SDformer: Transformer with Spectral Filter and Dynamic Attention for Multivariate Time Series Long-term Forecasting

Ziyu Zhou[1] , **Gengyu Lyu**[1*] , **Yiming Huang**[1] , **Zihao Wang**[1] , **Ziyu Jia**[2] and **Zhen Yang**[1]
[1]Faculty of Information Technology, Beijing University of Technology, Beijing, China
[2]Institute of Automation, Chinese Academy of Sciences, Beijing, China
{ziyuzhou30,rex.wangzihao}@gmail.com, huangyiming2002@126.com, jia.ziyu@outlook.com,
{lyugengyu,yangzhen}@bjut.edu.cn

Abstract

Transformer has gained widespread adoption in modeling time series due to the exceptional ability of its self-attention mechanism in capturing long-range dependencies. However, when processing time series data with numerous variates, the vanilla self-attention mechanism tends to distribute attention weights evenly and smoothly, causing row-homogenization in attention maps and further hampering time series forecasting. To tackle this issue, we propose an advanced Transformer architecture entitled SDformer, which designs two novel modules, Spectral-Filter-Transform (SFT) and Dynamic-Directional-Attention (DDA), and integrates them into the encoder of Transformer to achieve more intensive attention allocation. Specifically, the SFT module utilizes the Fast Fourier Transform to select the most prominent frequencies, along with a Hamming Window to smooth and denoise the filtered series data; The DDA module applies a specialized kernel function to the query and key vectors projected from the denoised data, concentrating this innovative attention mechanism more effectively on the most informative variates to obtain a sharper attention distribution. These two modules jointly enable attention weights to be more salient among numerous variates, which in turn enhances the attention's ability to capture multivariate correlations, improving the performance in forecasting. Extensive experiments on public datasets demonstrate its superior performance over other state-of-the-art models. Code is available at https://github.com/zhouziyu02/SDformer.

1 Introduction

Time series analysis holds significant value in a wide array of practical applications, including weather forecasting [Huang *et al.*, 2023], energy management [Dong *et al.*, 2023] and public opinion analysis [O'Connor *et al.*, 2010]. Recently, Transformer [Vaswani *et al.*, 2017] has fostered the modeling of long-range dependencies in sequential data, making them particularly suited for time series analysis [Wen *et al.*, 2023]. For instance, Pyraformer [Liu *et al.*, 2021] proposes a hierarchical pyramidal attention mechanism for Transformer architecture, which captures temporal dependencies at different ranges while ensuring its linear complexity in both time and memory. Autoformer [Wu *et al.*, 2021] introduces a seasonal-trend decomposition operation into Transformer, which incorporates an auto-correlation mechanism to achieve more precise series-wise correlations discovery.

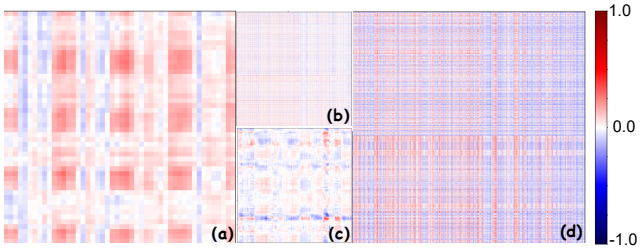

Figure 1: Pre-softmax attention maps from one layer in (a)PatchTST, (b)iTransformer, (c)Transformer and (d)SDformer. We also compute Gini coefficients for the attention matrices, which provide a clear indicator of the focus level of the attention mechanism across varieties. We observe that PatchTST, iTransformer and vanilla Transformer score **0.098**, **0.078** and **0.081** respectively, whereas SDformer achieves a higher value of **0.154**. This numeric disparity suggests that SDformer has a more concentrated attention distribution, indicating its improved ability to enhance focus on key variates in time series thereby alleviating the 'smooth' problem.

Although the above methods have achieved superior performance, they still suffer from some critical shortcomings. Especially, when modeling some datasets with a large number of variates (e.g., Traffic dataset with 862 variates), the vanilla self-attention mechanism often fails to effectively allocate attention weights among multiple variates, which is intuitively reflected in the homogenization between rows of the attention map. As shown in Figure 1(a)-(b), the color distribution of the attention map in iTransformer [Liu *et al.*, 2023] and vanilla Transformer [Vaswani *et al.*, 2017] is quite sparse with fewer intense areas, indicating a dispersed focus and an inability to prioritize critical inter-variate correlations.

To tackle this issue, we propose an innovative Transformer-based model named **SD**former, which integrates two strategic modular designations, **S**pectral-Filter-Transform (SFT) and

*Gengyu Lyu is the corresponding author.

Dynamic-Directional-Attention (DDA), to reallocate attention weights for increasing the heterogeneity of the attention map, further boosting the forecasting performance. Technically, in the SFT module, we utilize the Fast Fourier Transform to filter out insignificant frequencies including meaningless noise and fluctuations, accordingly preserving the essential features of the time series, such as continuity, periodicity and trend. Afterward, a bell-shaped Hamming Window is applied to the filtered data, which utilizes its spectral properties to minimize edge effects and enhance the smoothness of the series [Mottaghi-Kashtiban and Shayesteh, 2011]. The above two operations significantly improve data quality by effectively reducing noise, further facilitating better representation learning in subsequent modules. In the DDA module, we introduce a novel kernel function, equipped with dynamic parameters and directional weights, which is applied to Query (Q) and Key (K) simultaneously (where Q and K are linear projected from the smoothed series in the SFT) to bring similar Q-K pairs closer to their nearest axis, yielding higher attention scores, while distancing dissimilar Q-K pairs to their opposite axis, resulting in relatively lower scores. Such operation sharpens the distribution of attention weights across numerous variates, and consequently makes the self-attention mechanism more capable of identifying and prioritizing key inter-variate patterns, thereby achieving higher expressiveness and effectively mitigating the 'smooth' problem.

