DDM-Mixer: Dual-Domain Multi-scale Mixer for Time Series Forecasting

1st Yiqi Yu
School of Software Technology
Zhejiang University
Ningbo, China
yiqiyu@zju.edu.cn

Xuelin Cheng*
School of Software Technology
Zhejiang University
Ningbo, China
cxlin@zju.edu.cn

***Abstract*—Time series forecasting greatly benefits from synergistically fusing time-domain (local temporal dependencies) and frequency-domain (global periodic) representations. However, effectively integrating these spaces is often hindered by the "Spectral Geometric Mismatch" problem: when the historical input window and the desired output horizons differ, standard spectral features suffer from severe aliasing during mapping. To overcome this fundamental bottleneck, we propose the DDM-Mixer (Dual-Domain Multi-scale Mixer), utilizing a High-Density Spectral Interpolation strategy via Extended DFT to rigorously align mismatched spectral dimensions into a unified representation. Furthermore, to effectively distinguish true, predictable periodicity from chaotic stochastic noise and prevent overfitting to non-periodic elements, we introduce an adaptive Parameter-Free Information-Theoretic Gating Mechanism based on Spectral Entropy. Extensive experiments conducted across seven diverse real-world benchmark datasets demonstrate that DDM-Mixer consistently achieves state-of-the-art performance, showcasing superior robustness in complex forecasting scenarios.**

***Index Terms*—Time Series Forecasting(TSF), multi-scale decomposition, time domain, frequency domain, MLP.**

I. Introduction

Time series forecasting (TSF) is critical across diverse domains [1]–[3]. Deep learning solutions have evolved from Transformers utilizing advanced attention and decomposition [4]–[11] to simpler, robust linear, basis-expansion, and multi-scale architectures that mitigate temporal noise overfitting [12]–[21].

Inspired by multi-scale mixing [20], [21], we extend this paradigm to the frequency domain [22]–[24]. However, when forecasting horizons H differ from input lengths L, models suffer from a **Spectral Geometric Mismatch**, causing severe feature aliasing [24], [25].

To resolve this, we propose **DDM-Mixer (Dual-Domain Multi-scale Mixer)**, utilizing an Extended DFT-based **Time-Frequency Adapter (TFA)** for rigorous spectral alignment. Furthermore, unlike methods using harmonic energy ratios [26], we introduce a **Spectral Entropy-based** gating mechanism that prevents overfitting by dynamically weighing frequency features based on signal periodicity.

Our main contributions are:

- **Dual-Domain Multi-Scale Mixing:** A unified framework synergizing trend-seasonal and frequency branches.
- **Robust Spectral Alignment:** Resolves dimension mismatch ($L \neq H$) via Extended DFT interpolation.
- **Adaptive Entropy Gating:** An uncertainty-guided selector regulating frequency contributions.

*Xuelin Cheng is corresponding author.

II. Related Works

A. Decomposition-based Architectures

Decomposition fundamentally separates time series into trend and seasonal components. Deep learning approaches have advanced from simple moving averages [14] to adaptive, multi-granular strategies [27], [28]. Advanced architectures tightly integrate decomposition to progressively refine components [6], [7] or utilize hierarchical interpolation for multi-rate modeling [13]. Building on TimeMixer's [21] multi-scale mixing, our DDM-Mixer adopts this backbone but uniquely addresses the global periodicities often missed by purely time-domain decompositions.

B. Frequency-domain Forecasting

Frequency-domain methods capture global dependencies through spectral transforms. Early models selected subset frequency modes [7], while recent advances diversify spectral representations via Koopman theory [29], spectral graphs [30], complex-valued linear layers [25], and sparse sampling [31]. Despite these innovations, "frequency misalignment"—caused by differing input and output window lengths—remains a persistent challenge, hindering dense prediction tasks. DDM-Mixer addresses this fundamental gap by utilizing Extended DFT for rigorous spectral interpolation, explicitly aligning feature dimensions across varying time scales.

