

Amplifier: Bringing Attention to Neglected Low-Energy Components in Time Series Forecasting

Jingru Fei¹, Kun Yi², Wei Fan³, Qi Zhang⁴, Zhendong Niu^{1*}

¹Beijing Institute of Technology

²State Information Center of China

³University of Oxford

⁴Tongji University

{jingrufei, zniu}@bit.edu.cn, kunyi.cn@gmail.com, weifan.oxford@gmail.com, zhangqi_cs@tongji.edu.cn

Abstract

We propose an *energy amplification technique* to address the issue that existing models easily overlook low-energy components in time series forecasting. This technique comprises an *energy amplification block* and an *energy restoration block*. The energy amplification block enhances the energy of low-energy components to improve the model’s learning efficiency for these components, while the energy restoration block returns the energy to its original level. Moreover, considering that the energy-amplified data typically displays two distinct energy peaks in the frequency spectrum, we integrate the energy amplification technique with a seasonal-trend forecaster to model the temporal relationships of these two peaks independently, serving as the backbone for our proposed model, *Amplifier*. Additionally, we propose a *semi-channel interaction temporal relationship enhancement block* for Amplifier, which enhances the model’s ability to capture temporal relationships from the perspective of the *commonality* and *specificity* of each channel in the data. Extensive experiments on eight time series forecasting benchmarks consistently demonstrate our model’s superiority in both effectiveness and efficiency compared to state-of-the-art methods.

Code — <https://github.com/aikunyi/Amplifier>

1 Introduction

Time series forecasting holds significant importance in real-life applications, encompassing various fields such as financial markets (Yi et al. 2024d), weather forecasting (Yi et al. 2024a), traffic flow prediction (Yu, Yin, and Zhu 2018; Fan et al. 2022), energy planning (Yi et al. 2024b). Recently, the rapid advancement of deep learning has given rise to various models for time series forecasting, including RNN-based methods (e.g., LSTNet (Lai et al. 2018), DeepAR (Salinas et al. 2020)), TCN-based methods (e.g., SCINet (Liu et al. 2022a), TimesNet (Wu et al. 2023)), Transformer-based methods (e.g., PatchTST (Nie et al. 2023), iTransformer (Liu et al. 2024)), and Linear-based methods (e.g., DLinear (Zeng et al. 2023), RLinear (Li et al. 2023)), etc.

Although these deep learning methods have demonstrated competitive performance across various scenarios, they possess several inherent drawbacks in their ability to learn from

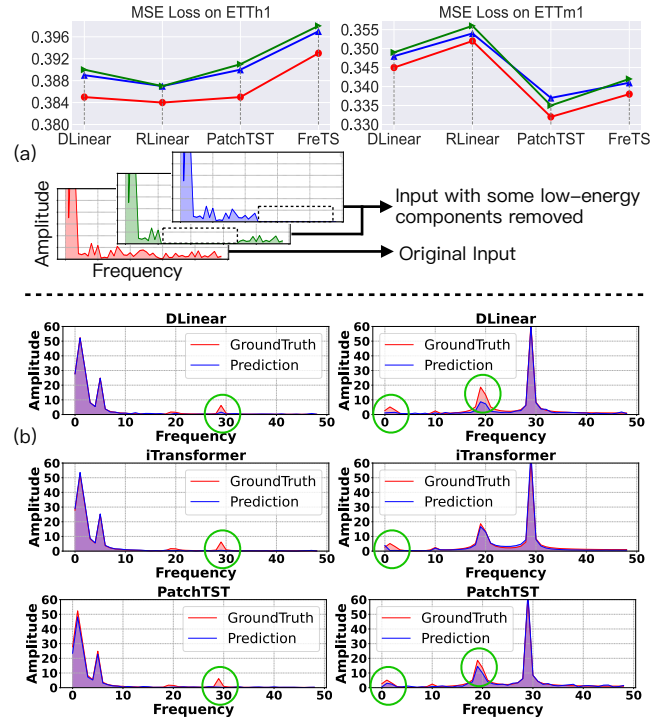


Figure 1: Analysis about low-energy components in time series forecasting. (a) *Indispensability*: Discarding low-energy components results in an increased MSE value. (b) *Dependence on energy magnitude*: The components that are ignored are consistently those with low energy, irrespective of their position within the frequency bands.

different energy components, and tend to **focus on learning high-energy components while neglecting low-energy components**, leading to an incomplete utilization of the informative aspects crucial for time series forecasting.

What are low-energy components? In signal processing and analysis, the energy of a signal refers to the total amount of signal strength or power (Lathi and Green 1998). Low-energy components in a signal refer to frequency components with smaller amplitudes within the frequency spectrum (Oppenheim 1999).

Low-energy components are indispensable. In contexts like signal compression or model simplification, low-energy components might be dismissed as unimportant or treated

*Corresponding author

as noise. However, in scenarios where subtle changes or background patterns are significant, low-energy components can hold critical details. For example, in weather forecasting, small meteorological shifts can accumulate over time, leading to dramatic changes in weather patterns. In financial markets, minor fluctuations can trigger larger trends or volatility, with high-frequency trading systems often capitalizing on these small changes to execute profitable trades. We also have empirically validated the indispensability of low-energy components in time series forecasting. As illustrated in Figure 1(a), directly filtering out low-energy components leads to increased MSE values across different types of models (PatchTST, RLinear, DLinear, FreTS), highlighting the crucial role of low-energy components in enhancing prediction accuracy.

Low-energy components are often overlooked. We generate two signals with different energy distributions (corresponding to the left and right of Figure 1(b)), and conduct iTransformer (Liu et al. 2024), PatchTST (Nie et al. 2023), and DLinear (Zeng et al. 2023) on them respectively. As detailed in Figure 1(b), the phenomenon of ignorance (rf. small green circles) always occurs with low-energy components, regardless of the frequency band in which they are located.

In this paper, to address the issue of low-energy components being neglected, we propose an **Energy Amplification Technique**, which comprises an *energy amplification block* and an *energy restoration block*. The energy amplification block is designed to boost the energy of low-energy components, enhancing the model’s learning efficiency for these components, while the energy restoration block brings the energy back to its original level. This technique can be applied as a general approach in other time series forecasting models to enhance their performance.

To better harness the potential of the energy amplification technique, we design a new model, **Amplifier**, based on the characteristics of the data after energy amplification. Specifically, since the data processed by the energy amplification block exhibits two energy peaks in the spectrum, it is crucial to separate these two peaks to prevent them from simultaneously dispersing the attention of a single module or layer. Thankfully, the trend and seasonal components obtained from the seasonal-trend decomposition (abbreviated as STD) correspond directly to these two energy peaks. And the widespread use of STD reflects its strong acceptance among researchers. Thus we integrate the energy amplification technique with a seasonal-trend forecaster (based on STD) to model the temporal relationships for these two peaks separately, forming the backbone of Amplifier. To further improve information utilization, we also develop a *semi-channel interaction temporal relationship enhancement block* (abbreviated as SCI block) as an optional built-in component for Amplifier to enhance the performance from the perspective of the *commonality* and *specificity* of each channel in time series data.

Our contributions can be summarized as follows:

- We identify a common issue in existing time series forecasting models, i.e., the neglect of low-energy components, which can lead to performance degradation.

- We propose an energy amplification technique to address the above issue, serving as a general method to enhance the performance of other foundational models.
- To better leverage the energy amplification technique, we combine this technique with a temporal relationship enhancement and a seasonal-trend forecaster to propose a novel model for time series forecasting, called Amplifier.
- Extensive experiments on 8 real-world datasets demonstrate that our Amplifier consistently outperforms state-of-the-art methods while offering higher efficiency.

2 Related Work

Deep Time Series Forecasting

With the rapid advancement of deep learning technology and the increasing importance of time series forecasting, various models have emerged in a flourishing and competitive landscape (Lim and Zohren 2021). Recent studies have proposed a series of upgraded transformer-based models for time series forecasting, such as LogTrans (Li et al. 2019), Informer (Zhou et al. 2021), and Pyraformer (Liu et al. 2021). Meanwhile, some models focus on functional improvements to the Transformer architecture, making it more suitable for time series forecasting tasks, including Autoformer (Wu et al. 2021), FEDformer (Zhou et al. 2022), PatchTST (Nie et al. 2023), and iTransformer (Liu et al. 2024). In addition to the Transformer architecture, lightweight and efficient MLP architectures have also been favored by many researchers. The brilliant debut of DLinear (Zeng et al. 2023) paved the way for MLP models to shine in the field of time series forecasting, leading to the emergence of numerous MLP models, such as: RLinear (Li et al. 2023), TiDE (Das et al. 2023), and SparseTSF (Lin et al. 2024).

