AdaDyTS: Dynamic Multi-Scale Spectral Decoupling and Time-Variant Inference for Time Series Forecasting

Jinmei Qi[1], Liwen Hu[1(✉)], Zhuo Dong[1], Qiyu Wu[3], Leyun Kong[4], Jinlai Zhang[2(✉)]

[1] School of Physics and Electronic Science, Changsha University of Science and Technology, Changsha, 410114, Hunan, China
[2] College of Mechanical and Vehicle Engineering, Changsha University of Science and Technology, Changsha, 410114, Hunan, China
[3] School of Computer Science and Technology, Changsha University of Science and Technology, Changsha, 410114, Hunan, China
[4] Elite Engineering School, Changsha University of Science and Technology, Changsha, 410114, Hunan, China

Abstract. Time series forecasting is fundamental to intelligent decision-making systems, enabling proactive planning and resource optimization across diverse application domains. However, the inherent complexity of real-world time series—including multi-scale temporal patterns, heterogeneous variable dependencies, and dynamic non-stationarity—poses significant challenges for existing forecasting models. Current approaches often suffer from high-frequency information attenuation in frequency-domain modeling, inadequate characterization of scale heterogeneity across variables, and limited capability to capture time-varying dynamics. To address these challenges, this paper introduces AdaDyTS, a unified knowledge-driven forecasting framework that synergistically integrates three complementary mechanisms: multi-scale frequency-domain interpolation decoupling via the Cascaded Spectral Residual Extractor (CSRE), dynamic morphological perception via the Dynamic Morphological Perception Unit (DMP-U), and time-variant state-space inference via the Time-Variant State-Space Module (TV-SS). CSRE separates low-frequency trends from high-frequency residuals through coarse-to-fine layer-wise self-reconstruction, preserving transient information that static filters typically attenuate. DMP-U employs deformable convolution guided by multi-expert attention to adaptively adjust receptive fields, enabling fine-grained modeling of local fluctuations and nonlinear distortions. TV-SS relaxes the conventional time-invariant parameter assumption, dynamically modulating state transition parameters to capture both short-term variations and long-term dependencies. Under a unified evaluation protocol across 13 benchmark datasets, AdaDyTS achieves average improvements of 4.35% in MSE and 4.31% in MAE over the AMD backbone, consistently outperforming state-of-the-art methods across long-horizon forecasting scenarios. The proposed framework demonstrates the effectiveness of integrating domain-specific knowledge—including spectral analysis, morphological feature extraction, and dynamic system modeling—within a unified deep learning architecture for enhanced predictive performance.

Keywords: Time Series Forecasting, Multi-scale frequency-domain Interpolation, Dynamic Morphological Perception, Cross-Channel Attention.

1 Introduction

Time series data constitute a cornerstone of intelligent systems, providing the temporal foundation for predictive analytics across critical application domains including energy grid management, transportation systems, environmental monitoring, and financial markets. The ability to accurately forecast future states from historical observations is essential for enabling proactive decision-making, optimizing resource allocation, and mitigating operational risks. As intelligent systems increasingly rely on data-driven predictions to automate and enhance complex processes, advancing the state-of-the-art in time series forecasting remains a paramount research challenge with direct practical implications.

The complexity of real-world time series stems from multiple interrelated factors that collectively challenge existing forecasting methodologies. First, time series exhibit multi-scale temporal patterns wherein dominant modes of variation can shift substantially across different temporal granularities and application contexts. Some sequences are characterized primarily by smooth longterm trends, while others exhibit pronounced short-term fluctuations and abrupt changes. Second, multi-scale information exhibits pronounced heterogeneity across variables, with traditional fixed-receptive-field mechanisms proving inadequate for capturing localized distortions and irregular patterns that vary across different regions of the temporal space. Third, real-world time series frequently demonstrate dynamic non-stationarity, wherein the underlying data-generating process evolves over time in ways that violate the stationarity assumptions underlying many traditional statistical models. These challenges are further compounded by the presence of complex inter-variable dependencies, particularly in multivariate settings where different time series exhibit correlated evolution patterns.

Contemporary time series forecasting has evolved substantially from traditional statistical approaches toward deep learning-based methodologies [27-29] that can automatically learn complex temporal representations from raw data. Early statistical methods such as ARIMA [1] and exponential smoothing [10] rely on strong prior assumptions regarding stationarity and linearity, limiting their applicability in scenarios characterized by nonlinear dynamics and distributional drift. The emergence of deep learning paradigms, particularly recurrent architectures such as LSTM [8] and hybrid models such as DeepAR [18], enabled substantial improvements in fitting complex nonlinear dependencies while reducing dependence on handcrafted feature engineering. More recently, the field has witnessed the rise of Transformer-based models [15,26,5,7] that excel at capturing long-range dependencies through attention mechanisms, as well as MLP-based approaches [23,20,4] that offer improved computational efficiency with competitive predictive accuracy. Additionally, convolutional architectures [21,19] have demonstrated effectiveness in extracting local patterns and multi-scale features from temporal data.

Despite significant advances, existing methods continue to face three fundamental bottlenecks that limit their predictive performance in complex real-world scenarios.

First, in frequency-domain modeling, high-frequency transient information is often attenuated by static filtering operations, leading to insufficient responsiveness to sudden changes and rapid fluctuations. This frequency-domain information loss degrades the model's ability to capture important local dynamics that manifest as high-frequency components in the spectral representation. Second, multi-scale modeling approaches frequently employ fixed-window or independent-channel strategies that struggle to adequately capture cross-variable coupling relationships. The inability to jointly model multiple temporal scales while respecting inter-variable dependencies results in incomplete representations that miss important shared information across variables. Third, fixed patch or fixed time-step configurations lack the adaptivity required to match the dynamic periodic structures encountered in real-world sequences. The temporal patterns of real-world phenomena evolve continuously, and models that assume static temporal partitions cannot flexibly adapt to these changes.

To address these limitations, this paper proposes AdaDyTS, a unified knowledge-driven forecasting framework that integrates three complementary mechanisms for enhanced temporal modeling. The proposed approach leverages domain-specific knowledge from signal processing, computer vision, and dynamic system modeling to construct a comprehensive solution that jointly addresses the identified bottlenecks. Given an input sequence, the model first performs multi-scale temporal decomposition through a cascaded spectral analysis pipeline. The resulting multi-scale representations are processed by CSRE, which separates low-frequency trend components from high-frequency residual components at different resolutions through coarse-to-fine layer-wise self-reconstruction. This spectral decoupling strategy preserves high-frequency information that static filters would otherwise attenuate. Concurrently, the decomposed temporal features undergo morphological enhancement through DMP-U, which employs deformable convolution guided by a multi-expert attention mechanism to adaptively adjust receptive fields and regional focus. The resulting local morphological features capture fine-grained patterns that conventional convolutions miss. Finally, TV-SS performs long-range recurrent modeling under a configurable time-variant parameterization scheme, jointly capturing short-term perturbations, long-term dependencies, and cross-variable interactions before prediction.

