phase* and *period* adequately. Firstly, we fold the sequence \tilde{s}_j according to the parameter τ_j :

$$\mathbf{S}_j = \text{Fold}_{\omega=\tau_j, v=\tau_j}(\text{Padding}(\tilde{s}_j)), \quad (7)$$

where $\text{Padding}(\cdot)$ is to lineup the sequence for folding $\text{Fold}_{\omega, v}(\cdot)$. $\mathbf{S}_j \in \mathbb{R}^{M \times (\iota_j/\tau_j) \times \tau_j}$ is 2D form of original 1D sequence, where each row is a period, and each column is a phase. Afterwards, linear mapping $W^{(intra)} \in \mathbb{R}^{d \times (\iota_j/\tau_j)}$, $W^{(inter)} \in \mathbb{R}^{d \times \tau_j}$ is applied to each period and phase separately, and position encoding $W_1^{(pos)} \in \mathbb{R}^{\tau_j}$, $W_2^{(pos)} \in \mathbb{R}^{(\iota_j/\tau_j)}$ is added to obtain the embedding of phases $\sigma_i \in \mathbb{R}^{M \times \tau_j \times d}$, and periods $\mu_i \in \mathbb{R}^{M \times (\iota_j/\tau_j) \times d}$:

$$\begin{aligned} \sigma_i &= W^{(intra)}\mathbf{S}_j + W_1^{(pos)}, \\ \mu_i &= W^{(inter)}\mathbf{S}_j^T + W_2^{(pos)}. \end{aligned} \quad (8)$$

Two-stage attention encoding layer. Transformer encoders have been widely acknowledged as versatile seq2seq approximators[26]. To capture both inter-period dependencies and intra-period dependencies effectively, we adopt a two-stage methodology. In the initial stage, self-attention encoders are deployed to independently model two dependencies:

$$\begin{aligned} \hat{\sigma}_j &= \text{MLP}^{intra}\left(\text{MSA}^{intra}(\sigma_j, \sigma_j, \sigma_j)\right), \\ \hat{\mu}_j &= \text{MLP}^{inter}\left(\text{MSA}^{inter}(\mu_j, \mu_j, \mu_j)\right). \end{aligned} \quad (9)$$

In the second stage, cross-attention encoders are employed to integrate these two types of dependencies:

$$\begin{aligned} \hat{\sigma}_j^{(mix)} &= \text{MLP}_1^{cross}\left(\text{MSA}_1^{cross}(\hat{\sigma}_j, \hat{\mu}_j, \hat{\mu}_j)\right), \\ \hat{\mu}_j^{(mix)} &= \text{MLP}_2^{cross}\left(\text{MSA}_2^{cross}(\hat{\mu}_j, \hat{\sigma}_j, \hat{\sigma}_j)\right), \end{aligned} \quad (10)$$

where $\text{MSA}(\cdot)$ is Multi-head dot-product attention layer and $\text{MLP}(\cdot)$ is the feed-forward layer(identical to MLP). For clarity, we omitted the residual connections and normalization layers between each layers. We will concatenate the output of the last layer for final encoding:

$$\mathbf{z}_j = [\hat{\sigma}_j^{(mix)}, \hat{\mu}_j^{(mix)}] \quad (11)$$

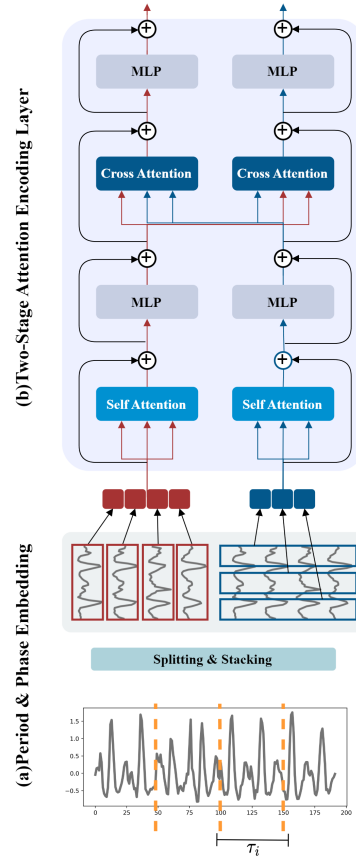


Figure 3: The period pattern recognition module.

