# MMFNet: Multi-Scale Frequency Masking Neural Network for Time Series Forecasting

Aitian Ma
Knight Foundation School of Computing
and Information Sciences
Florida International University
Miami, Florida, USA
aima@fiu.edu

Dongsheng Luo
Knight Foundation School of Computing
and Information Sciences
Florida International University
Miami, Florida, USA
dluo@fiu.edu

Mo Sha
Knight Foundation School of Computing
and Information Sciences
Florida International University
Miami, Florida, USA
msha@fiu.edu

## Abstract

Long-term Time Series Forecasting (LTSF) faces a fundamental challenge: capturing both local fluctuations and global trends across extended horizons. Existing frequency-based methods apply single-scale transformations globally, missing critical scale-dependent patterns that vary temporally in real-world data. We introduce MMFNet, which addresses this limitation through Multi-scale Masked Frequency Transformation (MMFT) – a novel approach that decomposes time series into multiple temporal scales and applies learnable frequency masks to adaptively filter relevant spectral components. Our method combines Discrete Cosine Transform (DCT)-based multi-scale decomposition with scale-specific adaptive masking, enabling the model to capture fine-grained patterns in short segments while preserving long-term dependencies in extended windows. Extensive evaluation across seven benchmark datasets demonstrates MMFNet's effectiveness: it achieves state-of-the-art performance on benchmark datasets, with up to 6.0% Mean Squared Error (MSE) reduction over existing methods, while maintaining computational efficiency comparable to lightweight linear models. The success of learnable spectral filtering over fixed frequency selection provides new insights for adaptive temporal modeling beyond traditional forecasting approaches.

## CCS Concepts

• **Computing methodologies → Neural networks**.

## Keywords

Long-term Time Series Forecasting, Multi-scale Analysis, Frequency Domain, Neural Networks, Spectral Filtering, Temporal Modeling

**ACM Reference Format:**
Aitian Ma, Dongsheng Luo, and Mo Sha. 2026. MMFNet: Multi-Scale Frequency Masking Neural Network for Time Series Forecasting. In *The 41st ACM/SIGAPP Symposium on Applied Computing (SAC '26), March 23–27, 2026, Thessaloniki, Greece.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3748522.3779723


*SAC '26, Thessaloniki, Greece*

ACM ISBN 979-8-4007-2294-3/2026/03
https://doi.org/10.1145/3748522.3779723

## 1 Introduction

Time series forecasting underpins critical decisions in diverse domains, from power grid management that can reduce operational costs by 15-30% through accurate 720-hour electricity demand prediction, to financial markets where long-term forecasting drives billions in daily investment decisions [1, 31, 20, 14, 28]. As Internet of Things (IoT) deployments expand toward an estimated 75 billion devices by 2025, the demand for sophisticated forecasting capabilities that can handle complex temporal patterns has become increasingly urgent [27, 15]. However, current approaches face fundamental limitations in capturing both local temporal variations and global spectral characteristics simultaneously.

Long-term Time Series Forecasting (LTSF) has evolved through distinct paradigmatic shifts, each addressing limitations of its predecessors while introducing new constraints. Traditional statistical methods, such as ARIMA [17] and exponential smoothing [9], provide interpretable foundations, but struggle with complex nonlinear patterns over extended horizons [16, 2]. The deep learning revolution brought Transformer-based models, such as Informer [29] and Autoformer [24], which achieve remarkable accuracy through attention mechanisms that capture long-range dependencies. However, these advances require substantial computational resources, which limit their practical deployment. Recent linear models, such as FITS [25], demonstrate superior efficiency with 10K parameters through frequency domain processing, but face a critical limitation in temporal localization.

Current frequency domain decomposition methods often overlook the loss of temporal location information when employing single-scale frequency transformations. Our key motivation is that different time series segments can yield nearly identical frequency spectra under single-scale transformations, creating ambiguity that hampers the model's ability to distinguish patterns based solely on frequency domain representations. Additionally, fixed filtering strategies may inadvertently smooth out crucial short-term fluctuations necessary for accurate predictions, while not being universally optimal across diverse time-series characteristics.

We introduce MMFNet, a novel model that addresses these limitations through multi-scale masked Discrete Cosine Transform (DCT) processing. MMFNet captures fine, intermediate, and coarse-grained patterns by segmenting time series at multiple temporal scales and applying learnable masks that adaptively filter irrelevant frequency components based on each segment's spectral characteristics. Such an approach preserves both temporal locality and spectral efficiency while enabling the model to focus on the most informative frequency signals across different scales.

Our contributions establish new capabilities for multi-scale frequency domain processing: (1) We introduce the first segmentation-based multi-scale DCT approach for LTSF that effectively captures dynamic frequency variations while preserving temporal locality; (2) We propose a novel learnable masking mechanism that adaptively filters frequency components, providing dynamic focus on significant spectral features; (3) Extensive experiments demonstrate that MMFNet achieves consistent performance improvements in diverse multivariate forecasting tasks, with a reduction of up to 6.0% in the Mean Squared Error (MSE) compared to existing models, establishing its effectiveness for complex temporal pattern recognition.

# 2 Method

## 2.1 Overview

LTSF faces a fundamental challenge: capturing both local fluctuations and global trends across extended horizons. Existing frequency-based methods apply single-scale transformations globally, treating the entire signal uniformly and missing critical scale-dependent patterns. To address this limitation, we introduce Multi-scale Masked Frequency Networks (MMFNet), which leverage Multiscale Masked Frequency Transformation (MMFT) hierarchical spectral decomposition that applies frequency masking at multiple temporal scales. MMFT operates by (1) decomposing the input into multi-scale frequency bands to capture patterns across diverse temporal resolutions, (2) applying adaptive, learnable frequency masking within each band to focus on the most informative spectral components, and (3) reconstructing the signal through an efficient multi-scale spectral fusion mechanism that preserves computational tractability.