C. Multi-scale Modeling

Multi-scale modeling captures complex, long-range temporal variations. This has been extensively explored through pyramidal or segmented Transformer attention [8], [9], adaptive multi-resolution pathways [32], and large-receptive-field CNNs [33]. However, these complex architectures often incur significant computational overhead. DDM-Mixer achieves efficient scale-aware modeling via an MLP-based multi-scale mixing structure. By avoiding heavy convolutional stacks, our framework disentangles microscopic noise from macroscopic trends while maintaining linear complexity for long-term forecasting.

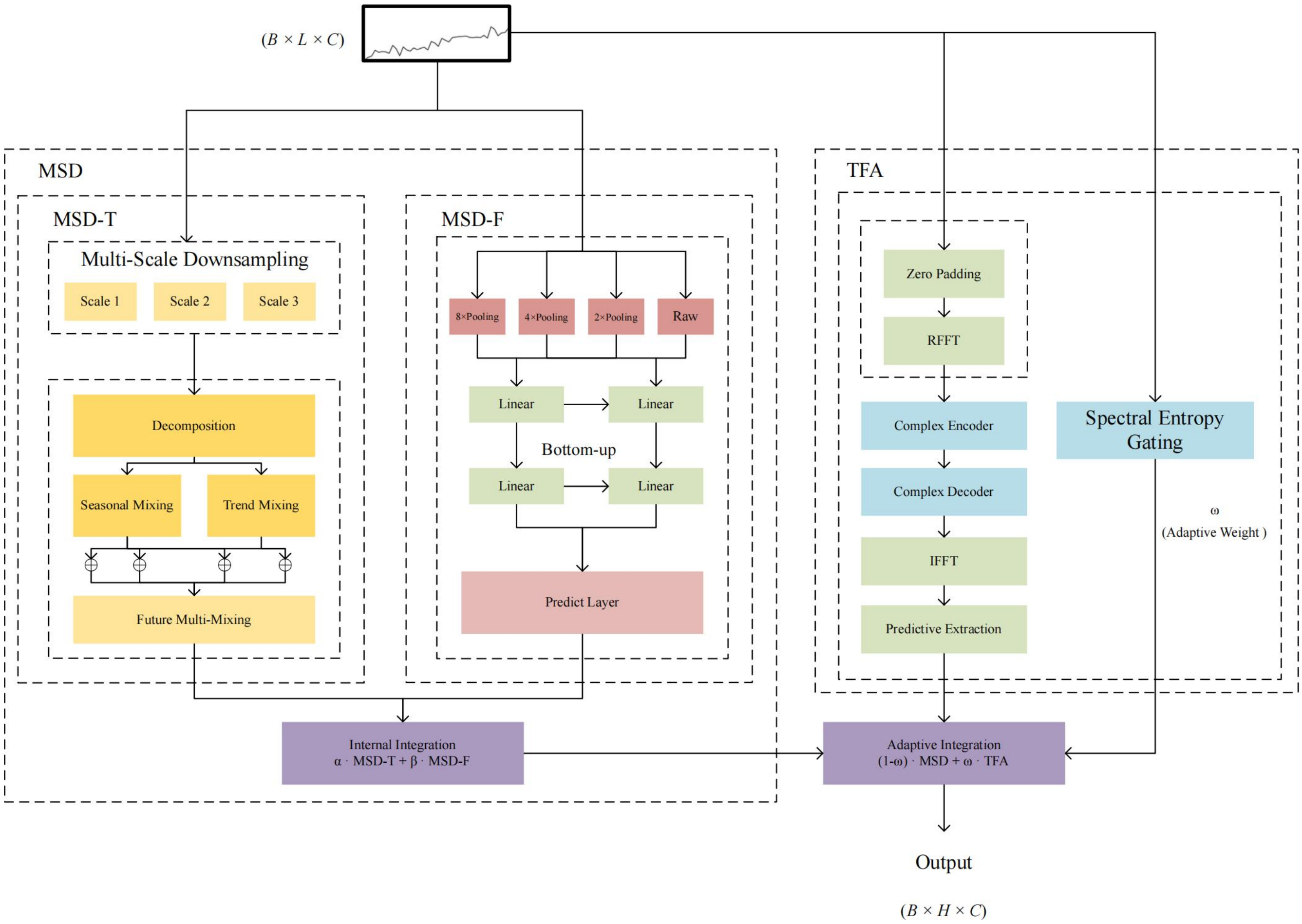


Fig. 1. The overall architecture of DDM-Mixer. The framework adopts a multi-scale backbone based on TimeMixer $(X^{(0)}, \ldots, X^{(M)})$. The time-domain features are refined by the Multi-Scale Decomposition (MSD) modules, while a parallel frequency-domain branch (Time-Frequency Adapter, TFA) utilizes Extended DFT to align spectral dimensions and capture global periodic dependencies.

III. Methodology

A. Overview

Time series forecasting inherently requires balancing two distinct types of information: local temporal variations and global periodic structures. To effectively tackle these intricate and often conflicting data characteristics, we propose DDM-Mixer (as illustrated in Fig. 1). This architecture is designed as a dual-domain framework that avoids the common pitfall of relying entirely on a single representation space. It integrates a Multi-Scale Decomposition (MSD) module dedicated to rigorous time-domain structural refinement, alongside a parallel Time-Frequency Adapter (TFA) module that explicitly models and captures long-range spectral dependencies. Recognizing that different sequences rely on these domains to varying degrees, the outputs from these two distinct branches are not simply aggregated; instead, they are dynamically fused via an adaptive harmonic gating mechanism driven by information theory, ensuring optimal feature utilization across diverse datasets.

B. Multi-Scale Decomposition (MSD)