Advancements in Frequency Domain Techniques

Recent studies have increasingly leveraged frequency techniques to enhance both the accuracy and efficiency of time series forecasting (Yi et al. 2023). FEDformer (Zhou et al. 2022) achieves a fast attention mechanism through low-rank approximated transformation in the frequency domain. FITS (Xu, Zeng, and Xu 2023) uses frequency domain interpolation to make predictions and essentially functions as a low-pass filter. FreTS (Yi et al. 2024c) retains the simplicity and efficiency of MLP architecture while incorporating the global perspective and energy aggregation characteristics of the frequency domain. Although the above methods demonstrate that frequency domain techniques hold great potential for time series forecasting tasks, they typically either treat components with different energy levels uniformly or focus exclusively on low-frequency components (high-energy parts) while discarding high-frequency components (low-energy parts). This oversight often leads to sub-optimal performance in time series forecasting.

More recently, a few works have begun to address these limitations. Fredformer (Piao et al. 2024) mitigates frequency bias in the Transformer architecture by proposing a framework that learns features equally across different frequency bands. However, Fredformer primarily explores the Transformer architecture and the improvements are specific

to that model. In contrast, in this paper, we propose a unified method applicable to the main paradigm networks, including Transformer and MLP, which employs energy amplification technique to bring model’s attention to neglected low-energy components in time series forecasting, fully leveraging all available data to achieve superior performance. Moreover, compared to Fredformer, Amplifier offers much faster training speeds and a significantly smaller parameter scale.

3 Preliminaries

In this section, we conduct a thorough analysis of the phenomenon where low-energy components are ignored in time series forecasting as mentioned in Introduction section.

Energy Energy can be referred to as a concept in signal processing that captures the overall strength or magnitude of a signal. Given a multivariate time series data $X \in \mathbb{R}^{C \times L}$ with the channel number of C and the look-back window size of L , for the i -th time series data $X^i \in \mathbb{R}^{1 \times L}$, its energy $\mathcal{E}(\mathcal{X}^i)$ can be defined as $\sum_{n=0}^{L-1} |\mathcal{X}[n]|^2$ where $\mathcal{X}[n] \in \mathbb{C}^L$ is the Discrete Fourier Transform (DFT) of X^i . The spectrum of real-world datasets often shows a clear distinction between high and low energy levels, as illustrated by the ETTm1 dataset in the right half of Figure 1(a). The left side of the spectrum represents high-energy frequency points, while the right side represents low-energy frequency points. We refer to a set of low-energy frequency points as low-energy components \mathcal{X}_L^i , and a set of high-energy frequency points as high-energy components \mathcal{X}_H^i , where $\mathcal{E}(\mathcal{X}_H^i) \gg \mathcal{E}(\mathcal{X}_L^i)$. Then, the \mathcal{X}^i can be expressed as the sum of these components: $\mathcal{X}^i = \{\mathcal{X}_H^i, \mathcal{X}_L^i\}$.

Time Series Forecasting For the multivariate time series input data $X \in \mathbb{R}^{C \times L}$ and its next τ data $Y \in \mathbb{R}^{C \times \tau}$, the time series forecasting task is to predict the next τ time-stamps values $\hat{Y} \in \mathbb{R}^{C \times \tau}$ based on the historic data X through a neural network F_θ parameterized by θ . Then, the loss in time domain can be formulated as $\mathcal{L}(Y, \hat{Y}; \theta) = \|Y - \hat{Y}\|_2^2$. Since the data can be divided into high-energy components and low-energy components in frequency domain, the above loss function can be rewritten as an equivalent form in the frequency domain as below:

$$\mathcal{L}(\mathcal{Y}, \hat{\mathcal{Y}}; \Theta) = \mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H) + \mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L), \quad (1)$$

where \mathcal{Y} and $\hat{\mathcal{Y}}$ is the DFT of Y and \hat{Y} respectively, and Θ denote parameters representing the characteristics in the frequency domain. To further investigate how the low-energy components are ignored in the learning process, we examine the loss function and parameter updates from the energy perspective, and we identify two main factors that lead to the neglect of low-energy components.

Theorem 1 *In the initial stage of network training, the loss of high-energy components $\mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H)$ occupies a significantly larger proportion of the overall loss compared to the loss of low-energy components $\mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L)$, that is:*

$$\frac{\mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H)}{\mathcal{L}(\mathcal{Y}, \hat{\mathcal{Y}}; \Theta)} \gg \frac{\mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L)}{\mathcal{L}(\mathcal{Y}, \hat{\mathcal{Y}}; \Theta)}. \quad (2)$$

We include the proof in Appendix A. It implies that in the initial stage of neural network training, the loss value of the high-energy components constitutes the majority of the total loss value. In other words, correcting the loss for the high-energy components can lead to a significant reduction in the total loss value and faster convergence from a global perspective. This guides the model to focus on reducing the error of the high-energy components during training, potentially leading to the neglect of the low-energy components.

Theorem 2 *Parameter updates are influenced by the energy of their corresponding components, meaning that the updates for parameters Θ_L related to low-energy components are much less efficient than those Θ_H for high-energy components, which can be expressed as:*

$$\frac{\mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H)}{\partial \Theta_H} \gg \frac{\mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L)}{\partial \Theta_L}. \quad (3)$$

The proof is shown in Appendix A. It indicates that during the training process, the parameters associated with low-energy components update more slowly and less efficiently than those associated with high-energy components. As a result, by the end of training, the parameters associated with low-energy components are further from their true values compared to those associated with high-energy components, indicating that these parameters have minimal participation during training, making them appear “ignored”.

4 Methodology

As mentioned earlier, previous work typically ignores low-energy components, which are essential for accurate time series forecasting. To address this issue, we propose an *energy amplification technique*, which can be applied as a general method to enhance the performance of existing forecasting models. Then, based on the technique, we propose a simple yet effective model, *Amplifier*, for time series forecasting.

Energy Amplification Technique

The energy amplification technique consists of the *energy amplification block* and the *energy restoration block*. The energy amplification block aims to increase the energy of low-energy components to improve the model’s learning efficiency for these components, while the energy restoration block restores the energy to its original level.

Energy Amplification Block To enable the model to learn both low-energy components and high-energy components without bias, we aim to equalize the energy between them. We achieve this by transferring high energy from the low-frequency region to the low-energy components located in the high-frequency region through spectrum flipping.

Energy can be represented by the square of its amplitude. For the multivariate time series input data $X \in \mathbb{R}^{C \times L}$, its energy can be calculated by:

$$\mathcal{E}(X) = \sum_{k=0}^{L-1} |\mathcal{X}[k]|^2, \quad (4)$$

where $\mathcal{X} \in \mathbb{C}^{C \times L}$ is the DFT of X and k is the frequency point. We shift high energy from the low-frequency region

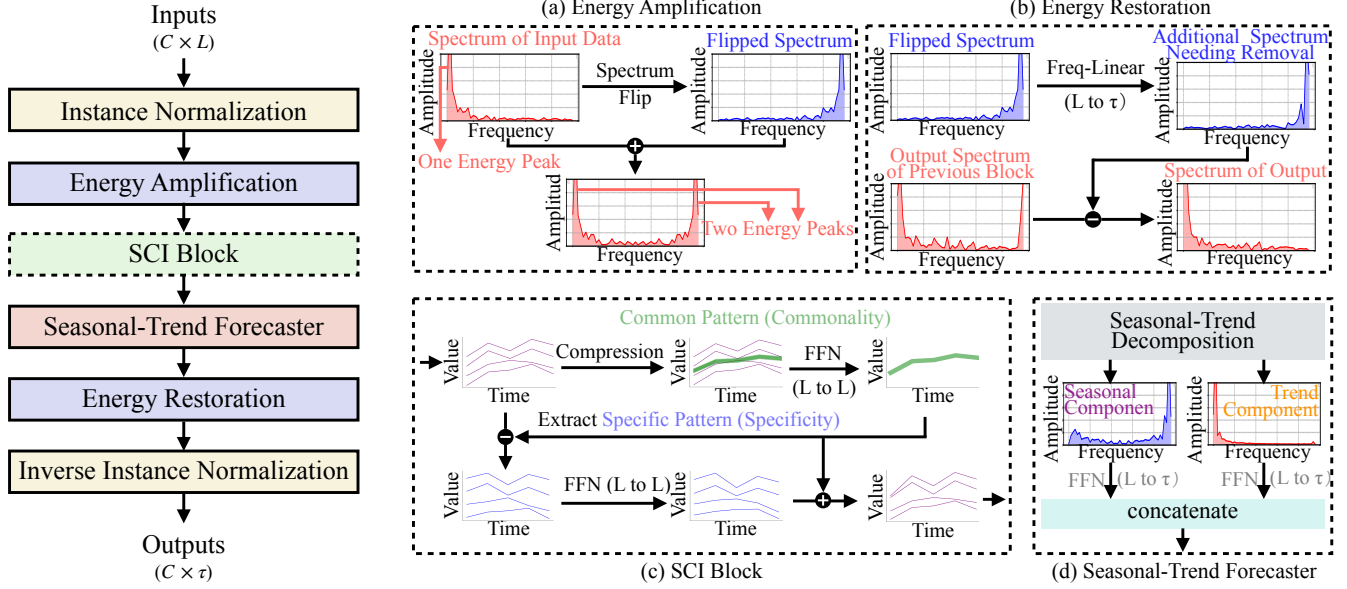


Figure 2: The overall architecture of Amplifier. (i) The energy amplification block aims to increase the energy of low-energy components, while energy restoration block performs the inverse operation of the energy amplification block; (ii) SCI block is employed to capture both the common temporal pattern and specific temporal pattern; (iii) Seasonal-Trend Forecaster is then utilized to decompose seasonal and trend information, and then make predictions.