The main contributions of this paper are summarized as follows:

- We propose AdaDyTS, a unified forecasting framework that jointly integrates multi-scale frequency-domain interpolation decoupling, dynamic morphological perception, and time-variant state-space inference,thereby enhancing high-frequency responsiveness, mitigating cross-variable entanglement, and improving adaptability to dynamically evolving periodic structures.
- We introduce three novel mechanisms: the Cascaded Spectral Residual Extractor (CSRE) for multi-scale frequency-domain interpolation decoupling,the Dynamic Morphological Perception Unit (DMP-U) for adaptive morphological feature extraction, and the Time-Variant State-Space Module (TV-SS) for dynamic long-range dependency modeling.
- We conduct comprehensive experiments on 13 benchmark datasets spanning diverse application domains, demonstrating that AdaDyTS consistently improves forecasting

accuracy and robustness over strong baseline methods, with average improvements of 4.35% in MSE and 4.31% in MAE relative to the AMD backbone.

2 Methodology

2.1 Overview of the Proposed AdaDyTS

The proposed AdaDyTS framework addresses the key limitations of existing time series forecasting methods through a unified architecture that synergistically integrates three knowledge-driven mechanisms. As illustrated in Figure 1, the framework comprises three core modules—CSRE, DMP-U, and TV-SS. This architectural design reflects the principle that effective temporal forecasting requires complementary modeling capabilities: spectral analysis for frequency-domain representations, morphological processing for local pattern extraction, and dynamic system modeling for long-range dependencies.

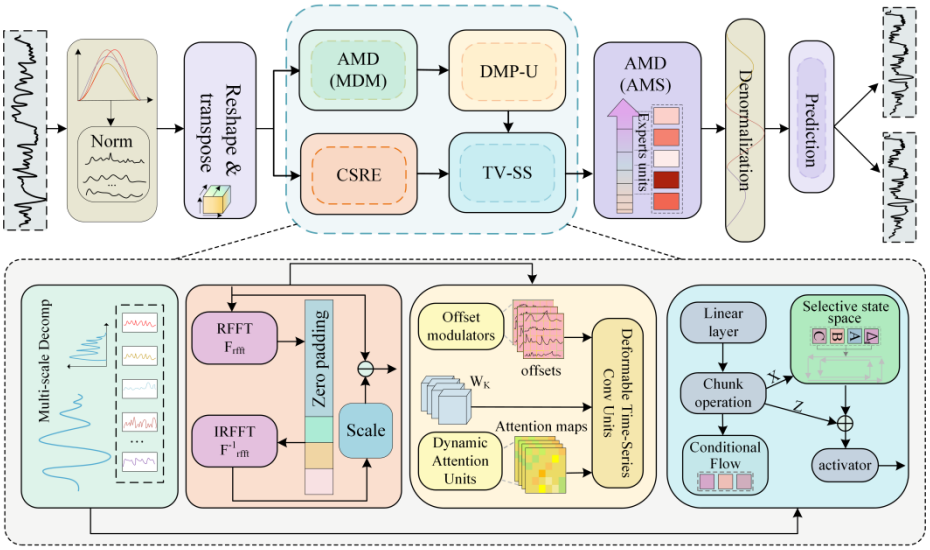

Fig. 1. Overall architecture of AdaDyTS. The model processes CSRE and DMP-U in parallel, fuses their outputs through adaptive weighting, and then performs time-variant state-space inference with TV-SS.

The design philosophy of AdaDyTS emphasizes knowledge integration from multiple domains. Spectral analysis principles motivate the CSRE module, which leverages frequency-domain interpolation to separate trend and residual components at multiple scales. Morphological image processing concepts inspire the DMP-U module, which adapts deformable convolution techniques for temporal data to enable adaptive receptive field adjustment. Dynamic system theory underlies the TV-SS module, which relaxes conventional time-invariant parameter assumptions to capture time-varying dynamics.

2.2 Cascaded Spectral Residual Extractor Module

To alleviate the loss of high-frequency information caused by filtering in frequency-domain modeling, we introduce the multi-scale frequency-domain interpolation decoupling mechanism and design CSRE. CSRE decomposes time-domain signals at different resolutions into a low-frequency backbone and multiple high-frequency residual components in the frequency domain through coarse-to-fine layer-wise self-reconstruction. This design extracts both long-term trends and local fluctuations from the time series, as illustrated in Fig. 2.

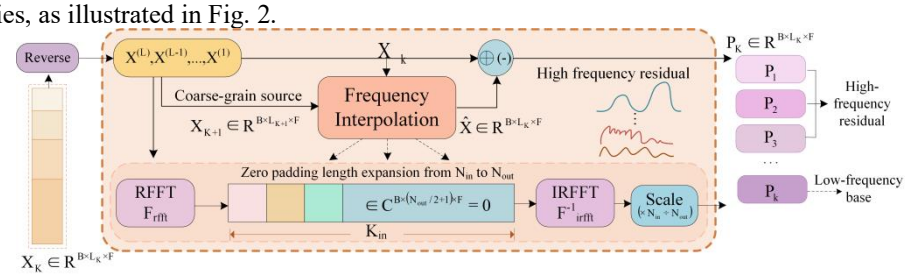

Fig. 2. Architecture of CSRE. Subtracting the interpolated multi-scale signal L from X(i) yields a high-frequency residual, separating the low-frequency backbone from multi-scale high-frequency components.

Given a set of multi-scale time series $X = \{X^{(0)}, X^{(1)}, \dots, X^{(L)}\}$, the temporal resolution decreases and the amount of high-frequency information becomes smaller as the level i increases. To enable coarse-to-fine layer-wise self-reconstruction, we reverse the sequence order to $X = \{X^{(L)}, X^{(L-1)}, \dots, X^{(0)}\}$and perform progressive decomposition.

At level i, the low-frequency representation $X^{(i-1)}$ from the previous level is first mapped to the current temporal scale T_i through the frequency-domain interpolation operator $\mathcal{L}(\cdot)$. Subtracting it from the original sequence $X^{(i)}$ yields the unexplained high-frequency residual component $F^{(i)}$:

$$F^{(i)} = X^{(i)} - \mathcal{L}\left(X^{(i-1)}; T_{i-1}, T_i\right) \tag{1}$$

By iterating this process across levels, we obtain the set of high-frequency residuals $F = \{F^{(0)}, F^{(1)}, \dots, F^{(L)}\}$ corresponding to band-pass components at different scales.Here, the frequency-domain interpolation operator $\mathcal{L}(\cdot)$resamples an input sequence X of length T to a target length T'. We first apply the real-valued fast Fourier transform $\mathcal{F}_r(\cdot)$ to obtain the original spectrum:

$$\hat{X} = \mathcal{F}_r(X), \quad \hat{X} \in \mathbb{C}^{B\times C\times(\lfloor T/2\rfloor+1)} \tag{2}$$

Next, zero-padding is applied to the high-frequency part of the spectrum, which is equivalent to sinc interpolation in the time domain, thereby extending the spectrum to the resolution corresponding to the target length without distortion:

$$\hat{X'}_{b,c,k} = \begin{cases} \hat{X}_{b,c,k}, 0 \le k \le \lfloor T/2\rfloor \\ 0, \lfloor T/2\rfloor < k \le \lfloor T'/2\rfloor \end{cases} \tag{3}$$

Finally, we apply the inverse Fourier transform $\mathcal{F}_r^{-1}(\cdot)$ to recover the extended spectrum in the time domain and scale the result by the length ratio to preserve amplitude consistency:

$$\mathcal{L}(X;T,T') = \frac{T}{T'} \cdot \mathcal{F}_r^{-1}(\hat{X}') \tag{4}$$

2.3 Dynamic Morphological Perception Unit Module

To alleviate multi-scale heterogeneity and the limited receptive field induced by fixed downsampling windows, we design DMP-U, as shown in Fig. 3.