3.4 Multi-Predictor Mixing

For each period sequence \tilde{s}_j , we obtained a unique encoding $\mathbf{z}_j \in \mathbb{R}^{M \times (\iota_j/\tau_j + \iota_j) \times d}$:

$$\mathbf{z}_j = \text{PPR}(\tilde{s}_j), j \in \{1, \dots, k+1\}, \quad (12)$$

Considering the distinction of each patterns, we use separate predictor for each encoding. Each predictor is linear layer $W \in \mathbb{R}^{(\iota_j/\tau_j + \iota_j) \times \eta_j}$. Then let each predict shorter sequences $\mathbf{y}_j \in \mathbb{R}^{M \times \eta_j}$ then use interpolation to complete them, to induce multi-scale hierarchical time series forecasts[1]. We get the final prediction y as following:

$$\begin{aligned} \mathbf{y}_j &= \text{Predictor}_j(\mathbf{z}_j), j \in \{1, \dots, k+1\}, \\ y &= \sum_{j=0}^{k+1} \text{Interpolate}_{\tau_j \rightarrow L}(\mathbf{y}_j), \end{aligned} \quad (13)$$

4 Experiments

Benchmarks. We conduct experiments of our model on eight main-stream benchmarks, including Weather, Traffic, Electricity, Solar-energy and 4 ETT datasets(ETTh1, ETTh2, ETTm1, ETTm2), which have been extensively used in previous works[28, 24, 29, 13, 11] and publicly available at [24]. Training/Validation/Test sets are zero-mean normalized with the mean and std of Training set. The Statistics of all benchmarks are gathered in table 1.

Table 1: Datasets Statistics

Datasets	Electricity	Traffic	Weather	Solar-Energy	ETTh1&ETTh2	ETTh1&ETTh2
Time-Series	321	862	21	137	7	7
Time-Points	26,304	17,544	52,696	52,560	69,680	17,420
Forecastability	0.77	0.68	0.75	0.33	0.46	0.46
Frequency	1 Hour	1 Hour	10 Minutes	10 Minutes	15 Minutes	1 Hour

Baselines. We selected eight popular State of The Art(SoTA) models as baselines, including *FEDformer*[29], *Autoformer*[24], *Informer*[28], *Non-Stationary*[12], *TimesNet*[23], *TimesMixer*[21], *Crossformer*[27] and *PatchTST*[13]. TimesMixer(MLP-based) is current SoTA baseline, and PatchTST is Transformer-based SoTA baseline.

Setup. We follow the experimental setup of mainstream methods [13]. The input length is set to 960 or 1680 most case, sometimes 336 for short-term forecasting, according to the multi-periodicity. And the prediction length is varied with $H = \{96, 192, 336, 720\}$. We utilize the Adam optimizer with Mean Squared Error($\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\mathbf{y} - \hat{\mathbf{y}})^2$) as the loss function and evaluate using Mean Absolute Error($\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\mathbf{y} - \hat{\mathbf{y}}|$) and MSE as metrics.

4.1 Main Results

All results are shown in Table 2. Other methods result employ a context window size of 96, and its cited from previous work[21]. our MTE utilizes an optimal window size ranging from 960 to 1680(some times 336), determined by the sequence’s periodicity. Certain methods may exhibit better performance with larger context window sizes, such as PatchTST. However, this enhancement comes at the cost of a quadratic increase in complexity as the window size expands (see Picture 4).

our MTE performs outstandingly in most cases, with an improvement of 11.43% compared to the MLP baseline TimeMixer and approximately 20.17% compared to the Transformer baseline PatchTST, in MSE. our MTE performs particularly well in long-term forecasting scenarios. For example, on large time series datasets such as Traffic, ECL, and Solar, when the prediction amplitude is 336 or above, the average improvement compared to MLP baseline is 15.13%, and the average improvement compared to Transformer baseline is 27.56%, and it shows an upward trend. The driving force for improvement lies in a larger context. In ETTm, Weather, and Solar Energy datasets, we used a context window of 960(10 days), while in ETTh, Traffic, and Electricity, the context window was 1680(10 weeks), which is about $10 \times$ to $20 \times$ more than previous work.