to the low-energy components located in the high-frequency region by flipping the spectrum, formulated as follows:

$$\mathcal{X}'[k] = \mathcal{X}[T - k], \quad (5)$$

where $\mathcal{X}' \in \mathbb{C}^{C \times L}$ refers to the flipped spectrum. By adding \mathcal{X} and \mathcal{X}' together and performing Inverse Discrete Fourier Transform (IDFT), we obtain the output X_{Amp} of this block as detailed below:

$$\begin{aligned} \mathcal{X}_{\text{Amp}} &= \mathcal{X} + \mathcal{X}', \\ X_{\text{Amp}} &= \text{IDFT}(\mathcal{X}_{\text{Amp}}). \end{aligned} \quad (6)$$

At this point, the original low-energy components have gained energy comparable to the high-energy components, effectively bringing attention to neglected low-energy components, as shown in the following equation:

$$\mathcal{E}_{\text{Amp}}[k] = \mathcal{E}_{\text{Amp}}[T - k]. \quad (7)$$

Energy Restoration Block The block is employed to remove the flipped spectrum added by the energy amplification block, serving as the inverse operation of energy amplification. First, we adjust the input length to align the prediction length by a frequency-domain linear operations as:

$$\mathcal{Y}' = \mathcal{X}'\mathcal{W} + \mathcal{B} \quad (8)$$

where $\mathcal{X}' \in \mathbb{C}^{C \times L}$ is the flipped spectrum, $\mathcal{W} \in \mathbb{C}^{L \times \tau}$ is a complex number weight matrix, $\mathcal{B} \in \mathbb{C}^{\tau}$ is a complex number bias, and $\mathcal{Y}' \in \mathbb{C}^{C \times \tau}$ denotes the additional spectrum needing removal. Then, we convert the output Y_{Amp} of the previous block from the time domain to the frequency domain, remove the additional spectrum \mathcal{Y}' , and perform domain conversion to obtain the final prediction result:

$$\begin{aligned} \mathcal{Y}_{\text{Amp}} &= \text{DFT}(Y_{\text{Amp}}), \\ \mathcal{Y} &= \mathcal{Y}_{\text{Amp}} - \mathcal{Y}', \\ \hat{Y} &= \text{IDFT}(\mathcal{Y}). \end{aligned} \quad (9)$$

For more explanation on the energy amplification technique, please refer to the Appendix B.

Amplifier

Given that the energy-amplified data usually exhibits two energy peaks, we combine the energy amplification technique with a seasonal-trend forecaster to model the temporal relationships for these two peaks separately, forming the backbone of our newly proposed model, Amplifier. The overall structure of Amplifier is illustrated in Figure 2, including four main blocks: energy amplification block, semi-channel interaction temporal relationship enhancement block (abbreviated as SCI block), seasonal-trend forecaster block, and energy restoration block. At the same time, we use Instance Normalization and its inverse operation to address non-stationarity in time series data. As the energy amplification technique has already been introduced in previous subsection, we describe the rest of the other blocks below.

SCI Block The data is composed of multiple channels of time series, usually following one common pattern, which we refer to as *commonality*. Excluding this commonality, the remaining parts represent each channel's specific pattern, which we refer to as *specificity*.

The commonality can be considered an abstract main channel, and it can be calculated as follow:

$$X_{\text{Com}} = \text{Compression}_C(X) \quad (10)$$

where $\text{Compression}_C : \mathbb{R}^C \mapsto \mathbb{R}^1$ contains two linear layers with intermediate LeakyReLU activation function and $X_{\text{Com}} \in \mathbb{R}^{1 \times L}$ is the commonality of $X \in \mathbb{R}^{C \times L}$. To further achieve the common temporal patterns, we employ a FFN and output the common pattern $X_{\text{Cp}} \in \mathbb{R}^{1 \times L}$.

Models	Amplifier		RLinear		DLinear		FreTS		FITS		Fredformer		iTransformer		PatchTST		Stationary		TimesNet	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	0.381	0.394	0.414	0.407	0.403	0.407	0.408	0.416	0.415	0.408	0.384	0.395	0.407	0.410	0.387	0.400	0.481	0.456	0.400	0.406
ETTm2	0.276	0.323	0.286	0.327	0.350	0.401	0.321	0.368	0.286	0.328	0.279	0.324	0.288	0.332	0.281	0.326	0.306	0.347	0.291	0.333
ETTh1	0.430	0.428	0.446	0.434	0.456	0.452	0.475	0.463	0.451	0.440	0.435	0.426	0.454	0.447	0.469	0.454	0.570	0.537	0.458	0.450
ETTh2	0.359	0.391	0.374	0.398	0.559	0.515	0.472	0.465	0.383	0.408	0.365	0.393	0.383	0.407	0.387	0.407	0.526	0.516	0.414	0.427
ECL	0.171	0.265	0.219	0.298	0.212	0.300	0.189	0.278	0.217	0.295	0.175	0.269	0.178	0.270	0.216	0.304	0.193	0.296	0.192	0.295
Exchange	0.361	0.402	0.378	0.417	0.354	0.414	0.329	0.409	0.353	0.399	0.374	0.408	0.360	0.403	0.367	0.404	0.461	0.454	0.416	0.443
Traffic	0.482	0.315	0.626	0.378	0.625	0.383	0.618	0.390	0.627	0.376	0.431	0.287	0.428	0.282	0.481	0.304	0.624	0.340	0.620	0.336
Weather	0.243	0.271	0.272	0.291	0.265	0.317	0.250	0.270	0.249	0.276	0.246	0.272	0.258	0.278	0.259	0.281	0.288	0.314	0.259	0.287
1 st Count	6	5	0	0	0	0	1	0	0	1	0	1	1	1	0	0	0	0	0	0

Table 1: Time series forecasting comparison. We set the lookback window size L as 96 and the prediction length as $\tau \in \{96, 192, 336, 720\}$. The best results are in red and the second best are blue. Results are averaged from all prediction lengths. Full results for all datasets are listed in Table 6 of Appendix C.

Then based on X_{Cp} we can obtain the specificity temporal patterns, which can be formulated as:

$$\begin{aligned} X_{Spc} &= X - X_{Cp}, \\ X_{Sp} &= \text{FFN}(X_{Spc}) \end{aligned} \quad (11)$$

where $X_{Sp} \in \mathbb{R}^{C \times L}$. Finally, we calculate the summation of X_{Cp} and X_{Sp} as the output $X_{Sci} \in \mathbb{R}^{C \times L}$ of SCI block.

Seasonal-Trend Forecaster Seasonal-trend decomposition as a basic function has been widely used in previous work (Wu et al. 2021; Zhou et al. 2022) for time series forecasting, and we also employ this in our model. Specifically, we first obtain the seasonal and trend components of $X_{Sci} \in \mathbb{R}^{C \times L}$ through the seasonal-trend decomposition as:

$$X_{Trend}^{Sci}, X_{Season}^{Sci} = \text{STD}(X_{Sci}). \quad (12)$$

Subsequently, we make predictions by applying two FFNs corresponding to the seasonal and trend components, which can be formulated as below:

$$\begin{aligned} Y_{Trend}^{Sci} &= \text{Trend-FFN}(X_{Trend}^{Sci}), \\ Y_{Season}^{Sci} &= \text{Season-FFN}(X_{Season}^{Sci}), \end{aligned} \quad (13)$$

where both Trend-FFN(\cdot) and Season-FFN(\cdot) contain two linear layers with intermediate LeakyReLU activation function. Then Y_{Trend}^{Sci} and Y_{Season}^{Sci} are concatenated as the output of the seasonal-trend forecaster block:

$$Y = Y_{Trend}^{Sci} + Y_{Season}^{Sci}. \quad (14)$$

5 Experiments

Experiments Setup

Datasets We conduct extensive experiments on eight popular datasets, including ETT datasets (Zhou et al. 2021), Electricity (Wu et al. 2021), Exchange (Lai et al. 2018), Traffic (Sen, Yu, and Dhillon 2019) and Weather (Wu et al. 2021). More dataset details are in Appendix C.

Baselines We select 10 highly regarded forecasting methods to serve as our benchmarks, including (i) Linear-based methods: RLinear (Li et al. 2023), DLinear (Zeng et al. 2023), SparseTSF (Lin et al. 2024); (ii) Frequency-based methods: FreTS (Yi et al. 2024c), FITS (Xu, Zeng, and Xu 2023); (iii) Transformer-based methods: Fredformer (Piao et al. 2024), iTransformer (Liu et al. 2024), PatchTST (Nie et al. 2023), Stationary (Liu et al. 2022b); and (iv) TCN-based methods: TimesNet (Wu et al. 2023).

Implementation Details All experiments in this study were carried out using PyTorch on one single NVIDIA RTX 3070 GPU with 8GB. We use Mean Squared Error (MSE) as the loss function and report the results using both MSE and Mean Absolute Error (MAE) as evaluation metrics.