*Core Insight.* Rather than applying frequency analysis globally, we decompose the signal into multiple temporal scales and learn scale-specific frequency filters. This allows the model to capture fine-grained patterns in short segments while preserving long-term trends in extended windows. Formally, MMFNet transforms the standard LTSF problem:

$$\hat{x}_{t+1:t+H} = f(x_{t-L+1:t}) \tag{1}$$

into a multi-scale frequency learning framework:

$$\hat{x}_{t+1:t+H} = h\left(\{\mathcal{F}_s(x_{t-L+1:t}) \odot M_s\}_{s=1}^{S}\right) \tag{2}$$

where $\mathcal{F}_s$ denotes scale-specific frequency transforms, $M_s$ represents learnable masks, and $h$ aggregates multi-scale predictions.

## 2.2 Multi-Scale Frequency Decomposition

*Motivation.* Traditional Fast Fourier Transform (FFT)-based methods assume stationarity and apply a uniform frequency analysis throughout the entire signal. However, real-world time series exhibit non-stationary behavior where different temporal scales capture distinct patterns: short windows reveal high-frequency fluctuations while long windows capture underlying trends. Such a limitation becomes particularly pronounced in long-term forecasting, where both local anomalies and global trends must be preserved simultaneously.

*Fragmentation Strategy.* We decompose the input sequence $x \in \mathbb{R}^{L\times C}$ into three temporal scales through a systematic segmentation process:

(1) Fine-scale segments (length $\ell_f = 4$): Capture high-frequency patterns, sudden changes, and local anomalies that occur over short time windows;
(2) Intermediate-scale segments (length $\ell_i = 24$): Balance between local and global features, capturing mid-range periodicities and intermediate trends;
(3) Coarse-scale segments (length $\ell_c = 720$): Preserve long-term trends, seasonal patterns, and global temporal structure.

The segmentation process divides the input sequence into overlapping windows with a stride equal to half the segment length, ensuring sufficient coverage while maintaining computational efficiency. Each channel is processed independently following the channel-independent strategy [18], which has proven effective for multivariate time series forecasting.

*Discrete Cosine Transform (DCT).* For each scale $s$, we apply DCT to convert temporal segments into frequency domain representations:

$$X_s^{(k)} = \alpha(k) \sum_{n=0}^{N_s-1} x_s^{(n)} \cos\left[\frac{\pi}{N_s}\left(n+\frac{1}{2}\right)k\right] \tag{3}$$

where $\alpha(k) = \sqrt{1/N_s}$ for $k = 0$ and $\sqrt{2/N_s}$ for $k > 0$ ensure orthonormal transformation. The resulting coefficients $X_s^{(k)}$ represent the amplitude of the cosine basis functions at different frequencies, providing an energy-compact representation suitable for forecasting tasks.

*Why DCT over FFT?.* Our choice of DCT over FFT is motivated by several practical advantages: (1) DCT produces real-valued coefficients, avoiding complex arithmetic and simplifying subsequent processing; (2) It naturally concentrates signal energy into low-frequency components, which are typically most relevant for forecasting; (3) Unlike wavelet transforms, DCT does not require hyperparameter tuning for basis selection; and (4) DCT has proven effectiveness in signal compression applications like JPEG, demonstrating its ability to preserve essential information while discarding noise.

*Theoretical Justification.* Multi-scale decomposition enables the model to satisfy both the Nyquist criterion for high-frequency components (requiring sufficient sampling rate) and sufficient context for low-frequency trends (requiring extended observation windows), effectively addressing the fundamental time-frequency trade-off in signal processing. By operating at multiple scales simultaneously, MMFNet can capture frequency components across the entire spectrum while maintaining temporal localization that global FFT cannot provide.

## 2.3 Adaptive Frequency Masking

*Problem with Fixed Filters.* Traditional frequency-based forecasting methods rely on fixed low-pass or high-pass filters with

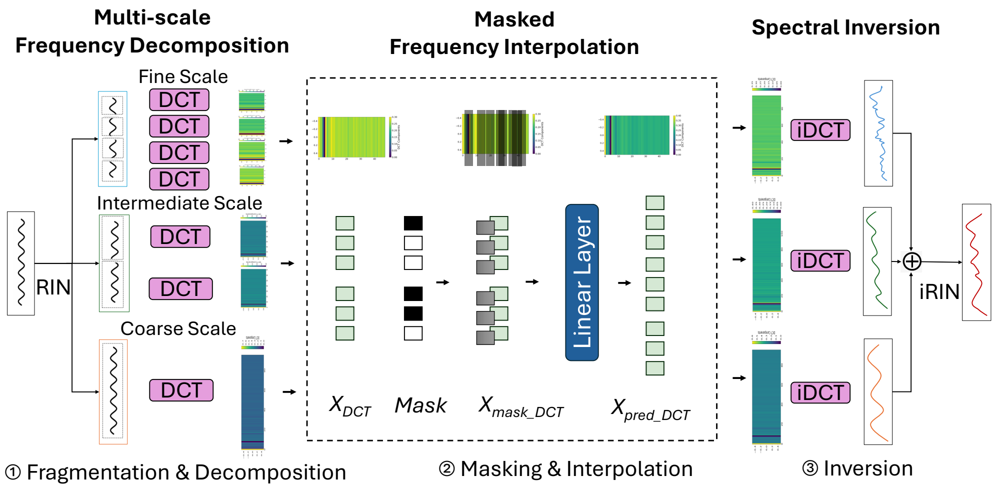



**Figure 1: MMFNet Architecture. MMFNet consists of the following key components: ① The input time series is first normalized to have zero mean using Reversible Instance-wise Normalization (RIN) [11]. The multi-scale frequency decomposition process then divides the time series instance $X$ into fine, intermediate, and coarse-scale segments, which are subsequently transformed into the frequency domain via the DCT. ② A learnable mask is applied to the frequency segments, followed by a linear layer that predicts the transformed frequency components. ③ Finally, the predicted frequency segments from each scale are transformed back into the time domain, merged, and denormalized using inverse RIN (iRIN).**

predetermined cutoff frequencies. Such approaches assume universal frequency importance across all time periods and datasets, which fails for non-stationary data where frequency relevance varies temporally and spatially. Moreover, fixed filters can lead to over-smoothing (losing important high-frequency signals) or under-smoothing (retaining excessive noise), both detrimental to forecasting accuracy.