As a result, Figure 1(c) visually shows more intense colors in the attention map compared with the other two maps on certain regions, indicating a stronger focus capacity on important variates. In general, these two novel components collectively overcome the limitation of the vanilla attention mechanism in Transformer when modeling time series with a large number of variates, enabling a stronger capacity for forecasting multivariate time series data. The contributions of our paper are summarized in three folds:

1. We propose a novel Transformer architecture (named SDformer) for long-term time series forecasting. To the best of our knowledge, it is **the first time** to address the problem of smooth attention distribution when modeling time series data with a large number of variates.
2. The Spectral-Filter-Transform and Dynamic-Directional Attention modules are designed for filtering pivotal frequency features and sharpening attention distribution respectively, which jointly enable the attention weights to be more salient among variates, further fostering the attention's ability to capture multivariate correlations and improve the final forecasting performance.
3. Extensive experiments on various datasets demonstrate the effectiveness of SDformer against other state-of-the-art methods. Especially, it achieves superior performance on some datasets with numerous variates, such as 11.6% forecasting error reduction on Traffic dataset.

2 Related Work

2.1 Forecasting with Special Attention Mechanism

In multivariate time series forecasting, several models focus on innovating attention mechanisms to enhance the forecasting performance. For example, Informer [Zhou *et al.*, 2021] and Reformer [Kitaev *et al.*, 2020] leverage ProbSparse and locality-sensitive hashing mechanisms in modeling long sequences data. Pyraformer [Liu *et al.*, 2021] employs a pyramidal attention module to capture multi-resolution temporal dependency, while ContiFormer [Chen *et al.*, 2023] merges attention with Neural ODEs for irregular time series. Moreover, PatchTST [Nie *et al.*, 2023] focuses on excavating patch-wise correlations using self-attention. All of these attention mechanisms are designed from the perspective of balancing between computational complexity and forecasting performance. To the best of our knowledge, none of the existing attention mechanisms effectively address the issue of overly smooth attention distribution in time series analysis.

2.2 Forecasting in the Frequency Domain

Except for the above attention innovation-based methods, various methods harness strategies in the frequency domain with different foundation models to improve forecasting results. SMF [Zhang *et al.*, 2017] decomposes time series into distinct frequency components for varied time horizon forecasts. StemGNN [Cao *et al.*, 2020] combines GNN and Discrete Fourier Transforms to examine both inter-series and intra-series correlations. CoST [Woo *et al.*, 2022a] and ATFN [Yang *et al.*, 2020] focus on seasonal-trend separation and dynamic periodic pattern capture. FreTS [Yi *et al.*, 2023] leverages its frequency-domain MLPs to explore global dependencies and enhance key frequency component learning. Although these models signify a trend towards exploring frequency domain strategies in time series analysis, they often directly utilize enhanced data for subsequent learning without considering potential noise. This oversight leads to the inadvertent amplification of noise components during frequency decomposition, compromising the forecasting accuracy.

2.3 Forecasting with Special Attention and Methods in the Frequency Domain

Based on the above two types of methods, some recent methods combine both innovative attention mechanisms and frequency domain analysis to achieve notable progress in time series forecasting. For example, Autoformer [Wu *et al.*, 2021] utilizes an Auto-Correlation mechanism for improved temporal pattern recognition. FEDformer [Zhou *et al.*, 2022] fuses Transformer architecture with seasonal-trend decomposition, exploiting frequency domain sparsity for thorough analysis. ETSformer [Woo *et al.*, 2022b] introduces exponential smoothing attention, enhancing both efficiency and interpretability. However, existing models overlook the amplification of noise during frequency decomposition and the challenge of overly smooth attention distribution. To the best of our knowledge, SDformer is the first model to tackle these overlooked issues by integrating the SFT and DDA into the Transformer architecture, which significantly enhances the accuracy of multivariate time series forecasting.