Main Results

Table 1 shows the average forecasting performance across four prediction lengths under a look-back window size of 96. Overall, our approach achieves leading performance on most datasets, securing 11 top-1 and 2 top-2 positions out of 16 in total across two metrics over eight datasets. We primarily attribute this to the energy amplification technique, which enhances the model’s capability to model low-energy components. However, we notice that the performance of Amplifier on the Traffic dataset is not outstanding. The reason is that for the Traffic dataset, which has strong periodicity, periodic information is typically represented by high-energy components. The energy amplification technique is designed to bring attention to neglected low-energy components, which may not be as beneficial for forecasting datasets with strong periodicity. To further assess the performance of our model with varying lookback window sizes, we conduct additional experiments using a lookback window size of 336. The results, presented in Table 3, indicate that as the lookback window size increases, Amplifier continues to demonstrate exceptional predictive performance.

Dataset	ETTh1		ETTm2		Weather	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE
w/o EAT	0.393	0.407	0.182	0.266	0.184	0.230
Amplifier	0.371	0.392	0.176	0.258	0.156	0.205
Boost	5.643%	3.613%	3.484%	2.973%	15.100%	10.923%

Table 2: Ablation experiments of the energy amplification technique within Amplifier. w/o EAT refers to the version of Amplifier without the energy amplification technique. The ‘Boost’ indicates the percentage of performance improvement after incorporating the energy amplification technique.

Model Analysis

Effectiveness Analysis of the Energy Amplification Technique In this part, we investigate the effectiveness of the energy amplification technique from two perspectives: the

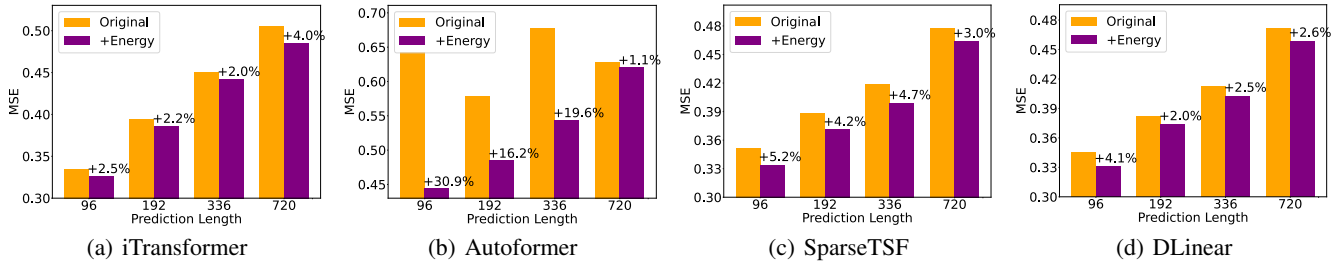


Figure 3: Ablation prediction results of the energy amplification technique in four representative models.

Models	Amplifier		SparseTSF		DLinear		iTransformer		TimsNet		
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	
ETTm1	96	0.288	0.343	0.306	0.347	0.300	0.345	0.303	0.357	0.335	0.380
	192	0.325	0.366	0.341	0.368	0.336	0.366	0.345	0.383	0.358	0.388
	336	0.364	0.388	0.373	0.385	0.372	0.390	0.382	0.405	0.406	0.418
	720	0.430	0.423	0.429	0.417	0.427	0.423	0.443	0.439	0.449	0.443
	Avg	0.352	0.380	0.362	0.379	0.359	0.381	0.368	0.396	0.387	0.407
ETTh1	96	0.371	0.395	0.393	0.407	0.384	0.405	0.402	0.418	0.398	0.418
	192	0.408	0.415	0.421	0.423	0.430	0.442	0.450	0.449	0.447	0.449
	336	0.396	0.416	0.401	0.425	0.447	0.448	0.479	0.470	0.493	0.468
	720	0.451	0.463	0.429	0.446	0.504	0.515	0.584	0.548	0.518	0.504
	Avg	0.407	0.422	0.409	0.425	0.441	0.453	0.479	0.471	0.464	0.460
Weather	96	0.150	0.203	0.177	0.227	0.175	0.235	0.164	0.216	0.172	0.220
	192	0.192	0.242	0.221	0.264	0.218	0.278	0.205	0.251	0.219	0.261
	336	0.241	0.280	0.267	0.296	0.263	0.314	0.256	0.290	0.280	0.306
	720	0.316	0.332	0.334	0.343	0.324	0.362	0.326	0.338	0.365	0.359
	Avg	0.225	0.264	0.250	0.283	0.245	0.297	0.238	0.274	0.259	0.287
Electricity	96	0.133	0.231	0.147	0.240	0.140	0.237	0.133	0.229	0.168	0.272
	192	0.156	0.252	0.158	0.251	0.154	0.250	0.156	0.251	0.184	0.289
	336	0.169	0.266	0.174	0.268	0.169	0.268	0.172	0.267	0.198	0.300
	720	0.199	0.296	0.212	0.299	0.204	0.300	0.209	0.304	0.220	0.320
	Avg	0.164	0.261	0.173	0.265	0.167	0.264	0.168	0.263	0.193	0.287
1 st Count	17	14	1	4	3	2	1	1	0	0	

Table 3: Time series forecasting comparison with the look-back window size L as 336 and the prediction length as $\tau \in \{96, 192, 336, 720\}$. The best results are in red and the second best are in blue. Avg means the average results from all four prediction lengths.

role it plays within Amplifier and the performance improvements it brings when integrated as a general technique into other foundational models. As shown in Table 2, we compare Amplifier and w/o EAT which is the version of Amplifier without the energy amplification technique (EAT), on the ETTh1, ETTm2, and Weather datasets. It demonstrates that the energy amplification technique can improve predictive performance by increasing the model’s attention to low-energy components. It’s worth noting that, on the weather dataset, the impact of using the energy amplification technique on prediction results is significant, affecting MSE and MAE by as much as 15.100% and 10.923%, respectively. As illustrated by the butterfly effect (Lorenz 1972), small changes in a weather model (such as a butterfly flapping its wings in Brazil) can trigger large-scale atmospheric changes far away (such as a hurricane in the United States) through a series of causal relationships. Effectively capturing and modeling these low-energy components can significantly improve the accuracy of weather predictions.

Besides, the energy amplification technique can not only function as a built-in module within Amplifier but also as a universal technology that can be integrated into other forecasting models, such as Transformer-based models: iTransformer and Autoformer (Wu et al. 2021); MLP-based models: SparseTSF (Lin et al. 2024) and DLinear (Zeng et al. 2023). As illustrated in Figure 3, on the ETTm1

Horizon	96		192		336		720		
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	
ECL	w/o SCI	0.156	0.247	0.167	0.256	0.189	0.278	0.230	0.312
	Amplifier	0.147	0.243	0.157	0.251	0.174	0.269	0.206	0.296
	Boost	5.769%	1.619%	5.988%	1.953%	7.839%	3.237%	10.318%	5.219%
Traffic	w/o SCI	0.500	0.328	0.477	0.305	0.509	0.326	0.544	0.345
	Amplifier	0.455	0.298	0.470	0.304	0.479	0.316	0.523	0.328
	Boost	9.000%	9.063%	1.426%	0.230%	5.838%	3.067%	3.807%	4.817%

Table 4: Ablation experiments for SCI block of Amplifier where w/o SCI refers to the version without the SCI block. The ‘Boost’ indicates the percentage of performance improvement after equipping with the SCI block.

dataset, Transformer-based models achieved improvements of 9.837% in MSE and 4.236% in MAE, while Linear-based models achieved improvements of 3.603% in MSE and 2.244% in MAE (the statistical values are in the sense of average). These results clearly demonstrate the effectiveness of the energy amplification technique, showing that it can significantly enhance the performance of foundational models in time series forecasting.

Effectiveness Analysis of SCI Block The SCI block enhances temporal relationship modeling by improving information utilization, specifically through the consideration of interactions between channels. We conduct ablation experiments on the role of the SCI block within Amplifier, as shown in Table 4, where w/o SCI refers to the version without the SCI block. We choose the Electricity and Traffic datasets with a huge number of channels for our experiments because the interactions between channels becomes more pronounced as the number of channels increases. Specifically, using the SCI block achieved improvements of 7.479% in MSE and 3.007% in MAE on the Electricity dataset, while showed improvements of 5.018% in MSE and 4.294% in MAE on the Traffic dataset. Therefore, the SCI block contributes to enhancing model performance by effectively leveraging interactions between channels.

Efficiency Analysis

The theoretical complexity of the Amplifier is $\mathcal{O}(L \log L)$. We conduct a comprehensive comparison of the forecasting performance, the scale of parameters, and the training speed of the following representative models, including SparseTSF, RLinear, DLinear, FITS, FreTS, and iTransformer. We choose the Weather dataset under the scenario of $L = 96$ and $\tau = 96$ for comparison.