The fixed sampling locations of conventional convolution are poorly aligned with phase shifts and local distortions in time series, whereas deformable convolution learns offsets to achieve dynamic alignment along the temporal axis[3]. Built on standard convolution, we further introduce a multi-branch dynamic convolution selection mechanism together with deformable sampling offset learning, enabling adaptive adjustment of both the receptive field and regional focus.The module consists of three branches: an offset prediction branch, a dynamic attention map branch, and a multi-branch dynamic convolution unit.

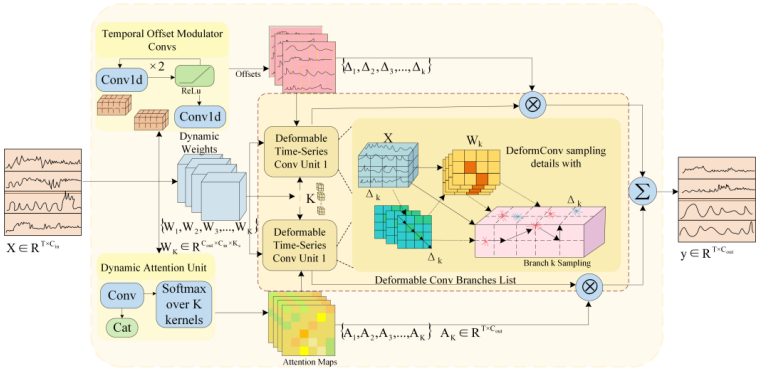

Fig. 3. Architecture of CSRE. Subtracting the interpolated multi-scale signal L from X(i) yields a high-frequency residual, separating the low-frequency backbone from multi-scale high-frequency components.

The offset prediction branch is composed of three convolutional layers and generates the receptive-field offsets for all dynamic convolution experts. To improve expressive power and dynamically capture dependencies across historical steps and correlated variables, nonlinear activations are applied after the first two convolutions. For an input **X**, the spatial offsets of the dynamic convolution unit are given by:

$$\Delta P = Conv_3(\sigma(Conv_2(\sigma(Conv_1(\mathbf{X}))))) \tag{5}$$

Here, $\sigma(\cdot)$ denotes the ReLU activation. The output $\Delta P \in \mathbb{R}^{K\times 2k^2\times V\times T}$, where k is the kernel size, each location corresponds to the two-dimensional offsets of $k \times k$ sampling points, and K is the number of dynamic convolution experts.The tensor ΔP is then split along the channel dimension into K parts, $\{\Delta P_1, \Delta P_2, \ldots, \Delta P_K\}$, each of which guides the deformation of one expert.

Conventional dynamic convolution usually assigns a scalar weight to each expert[2]. To enable finer-grained feature routing, we introduce a dynamic attention branch that generates an element-wise attention map through a feature mapping function

$\mathcal{F}_{attn}$.Specifically, the model first predicts an attention map with $K \times C_{out}$ channels, reshapes it, and then applies Softmax over the expert dimension for normalization:

$$A = Softmax(\mathcal{F}_{attn}(X)) \tag{6}$$

Here, A$\in \mathbb{R}^{K\times C_{out}\times H\times W}$. This guarantees that, for every spatial position and every channel of the output feature map, the weights of all experts sum to 1, namely $\sum_{k=1}^{K} A_{k,c,i,j} = 1$, leading to normalized and stable expert allocation.

In the multi-branch dynamic convolution unit, once the offset ΔP_k and attention map A_k of each expert are obtained, deformable convolution is applied to the input feature map X and the expert outputs are aggregated through weighted summation.Let the weight and bias of the k-th expert be W_k and B_k, respectively. The final output feature map $Y \in \mathbb{R}^{C_{out}\times H\times W}$ can be written as:

$$Y = \sum_{k=1}^{K} A_k \odot DConv(X, W_k, B_k, \Delta P_k) \tag{7}$$

where $\odot$ denotes element-wise multiplication. In this way, DMP-U equips different experts with heterogeneous receptive-field structures and adaptively fuses their responses at a fine-grained feature level, thereby providing more robust representations for subsequent temporal evolution forecasting.

2.4 Time-Variant State-Space Module

To enhance the model's sensitivity to dynamic scale variation across different time series, we propose the TV-SS module for long-range time-variant state-space inference, as illustrated in Fig.4.

Existing state-space models for time series forecasting usually assume time-invariant parameters, meaning that the state transition matrix A, input mapping B, and readout matrix C remain fixed throughout the sequence. However, real-world time series such as traffic flow, meteorological observations, and energy load often exhibit pronounced non-stationarity and stage-wise dynamic variation. Across different temporal intervals, the system evolution speed, input response strength, and state representation may all change. As a result, fixed-parameter SSMs struggle to handle both short-term rapid changes and long-term stable dependencies. To address this issue, we decouple the key control parameters in the state transition process. This design allows the model to switch smoothly between time-invariant and time-variant regimes while balancing representation power and optimization stability across input scenarios.

Given an input temporal feature $X \in \mathbb{R}^{B\times L\times F}$, the model first projects it into an expanded hidden space of dimension D_{inner}. It then splits the representation along the feature dimension into a backbone branch X and a gating branch Z, allowing long-term memory and local temporal evolution to be processed in parallel.

For the backbone branch X, the model captures short-range dependencies through an optional local one-dimensional causal convolution followed by a SiLU activation, namely $x_t = SiLU(Conv1D(X_t))$. This design accommodates differences in frequency structure across time series. When a sequence relies more on long-range global relationships, the convolution layer can be skipped and the nonlinear mapping is applied directly.

On top of this design, we relax the fixed input/output mapping assumption in conventional discrete state-space equations. Specifically, we extend the time step Δ_t, which controls sampling granularity, the parameter B_t, which controls input filtering, and the parameter C_t, which determines hidden-state reconstruction, into input-dependent dynamic parameters. This design enables dynamic input selection and time-variant state representation: when Δ_t is large, the system favors long-term smooth evolution, whereas smaller Δ_t values make it more responsive to rapid local variation.