Real-world time series exhibit complex, multi-granular patterns where short-term anomalies can obscure long-term trends. To capture these diverse dynamics, we project the input sequence $\mathbf{X}$ into $M+1$ temporal scales, denoted as $\mathbf{X}^{(m)}$ ($m \in \{0, \ldots, M\}$). This hierarchical projection equips the model with varying receptive fields. Each specific scale $\mathbf{X}^{(m)}$ is then processed through a parallel, dual-branch architecture designed to disentangle different structural aspects:

1) Trend-Seasonal Branch (MSD-T): Following [6], the MSD-T branch utilizes moving average blocks to cleanly separate $\mathbf{X}^{(m)}$ into its trend and seasonal components. This iteratively smooths out high-frequency noise to isolate the underlying low-frequency trajectory. Dedicated linear mixing layers then refine these components, structurally preserving the core trend essential for long-term prediction stability.

2) Frequency-Oriented Branch (MSD-F): While moving averages excel at trend extraction, they can inadvertently oversmooth the signal, filtering out crucial localized structural changes vital for precise forecasting. To complement the smoothed MSD-T features, the MSD-F branch applies a direct linear projection to the raw input at scale m. This acts as an unconstrained, high-fidelity feature extractor, actively retaining deterministic high-frequency variations that strict decomposition typically discards.

Outputs from these complementary branches are dynamically fused at each individual scale m using a learnable gating mechanism:

$$\hat{\mathbf{Y}}^{(m)} = \alpha_m \cdot \text{MSD-T}(\mathbf{X}^{(m)}) + \beta_m \cdot \text{MSD-F}(\mathbf{X}^{(m)}) \tag{1}$$

where α_m and β_m are learnable parameters that adaptively balance the smoothed, stable trend with raw structural features. Finally, the ultimate multi-scale time-domain representation is aggregated across all levels via $\hat{\mathbf{Y}}_{MSD} = \sum_{m=0}^{M} w_m \cdot \hat{\mathbf{Y}}^{(m)}$, yielding a robust feature map that achieves both microscopic noise isolation and macroscopic trend sensitivity.