From Figure 4, it clear to see that the scale of parameters and training speed of Amplifier are at a medium level. However, it’s noteworthy that both SparseTSF and FITS consider lightweight architecture as one of their main contributions,

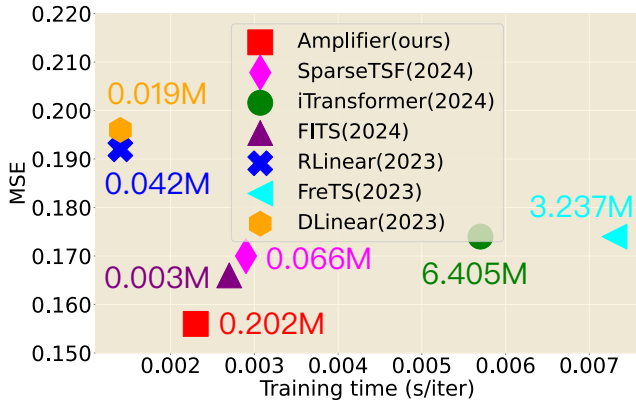


Figure 4: Model efficiency comparison in terms of MSE, the scale of parameters, and training speed.

whereas Amplifier does not have this as a primary goal in its model design. As for RLinear and DLinear, since Amplifier aims to enhance the model’s attention on low-energy components while avoiding any negative impact on modeling high-energy components, it requires a dedicated model component to handle low-energy components. This inevitably results in the Amplifier having a larger parameter size and longer training time compared to these two models.

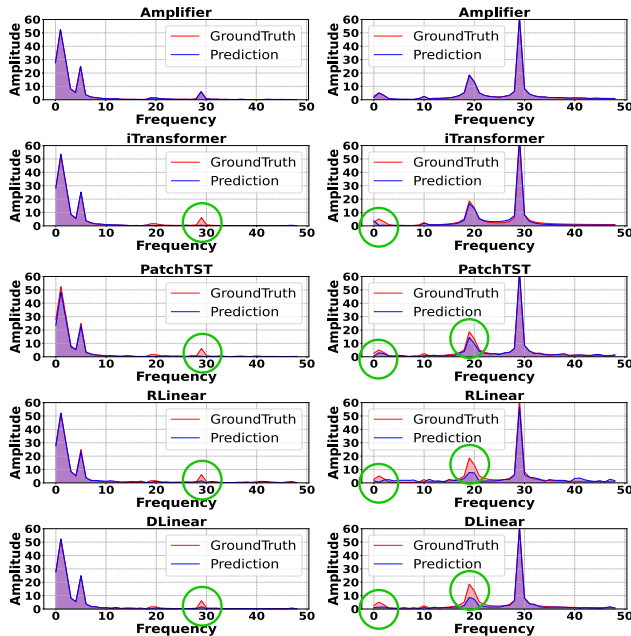


Figure 5: Prediction spectrums of both Amplifier and four SOTA models on two synthetic signals. The small green circles represent the neglected low-energy components.

Visualizations

Visualization Analysis about the Energy Amplification Technique To intuitively validate the modeling capability of our amplification technique in low energy components, we conduct experiments on two simple synthetic signals, and compare our model with the four representative models,

including iTransformer, PatchTST, RLinear, and DLinear. The results are shown in Figure 5, which demonstrates that Amplifier can handle low-energy components as effectively as high-energy components. Compared with our model, the four models neither can fully model the low-energy components, regardless of whether these low-energy components appear at high-frequency region or low-frequency region.

Visualization of Forecasting Results To visually compare the performance of Amplifier with state-of-the-art models, including iTransformer, FreTS, and DLinear, we present prediction showcases on the ETTm2 and Electricity datasets, and the results are shown in Figures 6 and 7. The red line represents the input sequence, the blue line represents the ground truth, and the yellow line represents the predicted value. Compared to these different types of state-of-the-art models, Amplifier provides the most accurate predictions of future series variations, demonstrating superior performance.

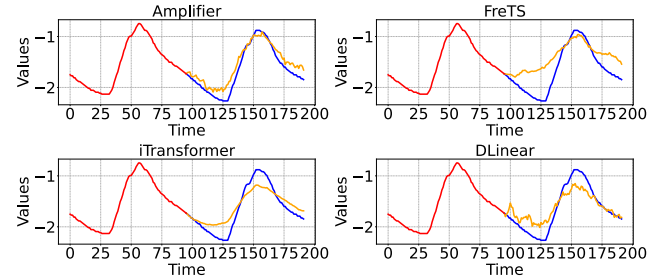


Figure 6: Visualization of prediction results on the ETTm2 dataset (96 → 96).

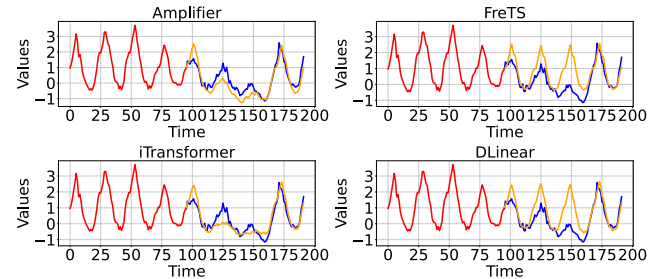


Figure 7: Visualization of prediction results on the Electricity dataset (96 → 96).

6 Conclusion

Considering that low-energy components is often overlooked in existing methods, we propose the energy amplification technique composed of the energy amplification block and energy restoration block. The core idea of this technique is to increase the model’s attention to low-energy components by flipping spectrum to amplify the energy of those components, thereby enhancing the model’s ability to process low-energy components. To better leverage the energy amplification technique, we design a novel model called Amplifier for time series forecasting. Comprehensive empirical experiments on eight real-world datasets have validated the superiority of our proposed model.

Acknowledgments

This research was funded by the National Natural Science Foundation of China, Grant No. 62272048.

References

- Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational Statistics Paris France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, 177–186. Springer.
- Das, A.; Kong, W.; Leach, A.; Mathur, S.; Sen, R.; and Yu, R. 2023. Long-term forecasting with tide: Time-series dense encoder. *arXiv preprint arXiv:2304.08424*.
- Fan, W.; Zheng, S.; Yi, X.; Cao, W.; Fu, Y.; Bian, J.; and Liu, T.-Y. 2022. DEPTS: Deep expansion learning for periodic time series forecasting. *arXiv preprint arXiv:2203.07681*.
- Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256. JMLR Workshop and Conference Proceedings.
- Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>.
- Hayes, M. H. 1996. *Statistical digital signal processing and modeling*. John Wiley & Sons.
- Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Lai, G.; Chang, W.-C.; Yang, Y.; and Liu, H. 2018. Modeling long- and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, 95–104.
- Lathi, B. P.; and Green, R. A. 1998. *Signal processing and linear systems*, volume 2. Oxford university press Oxford.
- Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.-X.; and Yan, X. 2019. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in neural information processing systems*, 32.
- Li, Z.; Qi, S.; Li, Y.; and Xu, Z. 2023. Revisiting long-term time series forecasting: An investigation on linear mapping. *arXiv preprint arXiv:2305.10721*.
- Lim, B.; and Zohren, S. 2021. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194): 20200209.
- Lin, S.; Lin, W.; Wu, W.; Chen, H.; and Yang, J. 2024. SparseTSF: Modeling Long-term Time Series Forecasting with 1k Parameters. *arXiv preprint arXiv:2405.00946*.
- Liu, M.; Zeng, A.; Chen, M.; Xu, Z.; Lai, Q.; Ma, L.; and Xu, Q. 2022a. Scinet: Time series modeling and forecasting with sample convolution and interaction. *Advances in Neural Information Processing Systems*, 35: 5816–5828.
- Liu, S.; Yu, H.; Liao, C.; Li, J.; Lin, W.; Liu, A. X.; and Dustdar, S. 2021. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International conference on learning representations*.
- Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2024. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *The Twelfth International Conference on Learning Representations*.
- Liu, Y.; Wu, H.; Wang, J.; and Long, M. 2022b. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in Neural Information Processing Systems*, 35: 9881–9893.
- Lorenz, E. N. 1972. Predictability: does the flap of a butterfly’s wings in Brazil set off a tornado in Texas? American Association for the Advancement of Science. In *139th meeting*, volume 29.
- Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *The Eleventh International Conference on Learning Representations*.
- Oppenheim, A. V. 1999. *Discrete-time signal processing*. Pearson Education India.
- Piao, X.; Chen, Z.; Murayama, T.; Matsubara, Y.; and Sakurai, Y. 2024. Fredformer: Frequency Debiased Transformer for Time Series Forecasting. *arXiv preprint arXiv:2406.09009*.
- Salinas, D.; Flunkert, V.; Gasthaus, J.; and Januschowski, T. 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191.
- Sen, R.; Yu, H.-F.; and Dhillon, I. S. 2019. Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. *Advances in neural information processing systems*, 32.
- Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *International Conference on Learning Representations*.
- Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34: 22419–22430.
- Xu, Z.; Zeng, A.; and Xu, Q. 2023. FITS: Modeling time series with 10k parameters. *arXiv preprint arXiv:2307.03756*.
- Yi, K.; Fei, J.; Zhang, Q.; He, H.; Hao, S.; Lian, D.; and Fan, W. 2024a. FilterNet: Harnessing Frequency Filters for Time Series Forecasting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Yi, K.; Zhang, Q.; Cao, L.; Wang, S.; Long, G.; Hu, L.; He, H.; Niu, Z.; Fan, W.; and Xiong, H. 2023. A Survey on Deep Learning based Time Series Analysis with Frequency Transformation. *CoRR*, abs/2302.02173.
- Yi, K.; Zhang, Q.; Fan, W.; He, H.; Hu, L.; Wang, P.; An, N.; Cao, L.; and Niu, Z. 2024b. FourierGNN: Rethinking multivariate time series forecasting from a pure graph perspective. *Advances in Neural Information Processing Systems*, 36.
- Yi, K.; Zhang, Q.; Fan, W.; Wang, S.; Wang, P.; He, H.; An, N.; Lian, D.; Cao, L.; and Niu, Z. 2024c. Frequency-domain MLPs are more effective learners in time series forecasting. *Advances in Neural Information Processing Systems*, 36.