Specifically, to adapt to different types of temporal features, we design configurable switches that independently control whether Δ_t, B_t, and C_t are time-variant. When a parameter is configured as time-variant, it is generated dynamically from the current input; otherwise, it degenerates to a global learnable static parameter.Taking Δ_t as an example, we use the Softplus function to guarantee positivity and avoid numerical instability during discretization:

$$\Delta_t = \begin{cases} Softplus(W_\Delta x_t + b_\Delta), if \ \ time\_variant \\ Softplus(b_\Delta), otherwise \end{cases} \tag{8}$$

Likewise, B_t and $C_t$can switch flexibly between input-dependent dynamic matrices ($\in \mathbb{R}^{B\times L\times D_{state}}$) and static parameters ($\in \mathbb{R}^{D_{inner}\times D_{state}}$).

During this process, the base template of the global state transition matrix is defined as $A = -exp(A_{\log})$, $where A_{\log} = \log(\text{Repeat}([1,2,\dots,D_{state}]))$ is replicated along the channel dimension and updated during training.For the input-dependent branch, the model generates the joint projection $[\hat{\Delta}_t, \hat{B}_t, \hat{C}_t] = W_x x_t$ from the backbone feature x_t, and then determines B_t and C_t according to the switch configuration. This mechanism allows a smooth transition between globally stable parameterization and input-adaptive parameterization.

After the dynamic control parameters are determined, the continuous state transition matrix A and the input matrix B_t are discretized into $\bar{A}_t$ and $\bar{B}_t$ under the zero-order hold assumption through Δ_t. The hidden state $h_t$is then updated recurrently along the temporal dimension:

$$\begin{aligned} \bar{A}_t &= exp(\Delta_t A) \\ \bar{B}_t &= (\Delta_t A)^{-1}(exp(\Delta_t A) - I) \cdot \Delta_t B_t \\ h_t &= \bar{A}_t h_{t-1} + \bar{B}_t x_t \end{aligned} \tag{9}$$

To ensure mathematical stability in the state evolution process, we strictly constrain the real parts of the eigenvalues of the continuous system matrix A to be negative, which guarantees that the spectral radius of the discrete transition matrix $\bar{A}_t$satisfies $\rho(\bar{A}_t) < 1$.

After state evolution, the output feature at the current time step is computed by the reconstruction matrix as $y_t = C_t h_t + D \odot x_t$. Here, we introduce an additional channel-wise learnable residual skip parameter $D \in \mathbb{R}^{D_{inner}}$ to mitigate gradient attenuation in deep recurrent networks while preserving the original input information.

Finally, the output Y of the state-space model is fused with the previously separated gating branch Z through element-wise multiplication, and a linear projection $O = W_{out}(Y \odot SiLU(Z))$ maps the result back to the target output dimension D_{out}. In this

way, TV-SS produces a time-variant long-range forecasting representation with strong generalization ability.

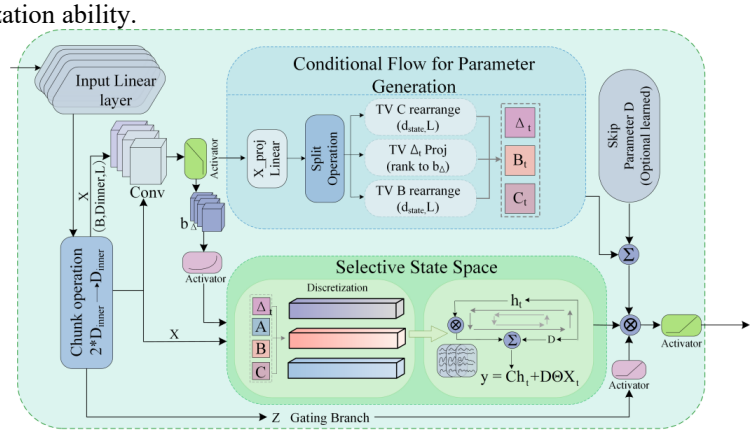

Fig. 4. Architecture of CSRE. Subtracting the interpolated multi-scale signal L from X(i) yields a high-frequency residual, separating the low-frequency backbone from multi-scale high-frequency components.

3 Experimental Results

3.1 Datasets

To verify the effectiveness and generalization ability of AdaDyTS, we conduct experiments on 13 public datasets, including Electricity (ECL), ETTh1, ETTh2,ETTm1, ETTm2, Exchange,Solar,Traffic,Weather,PEMS03,PEMS04,PEMS07,andPEMS08[12],[25],[22],[6]. These datasets cover diverse real-world scenarios, such as electricity load, meteorological observations, exchange-rate fluctuation, and traffic flow, and include both hourly and minute-level sampling frequencies. They also differ significantly in sequence length, variable dimensionality,and periodic structure, allowing a comprehensive evaluation of modeling ability under complex distributions and generalization across scenarios. Specifically, the first nine datasets are used for long-term forecasting with prediction horizons $\{96, 192, 336, 720\}$, whereas the last four PEMS datasets are used for short-term traffic forecasting with input length $L = 96$ and prediction horizon $T = 12$.Data organization, train/validation/test splits, and preprocessing procedures follow mainstream public benchmark settings[15],[22],[25].

3.2 Implementation Details

All experiments are implemented in PyTorch 2.0.1 with Python 3.10[17] and are run on a server equipped with eight NVIDIA Tesla P40 GPUs, each with 24 GB of memory. To ensure fairness, we re-evaluate all baselines under different input lengths L and report their best results. For the short-term forecasting task, the input length is fixed to $L = 96$. The learning rate is selected from $\{1 \times 10^{-3},\ 5 \times 10^{-4},\ 1 \times 10^{-4}\}$ according to

validation performance. The default batch size is 128 and is reduced to 64 or 32 for datasets with high feature dimensionality. The number of dynamic convolution experts is adjusted according to the scale and feature dimensionality of the input sequence. All model parameters are optimized using Adam[11]. During training, we adopt a joint objective consisting of mean squared error and MoE load-balancing regularization. During validation and testing, we report MSE and MAE, and the best checkpoint is selected according to validation MSE.

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{MoE}}, \tag{10}$$

Where $\mathcal{L}_{\text{MSE}}$ denotes the element-wise mean squared error, defined as

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{11}$$

where N is the total number of predicted elements in a batch.$\mathcal{L}_{\text{MoE}}$ comes from the load-balancing constraint of MoE gating assignment and regularizes the squared coefficient of variation of the expert-importance vector:

$$\mathcal{L}_{\text{MoE}} = \sum_{c=1}^{C}\lambda \cdot \frac{\text{Var}(\mathbf{I}_c)}{(\text{Mean}(\mathbf{I}_c))^2 + \epsilon}, \tag{12}$$

where $\mathbf{I}_c$ denotes the expert importance for the c-th variable, C is the number of variables, λ is the regularization weight, and ϵ is a numerical stability constant used to avoid division by zero.

3.3 Evaluation Metrics

We adopt mean squared error (MSE) and mean absolute error (MAE) as the unified evaluation metrics.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{13}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{14}$$

For both metrics, lower values indicate better forecasting performance. MSE amplifies the penalty on large prediction deviations through the quadratic term and is therefore effective for evaluating robustness to abrupt outliers and extreme distributions. In contrast, MAE treats all errors uniformly, is less sensitive to local outliers, and more directly reflects the average deviation over the overall data distribution.