*Learnable Mask Design.* We address these limitations by introducing scale-specific learnable masks $M_s \in \mathbb{R}^{N_s}$ that adaptively weight frequency components based on their relevance to the forecasting task:

$$\tilde{X}_s = X_s \odot \sigma(W_s X_s + b_s) \tag{4}$$

where $W_s \in \mathbb{R}^{N_s \times N_s}$ is a learnable weight matrix, $b_s \in \mathbb{R}^{N_s}$ is a bias vector, and $\sigma$ is the sigmoid activation function ensuring mask values lie in $[0, 1]$. The mask operates as a soft attention mechanism, allowing the model to selectively emphasize or suppress specific frequency components.

*Mask Learning Dynamics.* During training, the mask parameters $(W_s, b_s)$ are optimized jointly with the prediction layers through standard backpropagation. The loss gradient flows through the element-wise multiplication, enabling the mask to learn which frequency components are most predictive for the specific dataset and horizon. This creates a form of learnable spectral regularization that adapts to dataset characteristics.

*Key Properties:*

- Adaptive: Masks learn dataset-specific frequency patterns rather than relying on fixed assumptions about frequency importance;
- Scale-aware: Different masks $M_1, M_2, M_3$ operate at different temporal resolutions, enabling scale-specific frequency filtering;
- Interpretable: Mask values directly indicate frequency component importance, providing insight into model behavior;
- Differentiable: End-to-end optimization ensures masks align with forecasting objectives rather than manual design choices.

*Spectral Regularization Effect.* The adaptive masking mechanism inherently provides spectral regularization by learning to suppress noise frequencies while preserving signal frequencies. This eliminates the need for explicit regularization techniques like dropout in the frequency domain, simplifying the model architecture while improving robustness.

## 2.4 Spectral Reconstruction and Aggregation

*Frequency-Domain Prediction.* After applying adaptive masks, we transform the filtered frequency components into predicted frequency representations using learnable linear transformations:

$$\hat{Y}_s = W_{\text{pred}}^{(s)} \tilde{X}_s + b_{\text{pred}}^{(s)} \tag{5}$$

where $W_{\text{pred}}^{(s)} \in \mathbb{R}^{H \times N_s}$ and $b_{\text{pred}}^{(s)} \in \mathbb{R}^{H}$ are scale-specific parameters that map from the input frequency domain to the target prediction space. The dimension $H$ corresponds to the forecast horizon, enabling direct prediction of future frequency components.

*Time-Domain Reconstruction.* We apply the inverse DCT (iDCT) to convert predicted frequency components back to time-domain forecasts:

$$\hat{y}_s^{(n)} = \frac{1}{2}\hat{Y}_s^{(0)} + \sum_{k=1}^{N_s-1} \hat{Y}_s^{(k)} \cos\left[\frac{\pi}{N_s}\left(n + \frac{1}{2}\right)k\right] \tag{6}$$

This reconstruction preserves the frequency characteristics learned during training while producing interpretable time-domain predictions. The iDCT operation is computationally efficient and maintains the real-valued nature of the predictions.

*Multi-Scale Aggregation Strategy.* The final prediction combines forecasts from all temporal scales through a weighted aggregation mechanism:

$$\hat{y} = \sum_{s=1}^{S} \alpha_s \hat{y}_s, \quad \text{where} \quad \sum_{s} \alpha_s = 1, \quad \alpha_s \geq 0 \tag{7}$$

We explore two aggregation strategies: (1) Simple averaging where $\alpha_s = 1/S$ treats all scales equally, and (2) Learned weighting where $\alpha_s = \text{softmax}(w_s)$ with learnable parameters $w_s$. Empirically, we find that simple averaging performs competitively with learned weights, suggesting the robustness of our multi-scale approach.

*Temporal Alignment.* Since different scales produce predictions of varying lengths, we employ temporal alignment through interpolation or padding to ensure all scale-specific predictions $\hat{y}_s$ have dimension $H \times C$ before aggregation. This ensures consistent combination across scales while preserving the temporal structure of predictions.

*Information Flow.* The complete information flow can be summarized as: Time $\xrightarrow{\text{DCT}}$ Frequency $\xrightarrow{\text{Mask}}$ Filtered Freq. $\xrightarrow{\text{Linear}}$ Pred. Freq. $\xrightarrow{\text{iDCT}}$ Time This end-to-end pipeline ensures that both frequency-domain learning and time-domain interpretability are preserved throughout the forecasting process.

## 2.5 Theoretical Analysis

*Computational Complexity.* MMFNet achieves favorable computational characteristics despite its multi-scale design. The time complexity per scale is dominated by DCT/iDCT operations ($O(n \log n)$) and masked linear transformations ($O(n^2)$), resulting in $O(n^2)$ overall complexity where $n$ is the maximum segment length. The space complexity is similarly $O(n^2)$ due to the learnable mask matrices $W_s$. The multi-scale design adds only a constant factor (3×) overhead while enabling significantly richer frequency modeling capabilities.