Yi, K.; Zhang, Q.; He, H.; Shi, K.; Hu, L.; An, N.; and Niu, Z. 2024d. Deep Coupling Network For Multivariate Time Series Forecasting. *ACM Transactions on Information Systems*, 42(5): 1–28.

Yu, B.; Yin, H.; and Zhu, Z. 2018. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 3634–3640.

Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 11121–11128.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11106–11115.

Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; and Jin, R. 2022. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, 27268–27286. PMLR.

A Theoretical Proofs

To clarify the proof process, we first define the frequency domain representation of a time domain signal and introduce a lemma to assist the proof process.

Definition 1 *In the time domain, a signal $X(t)$ is a function of time t and can be represented as a sum of various frequency components (Hayes 1996), each with a specific amplitude and phase:*

$$\begin{aligned} X(t) &= \sum_f \mathcal{X}[f], \\ \mathcal{X}[f] &= \mathcal{A}_f e^{j\phi_f} = \mathcal{A}_f e^{j(\omega_f t + \psi_f)}, \end{aligned} \quad (15)$$

where $\phi_f = \omega_f t + \psi_f$ refers to the total phase, \mathcal{A}_f , ω_f ($\omega_f = 2\pi f$) and ψ_f represent the amplitude, angular frequency, and initial phase of a particular frequency component f , respectively.

Lemma 1 *Given that $c = e^{j\theta_1}$ and $d = e^{j\theta_2}$ are complex numbers with $\|c\| = \|d\| = 1$, then $\|ac - bd\|^2$ can be expanded as follows:*

$$\|ac - bd\|^2 = \|a\|^2 + \|b\|^2 - 2ab \cos(\theta_1 - \theta_2). \quad (16)$$

Proof of Lemma 1

$$\begin{aligned} \|ac - bd\|^2 &= (ac - bd)(\overline{ac} - \overline{bd}) \\ &= a\overline{c}a\overline{c} - a\overline{c}b\overline{d} - b\overline{d}a\overline{c} + b\overline{d}b\overline{d} \\ &= \|a\|^2 \|c\|^2 - a\overline{c}b\overline{d} - b\overline{d}a\overline{c} + \|b\|^2 \|d\|^2 \\ &= \|a\|^2 + \|b\|^2 - ab(\overline{c}d + \overline{d}c) \\ &= \|a\|^2 + \|b\|^2 - 2ab \cos(\theta_1 - \theta_2) \end{aligned}$$

Proved.

Theorem 1 *In the initial stage of network training, the loss of high-energy components $\mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H)$ occupies a significantly larger proportion of the overall loss compared to the loss of low-energy components $\mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L)$, that is:*

$$\frac{\mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H)}{\mathcal{L}(\mathcal{Y}, \hat{\mathcal{Y}}; \Theta)} \gg \frac{\mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L)}{\mathcal{L}(\mathcal{Y}, \hat{\mathcal{Y}}; \Theta)}. \quad (17)$$

Proof of Theorem 1

As we have mentioned in Preliminaries section, since the data can be divided into high-energy components and low-energy components, the overall loss can be composed of the individual losses of these two components as below:

$$\mathcal{L}(\mathcal{Y}, \hat{\mathcal{Y}}; \Theta) = \mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H) + \mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L). \quad (18)$$

Combine Eq (18) and Definition 1, $\mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H)$ and $\mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L)$ can be rewritten separately as follows:

$$\begin{aligned} \mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H) &= \|\mathcal{Y}_H - \hat{\mathcal{Y}}_H\|_2^2 \\ &= \|\mathcal{A}_H e^{j\phi_H} - \hat{\mathcal{A}}_H e^{j\hat{\phi}_H}\|_2^2, \end{aligned} \quad (19)$$

where the subscript H is an abbreviation for high-energy components. \mathcal{A}_H and ϕ_H represent the true values of the amplitude and the total phase of the high-energy components, while $\Theta_H = \{\hat{\mathcal{A}}_H, \hat{\phi}_H\}$ refers to the corresponding values learned by the neural network.

$$\begin{aligned} \mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L) &= \|\mathcal{Y}_L - \hat{\mathcal{Y}}_L\|_2^2 \\ &= \|\mathcal{A}_L e^{j\phi_L} - \hat{\mathcal{A}}_L e^{j\hat{\phi}_L}\|_2^2, \end{aligned} \quad (20)$$

where the subscript L is an abbreviation for low-energy components. \mathcal{A}_L and ϕ_L represent the true values of the amplitude and the total phase of the low-energy components, while $\Theta_L = \{\hat{\mathcal{A}}_L, \hat{\phi}_L\}$ refers to the corresponding values learned by the neural network.

Based on Lemma 1, Eq (19) and Eq (20) can be expanded as:

$$\begin{aligned} \mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H) &= \|\mathcal{A}_H e^{j\phi_H} - \hat{\mathcal{A}}_H e^{j\hat{\phi}_H}\|_2^2 \\ &= \|\mathcal{A}_H\|_2^2 + \|\hat{\mathcal{A}}_H\|_2^2 - \mathcal{A}_H \hat{\mathcal{A}}_H (e^{j\phi_H} e^{-j\hat{\phi}_H} + e^{-j\phi_H} e^{j\hat{\phi}_H}) \\ &= \|\mathcal{A}_H\|_2^2 + \|\hat{\mathcal{A}}_H\|_2^2 - \mathcal{A}_H \hat{\mathcal{A}}_H \Gamma_H, \end{aligned} \quad (21)$$

where $\Gamma_H = 2 \cos(\phi_H - \hat{\phi}_H)$.

$$\begin{aligned} \mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L) &= \|\mathcal{A}_L e^{j\phi_L} - \hat{\mathcal{A}}_L e^{j\hat{\phi}_L}\|_2^2 \\ &= \|\mathcal{A}_L\|_2^2 + \|\hat{\mathcal{A}}_L\|_2^2 - \mathcal{A}_L \hat{\mathcal{A}}_L (e^{j\phi_L} e^{-j\hat{\phi}_L} + e^{-j\phi_L} e^{j\hat{\phi}_L}) \\ &= \|\mathcal{A}_L\|_2^2 + \|\hat{\mathcal{A}}_L\|_2^2 - \mathcal{A}_L \hat{\mathcal{A}}_L \Gamma_L, \end{aligned} \quad (22)$$

where $\Gamma_L = 2 \cos(\phi_L - \hat{\phi}_L)$.

In the initial stages of neural network training, the parameters are randomly initialized and their values are typically small (Glorot and Bengio 2010; Goodfellow, Bengio, and Courville 2016). So based on Eq (21) and Eq (22), $\mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H)$ and $\mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L)$ can be approximately expressed in the following forms:

$$\mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H) \approx \|\mathcal{A}_H\|_2^2, \quad (23)$$

$$\mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L) \approx \|\mathcal{A}_L\|_2^2. \quad (24)$$

Since the amplitude of the high-energy components is much greater than that of the low-energy components, i.e. $\|\mathcal{A}_H\| \gg \|\mathcal{A}_L\|$, then $\|\mathcal{A}_H\|_2^2 \gg \|\mathcal{A}_L\|_2^2$. Combining Eq (23) and Eq (24), we obtain:

$$\mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H) \gg \mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L), \quad (25)$$

that is:

$$\frac{\mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H)}{\mathcal{L}(\mathcal{Y}, \hat{\mathcal{Y}}; \Theta)} \gg \frac{\mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L)}{\mathcal{L}(\mathcal{Y}, \hat{\mathcal{Y}}; \Theta)}. \quad (26)$$

Proved.