Since our tasks cover both long-horizon forecasting and short-term high-volatility scenarios, reporting both MSE and MAE provides a more comprehensive assessment of AdaDyTS in terms of accuracy and robustness. All final results are reported on the test set and averaged over five random seeds.

3.4 Main Results

We select three representative categories of forecasting models as baselines, covering the three mainstream paradigms of MLP, Transformer, and CNN. Specifically, the MLP family includes AMD[9], TimeMixer[20], MTS-Mixers[13], DLin-ear[23], and RLin-

ear[14]; the Transformer family includes PatchTST[16], iTransformer[15], Crossformer[24], FEDformer[26], and TiDE[4]; and the CNN family includes TimesNet[21] and MICN[19].

Tables 2 and 1 report the long-term and short-term forecasting results, respectively. To avoid conclusions that rely on occasional wins on individual datasets, we analyze the results from three perspectives: the number of wins, the average rank, and the relative improvement.

Table 1. Short-term forecasting results. The input sequence length is fixed to L = 96, and the prediction horizon is T = 12. Red values indicate the best results, and blue values indicate the second-best results.

Models	AdaDyTS		AMD		PatchTST		Cross-former		FEDformer		TimesNet		TiDE		DLinear		MTS-Mixers		RLinear	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PEMS03	0.089	0.201	0.084	0.198	0.099	0.216	0.090	0.203	0.126	0.251	0.085	0.192	0.178	0.305	0.122	0.243	0.117	0.232	0.126	0.236
PEMS04	0.091	0.2	0.083	0.198	0.105	0.224	0.098	0.218	0.138	0.262	0.087	0.195	0.219	0.340	0.148	0.272	0.129	0.267	0.138	0.252
PEMS07	0.08	0.16	0.074	0.180	0.095	0.150	0.094	0.200	0.109	0.225	0.082	0.181	0.173	0.304	0.115	0.242	0.134	0.278	0.118	0.235
PEMS08	0.142	0.198	0.093	0.206	0.168	0.232	0.165	0.214	0.173	0.273	0.112	0.212	0.227	0.343	0.154	0.276	0.186	0.286	0.133	0.247

In long-term forecasting (Table 2), AdaDyTS achieves the best MSE in 28 out of 36 dataset-horizon settings and the best MAE in 29 out of 36 settings, with average ranks of 1.33 for MSE and 1.26 for MAE. Compared with the strongest baseline under the same setting, AdaDyTS yields an average relative improvement of 1.23% in MSE and 2.42% in MAE. Compared with the AMD backbone, the average improvement reaches 4.35% in MSE and 4.31% in MAE, and AdaDyTS strictly outperforms AMD in 36/36 MSE settings and 35/36 MAE settings. Representative examples include Weather-336, where the MSE decreases from 0.240 to 0.200, corresponding to a 16.7% reduction, and ECL-720, where the MSE decreases from 0.193 to 0.181, corresponding to a 6.2% reduction. These results indicate that the proposed multi-scale frequency-domain interpolation decoupling and dynamic morphological perception mechanisms bring stable gains for medium-range and long-range dependency modeling as well as multi-scale fluctuation modeling. The aggregated advantage over the strongest baselines is further summarized in Table 3, while the non-ETT dataset comparison among the top-performing methods is visualized in Figs. 5 .

Table 2. Long-term forecasting results. To ensure a fair comparison, the input sequence length L is selected from $\{96, 192, 336, 512, 720\}$ by searching for the best value for each baseline. Red values indicate the best results, and blue values indicate the second-best results.