C. Time-Frequency Adapter (TFA)

Directly applying frequency-domain methods to dense forecasting causes a critical "Spectral Geometric Mismatch" when the historical look-back window L and the desired horizon H differ. Mapping an L-point spectrum directly to an H-point output forces the model across disjoint spectral grids, leading to severe feature aliasing.

To fundamentally resolve this, we implement an **Extended DFT** strategy. By zero-padding the encoded input sequence $\mathbf{X}_{enc}$ to a total length of $P = L + H$, we force a **High-Density Spectral Interpolation**:

$$\mathbf{S} = \mathcal{F}(\text{Concat}(\mathbf{X}_{enc}, \mathbf{0}_H)) \in \mathbb{C}^{P \times C} \tag{2}$$

This constructs an over-complete spectral basis that rigorously aligns historical data and future predictions, seamlessly bridging the dimensional gap [26].

Processing this spectrum requires preserving temporal shift information encoded in the spectral phase. Since real-valued layers would destructively entangle phase-amplitude relationships, we utilize a strictly Complex-valued Encoder:

$$\mathbf{H}_{freq} = \mathbf{W} \cdot \mathbf{S} + \mathbf{b}, \quad \mathbf{W} \in \mathbb{C}^{C \times C} \tag{3}$$

This network learns to modify amplitudes purely within the complex plane while preserving relative phase shifts. We stabilize deep representations using Complex Layer Normalization [25]. Finally, an Inverse FFT transforms the enhanced spectrum back into the time domain, yielding the globally aware prediction component $\hat{\mathbf{Y}}_{TFA}$.

D. Adaptive Harmonic Fusion

Empirical evidence shows that datasets exhibit profoundly varied spectral characteristics (as clearly delineated in Fig. 2).

For instance, energy grid traffic demonstrates stark periodicity, whereas weather patterns are far more chaotic. Consequently, applying a static fusion weight across all time series scenarios is inherently suboptimal and prone to overfitting non-periodic noise. To address this, we introduce an instance-wise **Spectral Entropy Gating Mechanism** designed to mathematically quantify the level of signal determinism versus stochasticity on the fly.

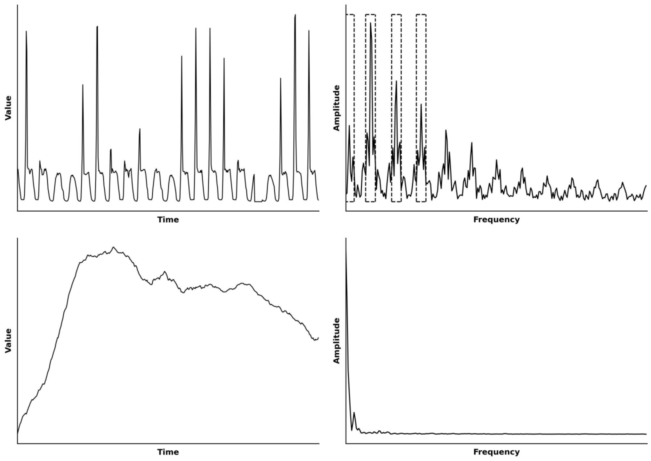


Fig. 2. Top: Periodic time series (Traffic dataset) showing obvious harmonic groups. Bottom: Non-periodic time series (Weather dataset) with no obvious harmonic structure.

First, we compute the normalized power density to estimate the probability distribution of frequencies: $p_k = \frac{|\mathbf{S}_k|^2}{\sum_i |\mathbf{S}_i|^2 + \epsilon}$. From this, we define the Spectral Entropy as $H = -\sum_k p_k \log p_k$, a metric deeply rooted in information theory. A spectrum dominated by clear harmonic groups (low entropy) indicates structured, predictable periodicity. Conversely, a flat spectrum (high entropy) strongly suggests disorder or noise.