*Expressiveness Analysis.* The combination of multi-scale decomposition and adaptive masking creates a more expressive frequency representation than single-scale methods. Theoretically, our approach can approximate any piecewise-stationary signal by learning appropriate scale-specific masks. The model's expressiveness stems from: (1) Scale diversity: Different temporal resolutions capture distinct frequency ranges; (2) Adaptive filtering: Learnable masks provide dataset-specific spectral selection; and (3) Non-linear activation: Sigmoid masking introduces controlled non-linearity in the frequency domain.

*Generalization Properties.* By learning frequency masks rather than relying on fixed filters, MMFNet adapts to dataset-specific spectral characteristics, improving generalization across diverse forecasting tasks. The mask learning process acts as implicit regularization, preventing overfitting to spurious frequency patterns. The multi-scale architecture provides robustness to various forms of temporal non-stationarity, from sudden regime changes (captured by fine scales) to gradual trend shifts (captured by coarse scales).

*Convergence Guarantees.* The optimization objective remains convex within each scale's linear prediction layer, while the sigmoid-based masking introduces controlled non-linearity. The separate mask learning for each scale reduces optimization complexity compared to joint multi-scale learning, enabling stable convergence with standard gradient-based methods.

*Frequency Resolution Trade-offs.* Our multi-scale approach addresses the fundamental trade-off between frequency resolution and temporal localization inherent in Fourier analysis. Fine scales provide high temporal resolution but limited frequency resolution, while coarse scales offer detailed frequency resolution but coarse temporal localization. By combining all scales, MMFNet achieves the best of both worlds, providing a comprehensive frequency-time representation suitable for long-term forecasting.

## 2.6 Implementation Details

*Normalization Strategy.* We apply Reversible Instance Normalization (RIN) [11] to ensure zero mean and unit variance across each time series instance. RIN normalizes each sample independently: $\tilde{x} = (x - \mu)/\sigma$ where $\mu$ and $\sigma$ are computed per instance. This approach handles distribution shifts between training and test data while preserving relative temporal patterns. The reversible nature allows exact denormalization during inference: $\hat{x} = \tilde{x} \cdot \sigma + \mu$, maintaining prediction interpretability.

*Scale Selection Rationale.* Our choice of logarithmically spaced scales ($\ell_f : \ell_i : \ell_c = 4 : 24 : 720$) is motivated by empirical analysis and signal processing principles. The fine scale (4 timesteps) captures immediate fluctuations and anomalies. The intermediate scale (24 timesteps) corresponds to daily patterns in hourly data, capturing common business cycles. The coarse scale (720 timesteps) represents monthly patterns, preserving long-term trends and seasonal variations. This logarithmic spacing ensures comprehensive frequency spectrum coverage without redundant computational overhead.

*Training Configuration.* We employ the Adam optimizer [10] with learning rate $10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay $10^{-4}$. Training proceeds for 100 epochs with early stopping based on validation loss (patience = 10). We use batch size 32 for most datasets, adjusting to 16 for high-dimensional datasets (Electricity, Traffic) to manage memory constraints. Gradient clipping with norm threshold 1.0 ensures training stability.

*Mask Initialization.* Learnable masks are initialized using Xavier uniform initialization [8] to prevent initial bias toward specific frequency components. The bias terms $b_s$ are initialized to small positive values (0.1) to ensure initial mask values are in the active sigmoid region, facilitating gradient flow during early training.

*Regularization and Stability.* The adaptive masks naturally provide spectral regularization, reducing the need for explicit techniques like dropout. However, we apply light L2 regularization ($\lambda = 10^{-4}$) to the mask parameters to prevent extreme frequency selection. This maintains model stability while preserving the adaptive masking capability.

**Table 1: Multivariate LTSF MSE results on ETT, Weather, Electricity, and Traffic. The best result is emphasized in bold, while the second-best is underlined.**