Models		AdaDyTS		AMD		TimeMixer		PatchTST		iTransformer		Crossformer		FEDformer		TimesNet		MICN		DLinear		MTS-Mixers	
Data	Met-ric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	**0.137**	**0.186**	0.145	0.197	0.147	0.197	0.149	0.198	0.174	0.214	0.153	0.217	0.238	0.314	0.172	0.220	0.161	0.226	0.152	0.237	0.156	0.206
	192	**0.180**	**0.223**	0.187	0.238	0.189	0.239	0.194	0.241	0.221	0.254	0.197	0.269	0.275	0.329	0.219	0.261	0.220	0.283	0.220	0.282	0.199	0.248
	336	**0.200**	**0.260**	0.240	0.280	0.241	0.280	0.245	0.282	0.278	0.296	0.252	0.311	0.339	0.377	0.280	0.306	0.275	0.328	0.265	0.319	0.249	0.291
	720	**0.300**	**0.320**	0.315	0.330	0.310	0.330	0.314	0.334	0.358	0.349	0.318	0.363	0.389	0.409	0.365	0.359	0.311	0.356	0.323	0.362	0.336	0.343
ETTh1	96	**0.355**	**0.387**	0.369	0.397	0.361	0.390	0.370	0.399	0.386	0.405	0.386	0.429	0.376	0.415	0.384	0.402	0.396	0.427	0.375	0.399	0.372	0.395
	192	**0.400**	**0.410**	0.401	0.416	0.409	0.414	0.413	0.421	0.441	0.436	0.416	0.441	0.423	0.446	0.457	0.436	0.430	0.453	0.405	0.416	0.416	0.426
	336	**0.399**	**0.406**	0.418	0.427	0.430	0.429	0.422	0.436	0.487	0.458	0.440	0.461	0.444	0.462	0.491	0.469	0.433	0.458	0.439	0.443	0.455	0.449
	720	**0.411**	**0.423**	0.434	0.454	0.445	0.450	0.447	0.466	0.503	0.491	0.519	0.523	0.469	0.492	0.541	0.500	0.474	0.472	0.490	0.475	0.475	0.472
ETTh2	96	**0.264**	**0.324**	0.274	0.337	0.271	0.330	0.274	0.337	0.297	0.349	0.628	0.563	0.332	0.374	0.340	0.374	0.289	0.357	0.289	0.353	0.307	0.354
	192	0.370	0.372	0.375	0.383	**0.317**	**0.340**	0.339	0.379	0.380	0.400	0.703	0.624	0.407	0.446	0.402	0.414	0.438	0.357	0.438	0.403	0.374	0.399
	336	0.358	0.394	0.375	0.411	0.332	0.396	**0.329**	**0.380**	0.428	0.432	0.827	0.675	0.400	0.447	0.452	0.452	0.417	0.452	0.448	0.465	0.398	0.432
	720	0.397	0.411	0.402	0.438	**0.342**	**0.408**	0.379	0.422	0.427	0.445	1.181	0.840	0.412	0.469	0.462	0.468	0.426	0.473	0.605	0.551	0.463	0.456
ETTm1	96	**0.264**	**0.328**	0.284	0.339	0.291	0.340	0.290	0.342	0.334	0.388	0.316	0.373	0.326	0.390	0.338	0.375	0.314	0.360	0.299	0.343	0.314	0.358
	192	**0.300**	**0.342**	0.322	0.362	0.327	0.365	0.332	0.369	0.377	0.391	0.377	0.411	0.365	0.415	0.371	0.387	0.359	0.387	0.335	0.365	0.354	0.386
	336	**0.342**	**0.361**	0.360	0.380	0.360	0.381	0.366	0.392	0.426	0.420	0.431	0.442	0.391	0.425	0.410	0.390	0.398	0.413	0.369	0.386	0.384	0.405
	720	**0.412**	**0.402**	0.421	0.416	0.415	0.417	0.416	0.420	0.491	0.459	0.600	0.547	0.446	0.458	0.478	0.450	0.459	0.464	0.425	0.421	0.427	0.432
ETTm2	96	**0.154**	0.258	0.167	0.258	0.164	**0.254**	0.166	0.256	0.180	0.284	0.421	0.461	0.180	0.271	0.187	0.267	0.178	0.273	0.167	0.260	0.177	0.259
	192	**0.211**	**0.276**	0.221	0.294	0.223	0.295	0.223	0.296	0.250	0.309	0.503	0.519	0.252	0.318	0.249	0.309	0.245	0.316	0.224	0.303	0.241	0.303
	336	**0.252**	**0.317**	0.270	0.327	0.279	0.330	0.274	0.329	0.311	0.348	0.611	0.580	0.324	0.364	0.321	0.351	0.295	0.350	0.281	0.342	0.297	0.338
	720	**0.343**	**0.372**	0.356	0.382	0.359	0.383	0.362	0.389	0.412	0.407	0.996	0.750	0.410	0.420	0.497	0.403	0.409	0.397	0.396	0.398	0.396	0.398
ECL	96	**0.121**	**0.214**	0.129	0.224	0.129	0.224	0.129	0.222	0.148	0.240	0.187	0.283	0.186	0.302	0.168	0.272	0.159	0.267	0.153	0.237	0.141	0.243
	192	**0.138**	0.226	0.147	0.238	0.140	**0.220**	0.147	0.240	0.162	0.253	0.258	0.330	0.197	0.311	0.184	0.289	0.168	0.279	0.152	0.249	0.163	0.261
	336	**0.154**	**0.245**	0.160	0.253	0.161	0.255	0.163	0.259	0.178	0.269	0.323	0.369	0.213	0.328	0.198	0.300	0.196	0.308	0.169	0.267	0.176	0.277
	720	**0.181**	**0.280**	0.193	0.286	0.194	0.287	0.197	0.290	0.225	0.317	0.401	0.423	0.233	0.344	0.220	0.320	0.203	0.312	0.233	0.344	0.212	0.308
Exchange	96	**0.079**	**0.199**	0.083	0.201	0.086	0.204	0.093	0.214	0.086	0.206	0.186	0.346	0.136	0.276	0.107	0.234	0.102	0.235	0.081	0.203	0.083	0.201
	192	0.165	**0.288**	0.171	0.293	0.176	0.298	0.192	0.312	0.177	0.299	0.467	0.522	0.256	0.369	0.226	0.344	0.172	0.316	**0.157**	0.293	0.174	0.296
	336	0.289	**0.386**	0.309	0.402	0.345	0.426	0.350	0.432	0.331	0.417	0.783	0.721	0.426	0.464	0.367	0.448	**0.272**	0.407	0.305	0.414	0.336	0.417
	720	0.742	0.620	0.750	0.652	0.848	0.692	0.911	0.716	0.847	0.691	1.367	0.943	1.090	0.800	0.964	0.746	0.714	0.658	**0.643**	**0.601**	0.900	0.715
Traffic	96	**0.358**	**0.246**	0.366	0.259	0.380	0.249	0.380	0.249	0.395	0.268	0.512	0.290	0.576	0.359	0.593	0.321	0.508	0.301	0.410	0.282	0.462	0.332
	192	**0.374**	0.260	0.381	0.265	0.379	**0.256**	0.375	0.350	0.417	0.276	0.523	0.297	0.610	0.380	0.617	0.336	0.536	0.315	0.423	0.287	0.485	0.354
	336	0.391	**0.259**	0.397	0.269	**0.385**	0.270	0.392	0.264	0.437	0.283	0.530	0.300	0.637	0.392	0.629	0.336	0.525	0.310	0.436	0.296	0.498	0.360
	720	**0.420**	**0.280**	0.429	0.292	0.430	0.281	0.432	0.286	0.467	0.302	0.573	0.313	0.621	0.375	0.640	0.350	0.571	0.323	0.466	0.315	0.529	0.370
Solar	96	0.169	**0.216**	0.175	0.228	**0.167**	0.220	0.224	0.278	0.203	0.237	0.181	0.240	0.201	0.304	0.219	0.314	0.188	0.252	0.289	0.337	0.284	0.325
	192	**0.184**	**0.225**	0.186	0.235	0.187	0.249	0.253	0.298	0.233	0.261	0.196	0.252	0.237	0.337	0.291	0.322	0.215	0.280	0.319	0.397	0.307	0.362
	336	**0.188**	**0.210**	0.200	0.246	0.200	0.258	0.273	0.306	0.246	0.273	0.216	0.243	0.254	0.326	0.264	0.332	0.222	0.267	0.352	0.415	0.333	0.384
	720	**0.200**	**0.234**	0.203	0.249	0.215	0.250	0.272	0.308	0.249	0.275	0.220	0.256	0.280	0.397	0.280	0.363	0.226	0.264	0.356	0.412	0.335	0.383

In short-term traffic forecasting (Table 1), AdaDyTS does not achieve the best MSE, as the best MSE in all four settings is obtained by AMD. Nevertheless, AdaDyTS remains competitive in MAE, sharing the best average rank of 2.25 with AMD and TimesNet, and it achieves the best MAE on PEMS08 (0.198), which is 3.9% lower than AMD.

In addition, the non-optimal cases in long-term forecasting are mainly concentrated on ETTh2 at medium-to-long horizons (192/336/720) and Exchange-720, where the MSE of AdaDyTS is 0.742 compared with 0.643 for DLinear. This suggests that the current method still has room to improve its control over extreme errors, especially in highly non-stationary and heavy-tailed scenarios where MSE is more sensitive. Overall, the main strength of AdaDyTS lies in its consistent superiority on long-term forecasting and its cross-dataset generalization, whereas short-term, high-volatility traffic forecasting remains an important direction for further optimization. From the perspective of robustness, Table 4 shows that AdaDyTS achieves the smallest ΔMSE on Exchang e and ECL and ties the best result on Weather, indicating stronger stability as the prediction horizon increases.

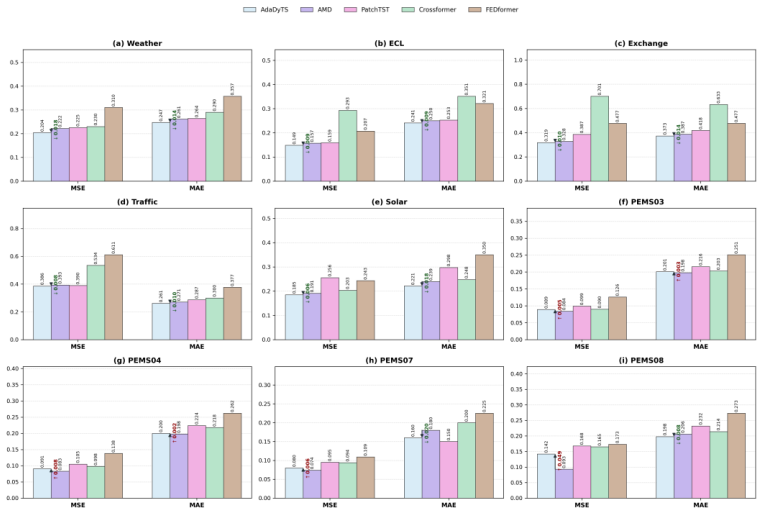

Fig. 5. Performance comparison with SOTA methods: (a) Weather, (b) ECL, (c) Exchange, (d) Traffic, (e) Solar, and (f)-(i) PEMS03, PEMS04, PEMS07, and PEMS08.