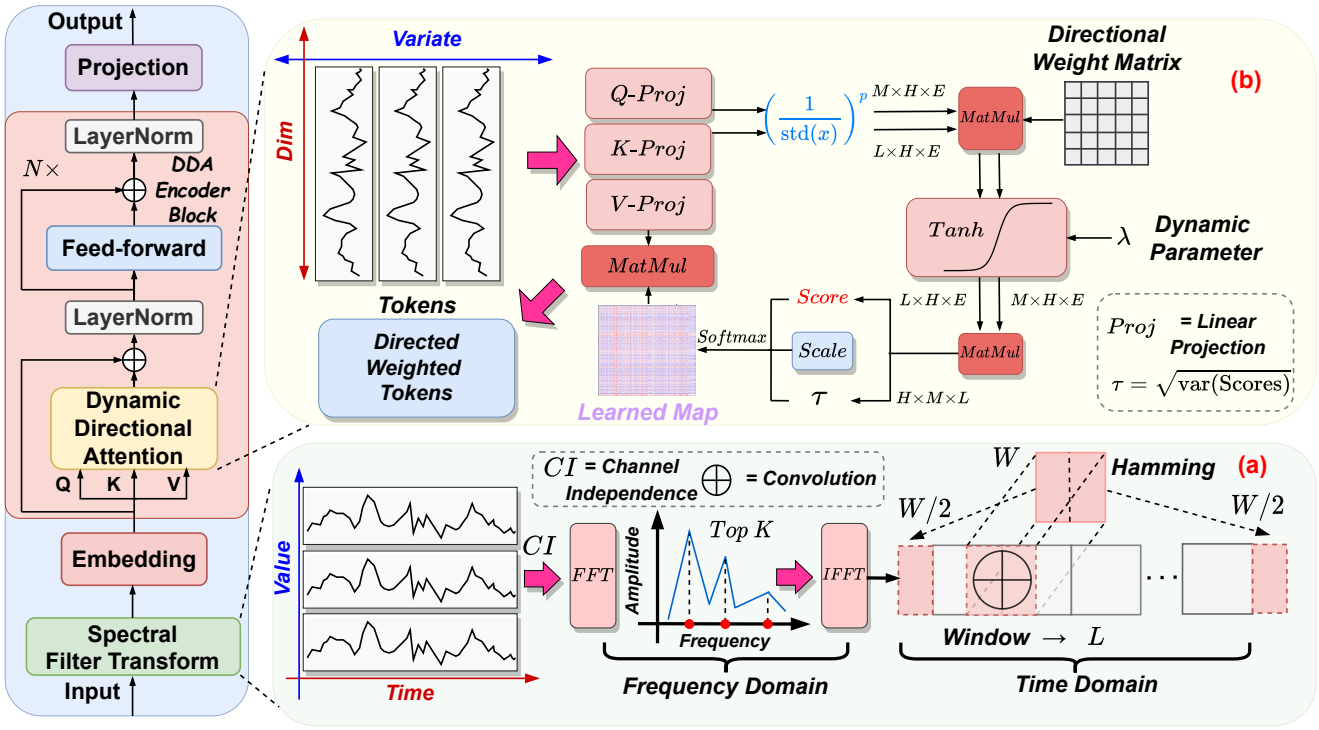

Figure 2: The framework of our proposed SDformer: (a) The Spectral-Filter-Transform (SFT) is a novel preprocessing method for time series, enhancing data analysis by retaining key frequency components. (b) The Dynamic-Directional-Attention (DDA) mechanism is an improvement over the vanilla self-attention, as it dynamically modulates the closeness of Q-K pairs according to their similarity.

3 SDformer

3.1 Problem Formulation

Formally speaking, we denote a multivariate time series by $[X_1, X_2, \ldots, X_T] \in \mathbb{R}^{N \times T}$, where each $X_t \in \mathbb{R}^N$ corresponds to the observations of N variates at the t-th timestamp for T total timestamps. For any given time t, the input to the model is a window of the preceding L observations, designated as $\mathbf{X_t} = [X_{t-L+1}, X_{t-L+2}, \ldots, X_t] \in \mathbb{R}^{N \times L}$. The forecasting objective at time t aims to predict the subsequent τ values, represented as $\mathbf{Y_t} = [X_{t+1}, X_{t+2}, \ldots, X_{t+\tau}] \in \mathbb{R}^{N \times \tau}$. The forecasting model, denoted by f_θ, utilizes the historical data $\mathbf{X_t}$ to estimate the future values $\hat{\mathbf{Y}}_t$, such that the forecast is given by $\hat{\mathbf{Y}}_t = f_\theta(\mathbf{X_t})$.

3.2 Overall Architecture

Figure 2 shows the overall framework of SDformer including the Spectral-Filter-Transform (SFT) module, the Embedding operation, the stacked Dynamic-Directional-Attention (DDA) Encoder Blocks module and the Projection operation. Technically, the input time series is denoised in the SFT, and its variates are independently embedded into separate tokens [Liu *et al.*, 2023]. These tokens are then fed into stacked DDA Encoder Blocks to extract complex representation, where the DDA achieves improved inter-variate correlations discovery through its unique kernel function, while layer normalization and feed-forward network with residual connections are utilized to learn temporal dependencies and alleviate series non-stationarization. Finally, the representations are decoded for final results via the Projection operation. Since our model focuses on addressing the insufficiency of attention mechanism in modeling time series with numerous variates, in this paper, we chiefly introduce the designed SFT and DDA, and explain how they discover effective multivariate correlations.

Algorithm 1 The Spectral-Filter-Transform module

Input: Time series $X \in \mathbb{R}^{T \times N}$, Length T, Variates N
Output: Denoised and smoothed series $X_h \in \mathbb{R}^{T \times N}$
Initialization: A Hamming Window w_n sized w; the number of top frequency components k.
for $n = 1$ **to** N **do**
  $X_{f_n} = \text{FFT}(x_n)$ {Fast Fourier Transform}
  $X_{f_{k_n}} = \text{TopK}(X_{f_n}, k)$ {Select k frequencies}
  $x_{if_n} = \text{IFFT}(X_{f_{k_n}})$ {Convert to the time domain}
  $x_{p_n} = \text{Reflective Padding}(x_{if_n}, w_n)$
  $x_{h_n} = \text{Applying Window}(x_{p_n}, w_n)$
end for
$X_h = \text{Concat}(x_{h_n})$ {Concatenate N univariate series}
Return: X_h

3.3 Spectral-Filter-Transform (SFT)

The Spectral-Filter-Transform (SFT) module plays a pivotal role in denoising and smoothing the input time series data at the beginning of our model. It achieves this data augmentation through a two-stage cross-domain process as illustrated in Algorithm 1: **Frequency Domain Denoising** (step 5, 6 and

7) and **Time Domain Smoothing** (step 8 and 9). It is noteworthy that, before conducting the frequency domain denoising, we treat multivariate time series as multiple univariate time series (inspired by the 'channel-independence' designation [Nie *et al.*, 2023]), processing them individually in the SFT and eventually concatenating all the processed univariate series for subsequent operations.