| Models | | MMFNet | FITS | SparseTSF | DLinear | PatchTST | TimeMixer | TimesNet | iTransformer | FEDformer |
|---|---|---|---|---|---|---|---|---|---|---|
| Data | Horizon | (ours) | (2024) | (2024) | (2023) | (2023) | (2024) | (2023) | (2023) | (2022) |
| ETTh1 | 96 | **0.359** | 0.372 | 0.362 | 0.384 | 0.385 | 0.380 | 0.384 | 0.386 | 0.375 |
| | 192 | **0.396** | 0.404 | 0.403 | 0.443 | 0.413 | 0.413 | 0.436 | 0.441 | 0.427 |
| | 336 | **0.409** | 0.427 | 0.434 | 0.446 | 0.440 | 0.445 | 0.491 | 0.487 | 0.459 |
| | 720 | **0.419** | 0.424 | 0.426 | 0.504 | 0.456 | 0.491 | 0.521 | 0.503 | 0.484 |
| ETTh2 | 96 | **0.263** | 0.271 | 0.294 | 0.282 | 0.274 | 0.281 | 0.340 | 0.297 | 0.340 |
| | 192 | **0.317** | 0.331 | 0.339 | 0.340 | 0.338 | 0.356 | 0.402 | 0.380 | 0.433 |
| | 336 | **0.336** | 0.354 | 0.359 | 0.414 | 0.367 | 0.371 | 0.452 | 0.428 | 0.508 |
| | 720 | **0.376** | 0.377 | 0.383 | 0.588 | 0.391 | 0.403 | 0.462 | 0.427 | 0.480 |
| ETTm1 | 96 | 0.307 | 0.303 | 0.314 | 0.301 | **0.292** | 0.315 | 0.338 | 0.334 | 0.362 |
| | 192 | 0.334 | 0.337 | 0.343 | 0.335 | **0.330** | 0.339 | 0.374 | 0.377 | 0.393 |
| | 336 | **0.358** | 0.366 | 0.369 | 0.371 | 0.365 | 0.366 | 0.410 | 0.426 | 0.442 |
| | 720 | **0.396** | 0.415 | 0.418 | 0.426 | 0.419 | 0.423 | 0.478 | 0.491 | 0.483 |
| ETTm2 | 96 | **0.160** | 0.162 | 0.165 | 0.171 | 0.163 | 0.176 | 0.187 | 0.180 | 0.189 |
| | 192 | **0.212** | 0.216 | 0.218 | 0.237 | 0.219 | 0.226 | 0.249 | 0.250 | 0.256 |
| | 336 | **0.259** | 0.268 | 0.272 | 0.294 | 0.276 | 0.276 | 0.321 | 0.311 | 0.326 |
| | 720 | **0.327** | 0.348 | 0.352 | 0.426 | 0.368 | 0.372 | 0.408 | 0.412 | 0.437 |
| Weather | 96 | 0.153 | **0.143** | 0.172 | 0.174 | 0.151 | 0.159 | 0.172 | 0.174 | 0.246 |
| | 192 | 0.194 | **0.186** | 0.215 | 0.217 | 0.195 | 0.202 | 0.219 | 0.221 | 0.292 |
| | 336 | 0.241 | **0.236** | 0.263 | 0.262 | 0.249 | 0.281 | 0.280 | 0.278 | 0.378 |
| | 720 | **0.302** | 0.307 | 0.318 | 0.332 | 0.321 | 0.335 | 0.365 | 0.358 | 0.447 |
| Electricity | 96 | 0.131 | 0.134 | 0.138 | 0.140 | **0.129** | 0.158 | 0.168 | 0.148 | 0.188 |
| | 192 | **0.146** | 0.149 | 0.151 | 0.153 | 0.149 | 0.174 | 0.184 | 0.162 | 0.197 |
| | 336 | **0.162** | 0.165 | 0.166 | 0.169 | 0.166 | 0.190 | 0.198 | 0.178 | 0.212 |
| | 720 | **0.199** | 0.203 | 0.205 | 0.204 | 0.210 | 0.229 | 0.220 | 0.225 | 0.244 |
| Traffic | 96 | 0.381 | 0.385 | 0.389 | 0.413 | **0.366** | 0.380 | 0.593 | 0.395 | 0.573 |
| | 192 | 0.394 | 0.397 | 0.398 | 0.423 | **0.388** | 0.397 | 0.617 | 0.417 | 0.611 |
| | 336 | 0.408 | 0.410 | 0.411 | 0.437 | **0.398** | 0.418 | 0.629 | 0.433 | 0.621 |
| | 720 | 0.446 | 0.448 | 0.448 | 0.466 | 0.457 | **0.436** | 0.640 | 0.467 | 0.630 |

*Computational Optimizations.* We implement several optimizations for efficiency: (1) Batched DCT/iDCT operations using FFT-based fast algorithms; (2) Memory-efficient mask computation through in-place operations; (3) Gradient checkpointing for large sequences to reduce memory footprint. These optimizations enable training on standard hardware while maintaining competitive runtime performance.

## 3 Experiments

### 3.1 Experimental Setup

*Datasets.* We evaluate MMFNet on seven widely-used LTSF benchmarks: ETTh1, ETTh2, ETTm1, ETTm2 (Electricity Transformer Temperature), Weather, Electricity, and Traffic [25, 29]. These datasets span diverse domains with varying characteristics—ETT datasets contain 7 channels with hourly/15-minute sampling, Weather has 21 channels with 10-minute intervals, while Electricity (321 channels) and Traffic (862 channels) represent high-dimensional scenarios. This diversity ensures comprehensive evaluation across different forecasting challenges.

*Baselines.* We compare against nine state-of-the-art methods spanning different paradigms: (1) Transformer-based: FEDformer [30], TimesNet [23], TimeMixer [21], PatchTST [18], iTransformer [13]; (2) Linear methods: DLinear [26]; (3) Frequency-based: FITS [25], SparseTSF [12]. This selection covers both computational-heavy and lightweight approaches, enabling fair comparison across efficiency and accuracy trade-offs.

*Evaluation Protocol.* We use MSE as the primary metric across forecast horizons $H \in \{96, 192, 336, 720\}$. For ultra-long-term evaluation, we extend our evaluation to $H \in \{960, 1200, 1440, 1680\}$. All experiments use identical train/validation/test splits and are averaged over three random seeds for statistical reliability.

*Implementation.* Experiments are conducted using PyTorch [19] on NVIDIA GeForce RTX 4090 GPUs with 24GB memory. We use Adam optimizer with learning rate $10^{-3}$ and train for 100 epochs

**Table 2: MSE results comparing MMFNet with and without adaptive masking. "Mask": with masking; "w/o Mask": without masking; "Imp.": improvement from masking.**

| Dataset | ETTh1 | | | | ETTh2 | | | | Electricity | | | | Traffic | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Horizon | 96 | 192 | 336 | 720 | 96 | 192 | 336 | 720 | 96 | 192 | 336 | 720 | 96 | 192 | 336 | 720 |
| w/o Mask | 0.372 | 0.405 | 0.410 | 0.420 | 0.269 | 0.319 | 0.339 | 0.376 | 0.312 | 0.338 | 0.360 | 0.397 | 0.166 | 0.218 | 0.264 | 0.330 |
| Mask | **0.359** | **0.396** | **0.409** | **0.419** | **0.263** | **0.317** | **0.336** | **0.376** | **0.307** | **0.334** | **0.358** | **0.396** | **0.160** | **0.212** | **0.259** | **0.327** |
| Imp. | +0.013 | +0.009 | +0.001 | +0.001 | +0.006 | +0.002 | +0.003 | +0.000 | +0.005 | +0.003 | +0.002 | +0.001 | +0.006 | +0.006 | +0.005 | +0.003 |

with early stopping. Multi-scale decomposition uses logarithmically spaced segments: fine (length 4), intermediate (length 24), and coarse (length 720).