Table 2: Long Time Series Forecasting Results.

Models	MTE (Ours)		TimeMix (2024)		PatchTST (2023)		TimesNet (2023)		Crossformer (2023)		FEDformer (2022)		Stationary (2022)		Autoformer (2021)		Informer (2021)		
	Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	0.150	0.202	0.163	0.209	0.186	0.227	0.172	0.220	0.195	0.271	0.217	0.296	0.173	0.223	0.266	0.336	0.300	0.384
	192	0.194	0.245	0.208	0.250	0.234	0.265	0.219	0.261	0.209	0.277	0.276	0.336	0.245	0.285	0.307	0.367	0.598	0.544
	336	0.241	0.283	0.251	0.287	0.284	0.301	0.246	0.337	0.273	0.332	0.339	0.380	0.321	0.338	0.359	0.395	0.578	0.523
	720	0.302	0.331	0.339	0.341	0.356	0.349	0.365	0.359	0.379	0.401	0.403	0.428	0.414	0.410	0.419	0.428	1.059	0.741
Solar	96	0.167	0.240	0.189	0.259	0.265	0.323	0.373	0.358	0.232	0.302	0.286	0.341	0.321	0.380	0.456	0.446	0.287	0.323
	192	0.180	0.248	0.222	0.283	0.288	0.332	0.397	0.376	0.371	0.410	0.291	0.337	0.346	0.369	0.588	0.561	0.297	0.341
	336	0.186	0.253	0.231	0.292	0.301	0.339	0.420	0.380	0.495	0.515	0.354	0.416	0.357	0.387	0.595	0.588	0.367	0.429
	720	0.197	0.277	0.223	0.285	0.295	0.336	0.420	0.381	0.526	0.542	0.380	0.437	0.375	0.424	0.733	0.633	0.374	0.431
ECL	96	0.130	0.229	0.153	0.247	0.190	0.296	0.168	0.272	0.219	0.314	0.193	0.308	0.169	0.273	0.201	0.317	0.274	0.368
	192	0.148	0.247	0.166	0.256	0.199	0.304	0.184	0.322	0.231	0.322	0.201	0.315	0.182	0.286	0.222	0.334	0.296	0.386
	336	0.165	0.263	0.185	0.277	0.217	0.319	0.198	0.300	0.246	0.337	0.214	0.329	0.200	0.304	0.231	0.443	0.300	0.394
	720	0.197	0.290	0.225	0.310	0.258	0.352	0.220	0.320	0.280	0.363	0.246	0.355	0.222	0.321	0.254	0.361	0.373	0.439
Traffic	96	0.361	0.255	0.462	0.285	0.526	0.347	0.593	0.321	0.644	0.429	0.587	0.366	0.612	0.338	0.613	0.388	0.719	0.391
	192	0.380	0.266	0.473	0.296	0.522	0.332	0.617	0.336	0.665	0.431	0.604	0.373	0.613	0.340	0.616	0.382	0.696	0.379
	336	0.392	0.272	0.498	0.296	0.517	0.334	0.629	0.336	0.674	0.420	0.621	0.383	0.618	0.328	0.622	0.337	0.777	0.420