The dynamic fusion weight $\omega \in [0, 1]$ is then computed using a robust nonlinear transition function:

$$\omega = \text{Sigmoid}\left(\alpha \cdot (H_{max} - H)\right) \tag{4}$$

where H_{max} is the theoretical maximum entropy (representing pure white noise) and α serves as a tunable scaling factor controlling the gate's sensitivity. Finally, the ultimate prediction dynamically fuses the dual branches based on this uncertainty metric:

$$\hat{\mathbf{Y}} = (1 - \omega) \cdot \hat{\mathbf{Y}}_{MSD} + \omega \cdot \hat{\mathbf{Y}}_{TFA} \tag{5}$$

This formulation ensures the model aggressively leverages global spectral context for periodic sequences while gracefully decaying to rely on the MSD backbone for trend-dominated, chaotic data, a behavior fully validated by the dynamic weight distributions plotted in Fig. 4.

IV. Experiments

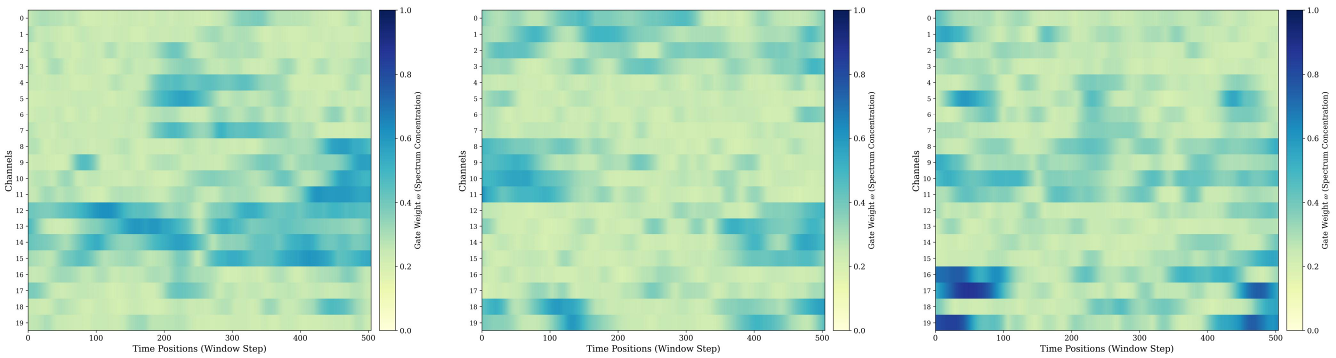
A. Experimental Setup

We evaluate DDM-Mixer on seven real-world datasets: ETT (ETTh1, ETTh2, ETTm1, ETTm2), Weather, Electricity, and Traffic [5], [6], [18]. Performance is measured using MSE and MAE [6]. Following standard protocols, the look-back window is fixed at $L = 96$, and prediction horizons are $H \in \{96, 192, 336, 720\}$ [14], [21]. We benchmark against our backbone TimeMixer [21], MLP-based DLinear [14], frequency-domain FreTS [24], Transformer-based PatchTST

TABLE I

MULTI-VARIABLE LONG-TERM SEQUENCE PREDICTION ON THE SEVEN COMMON DATASETS WITH INPUT LENGTH OF L = 96 AND PREDICTION LENGTH OF $H \in \{96, 192, 336, 720\}$. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD RED, AND THE SECOND-BEST RESULTS ARE HIGHLIGHTED IN BOLD BLUE.