Theorem 2 *Parameter updates are influenced by the energy of their corresponding components, meaning that the updates for parameters Θ_L related to low-energy components are much less efficient than those Θ_H for high-energy components, which can be expressed as:*

$$\frac{\mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H)}{\partial \Theta_H} \gg \frac{\mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L)}{\partial \Theta_L}. \quad (27)$$

Proof of Theorem 2

Firstly, according to Definition 1, $\mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H)$ and $\mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L)$ can be expanded in detail as follows:

$$\begin{aligned} \mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H) &= \left\| \mathcal{Y}_H - \hat{\mathcal{Y}}_H \right\|_2^2 \\ &= \left\| \mathcal{A}_H e^{j(\omega_H t + \psi_H)} - \hat{\mathcal{A}}_H e^{j(\hat{\omega}_H t + \hat{\psi}_H)} \right\|_2^2, \end{aligned} \quad (28)$$

$$\begin{aligned} \mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L) &= \left\| \mathcal{Y}_L - \hat{\mathcal{Y}}_L \right\|_2^2 \\ &= \left\| \mathcal{A}_L e^{j(\omega_L t + \psi_L)} - \hat{\mathcal{A}}_L e^{j(\hat{\omega}_L t + \hat{\psi}_L)} \right\|_2^2, \end{aligned} \quad (29)$$

where $\hat{\mathcal{A}}_H$, $\hat{\omega}_H$, and $\hat{\psi}_H$ refer to the parameters related to the high-energy components while $\hat{\mathcal{A}}_L$, $\hat{\omega}_L$, and $\hat{\psi}_L$ represent the parameters associated with the low-energy components.

Then based on the common time-domain parameter update algorithms (Bottou 2010), we derive the parameter update algorithms in the frequency domain:

$$\begin{aligned} \hat{\mathcal{A}}_f^{i+1} &= \hat{\mathcal{A}}_f^i - \eta \frac{\partial \mathcal{L}}{\partial \hat{\mathcal{A}}_f^i} \\ &= \hat{\mathcal{A}}_f^i - \eta e^{j(\hat{\omega}_f^i t + \hat{\psi}_f^i)}, \end{aligned} \quad (30)$$

$$\begin{aligned} \hat{\omega}_f^{i+1} &= \hat{\omega}_f^i - \eta \frac{\partial \mathcal{L}}{\partial \hat{\omega}_f^i} \\ &= \hat{\omega}_f^i - \eta \hat{\mathcal{A}}_f^i t e^{j(\hat{\omega}_f^i t + \hat{\psi}_f^i)}, \end{aligned} \quad (31)$$

$$\begin{aligned} \hat{\psi}_f^{i+1} &= \hat{\psi}_f^i - \eta \frac{\partial \mathcal{L}}{\partial \hat{\psi}_f^i} \\ &= \hat{\psi}_f^i - \eta \hat{\mathcal{A}}_f^i e^{j(\hat{\omega}_f^i t + \hat{\psi}_f^i)}, \end{aligned} \quad (32)$$

where subscript f denotes the value at frequency point f , superscript i represents the i -th update iteration, and η refers to the learning rate. Since the amplitude $\hat{\mathcal{A}}_L$ of low-energy components is much smaller than that of high-energy components, the parameters $\hat{\omega}_L$ and $\hat{\psi}_L$ related to low-energy components are updated to a much lesser extent each iteration compared to those related to high-energy components,

according to Eq (31) and Eq (32). At the same time, the inefficient updates of parameters $\hat{\omega}_L$ and $\hat{\psi}_L$ also hinder the effective update of parameter $\hat{\mathcal{A}}_L$ in Eq (30).

In summary, the update efficiency of parameters $\Theta_L = \{\hat{\mathcal{A}}_L, \hat{\omega}_L, \hat{\psi}_L\}$ related to low-energy components is significantly lower than that of the parameters related to high-energy components, i.e.,

$$\frac{\mathcal{L}(\mathcal{Y}_H, \hat{\mathcal{Y}}_H; \Theta_H)}{\partial \Theta_H} \gg \frac{\mathcal{L}(\mathcal{Y}_L, \hat{\mathcal{Y}}_L; \Theta_L)}{\partial \Theta_L}. \quad (33)$$

Proved.

B More Explanation of Energy Amplification Technique

Our proposed energy amplification technique, which amplifies energy through spectrum flipping, has the following advantages:

Simplicity: It is straightforward to implement and requires no additional feature engineering.

Equal Energy Levels: Flipping the spectrum naturally allows low-energy components to attain the same energy levels as high-energy components, enabling the model to treat both equally.

Adaptability: The frequency-domain linear operation in Equation 8 not only adjusts the input length to match the prediction length but also fundamentally performs frequency-domain interpolation. This interpolation enables the model to learn amplitude scaling, dynamically adjusting the flipped spectrum adaptively for different datasets and forecasting scenarios.

C Supplementary Details of the Experiments

Experiment Settings. We follow the same data processing and train-validation-test set split protocol employed in iTransformer (Liu et al. 2024). All the experiments are implemented in PyTorch 2.0.1 and conducted on a single NVIDIA RTX 3070 GPU with 8GB. We utilize ADAM (Kingma and Ba 2014) with an initial learning rate in $\{5 \times 10^{-3}, 10^{-2}, 2 \times 10^{-2}\}$ and L2 loss for the model optimization. The training epochs is fixed to 10 with batch size in $\{128, 256\}$. The hidden size of FFNs is set from $\{128, 256, 512\}$ with intermediate LeakyReLU activation function. The codes have been uploaded as supplementary and will be publicly available soon.

Datasets. We evaluate the performance of our proposed Amplifier on eight popular datasets, including ETT, Electricity, Exchange, Traffic, and Weather datasets.

The ETT (Zhou et al. 2021) datasets contain two visions of the sub-dataset: ETTh and ETTm, which were collected from electricity transformers at intervals of 15 minutes and 1 hour, respectively, between July 2016 and July 2018.

The Electricity (Wu et al. 2021) dataset records the hourly electricity consumption of 321 clients from 2012 to 2014.

The Exchange (Lai et al. 2018) dataset collects panel data of daily exchange rates of eight different countries including Singapore, Australia, British, Canada, Switzerland, China, Japan, and New Zealand ranging from 1990 to 2016.

The Traffic (Sen, Yu, and Dhillon 2019) dataset contains hourly traffic data measured by 862 sensors on San Francisco Bay area freeways since January 1, 2015.

The Weather (Wu et al. 2021) dataset includes 21 meteorological factors collected every 10 minutes from the Weather Station of the Max Planck Biogeochemistry Institute in 2020. The data sampling interval is every 10 minutes.

The details of these datasets are presented in Table 5.

Datasets	ETTm1(2)	ETTh1(2)	Electricity	Exchange	Traffic	Weather
Channels	7	7	321	8	862	21
Frequency	15min	Hourly	Hourly	Daily	Hourly	10min
Timesteps	69680	17420	26304	7588	17544	52696
Information	Electricity	Electricity	Electricity	Economy	Traffic	Weather

Table 5: Summary of datasets.

Baselines. We choose ten well-acknowledged and state-of-the-art models for comparison to evaluate the effectiveness of our proposed Amplifier for time series forecasting, including MLP-based models, Frequency-based models, Transformer-based models, and TCN-based models. We introduce these models as below:

SparseTSF (Lin et al. 2024) marks a significant milestone in advancing lightweight models for long-term time series forecasting, based on the Cross-Period Sparse Forecasting technique. Code is available at this repository: <https://github.com/lss-1138/SparseTSF>.

RLinear (Li et al. 2023) employs linear mapping in long-term time series forecasting with RevIN (reversible normalization) and CI (Channel Independent) improve overall forecasting performance. Code is available at this repository: <https://github.com/plumprc/RTSF>.

DLinear (Zeng et al. 2023) utilizes a simple yet effective one-layer linear model to capture temporal relationships. Code is available at this repository: <https://github.com/curelab/LTSF-Linear>.

FreTS (Yi et al. 2024c) explores a novel direction and make a new attempt to apply frequency-domain MLPs for time series forecasting, benefiting from global view and energy compaction. Code is available at this repository: <https://github.com/aikunyi/FreTS>.

FITS (Xu, Zeng, and Xu 2023) is a lightweight yet powerful model for time series analysis, essentially functioning as a low-pass filter. Code is available at this repository: <https://github.com/VEWOXIC/FITS>.

Fredformer (Piao et al. 2024) addresses frequency bias in the Transformer architecture by introducing a framework that learns features equally across different frequency bands. Code is available at this repository: <https://github.com/chenzRG/Fredformer>.

iTransformer (Liu et al. 2024) applies the attention and feed-forward network on the inverted dimensions and regards independent series as variate tokens. Code is available at this repository: <https://github.com/thuml/iTransformer>.

PatchTST (Nie et al. 2023) introduces an efficient design for Transformer-based models in time series forecasting by incorporating two essential components: patching and a channel-independent structure. Code is available at this repository: <https://github.com/yuqinie98/PatchTST>.

Stationary (Liu et al. 2022b) proposes an effective approach to enhance series stationarity while updating the internal mechanism to reintegrate non-stationary information, thereby improving both data predictability and the model’s predictive performance. Code is available at this repository: https://github.com/thuml/Nonstationary_Transformers.

TimesNet (Wu et al. 2023) unravels complex temporal variations through a modular architecture, capturing both intraperiod and interperiod variations by converting the 1D time series into a collection of 2D tensors across multiple periods. Code is available at this repository: <https://github.com/thuml/TimesNet>.