	720	0.430	0.295	0.506	0.313	0.552	0.352	0.640	0.350	0.683	0.424	0.626	0.382	0.653	0.355	0.660	0.408	0.864	0.472
ETTh1	96	0.373	0.399	0.375	0.400	0.460	0.447	0.384	0.402	0.423	0.448	0.395	0.424	0.513	0.491	0.449	0.459	0.865	0.713
	192	0.406	0.417	0.429	0.421	0.512	0.477	0.436	0.429	0.471	0.474	0.469	0.470	0.534	0.504	0.500	0.482	1.008	0.792
	336	0.420	0.426	0.484	0.458	0.546	0.496	0.638	0.469	0.570	0.546	0.530	0.499	0.588	0.535	0.521	0.496	1.107	0.809
	720	0.432	0.451	0.498	0.482	0.544	0.517	0.521	0.500	0.653	0.621	0.598	0.544	0.643	0.616	0.514	0.512	1.181	0.865
ETTh2	96	0.277	0.337	0.289	0.341	0.308	0.355	0.340	0.374	0.745	0.584	0.358	0.397	0.476	0.458	0.346	0.388	3.755	1.525
	192	0.338	0.378	0.372	0.392	0.393	0.405	0.402	0.414	0.877	0.656	0.429	0.439	0.512	0.493	0.456	0.452	5.602	1.931
	336	0.327	0.382	0.386	0.414	0.427	0.436	0.452	0.452	1.043	0.731	0.496	0.487	0.552	0.551	0.482	0.486	4.721	1.835
	720	0.376	0.419	0.412	0.434	0.436	0.450	0.462	0.468	1.104	0.763	0.463	0.474	0.562	0.560	0.515	0.511	3.647	1.625
ETTm1	96	0.288	0.348	0.320	0.357	0.352	0.374	0.338	0.375	0.404	0.426	0.379	0.419	0.386	0.398	0.505	0.475	0.672	0.571
	192	0.332	0.370	0.361	0.381	0.390	0.393	0.374	0.387	0.450	0.451	0.426	0.441	0.459	0.444	0.553	0.496	0.795	0.669
	336	0.361	0.387	0.390	0.404	0.421	0.414	0.410	0.411	0.532	0.515	0.445	0.459	0.495	0.464	0.621	0.537	1.212	0.871
	720	0.400	0.408	0.454	0.441	0.462	0.449	0.478	0.450	0.666	0.589	0.543	0.490	0.585	0.516	0.671	0.561	1.166	0.823
ETTm2	96	0.164	0.252	0.175	0.258	0.183	0.270	0.187	0.267	0.287	0.366	0.203	0.287	0.192	0.274	0.255	0.339	0.365	0.453
	192	0.218	0.298	0.237	0.299	0.255	0.314	0.249	0.309	0.414	0.492	0.269	0.328	0.280	0.339	0.281	0.340	0.533	0.563
	336	0.265	0.330	0.298	0.340	0.309	0.347	0.321	0.351	0.597	0.542	0.325	0.366	0.334	0.361	0.339	0.372	1.363	0.887
	720	0.331	0.374	0.391	0.396	0.412	0.404	0.408	0.403	1.730	1.042	0.421	0.415	0.417	0.413	0.433	0.432	3.379	1.338