Models	Metric	DDM-Mixer		TimeMixer		PatchTST		TimesNet		MICN		DLinear		FreTS	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.349**	**0.387**	**0.361**	**0.390**	0.370	0.399	0.384	0.402	0.396	0.427	0.375	0.399	0.395	0.407
	192	**0.401**	**0.416**	0.409	**0.414**	0.413	0.421	0.457	0.436	0.430	0.453	**0.405**	**0.416**	0.490	0.477
	336	**0.418**	**0.421**	0.430	**0.429**	**0.422**	0.436	0.491	0.469	0.433	0.458	0.439	0.443	0.510	0.480
	720	**0.436**	**0.464**	**0.445**	**0.460**	0.447	0.466	0.521	0.500	0.474	0.508	0.472	0.490	0.568	0.538
ETTh2	96	**0.257**	**0.335**	**0.271**	**0.330**	0.274	0.337	0.340	0.374	0.289	0.357	0.289	0.353	0.332	0.387
	192	**0.301**	**0.386**	**0.317**	0.402	0.339	**0.379**	0.402	0.414	0.409	0.438	0.383	0.418	0.451	0.457
	336	**0.296**	**0.372**	0.332	0.396	**0.329**	**0.380**	0.452	0.452	0.417	0.452	0.448	0.465	0.466	0.473
	720	**0.312**	**0.388**	**0.342**	**0.408**	0.379	0.422	0.462	0.468	0.426	0.473	0.605	0.551	0.485	0.471
ETTm1	96	0.294	0.349	**0.291**	**0.340**	**0.290**	**0.342**	0.338	0.375	0.314	0.360	0.299	0.343	0.337	0.374
	192	**0.326**	0.369	**0.327**	**0.365**	0.332	0.369	0.371	0.387	0.359	0.387	0.335	**0.365**	0.382	0.398
	336	**0.360**	**0.385**	**0.360**	**0.381**	**0.366**	0.392	0.410	0.411	0.398	0.413	0.369	0.386	0.420	0.423
	720	0.417	**0.419**	**0.415**	**0.417**	**0.416**	0.420	0.478	0.450	0.459	0.464	0.425	0.421	0.490	0.471
ETTm2	96	**0.156**	**0.224**	**0.164**	**0.254**	0.166	0.256	0.187	0.267	0.178	0.273	0.167	0.260	0.186	0.275
	192	**0.211**	**0.284**	**0.223**	**0.295**	**0.223**	0.296	0.249	0.309	0.245	0.316	0.224	0.303	0.259	0.323
	336	**0.260**	**0.307**	0.279	0.330	**0.274**	**0.329**	0.321	0.351	0.295	0.350	0.281	0.342	0.349	0.386
	720	**0.356**	**0.372**	**0.359**	**0.383**	0.362	0.385	0.497	0.403	0.389	0.409	0.397	0.421	0.559	0.511
Weather	96	**0.128**	**0.177**	**0.147**	**0.197**	0.149	0.198	0.172	0.220	0.161	0.226	0.152	0.237	0.171	0.227
	192	**0.167**	**0.216**	**0.189**	**0.239**	0.194	0.241	0.219	0.261	0.220	0.283	0.220	0.282	0.218	0.280
	336	**0.230**	**0.260**	**0.241**	**0.280**	0.245	0.282	0.280	0.306	0.275	0.328	0.265	0.319	0.265	0.317
	720	**0.296**	**0.316**	**0.310**	**0.330**	0.314	0.334	0.365	0.359	0.311	0.356	0.323	0.362	0.326	0.351
Electricity	96	**0.119**	**0.215**	**0.129**	0.224	**0.129**	**0.222**	0.168	0.272	0.159	0.267	0.153	0.237	0.171	0.260
	192	**0.127**	**0.228**	**0.140**	**0.220**	0.147	0.240	0.184	0.289	0.168	0.279	0.152	0.249	0.177	0.268
	336	**0.163**	**0.253**	**0.161**	**0.255**	**0.163**	0.259	0.198	0.300	0.196	0.308	0.169	0.267	0.190	0.284
	720	**0.188**	**0.289**	**0.194**	**0.287**	0.197	0.290	0.220	0.320	0.203	0.312	0.233	0.344	0.228	0.316
Traffic	96	**0.341**	**0.238**	**0.360**	**0.249**	**0.360**	**0.249**	0.593	0.321	0.508	0.301	0.410	0.282	0.391	0.867
	192	**0.361**	**0.243**	0.379	**0.256**	**0.375**	0.350	0.617	0.336	0.536	0.315	0.423	0.287	0.379	0.869
	336	**0.373**	**0.267**	**0.385**	0.270	0.392	**0.264**	0.629	0.336	0.525	0.310	0.436	0.296	0.420	0.881
	720	**0.409**	0.288	**0.430**	**0.281**	0.432	**0.286**	0.640	0.350	0.571	0.323	0.466	0.315	0.472	0.896