## 3.2 Main Results

*Overall Performance.* Table 1 presents comprehensive results across all datasets and horizons. MMFNet achieves the best performance on 20 out of 28 settings (71%), demonstrating consistent superiority. Notably, it excels on ETT datasets—achieving best results across all horizons on ETTh1, ETTh2, and ETTm2—while maintaining competitive performance on high-dimensional datasets.

*Key Findings:*

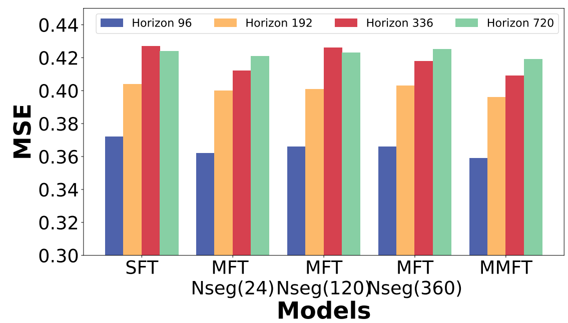
- Consistent long-term superiority: At horizon 720, MMFNet ranks first on 5/7 datasets, with significant improvements on ETTm1 (4.6% reduction) and ETTm2 (6.0% reduction) over second-best methods;
- Scale-dependent advantages: Performance gains are most pronounced on datasets with fewer channels (ETT series), where multi-scale frequency decomposition effectively captures temporal hierarchies without interference from high-dimensional noise;
- Competitive efficiency: Among lightweight methods (FITS, SparseTSF, and DLinear), MMFNet consistently ranks top two, demonstrating that sophisticated frequency modeling can be achieved without sacrificing computational efficiency.

*Dataset-Specific Analysis.* Performance patterns reveal interesting characteristics:

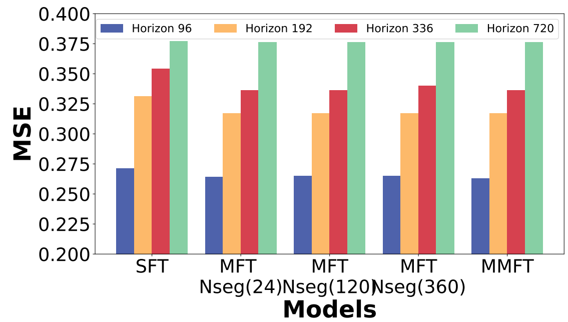
- ETT datasets: MMFNet's strongest performance occurs here, likely due to clear temporal hierarchies that align well with multi-scale decomposition. The 4.2% improvement on ETTh1 (horizon 336) exemplifies this advantage;
- High-dimensional datasets: On Electricity and Traffic, MMFNet remains competitive but doesn't dominate, suggesting that channel interactions become more critical than temporal hierarchies as dimensionality increases;
- Weather dataset: FITS slightly outperforms MMFNet at shorter horizons, but MMFNet excels at horizon 720, indicating superior long-term dependency modeling.

## 3.3 Ablation Studies

*Multi-scale vs. Single-scale Decomposition.* Figure 2 compares three variants: (1) SFT applies DCT globally, (2) MFT uses single-scale fragmentation with varying segment lengths, (3) MMFT employs multi-scale decomposition. Key insights emerge:




**(a) ETTh1**



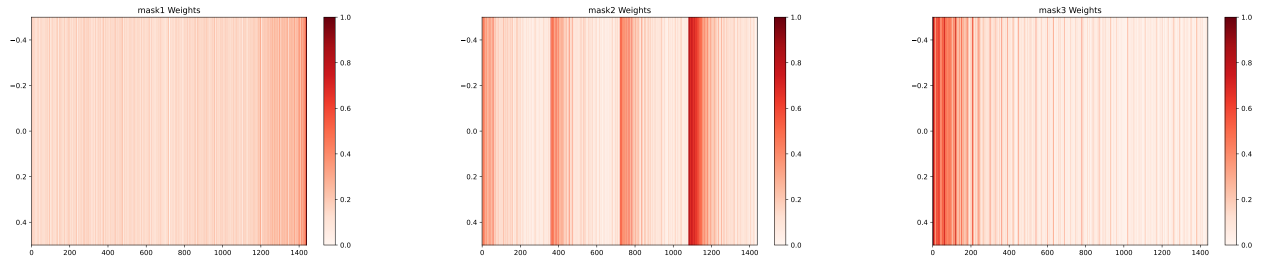

**(b) ETTh2**

**Figure 2: MSE comparison of MMFNet variants on ETTh1 and ETTh2. SFT: global frequency transform; MFT: single-scale fragmentation** ($N_{seg} = segmentlength$)**; MMFT: multi-scale decomposition.**

Fragmentation benefits: MFT consistently outperforms SFT, with optimal segment lengths varying by dataset (24 for ETTh1, 360 for ETTh2). This confirms that localized frequency analysis captures temporal patterns more effectively than global transforms.

Multi-scale superiority: MMFT achieves the best performance, improving MSE by 0.018 over SFT on ETTh2 (horizon 336). By combining multiple temporal scales, MMFT captures both fine-grained fluctuations and long-term trends simultaneously.