Frequency Domain Denoising

Denoising is crucial in time series analysis to enhance forecasting accuracy by mitigating the impact of noise on the identification of periodic or trend patterns. In this paper, we propose a new denoising strategy that converts the time series into its corresponding frequency domain to alleviate the random environmental noise and white noise in the time series. Specifically, we first utilize the Fast Fourier Transform (FFT) to convert a univariate time series $x \in \mathbb{R}^T$ with length T into the frequency domain X_f, i.e.,

$$X_f = \sum_{t=0}^{T-1} x[t] e^{-\frac{2\pi i}{T} kt}, \quad k = 0, \ldots, T-1. \tag{1}$$

Then, we retain the highest k frequency components to filter out insignificant frequencies by $X_{f_k} = \text{TopK}(X_f, k)$, where $\text{TopK}(X_f, k)$ selects the k largest amplitudes and k is the hyper-parameter. Afterward, the retained frequency spectrum X_{f_k} is transformed back into its corresponding time domain x_{if} by adopting the Inverse Fast Fourier Transform (IFFT) for further analysis in subsequent operations, i.e.,

$$x_{if} = \frac{1}{T} \sum_{k=0}^{T-1} X_{f_k} e^{\frac{2\pi i}{T} kt}, \quad t = 0, \ldots, T-1. \tag{2}$$

Time Domain Smoothing

After obtaining the denoised time series x_{if}, we further process the time series by implementing smoothing techniques to mitigate spectral leakage effects [Mottaghi-Kashtiban and Shayesteh, 2011], where a convolution operation by windowing function is useful to smooth transitions at the boundaries and reduce discontinuities. Specifically, we first define a bell-shaped Hamming Window with the size of w (w is an even number) and a window function

$$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{w}\right), \quad n = 1, \ldots, w, \tag{3}$$

where n indexes the sample points within the window. Subsequently, the denoised series x_{if} from Eq.2 is padded reflectively to ensure its length matches the window size w by

$$x_p[n] = \begin{cases} x_{if}\left[\frac{w}{2} - n\right], & 1 \le n \le \frac{w}{2} \\ x_{if}[n - \frac{w}{2}], & \frac{w}{2} < n \le N + \frac{w}{2} \\ x_{if}[N + w - n], & N + \frac{w}{2} < n \le N + w \end{cases} \tag{4}$$

According to Eq.4, the first and last $\frac{w}{2}$ elements of the filtered series x_{if} are mirrored and appended at the beginning and end of the series respectively. Such operation effectively extends the series $x[n]$ to a new length of $N + w$, where N is the original length of x_{if}, thereby adapting to the subsequent convolution smoothing operation, i.e.,

$$x_h[t] = \frac{\sum_{n=1}^{w} x_p[t+n] \cdot w[n]}{\sum_{n=1}^{w} w[n]}, \quad t = 1, \ldots, T, \tag{5}$$

where the smoothed series $x_h[t]$ is calculated by taking a weighted average of the series at each time point t and T denotes the total length.

After the above operations, we concatenate all x_h corresponded to N variates to obtain a multivariate time series $X \in \mathbb{R}^{T \times N}$ with length T and the number of variates N, where the X maintains the same shape as the input sequence of the SFT, simultaneously achieving both denoising and smoothing effectively.

3.4 Dynamic-Directional-Attention (DDA)

To capture more reliable multivariate correlations, we introduce a new Dynamic-Directional-Attention (DDA) mechanism to calculate attention weights for more effective attention distribution. The DDA mechanism operates on the principle of dynamically reorienting and scaling the Query $\mathbf{Q} \in \mathbb{R}^{M \times H \times E}$ and Key $\mathbf{K} \in \mathbb{R}^{L \times H \times E}$, where M and L are sequence lengths, H is the number of heads, and E denotes dimensions per head. The attention score within each head of the DDA is formulated as:

$$Score(\mathbf{Q}_i, \mathbf{K}_j) = \phi_p(\mathbf{Q}_i)\phi_p(\mathbf{K}_j)^T, \tag{6}$$

where $\phi_p(x) = f_p(\tan(x))$ is a specially designed function that is applied to both the query and key simultaneously. Here, $\phi_p(x)$ is defined by the composition of a nonlinear mapping $tan(x)$ and a special kernel function $f_p(x) = x \cdot w_{\text{dir}} \cdot (\text{std}(x))^{-p} \cdot \lambda_{\text{dyn}}$, where p is the element-wise power, w_{dir} and λ_{dyn} are learnable parameters representing directional weight and dynamic parameter, respectively. These two parameters contribute to the 'dynamic' effect of our proposed DDA. Moreover, $\text{std}(x)$ represents the standard deviation of the input x.