4.2 Ablation

We conducted detailed ablation study on all modules from the model, the main results are shown in Table 3. From aba3, it can be seen that the Intra-period module works fine on its own when in short-term forecasting scenario. This confirms the previous argument that the small token model focuses more on local variations, since each token in the Intra-period module is actually a time point. However, in long-term forecasting scenarios, it cannot continue to be competent, which led us to the inter-period module. From aba4, the cross-period module itself performs mediocorely. The fact that Inter-period module is actually a standard paradigm of Transformer, encourages us to explore innovative structures beyond conventional framework. On the other hand, from aba5, it can be seen that the decomposition module itself can significantly improve the performance of long-term forecasting. Simply replacing the subsequent PPR module, it can be used for any method. Lastly, aba2 and aba3 prove that complex neural networks cannot function properly without the guidance of a coherent data flow structure.

Table 3: Ablation Results

components	MP-SeriesDecomp	Intra-Period	Inter-Period	Cross-Attn	ETTh1-96		ETTh1-192		ETTh1-336		ETTh1-720	
					MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
baseline	1	1	1	1	0.373	0.399	0.406	0.417	0.42	0.426	0.432	0.451
aba1	0	1	1	1	0.386	0.411	0.418	0.434	0.431	0.448	0.482	0.486
aba2	0	1	1	0	0.38	0.407	0.413	0.429	0.431	0.449	0.483	0.487
aba3	0	1	0	0	0.382	0.41	0.407	0.425	0.42	0.441	0.463	0.475
aba4	0	0	1	0	0.4	0.428	0.426	0.444	0.438	0.457	0.473	0.486
aba5	1	1	0	0	<u>0.377</u>	<u>0.407</u>	<u>0.411</u>	<u>0.427</u>	<u>0.441</u>	<u>0.451</u>	<u>0.452</u>	<u>0.471</u>

4.3 Efficiency Analysis

In long input scenarios, the most concerning issue is perhaps the efficiency of the model. Therefore, we conducted a detailed analysis of usability to the model by examining both its computational complexity and its real-world time and memory cost.

Complexity. We also analyzed the computational complexity, as detailed in Table 4. As a benchmark, the computational complexity of a Transformer scales quadratically with the input length L . Early methods, such as Informer, Autoformer, and Fedformer, process all sequence dimensions

simultaneously within a single token, making their complexity unaffected by the input sequence dimension M . This approach neglects inter-channel differences, leading to suboptimal performance. In contrast, similar to other recent methods, our MTE adopts a channel-independent strategy to handle multivariate sequences.

Furthermore, we observed that FEDformer exhibits a very low complexity, which scale linearly with the input length. However, it introduces FFT and RFFT at each layer to achieve this reduction in complexity, which significantly slows down the actual processing speed (we only compute FFT once at the beginning). Crossformer and PatchTST employ segmented embedding, resulting in a complexity that is quadratic with respect to the number of tokens (L_{seg} , where p is the token size and s is the stride), which is significantly lower than the quadratic complexity relative to the input length. our MTE broadly classified to this approach but is correlated to the maximum number of periods ρ or the maximum number of period length p_{max} . Due to pooling and truncation, these values are relatively small and do not increase significantly with the input length, thereby keeping the overall cost manageable.

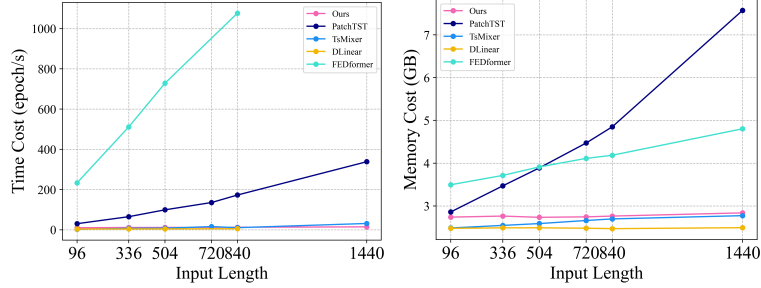


Figure 4: Time cost and memory cost (ETTm2-96).

Efficiency. Based on the above analysis, lower complexity on paper may not necessarily lead to proportionally lower costs. As a supplementary analysis to the complexity assessment, we conducted experiments to measure the training time and memory usage (nvidia-smi) of different methods. Using the ETTm2 dataset with a fixed prediction length of 96, the results are depicted in Figure 4. our MTE demonstrates lower time and memory costs compared to Transformer-based approaches, and is approximately on par with linear and MLP models. This indicates that our MTE remains more practical compared to other Transformer models even in the case of 10 times increased the input sizes.