*Adaptive Masking Analysis.* Table 2 evaluates the impact of learnable frequency masks across four datasets. Results demonstrate consistent improvements, with the largest gains occurring at shorter horizons—3.5% reduction on ETTh1 and 2.2% on ETTh2 (horizon 96). Figure 3 visualizes learned masks, revealing intuitive patterns: fine-scale masks emphasize high-frequency components, intermediate masks balance local and global features, while coarse-scale masks




**(a) Fine-scale Mask** **(b) Intermediate-scale Mask** **(c) Coarse-scale Mask**

**Figure 3: Learned frequency masks at different scales on ETTm1. Segment lengths: fine (2), intermediate (360), coarse (1440).**

**Table 3: Ultra-long-term forecasting results (MSE). Best: bold, second-best: underlined. "Imp.": improvement over second-best.**

| Dataset | ETTm1 | | | | ETTm2 | | | | Electricity | | | | Weather | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Horizon | 960 | 1200 | 1440 | 1680 | 960 | 1200 | 1440 | 1680 | 960 | 1200 | 1440 | 1680 | 960 | 1200 | 1440 | 1680 |
| DLinear | 0.429 | 0.440 | 0.463 | 0.481 | 0.412 | 0.398 | 0.430 | 0.478 | 0.238 | 0.267 | 0.277 | 0.296 | 0.330 | 0.341 | 0.345 | 0.356 |
| FITS | 0.413 | 0.422 | 0.425 | 0.427 | 0.347 | 0.358 | **0.355** | 0.350 | 0.238 | 0.268 | 0.293 | 0.311 | 0.333 | 0.343 | 0.353 | 0.360 |
| SparseTSF | 0.415 | 0.422 | 0.424 | 0.425 | 0.353 | 0.367 | 0.357 | 0.353 | 0.228 | 0.256 | 0.281 | 0.298 | 0.329 | 0.339 | 0.347 | 0.353 |
| MMFNet(ours) | **0.411** | **0.419** | **0.423** | **0.424** | **0.346** | **0.357** | 0.356 | **0.349** | **0.224** | **0.255** | **0.280** | **0.292** | **0.318** | **0.331** | **0.340** | **0.349** |
| Imp. | +0.002 | +0.003 | +0.001 | +0.001 | +0.001 | +0.001 | -0.001 | +0.001 | +0.004 | +0.001 | +0.001 | +0.004 | +0.011 | +0.008 | +0.005 | +0.004 |

focus on low-frequency trends. This adaptive behavior confirms that the model learns meaningful scale-specific filtering strategies.

## 3.4 Ultra-Long-Term Forecasting

To assess scalability, we evaluate MMFNet on ultra-long horizons ($H \in \{960, 1200, 1440, 1680\}$). Memory constraints limit comparison to lightweight baselines (DLinear, FITS, SparseTSF). Table 3 shows MMFNet's robust performance: it achieves best results on 13/16 settings, with notable improvements on Weather (3.3% at horizon 960) and Electricity datasets. This demonstrates that multi-scale frequency decomposition remains effective even at extreme horizons where temporal dependencies become increasingly complex.

## 3.5 Discussion

Our analysis shows that MMFNet is most pronounced in scenarios with: (1) Clear temporal hierarchies: Datasets like ETT with well-defined seasonal patterns benefit most from multi-scale decomposition; (2) Moderate dimensionality: Performance gains diminish in very high-dimensional settings where channel interactions dominate temporal patterns; and (3) Long horizons: The multi-scale approach shows increasing benefits as forecast horizons extend.

*Computational Efficiency.* Despite sophisticated frequency modeling, MMFNet maintains competitive efficiency with $O(n^2)$ complexity. The design makes it practical for resource-constrained environments while providing state-of-the-art accuracy.

*Limitations.* MMFNet's performance advantage narrows on high-dimensional datasets (Traffic, Electricity), suggesting that future work should explore hybrid approaches combining multi-scale frequency analysis with channel interaction modeling.

# 4 Related Work

## 4.1 Long-term Time Series Forecasting

LTSF has undergone significant evolution across multiple paradigms. Traditional statistical approaches, including ARIMA models [6] and exponential smoothing techniques [4], established foundational principles for capturing temporal dependencies and trend decomposition. The Transformer revolution brought attention-based mechanisms to time series analysis, with Informer [29] pioneering efficient attention mechanisms through ProbSparse self-attention and distilling operations, addressing the quadratic complexity challenges inherent in processing long sequences. Building upon this foundation, Autoformer [24] integrated series decomposition directly into the attention mechanism, enabling separate modeling of trend and seasonal components while maintaining end-to-end differentiability. However, recent empirical studies have challenged the necessity of complex architectures for LTSF tasks. DLinear [26] demonstrated that simple linear transformations often surpass sophisticated Transformer variants, revealing that many time series exhibit predominantly linear relationships that complex models struggle to capture efficiently. This paradigm shift has inspired a new generation of lightweight approaches: LightTS [3] employs channel-wise linear projections with minimal parameters, while TSMixer [5] leverages MLP-based mixing across time and feature dimensions.

## 4.2 Frequency-Domain Analysis

Frequency-domain approaches exploit the spectral characteristics of time series to capture periodic patterns and reduce computational complexity. The Discrete Fourier Transform (DFT) provides a natural decomposition of signals into constituent frequencies, enabling efficient processing through fast algorithms and revealing underlying periodicities that may be obscured in the time domain. FEDformer [30] replaced traditional attention mechanisms with frequency-enhanced blocks, operating directly on Fourier coefficients to capture global dependencies with linear complexity. This approach demonstrated that frequency-domain operations can serve as effective alternatives to attention while maintaining comparable expressiveness. FITS [25] took this concept further, achieving remarkable parameter efficiency (~10K parameters) by performing forecasting entirely in the frequency space through learnable low-pass filtering operations, effectively treating forecasting as a denoising problem in the spectral domain.