In addition, the DDA introduces a dynamic scaling factor τ to calculate the attention weight $\mathbf{A}$, which is formulated as:

$$\mathbf{A} = \text{Dropout}\left(\text{Softmax}\left(\frac{scale \cdot Score}{\tau}\right)\right), \tag{7}$$

where $\tau = \sqrt{\text{var}(Score)}$ dynamically normalizes the scores, and $\text{var}(Score)$ calculates the variance of score, *scale* is a pre-softmax scaling factor. A dropout procedure follows the softmax to regularize the attention weights $\mathbf{A}$, reducing overfitting by randomly zeroing a portion of the weights during the training phase. Finally, the $\mathbf{Output}$ of this module is computed as a weighted sum of the value matrix:

$$\mathbf{Output} = \sum_s \mathbf{A} \cdot \mathbf{V}. \tag{8}$$

Analysis on the Kernel Function

The essence of the DDA module lies in its specialized kernel function f_p, which effectively sharpens the contrast in the attention distribution for all the Q-K pairs, so that the issue

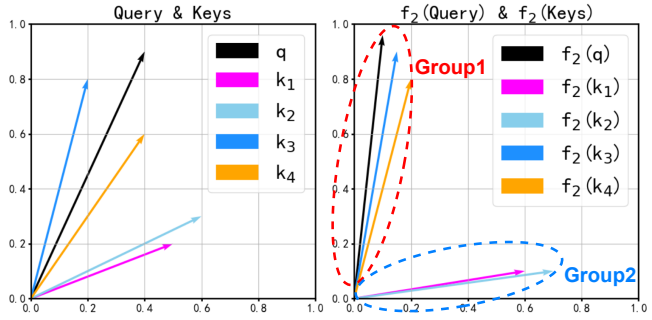

Figure 3: The vectors include a query q before applying the kernel function, and a query $f_2(q)$ after applying the kernel function. They also include keys k_1 to k_4 before the application of the kernel function, and keys $f_2(k_1)$ to $f_2(k_4)$ after its application.

of smooth attention weight distribution can be alleviated. To illustrate its effect more clearly, Figure 3 provides an example demonstrating the impact of f_p. Essentially, f_p draws each vector closer to its nearest axis. The parameter p is crucial here, as it determines the extent of this vector reorientation. This process aids in categorizing the vectors into distinct groups based on their proximity to specific axes as illustrated in the two groups in Figure 3. As a result, the similarities of Q-K pairs within the same group are enhanced, while those within different groups are weakened. Here, 'similarity' represents the attention score; therefore, the amplified similarity distribution results in a sharper attention weight distribution.

4 Experiments

4.1 Experimental Setting

Datasets. We employ seven real-world time series benchmarks for comparative study, including Weather, Exchange, Traffic, Illness, Electricity and ETT (2 subsets)[1].

Baselines. We compare SDformer with eleven state-of-the-art models from three categories, including (1) Transformer-based models: iTransformer [Liu *et al.*, 2023], PatchTST [Nie *et al.*, 2023], FEDformer [Zhou *et al.*, 2022], Autoformer [Wu *et al.*, 2021]. (2) TCN-based model: TimesNet [Wu *et al.*, 2023] and (3) MLP-based models: DLinear [Zeng *et al.*, 2023].

Evaluation Metrics. Two widely-used evaluation metrics including Mean Squared Error (MSE) and Mean Absolute Error (MAE) are employed for quantifying the accuracy of predictions among all comparing methods.

Implementation Details. Our experiments are conducted using PyTorch on a single NVIDIA RTX3090 24GB GPU. For model optimization, we employ the ADAM optimizer [Kingma and Ba, 2014] to optimize the L2 loss, selecting an initial learning rate from $\{1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}\}$.

4.2 Experimental Results

Table 1 summarizes the notable superiority of SDformer in long-term forecasting, particularly on some datasets with numerous variates, e.g., Traffic dataset. Specifically, SDformer has an average decrease of 15% and 8.5% in MSE and MAE over the previous SOTA iTransformer [Liu *et al.*, 2023], highlighting its enhanced ability in modeling high-dimensional time series data as well as capturing reliable multivariate correlations. Besides, SDformer also exhibits its superior ability across other datasets. For instance, it outperforms another SOTA DLinear [Zeng *et al.*, 2023] by 17% in MSE and MAE on ETTm2 dataset. In a nutshell, the enhanced ability to handle complex multivariate correlations makes SDformer particularly suitable for tackling the intricacies of multivariate time series forecasting challenges.

[1]Datasets are provided in Autoformer [Wu *et al.*, 2021].

4.3 Model Analysis

The Effectiveness of the SFT Module

Figure 4 presents two comparative case analyses on the Spectral-Filter-Transform (SFT), where the solid lines represent the post-SFT series and the corresponding dashed lines depict the pre-SFT series. The solid lines exhibit fewer fluctuations, indicating a conspicuous attenuation of noise and an enhanced smoothness compared to their dashed counterparts. This denoising and smoothing effect is pivotal for the subsequent extraction of semantic temporal patterns like trends and continuity, and is also essential for discovering multivariate correlations, further benefiting precise series forecasting.