Table 3. Overall performance improvement of AdaDyTS over the strongest baseline. The values are averaged across four prediction horizons (96, 192, 336, and 720). The best results are highlighted in bold.

Dataset	Metric	Best Baseline Model	Best Baseline Avg.	AdaDyTS Avg.	Improvement
Weather	MSE	TimeMixer	0.222	**0.204**	**+8.11%**
	MAE	MTS-Mixers	0.272	**0.247**	**+9.19%**
ETTh1	MSE	TimeMixer	0.411	**0.391**	**+4.86%**
	MAE	TimeMixer	0.421	**0.406**	**+3.56%**
ETTm2	MSE	MTS-Mixers	0.277	**0.240**	**+13.3%**
	MAE	TimeMixer	0.315	**0.305**	**+3.17%**
Exchange	MSE	TimeMixer	0.364	**0.318**	**+12.6%**

	MAE	DLinear	0.377	**0.373**	**+1.06%**

Table 4. Robustness analysis across different prediction horizons. The table shows the change in mean squared error (*Δ* MSE) as the prediction length increases from 96 to 720. A smaller *Δ* indicates better robustness.

Models	Weather(MSE) L = 96 → 720	Δ	Exchange(MSE) L = 96 → 720	Δ	ECL (MSE) L = 96 → 720	Δ
AdaDyTS	0.137→ 0.300	**+0.163**	0.079 → 0.742	**+0.663**	0.121→0.181	**+0.060**
AMD	0.145→ 0.315	+0.170	0.083 → 0.750	+0.667	0.129 → 0.193	+0.064
TimeMixer	0.147→ 0.310	**+0.163**	0.086 → 0.848	+0.762	0.129 → 0.194	+0.065
PatchTST	0.149→ 0.314	+0.165	0.093 → 0.911	+0.818	0.129 → 0.197	+0.068
iTransformer	0.174→ 0.358	+0.184	0.086 → 0.847	+0.761	0.148 → 0.225	+0.077
DLinear	0.152→ 0.323	+0.171	0.081 → 0.643	+0.562	0.153 → 0.233	+0.080

3.5 Ablation Experiments

To evaluate the individual contribution of each component, we perform a systematic ablation study on CSRE, DMP-U, and TV-SS. Considering that different datasets exhibit distinct spectral structures, periodicity patterns, and variable coupling characteristics, we select Weather, Exchange, ETTh1, and PEMS04 as representative benchmarks. For Weather, Exchange, and ETTh1, we report the average results over prediction horizons {96, 192, 336, 720}. For PEMS04, the prediction horizon is fixed at $T = 12$. Starting from the full model, we remove one module at a time (CSRE, DMP-U, or TV-SS) and compare the resulting variants with the complete model. We also include the AMD backbone without the three proposed enhancements as an additional baseline. The absolute ablation results are shown in Table 5, the relative performance changes are summarized in Table 6, and the overall trends are visualized in Fig. 6.

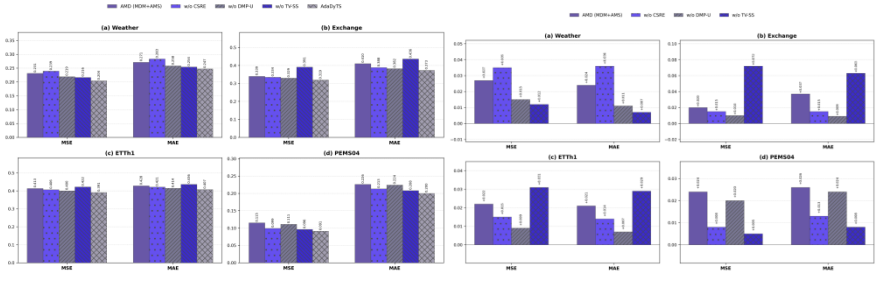

Fig. 6. Visualization of the ablation results. The right panel shows the error variation of each variant relative to the full model (AdaDyTS), while the left panel compares MSE and MAE across variants on different datasets.

Importance of CSRE. The variant without CSRE exhibits clear error increases on all four datasets, and the degradation is especially obvious on Weather, which contains relatively rich high-frequency signals. This observation indicates that static decomposition strategies struggle to preserve high-frequency residual information adequately. Through dynamic weighting and progressive reconstruction, CSRE strengthens high-frequency details while retaining low-frequency trends, thereby providing more complete multi-scale spectral input for the downstream modules.

Role of DMP-U in modeling heterogeneity. Removing DMP-U weakens the model's ability to fit abrupt intervals and locally distorted segments, and the degradation is especially pronounced on PEMS04, where local deformation and variable coupling are more complex. The main reason is that fixed-receptive-field convolution cannot adapt well to complex morphology variation. By combining deformable dynamic convolution with multi-expert attention, DMP-U adaptively adjusts sampling positions and response strength according to local patterns, thereby improving fine-grained feature extraction.

Effect of TV-SS on dynamic evolution modeling. The variant without TV-SS deteriorates more noticeably on Exchange and ETTh1, where long-range dependencies are stronger, indicating that fixed time steps and static state transitions are insufficient for describing the dynamic evolution of non-stationary sequences. Through learnable time-variant parameters, TV-SS adaptively adjusts both time steps and state mappings, strengthening the joint modeling of short-term and long-term dependencies.

Overall, the ablation study shows that CSRE, DMP-U, and TV-SS are not merely additive components. They correspond to three complementary levels of modeling, namely multi-scale frequency-domain interpolation decoupling, dynamic morphological perception, and time-variant state-space inference, and together they improve the overall performance of AdaDyTS.

Table 5. Ablation results (lower is better). For Weather, Exchange, and ETTh1, the reported values are averaged over $T \in \{96,192,336,720\}$, whereas the PEMS04 results correspond to $T = 12$.

Variant	CSRE	DMP-U	TV-SS	Weather		Exchange		ETTh1		PEMS04	
				MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓
AMD (MDM+AMS)	×	×	×	0.231	0.271	0.339	0.410	0.413	0.428	0.115	0.226
w/o CSRE	×	✓	✓	0.239	0.283	0.334	0.388	0.406	0.421	0.099	0.213
w/o DMP-U	✓	×	✓	0.219	0.258	0.329	0.382	0.400	0.414	0.111	0.224
w/o TV-SS	✓	✓	×	0.216	0.254	0.391	0.436	0.422	0.436	0.096	0.208
AdaDyTS (full)	✓	✓	✓	**0.204**	**0.247**	**0.319**	**0.373**	**0.391**	**0.407**	**0.091**	**0.200**

Table 6. Relative performance change with respect to AdaDyTS (full). Positive values indicate performance degradation, whereas negative values indicate improvement. The error is defined as $\Delta(\%) = \frac{Variant - Full}{Full} \times 100$.