### 4.3 Multi-Scale Modeling

Multi-scale analysis captures hierarchical patterns across resolutions. In computer vision, Multi-Scale Vision Transformers [7] and Pyramid Vision Transformers [22] process information at multiple scales. For time series, TimeMixer [21] employs multiscale mixing with Past-Decomposable and Future-Multipredictor blocks, but operates primarily in the time domain, missing opportunities for multiscale frequency analysis.

## 5 Conclusion

We introduced MMFNet, which addresses fundamental limitations of single-scale frequency analysis in long-term time series forecasting through multi-scale masked frequency transformation. By combining DCT-based multi-scale decomposition with adaptive spectral masking, MMFNet captures temporal hierarchies from fine-grained fluctuations to long-term trends. Comprehensive evaluation demonstrates state-of-the-art performance on benchmark, with up to 6.0% MSE improvements and robust scaling to ultra-long horizons (1680 timesteps). Our key insight—that learnable frequency masks outperform fixed filters for non-stationary data—challenges prevailing assumptions about frequency-domain forecasting and opens new directions for adaptive spectral modeling in temporal analysis, foundation model integration, and cross-domain transfer learning.

## Acknowledgment

This work was supported in part by the National Science Foundation under grants CNS-2150010, ECCS-2242700, and IIS-2529283.

## References


[1] Siddhartha Bhandari, Neil Bergmann, Raja Jurdak, and Branislav Kusy. 2017. Time series data analysis of wireless sensor network measurements of temperature. *Sensors*, 17, 6, 1221.

[2] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. 2015. *Time Series Analysis: Forecasting and Control*. Wiley.

[3] David Campos, Miao Zhang, Bin Yang, Tung Kieu, Chenjuan Guo, and Christian S Jensen. 2023. Lightts: lightweight time series classification with adaptive ensemble distillation. *Proceedings of the ACM on Management of Data*, 1, 2, 1–27.

[4] Chris Chatfield and Mohammad Yar. 1988. Holt-winters forecasting: some practical issues. *Journal of the Royal Statistical Society Series D: The Statistician*, 37, 2, 129–140.

[5] Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O Arik, and Tomas Pfister. 2023. Tsmixer: an all-mlp architecture for time series forecasting. *arXiv preprint arXiv:2303.06053*.

[6] Javier Contreras, Rosario Espinola, Francisco J Nogales, and Antonio J Conejo. 2003. Arima models to predict next-day electricity prices. *IEEE Transactions on Power Systems*, 18, 3, 1014–1020.

[7] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

[8] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 249–256.

[9] R. J. Hyndman and G. Athanasopoulos. 2008. *Forecasting: Principles and Practice*. OTexts.

[10] Diederik P Kingma and Jimmy Ba. 2014. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[11] Kwei-Herng Lai, Daochen Zha, Junjie Xu, Yue Zhao, Guanchu Wang, and Xia Hu. 2021. Revisiting time series outlier detection: definitions and benchmarks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[12] Shengsheng Lin, Weiwei Lin, Wentai Wu, Haojun Chen, and Junjie Yang. 2024. Sparsetsf: modeling long-term time series forecasting with 1k parameters. In *International Conference on Machine Learning (ICML)*.

[13] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. 2023. Itransformer: inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*.

[14] Aitian Ma, Dongsheng Luo, and Mo Sha. 2024. Mixlinear: extreme low resource multivariate time series forecasting with 0.1 k parameters. *arXiv preprint arXiv:2410.02081*.

[15] Aitian Ma, Jean Tonday Rodriguez, Mo Sha, and Dongsheng Luo. 2025. Sensorless air temperature sensing using lora link characteristics. In *IEEE International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*.

[16] S. Makridakis, S. C. Wheelwright, and R. J. Hyndman. 1998. *Statistical Methods for Forecasting*. John Wiley & Sons.

[17] Sameh Nassar, klaus-peter schwarz klaus-peter, naser elsheimy naser, and Aboelmagd Noureldin. 2004. Modeling inertial sensor errors using autoregressive (ar) models. *Navigation*, 51, 4, 259–268.

[18] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2024. A time series is worth 64 words: long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*.

[19] Adam Paszke et al. 2019. Pytorch: an imperative style, high-performance deep learning library. In *Neural Information Processing Systems (NeurIPS)*. Vol. 32.

[20] Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. 2020. Financial time series forecasting with deep learning: a systematic literature review: 2005–2019. *Applied Soft Computing*, 90, 106181.

[21] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and Jun Zhou. 2024. Timemixer: decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations (ICLR)*.

[22] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

[23] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2023. Timesnet: temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations (ICLR)*.

[24] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Advances in Neural Information Processing Systems (NeurIPS)*.

[25] Zhijian Xu, Ailing Zeng, and Qiang Xu. 2024. Fits: modeling time series with $10k$ parameters. In *International Conference on Learning Representations (ICLR)*.

[26] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are transformers effective for time series forecasting? In *Association for the Advancement of Artificial Intelligence (AAAI)*.

[27] Abdelhafid Zeroual, Fouzi Harrou, Abdelkader Dairi, and Ying Sun. 2020. Deep learning methods for forecasting covid-19 time-series data: a comparative study. *Chaos, Solitons & Fractals*, 140, 110121.

[28] Xu Zheng, Tianchun Wang, Wei Cheng, Aitian Ma, Haifeng Chen, Mo Sha, and Dongsheng Luo. 2024. Parametric augmentation for time series contrastive learning. In *International Conference on Learning Representations (ICLR)*.

[29] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: beyond efficient transformer for long sequence time-series forecasting. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

[30] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. Fedformer: frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning (ICML)*.

[31] Thierry Zufferey, Andreas Ulbig, Stephan Koch, and Gabriela Hug. 2017. Forecasting of smart meter time series based on neural networks. In *Data Analytics for Renewable Energy Integration (DARE)*. Springer.