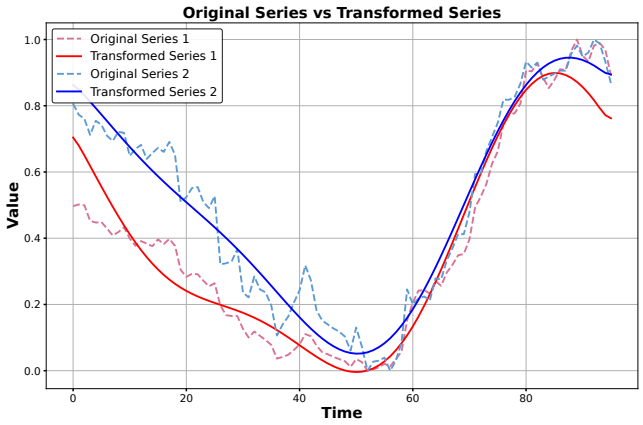

Figure 4: Comparison of original and transformed time series before and after the Spectral-Filter-Transform module. Two cases are selected from Traffic dataset and colored in red and blue.

The Effectiveness of the DDA Module

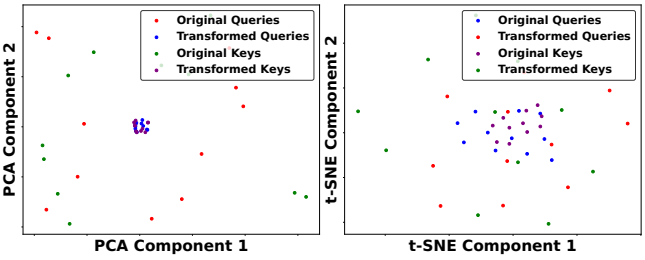

Figure 5: Query and Key shows a more intensive distribution after applying the kernel function. Visualization after PCA and t-SNE.

Methods	SDformer		iTransformer		DLinear		PatchTST		TimesNet		FEDformer		Autoformer	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
$ETTm2$	0.288	0.332	0.291	0.334	0.350	0.401	**0.255**	**0.327**	0.291	0.333	0.305	0.349	0.327	0.371
$ETTh2$	**0.378**	**0.401**	0.384	0.407	0.559	0.515	0.380	0.406	0.414	0.427	0.437	0.449	0.450	0.459
$Weather$	**0.258**	**0.278**	0.261	0.281	0.265	0.317	0.354	0.348	0.259	0.287	0.309	0.360	0.338	0.382
ECL	**0.176**	0.269	0.180	**0.261**	0.212	0.300	0.204	0.291	0.192	0.295	0.214	0.327	0.227	0.338
$Exchange$	0.356	**0.404**	0.365	0.407	**0.354**	0.414	0.362	0.404	0.416	0.443	0.519	0.429	0.613	0.539
$Traffic$	**0.408**	**0.278**	0.423	0.282	0.625	0.383	0.480	0.304	0.620	0.336	0.610	0.376	0.628	0.379
ILI	2.066	0.915	2.212	0.930	4.398	1.422	**1.443**	**0.797**	2.139	0.931	2.847	1.144	3.006	1.161

Table 1: Long-term forecasting results with forecasting lengths $O \in \{24, 36, 48, 60\}$ for ILI, and $O \in \{96, 192, 336, 720\}$ for other datasets and fixed lookback length $I = 96$. Results are averaged from various prediction lengths. **Red**: best, **Blue**: second best.

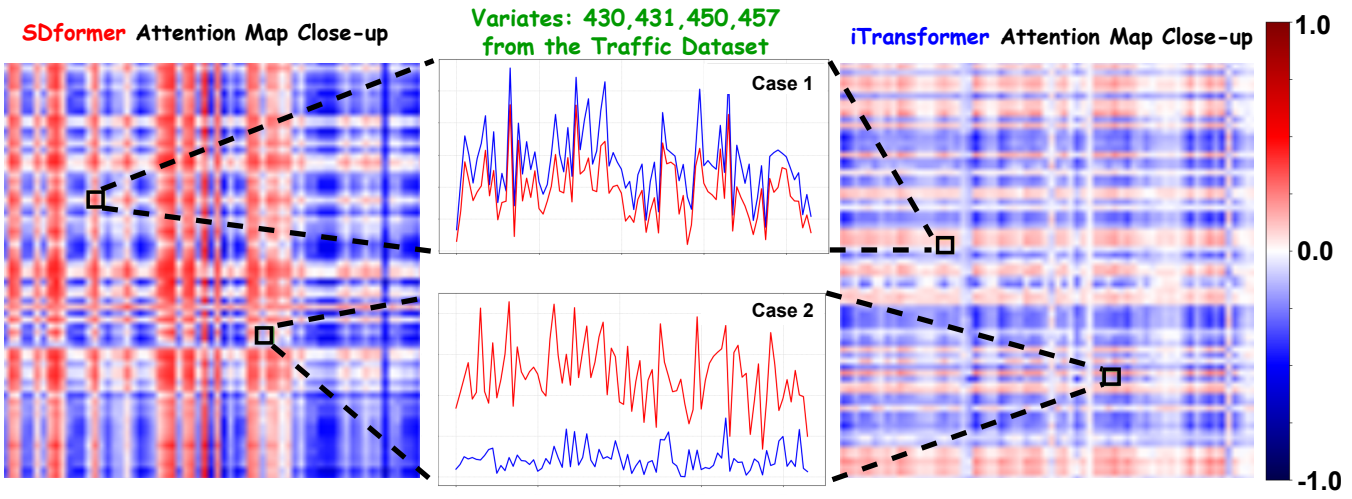

Figure 6: The visualization of attention maps for two sets of time series from the Traffic dataset in Case 1 and Case 2. It highlights specific points within the learned score maps that represent different degrees of similarity of variates. Due to the large size of the attention maps (i.e. 866×866), a close-up section of the entire map is captured to more clearly showcase the color differences of individual 'pixels' in each map.