(a) Sequence a

(b) Sequence b

(c) Sequence c

Fig. 3. Visualization of the adaptive gate weight ω across multivariate channels on synthetic series with diverse patterns.

[10], and other advanced models like TimesNet [18] and MICN [19].

B. Experimental Results

Models were implemented in PyTorch and trained on a single Nvidia 3090 GPU using the Adam optimizer (batch size 128, learning rate 10^{-4}). Following TimeMixer [21], inputs were fixed at $L = 96$.

As shown in Table I, DDM-Mixer achieves competitive or state-of-the-art (SOTA) performance on 6 out of 7 datasets. Table II highlights improvements against TimeMixer and TimesNet. Notably, DDM-Mixer achieves substantial gains on datasets with prominent periodic patterns, such as Traffic

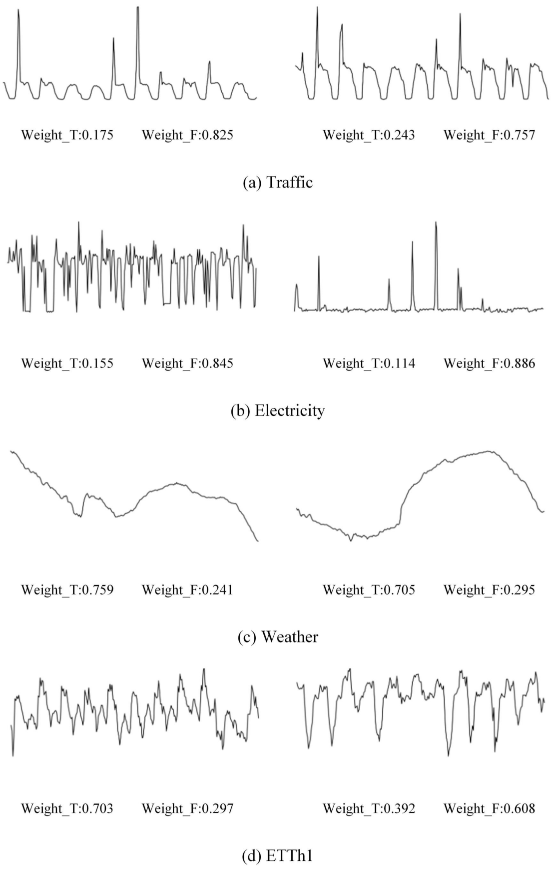


Fig. 4. Examples of weights allocation for distinct time series from 4 different datasets.