Full Results. The full multivariate time series forecasting results of Amplifier are presented in Table 6, along with extensive evaluations of competitive counterparts. We compare extensive competitive models under different prediction lengths following the setting of iTransformer (Liu et al. 2024) and Fredformer (Piao et al. 2024). For Amplifier, FreTS, and FITS we report the forecasting performance under five runs.

Models	Amplifier		RLinear		DLinear		FreTS		FITS		Fredformer		iTransformer		PatchTST		Stationary		TimesNet		
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	
ETTm1	96	0.316	0.355	0.355	0.376	0.345	0.372	0.335	0.372	0.355	0.375	0.326	0.361	0.334	0.368	0.329	0.367	0.386	0.398	0.338	0.375
	192	0.361	0.381	0.391	0.392	0.380	0.389	0.388	0.401	0.392	0.393	0.363	0.380	0.377	0.391	0.367	0.385	0.459	0.444	0.374	0.387
	336	0.393	0.402	0.424	0.415	0.413	0.413	0.421	0.426	0.424	0.414	0.395	0.403	0.426	0.420	0.399	0.410	0.495	0.464	0.410	0.411
	720	0.455	0.440	0.487	0.450	0.474	0.453	0.486	0.465	0.487	0.449	0.453	0.438	0.491	0.459	0.454	0.439	0.585	0.516	0.478	0.450
	Avg	0.381	0.394	0.414	0.407	0.403	0.407	0.408	0.416	0.415	0.408	0.384	0.395	0.407	0.410	0.387	0.400	0.481	0.456	0.400	0.406
ETTm2	96	0.176	0.258	0.182	0.265	0.193	0.292	0.189	0.277	0.183	0.266	0.177	0.259	0.180	0.264	0.175	0.259	0.192	0.274	0.187	0.267
	192	0.239	0.300	0.246	0.304	0.284	0.362	0.258	0.326	0.247	0.305	0.241	0.300	0.250	0.309	0.241	0.302	0.280	0.339	0.249	0.309
	336	0.297	0.338	0.307	0.342	0.369	0.427	0.343	0.390	0.307	0.342	0.302	0.340	0.311	0.348	0.305	0.343	0.334	0.361	0.321	0.351
	720	0.393	0.396	0.407	0.398	0.554	0.522	0.495	0.480	0.407	0.399	0.397	0.396	0.412	0.407	0.402	0.400	0.417	0.413	0.408	0.403
	Avg	0.276	0.323	0.286	0.327	0.350	0.401	0.321	0.368	0.286	0.328	0.279	0.324	0.288	0.332	0.281	0.326	0.306	0.347	0.291	0.333
ETTth1	96	0.371	0.392	0.386	0.395	0.386	0.400	0.395	0.407	0.386	0.396	0.373	0.392	0.386	0.405	0.414	0.419	0.513	0.491	0.384	0.402
	192	0.426	0.422	0.437	0.424	0.437	0.432	0.448	0.440	0.436	0.423	0.433	0.420	0.441	0.436	0.460	0.445	0.534	0.504	0.436	0.429
	336	0.448	0.434	0.479	0.446	0.481	0.459	0.499	0.472	0.478	0.444	0.470	0.437	0.487	0.458	0.501	0.466	0.588	0.535	0.491	0.469
	720	0.476	0.464	0.481	0.470	0.519	0.516	0.558	0.532	0.502	0.495	0.467	0.456	0.503	0.491	0.500	0.488	0.643	0.616	0.521	0.500
	Avg	0.430	0.428	0.446	0.434	0.456	0.452	0.475	0.463	0.451	0.440	0.435	0.426	0.454	0.447	0.469	0.454	0.570	0.537	0.458	0.450
ETTth2	96	0.279	0.337	0.288	0.338	0.333	0.387	0.309	0.364	0.295	0.350	0.293	0.342	0.297	0.349	0.302	0.348	0.476	0.458	0.340	0.374
	192	0.359	0.389	0.374	0.390	0.477	0.476	0.395	0.425	0.381	0.396	0.371	0.389	0.380	0.400	0.388	0.400	0.512	0.493	0.402	0.414
	336	0.377	0.406	0.415	0.426	0.594	0.541	0.462	0.467	0.426	0.438	0.382	0.409	0.428	0.432	0.426	0.433	0.552	0.551	0.452	0.452
	720	0.420	0.432	0.420	0.440	0.831	0.657	0.721	0.604	0.431	0.446	0.415	0.434	0.427	0.445	0.431	0.446	0.562	0.560	0.462	0.468
	Avg	0.359	0.391	0.374	0.398	0.559	0.515	0.472	0.465	0.383	0.408	0.365	0.393	0.383	0.407	0.387	0.407	0.526	0.516	0.414	0.427
ECL	96	0.147	0.243	0.201	0.281	0.197	0.282	0.176	0.258	0.200	0.278	0.147	0.241	0.148	0.240	0.181	0.270	0.169	0.273	0.168	0.272
	192	0.157	0.251	0.201	0.283	0.196	0.285	0.175	0.262	0.200	0.280	0.165	0.258	0.162	0.253	0.188	0.274	0.182	0.286	0.184	0.289
	336	0.174	0.269	0.215	0.298	0.209	0.301	0.185	0.278	0.214	0.295	0.177	0.273	0.178	0.269	0.204	0.293	0.200	0.304	0.198	0.300
	720	0.206	0.296	0.257	0.331	0.245	0.333	0.220	0.315	0.255	0.327	0.213	0.304	0.225	0.317	0.246	0.324	0.222	0.321	0.220	0.320
	Avg	0.171	0.265	0.219	0.298	0.212	0.300	0.189	0.278	0.217	0.295	0.175	0.269	0.178	0.270	0.216	0.304	0.193	0.296	0.192	0.295
Exchange	96	0.083	0.202	0.093	0.217	0.088	0.218	0.091	0.217	0.084	0.203	0.084	0.202	0.086	0.206	0.088	0.205	0.111	0.237	0.107	0.234
	192	0.175	0.297	0.184	0.307	0.176	0.315	0.175	0.310	0.177	0.298	0.183	0.302	0.177	0.299	0.176	0.299	0.219	0.335	0.226	0.344
	336	0.328	0.414	0.351	0.432	0.313	0.427	0.334	0.434	0.321	0.410	0.335	0.418	0.331	0.417	0.301	0.397	0.421	0.476	0.367	0.448
	720	0.858	0.696	0.886	0.714	0.839	0.695	0.716	0.674	0.828	0.685	0.893	0.711	0.847	0.691	0.901	0.714	1.092	0.769	0.964	0.746
	Avg	0.361	0.402	0.378	0.417	0.354	0.414	0.329	0.409	0.353	0.399	0.374	0.408	0.360	0.403	0.367	0.404	0.461	0.454	0.416	0.443
Traffic	96	0.455	0.298	0.649	0.389	0.650	0.396	0.593	0.378	0.651	0.391	0.406	0.277	0.395	0.268	0.462	0.295	0.612	0.338	0.593	0.321
	192	0.470	0.316	0.601	0.366	0.598	0.370	0.595	0.377	0.602	0.363	0.426	0.290	0.417	0.276	0.466	0.296	0.613	0.340	0.617	0.336
	336	0.479	0.316	0.609	0.369	0.605	0.373	0.609	0.385	0.609	0.366	0.432	0.281	0.433	0.283	0.482	0.304	0.618	0.328	0.629	0.336
	720	0.523	0.328	0.647	0.387	0.645	0.394	0.673	0.418	0.647	0.385	0.463	0.300	0.467	0.302	0.514	0.322	0.653	0.355	0.640	0.350
	Avg	0.482	0.315	0.626	0.378	0.625	0.383	0.618	0.390	0.627	0.376	0.431	0.287	0.428	0.282	0.481	0.304	0.624	0.340	0.620	0.336
Weather	96	0.156	0.204	0.192	0.232	0.196	0.255	0.174	0.208	0.166	0.213	0.163	0.207	0.174	0.214	0.177	0.218	0.173	0.223	0.172	0.220
	192	0.209	0.249	0.240	0.271	0.237	0.296	0.219	0.250	0.213	0.254	0.211	0.251	0.221	0.254	0.225	0.259	0.245	0.285	0.219	0.261
	336	0.264	0.290	0.292	0.307	0.283	0.335	0.273	0.290	0.269	0.294	0.267	0.292	0.278	0.296	0.278	0.297	0.321	0.338	0.280	0.306
	720	0.343	0.342	0.364	0.353	0.345	0.381	0.334	0.332	0.346	0.343	0.343	0.341	0.358	0.347	0.354	0.348	0.414	0.410	0.365	0.359
	Avg	0.243	0.271	0.272	0.291	0.265	0.317	0.250	0.270	0.249	0.276	0.246	0.272	0.258	0.278	0.259	0.281	0.288	0.314	0.259	0.287
1 st Count	28	25	0	0	0	0	3	3	0	1	6	12	3	5	1	1	0	0	0	0	

Table 6: Full results of eight datasets, with the best results are in red and the second best are blue. We set the lookback window size L as 96 and the prediction length as $\tau \in \{96, 192, 336, 720\}$. Avg means the average results from all four prediction lengths.