In order to validate the effectiveness of the Dynamic-Directional-Attention (DDA) module, especially its special kernel function f_p, we present a comparative visualization in Figure 5 to show the Query and Key's distribution before and after applying the kernel function. Given that the Query and Key are tensors with high dimensions, we use two common dimension reduction methods (PCA and t-SNE) to map them into 2D space as scatters for intuitive comprehension. In the left figure, the post-function Query and Key scatters show more intensive grouping compared to their pre-function state, which is echoed in the right figure, where the post-function scatters are similarly more concentrated. The intensiveness indicates the kernel function effectively encloses Query and Key that are similar, potentially leading to more distinct attention distribution among variates.

The Joint Evaluation of the SFT and DDA Module

We also evaluate the effectiveness of our proposed SFT and DDA modules from a joint perspective on the learned multivariate attention map in Figure 6. Specifically, we randomly select two sets of time series called Case 1 and Case 2, where it is obvious that the similarity between the two series in Case 1 is higher than that in Case 2. Afterward, we compare the regions in attention maps that correspond to Case 1 for SDformer and iTransformer, which reveals that SDformer assigns a higher attention score to similar series than that in iTransformer, whereas the attention score in SDformer is relatively lower for the dissimilar series in Case 2. Such discrepancy indicates that SFT and DDA jointly increase the attention score for similar Q-K pairs and reduce it for dissimilar ones. Such a result demonstrates the capability of SDformer for more distinctive and uneven attention allocation among numerous variates, enabling more effective multivariate correlations discovery.

To further elaborate on the effectiveness of our proposed SFT and DDA in solving the smooth attention distribution, we compute the Gini coefficients and ranks of the attention matrices for each layer in SDformer and iTransformer, which are summarized in Table 2. These two metrics are used to assess the model's capability to prioritize significant inter-variate correlations by analyzing the distribution of attention weights, where higher Gini coefficients and ranks indicate

Methods	iTransformer		SDformer	
Metric	Gini	Rank	Gini	Rank
Layer 1	0.078	260	**0.154**	**344**
Layer 2	0.086	281	**0.244**	**375**
Layer 3	0.104	302	**0.223**	**365**
Layer 4	0.095	296	**0.268**	**459**

Table 2: Comparison of the Gini coefficients and the ranks of attention matrix from iTransformer and SDformer in each Encoder layer, where we separately train the two models with four encoder layers.

a more uneven distribution [Han *et al.*, 2023]. As a result, SDformer consistently shows higher Gini coefficients across all layers than iTransformer, highlighting the more concentrated attention weights distribution. Furthermore, the rank in SDformer reaches 459 in the fourth layer, surpassing iTransformer's 296, which suggests a greater diversity in feature representation. Collectively, these quantitative results corroborate the effectiveness of SFT and DDA in overcoming attention matrix homogeneity, illustrating their efficiency in identifying key variates to capture multivariate correlations.

4.4 Further Analysis

Ablation Study

We conduct ablation studies on Weather and ETTm2 datasets to validate the indispensability of our proposed SFT and DDA module from three perspectives: (1) 'proposed' represents the intact SDformer proposed in this paper. (2) wo/DDA: the Dynamic-Directional-Attention in SDformer is replaced with the vanilla self-attention. (3) wo/SFT: the Spectral-Filter-Transform module is removed. Specifically, Figure 7 illustrates the results of the ablation studies, where removing any modular component will lead to performance degradation (higher MSE). Such results suggest that without the SFT, the model becomes more prone to interference from environmental or white noise, making it difficult to accurately capture the temporal patterns in time series data, which negatively impacts the forecasting accuracy. Additionally, we observe that replacing our DDA with the vanilla self-attention also results in a higher forecasting error, indicating that vanilla self-attention is less effective than our Dynamic-Directional-Attention in modeling multivariate time series with numerous variates. In summary, these two components collectively improve the model's performance in time series forecasting.

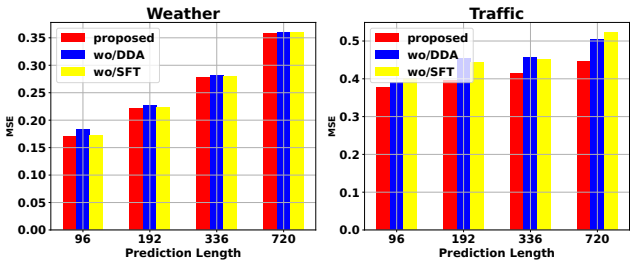

Figure 7: Ablation studies on Weather and Traffic dataset.

Hyperparameter Sensitivity

The hyperparameter analysis in Figure 8 examines the effects of top-k values and window size in the SFT module on Exchange dataset. We adjust the window size while keeping the top-k constant, and vice versa. We find that a larger top-k value correlates with lower MSE, suggesting that retaining more frequency components is crucial for capturing complex temporal dynamics in long-term forecasting. Conversely, a smaller selection of top-k values can distort key temporal features of the series, e.g., periodicity and trends. The analysis also reveals that the larger window size of the Hamming Window in the SFT module leads to higher MSE, implying that the huge window may distort the inherent characteristics of the input time series, especially in our small lookback window with 96 lengths. In summary, given the modest variations in MSE with changes in both the top-k values and window size in the SFT module, SDformer exhibits robustness, maintaining stable performance despite alterations in these hyperparameters.