Table 4: Complexity

Model	Complexity
Transformer	$O(L^2)$
Informer	$O(L \log L)$
Autoformer	$O(L \log L)$
FEDformer	$O(L)$
Crossformer	$O\left(M \left(\frac{L}{L_{seg}}\right)^2\right)$
PatchTST	$O\left(M \left(\left\lfloor \frac{L-p}{s} \right\rfloor + 2\right)^2\right)$
Ours	$O\left(M \max(\rho, p_{max})^2\right)$

Table 5: Sensitivity Results

Dataset	metric	ETTm2-96		ETTm2-96		ETTm2-720		ETTm2-720	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
L	96	0.296	0.342	0.184	0.184	0.415	0.433	0.397	0.399
	336	0.283	0.341	0.171	0.258	0.378	0.419	0.374	0.391
	960	0.283	0.339	0.165	0.254	0.416	0.448	0.337	0.373
	1680	0.279	0.34	0.164	0.252	0.417	0.456	0.331	0.374
ρ	4	0.274	0.347	0.167	0.259	0.403	0.449	0.343	0.381
	8	0.272	0.342	0.170	0.266	0.411	0.452	0.345	0.383
	16	0.276	0.347	0.164	0.260	0.417	0.456	0.331	0.374
	32	0.278	0.350	0.175	0.268	0.410	0.451	0.330	0.371

Hyper-parameter Sensitivity Analysis. our MTE sometimes encounters high errors with long inputs (Table 5 ETTm2-720), and we will carefully analyze the reasons for this phenomenon in Section 4.4. The sensitivity of the maximum number of periods ρ was tested with a fixed input length of 1680, showing no significant impact on performance, which is typically set to 16.

4.4 Model Interpretation Analysis

Output of our model can be expanded to observe the prediction results of each period component, providing a way to understand the behavior patterns of the model, as pattern of each component exhibits sufficient regularity. As shown in Figure 5a, when there are no outliers in the context, everything works ideally. However, our MTE is sensitive to outliers, and a long context window

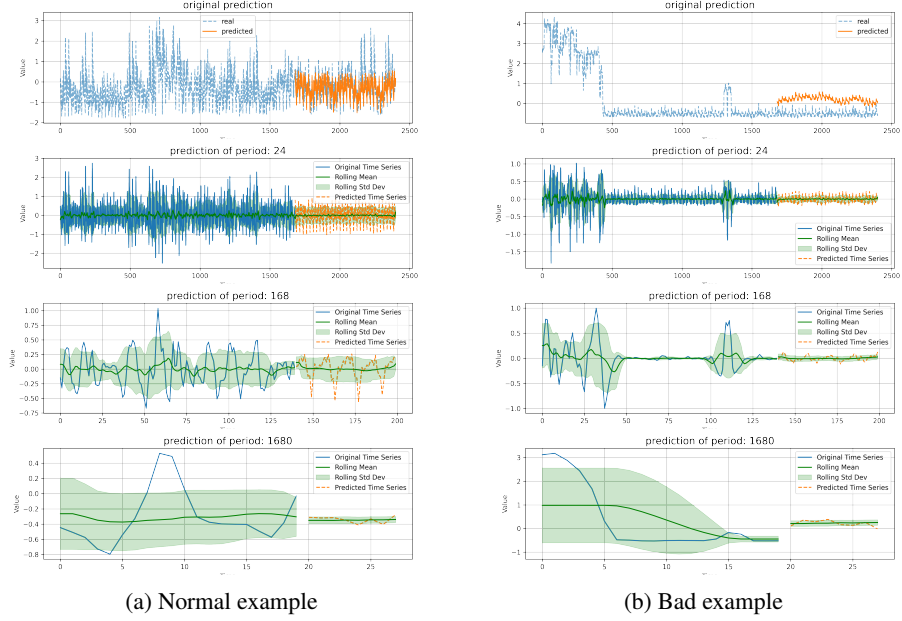


Figure 5: Expand Forecasting Results.

further exacerbates this issue (Picture 5b). In the prediction of period 168 of the bad case, the variance of the predicted values did not follow the expected decreasing trend but remained relatively high, which we refer to as *variance shift*. Similarly, in the prediction of period 1680, *mean shift* was observed. These issues reflect the same problem in current deep learning-based time series forecasting methods, which *focus on overall statistical significance patterns rather than temporal variation*. Maybe introducing higher order derivatives pattern recognition may help this problem, which will leaving to the future exploration.