Variant	Weather		Exchange		ETTh1		PEMS04	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
AMD (baseline)	13.24	9.72	6.27	9.92	5.63	5.16	**26.37**	**13.00**
w/o CSRE	**17.16**	**14.57**	4.70	4.02	3.84	3.44	8.79	6.50
w/o DMP-U	7.35	4.45	3.13	2.41	2.30	1.72	21.98	12.00
w/o TV-SS	5.88	2.83	**22.57**	**16.89**	**7.93**	**7.13**	5.49	4.00

4 Conclusion

This paper proposes AdaDyTS, a unified knowledge-driven forecasting framework that integrates multi-scale frequency-domain interpolation decoupling, dynamic morphological perception, and time-variant state-space inference to address key limitations in existing time series forecasting methods. The proposed framework systematically addresses three fundamental challenges: insufficient preservation of high-frequency information in spectral modeling, inadequate characterization of multi-scale heterogeneity across variables, and limited capability to capture time-varying dynamics in non-stationary sequences. Comprehensive experiments on 13 benchmark datasets demonstrate that AdaDyTS delivers stable and competitive forecasting performance across diverse application scenarios. The method achieves average improvements of 4.35% in MSE and 4.31% in MAE over the AMD backbone, with particularly strong gains on datasets characterized by complex periodic structures and pronounced non-stationarity. The ablation studies validate the complementary contributions of each proposed module, confirming that CSRE, DMP-U, and TV-SS address distinct aspects of the temporal modeling challenge.

Despite these strengths, several limitations remain opportunities for future work. First, the model exhibits reduced tolerance in scenarios with extreme non-stationarity and heavy-tailed fluctuations, where MSE-based optimization may not adequately capture the prediction uncertainty. Second, for large-scale time series, the dynamic routing and time-variant state-space inference introduce computational overhead that constrains inference efficiency. Future research directions include incorporating distributionally robust optimization and extreme-value-aware constraints to improve robustness under frequent anomalies, exploring sparse routing and low-rank state updates to reduce inference cost while maintaining accuracy, and extending the framework to probabilistic forecasting and uncertainty estimation for enhanced practical utility in decision-critical applications.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (No. 62403076 and 52472399), the Humanities and Social Science Fund of Ministry of Education (No. 24YJCZH416).

References

1. Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control. John Wiley & Sons, 5th edn. (2015)
2. Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z.: Dynamic convolution: Attention over convolution kernels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11030–11039 (June 2020)
3. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 764–773 (Oct 2017).
4. Das, A., Kong, W., Leach, A., Mathur, S., Sen, R., Yu, R.: Long-term forecasting with tiDE: Time-series dense encoder. Transactions on Machine Learning Research (2023)
5. Gu, R., Ding, Y., Li, J., Ding, Y., Sang, W., Huo, X., Qin, X., Ji, Y.: FM-LLM:A frequency-enhanced mixture-of-experts framework for adapting LLMs to time series forecasting. Knowledge-Based Systems 341, 115776 (2026).
6. Guo, S., Lin, Y., Feng, N., Song, C., Wan, H.: Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 922–929 (2019).
7. He, P., Gan, Y., Liu, Y., Lin, R., Zhou, G., Xie, J., Liu, Y., Liu, Q.: AsynFormer: Transformer capturing asynchronous cross-variate dependencies for efficient multivariate time series forecasting. Knowledge-Based Systems **339**, 115557 (2026).
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9(8), 1735–1780 (1997).
9. Hu, Y., Liu, P., Zhu, P., Cheng, D., Dai, T.: Adaptive multi-scale decomposition framework for time series forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 17359–17367 (2025).
10. Hyndman, R.J., Koehler, A.B., Snyder, R.D., Grose, S.: A state space framework for automatic forecasting using exponential smoothing methods. International Journal of Forecasting **18**(3), 439–454 (2002).
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015).
12. Lai, G., Chang, W.C., Yang, Y., Liu, H.: Modeling long- and short-term temporal patterns with deep neural networks. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 95–104 (2018).
13. Li, Z., Li, X., Rao, Z., Pan, L., Xu, Z.: Mts-mixers: Multivariate time series forecasting via factorized temporal and channel mixing. In: International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2025).
14. Li, Z., Qi, S., Li, Y., Xu, Z.: Revisiting long-term time series forecasting: An investigation on linear mapping. arXiv preprint arXiv:2305.10721 (2023).
15. Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., Long, M.: itransformer: Inverted transformers are effective for time series forecasting. In: The Twelfth International Conference on Learning Representations (2024).

16. Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: Long-term forecasting with transformers. In: The Eleventh International Conference on Learning Representations (2023)
17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T.,Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E.Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32 (NeurIPS 2019). vol. 32, pp. 8024–8035.
18. Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T.: DeepAR: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting **36**(3), 1181–1191 (2020).
19. Wang, H., Peng, J., Huang, F., Wang, J., Chen, J., Xiao, Y.: MICN: Multi-scale local and global context modeling for long-term series forecasting. In: The Eleventh International Conference on Learning Representations (2023)
20. Wang, S., Wu, H., Shi, X., Hu, T., Luo, H., Ma, L., Zhang, J.Y., ZHOU, J.:Timemixer: Decomposable multiscale mixing for time series forecasting. In: The Twelfth International Conference on Learning Representations (2024)
21. Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., Long, M.: Timesnet: Temporal 2d-variation modeling for general time series analysis. In: International Conference on Learning Representations (2023).
22. Wu, H., Xu, J., Wang, J., Long, M.: Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 22419–22430. Curran Associates, Inc. (2021)
23. Zeng, A., Chen, M., Zhang, L., Xu, Q.: Are transformers effective for time series forecasting? In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 11121–11128 (2023).
24. Zhang, Y., Yan, J.: Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In: The Eleventh International Conference on Learning Representations (2023).
25. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W.: Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 11106–11115 (2021).
26. Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., Jin, R.: FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 27268–27286. PMLR (17–23 Jul 2022)
27. Xu, S., Jin, R., Zhou, H., Yue, B., Qiao, G., Tai, Y., ... & Liu, G. (2026). From Reaction to Anticipation: Proactive Failure Recovery through Agentic Task Graph for Robotic Manipulation. arXiv preprint arXiv:2605.11951.
28. Xu, S., Liu, G., Kharrat, T., Luo, Y., Aloulou, M., Peña, J. L., ... & Zheng, W. S. (2026). TacticGen: Grounding Adaptable and Scalable Generation of Football Tactics. arXiv preprint arXiv:2604.18210.
29. Xu, S., Yue, B., Zha, H., & Liu, G. (2025, May). A distributional approach to uncertainty-aware preference alignment using offline demonstrations. In International Conference on Learning Representations (Vol. 2025, pp. 3024-3049).