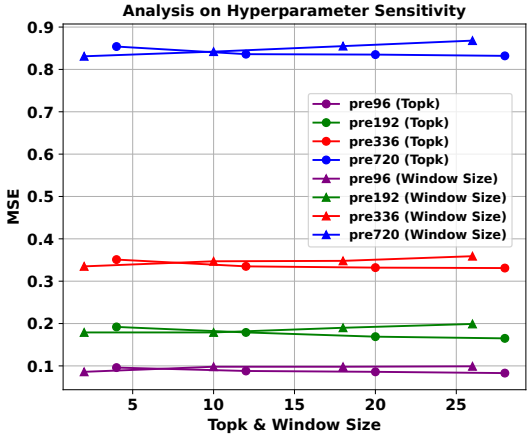

Figure 8: Analysis of the top-k and the window size of Hamming Window in the SFT module on Exchange dataset.

5 Conclusion

This paper proposes SDformer, a novel Transformer-based architecture that concentrates on tackling the dispersed distribution of attention weights in Transformer when modeling time series with numerous variates. Specifically, we introduce the Spectral-Filter-Transform module to denoise and enhance the smoothness of time series. Besides, we propose the Dynamic-Directional-Attention module to sharpen the distribution of attention weights on the most informative variates. These innovations jointly boost the ability of SDformer to discern and utilize multivariate correlations. Experiments across various datasets underscore its efficiency and accuracy, marking a notable advancement in long-term forecasting tasks. Future research will explore the scalability of SDformer on large-scale real-world datasets and deploy it to support public use.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2023YFB3107100), the National Natural Science Foundation of China (No. 62306020), the China Postdoctoral Science Foundation (No. 2022M720320), the Beijing Postdoctoral Science Foundation (No. 2023-zz-78), the Major Research Plan of National Natural Science Foundation of China (No. 92167102).

References

[Cao *et al.*, 2020] Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, et al. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in Neural Information Processing Systems*, 33:17766–17778, 2020.

[Chen *et al.*, 2023] Yuqi Chen, Kan Ren, Yansen Wang, Yuchen Fang, Weiwei Sun, and Dongsheng Li. Contiformer: Continuous-time transformer for irregular time series modeling. *Advances in Neural Information Processing Systems*, pages 1–33, 2023.

[Dong *et al.*, 2023] Jiaxiang Dong, Haixu Wu, Haoran Zhang, Li Zhang, Jianmin Wang, and Mingsheng Long. Simmtm: A simple pre-training framework for masked time-series modeling. In *Advances in Neural Information Processing Systems*, pages 1–30, 2023.

[Han *et al.*, 2023] Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention. In *IEEE/CVF International Conference on Computer Vision*, pages 5961–5971, 2023.

[Huang *et al.*, 2023] Yiming Huang, Ziyu Zhou, Zihao Wang, et al. Timesnet-pm2. 5: Interpretable timesnet for disentangling intraperiod and interperiod variations in pm2. 5 prediction. *Atmosphere*, 14(11):1604, 2023.

[Kingma and Ba, 2014] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, pages 1–15, 2014.

[Kitaev *et al.*, 2020] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

[Liu *et al.*, 2021] Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, et al. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International Conference on Learning Representations*, pages 1–20, 2021.

[Liu *et al.*, 2023] Liu, Tengge Hu, Haoran Zhang, Haixu Wu, et al. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, pages 1–24, 2023.

[Mottaghi-Kashtiban and Shayesteh, 2011] Mahdi Mottaghi-Kashtiban and Mahrokh G Shayesteh. New efficient window function, replacement for the hamming window. *IET Signal Processing*, 5(5):499–505, 2011.

[Nie *et al.*, 2023] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, pages 1–24, 2023.

[O'Connor *et al.*, 2010] Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the international AAAI conference on web and social media*, volume 4, pages 122–129, 2010.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:6000–6010, 2017.

[Wen *et al.*, 2023] Qingsong Wen, Tian Zhou, Chaoli Zhang, et al. Transformers in time series: A survey. In *International Joint Conference on Artificial Intelligence*, pages 6778–6786, 2023.

[Woo *et al.*, 2022a] Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, et al. CoST: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. In *International Conference on Learning Representations*, pages 1–18, 2022.

[Woo *et al.*, 2022b] Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. Etsformer: Exponential smoothing transformers for time-series forecasting. *arXiv preprint arXiv:2202.01381*, pages 1–18, 2022.

[Wu *et al.*, 2021] 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.

[Wu *et al.*, 2023] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, pages 1–23, 2023.

[Yang *et al.*, 2020] Zhangjing Yang, Weiwu Yan, Xiaolin Huang, et al. Adaptive temporal-frequency network for time-series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 34(4):1576–1587, 2020.

[Yi *et al.*, 2023] Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, et al. Frequency-domain MLPs are more effective learners in time series forecasting. In *Advances in Neural Information Processing Systems*, pages 1–24, 2023.

[Zeng *et al.*, 2023] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *AAAI Conference on Artificial Intelligence*, page 11121–11128, 2023.

[Zhang *et al.*, 2017] Liheng Zhang, Charu Aggarwal, and Guo-Jun Qi. Stock price prediction via discovering multi-frequency trading patterns. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2141–2149, 2017.

[Zhou *et al.*, 2021] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115, 2021.

[Zhou *et al.*, 2022] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, et al. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, pages 27268–27286, 2022.