5 Conclusion

Neural networks with a fixed token size tend to focus on temporal pattern at a dominant scale, which is usually smaller. Focusing on minor scale changes further leads to a smaller context window to prevent overfitting. Since the token size essentially determines the scale to which the model is focusing, the design of the embedding layer is crucial. Therefore, unlike previous work, the decoupling method and multi-scale recognition module proposed in this paper revolve around how to reasonably embed the time series, rather than focusing on the encoding layers. On the other hand, we have noticed that in the self-supervised learning methods of Transformers for natural language processing and image recognition tasks, a patch or token should have some semantic meaning. So in time series, we naturally thought of periodicity and used the pattern of a period as a token. By combining Fourier analysis of the multi-periodicity of sequences, we have implemented a neural network that processes multiple tokens in parallel, thereby enabling multi-scale processing also long context.

Our work altered the behavioral pattern of time series forecasting methods enabling the forecasted results to exhibit rich temporal patterns rather than a simple repetition of a specific scale variation. We hope these findings will inspire future work.

References

- [1] Cristian Challu, Kin G Olivares, Boris N Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. Nhits: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6989–6997, 2023. 1, 2, 3, 4
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2
- [3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2
- [4] 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*, pages 459–469, 2023. 2, 3
- [5] Alberto Gasparin, Slobodan Lukovic, and Cesare Alippi. Deep learning for time series forecasting: The electric load case. *CAAI Transactions on Intelligence Technology*, 7(1):1–25, 2022. 3
- [6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2
- [7] Rob J Hyndman and Shu Fan. Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems*, 25(2):1142–1153, 2009. 1
- [8] 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. 1, 2
- [9] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020. 2
- [10] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyu Zhou, Wenhui Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in neural information processing systems*, 32, 2019. 2
- [11] 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. 4
- [12] Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in Neural Information Processing Systems*, 35:9881–9893, 2022. 1, 2, 4
- [13] 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. 1, 1, 2, 4, 4
- [14] Eduardo Ogasawara, Leonardo C Martinez, Daniel De Oliveira, Geraldo Zimbrão, Gisele L Pappa, and Marta Mattoso. Adaptive normalization: A novel data normalization approach for non-stationary time series. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2010. 1
- [15] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*, 2019. 2
- [16] Alaa Sagheer and Mostafa Kotb. Time series forecasting of petroleum production using deep lstm recurrent networks. *Neurocomputing*, 323:203–213, 2019. 2
- [17] Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied soft computing*, 90:106181, 2020. 1

- [18] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. A comparison of arima and lstm in forecasting time series. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 1394–1401. IEEE, 2018. 2
- [19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2
- [20] Eleni I Vlahogianni, Matthew G Karlaftis, and John C Golias. Short-term traffic forecasting: Where we are and where we’re going. *Transportation Research Part C: Emerging Technologies*, 43:3–19, 2014. 1
- [21] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and JUN ZHOU. Timemixer: Decomposable multiscale mixing for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 1, 2, 4, 4.1
- [22] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*, 2022. 1, 2
- [23] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The eleventh international conference on learning representations*, 2022. 1, 2, 3.1, 4
- [24] 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. 1, 2, 3.2, 4, 4
- [25] Changtian Ying, Weiqing Wang, Jiong Yu, Qi Li, Donghua Yu, and Jianhua Liu. Deep learning for renewable energy forecasting: A taxonomy, and systematic literature review. *Journal of Cleaner Production*, 384:135414, 2023. 1
- [26] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*, 2019. 1, 3.3
- [27] Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2022. 1, 2, 3, 4
- [28] 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, pages 11106–11115, 2021. 2, 4, 4
- [29] 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*, pages 27268–27286. PMLR, 2022. 1, 2, 4, 4