TABLE II
COMPARISON OF RESULT IMPROVEMENT OF DDM-MIXER RELATIVE TO TIMEMIXER AND TIMESNET

Datasets	TimeMixer		TimesNet	
	MSE	MAE	MSE	MAE
ETTh1	$\uparrow$ 2.52%	$\uparrow$ 0.32%	$\uparrow$ 13.43%	$\uparrow$ 6.81%
ETTh2	$\uparrow$ 7.61%	$\uparrow$ 3.15%	$\uparrow$ 29.59%	$\uparrow$ 13.88%
ETTm1	$\downarrow$ 0.43%	$\downarrow$ 1.56%	$\uparrow$ 13.40%	$\uparrow$ 7.76%
ETTm2	$\uparrow$ 6.05%	$\uparrow$ 8.32%	$\uparrow$ 23.21%	$\uparrow$ 12.98%
Weather	$\uparrow$ 7.44%	$\uparrow$ 7.36%	$\uparrow$ 20.75%	$\uparrow$ 15.44%
Electricity	$\uparrow$ 4.97%	$\uparrow$ 1.23%	$\uparrow$ 20.83%	$\uparrow$ 16.29%
Traffic	$\uparrow$ 4.50%	$\uparrow$ 2.41%	$\uparrow$ 40.13%	$\uparrow$ 24.35%
Average	**$\uparrow$ 4.67%**	**$\uparrow$ 3.03%**	**$\uparrow$ 23.05%**	**$\uparrow$ 13.93%**

Note: The red arrows ($\uparrow$) denote performance improvement, and blue arrows ($\downarrow$) denote performance decline.

and Electricity, validating our frequency branch. Even on datasets where TimeMixer excels (e.g., ETTm1), DDM-Mixer maintains comparable stability.

Furthermore, DDM-Mixer demonstrates superior robustness compared to PatchTST [10], particularly in long-term scenarios ($H = 720$) on ETTh1 and ETTm2. Unlike FreTS [24], which struggles with non-periodic data like Weather, our entropy-based dynamic gating intelligently prevents forcing

TABLE III
ABLATION STUDY, WEATHER

Horizon		w/o MSD & TFA		w/o MSD		w/o TFA		DDM-Mixer	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	0.147	0.197	0.132	0.181	0.141	0.189	**0.128**	**0.177**
	192	0.189	0.239	0.172	0.220	0.181	0.237	**0.167**	**0.216**
	336	0.241	0.280	0.235	0.265	0.239	0.275	**0.230**	**0.260**
	720	0.310	0.330	0.301	0.320	0.306	0.325	**0.296**	**0.316**

TABLE IV
COMPARISON OF DIFFERENT GATING MECHANISMS ON WEATHER DATASET (MSE)

Horizon	Fixed	Learnable	Energy	**Entropy (Ours)**
96	0.138	0.135	0.132	**0.128**
192	0.178	0.174	0.171	**0.167**
336	0.242	0.238	0.235	**0.230**
720	0.308	0.304	0.301	**0.296**
Avg	0.217	0.213	0.210	**0.205**

TABLE V
ABLATION STUDY, TRAFFIC

Horizon	Standard DFT		TFA		Improvement%	
	MSE	MAE	MSE	MAE	MSE	MAE
96	0.353	0.249	0.341	0.238	3.4$\uparrow$	4.4$\uparrow$
192	0.371	0.254	0.361	0.243	2.7$\uparrow$	4.3$\uparrow$
336	0.384	0.279	0.373	0.267	2.9$\uparrow$	4.3$\uparrow$
720	0.422	0.296	0.409	0.288	3.1$\uparrow$	2.7$\uparrow$

Note: The red arrows ($\uparrow$) denote performance improvement (lower error is better).

frequency analysis on data lacking harmonic structure, ensuring consistent accuracy across diverse datasets.

C. Ablation Study

1) Effectiveness of Components: Table III presents ablations on the Weather dataset. Removing either the MSD or TFA module degrades performance compared to the full DDM-Mixer, emphasizing their complementary roles, particularly in long horizons where global context is crucial.

2) Analysis of Gating Mechanism: We compared our **Spectral Entropy Gating** against three variants: (1) Fixed Weight ($\omega = 0.5$); (2) Learnable Scalar; and (3) Energy Gate based on harmonic ratio [26]. Table IV demonstrates that our entropy-based approach outperforms all alternatives by effectively distinguishing periodicity from noise.

3) Parameter Sensitivity and Spectral Alignment: The scaling factor α in Eq. (6) remains robust within a reasonable range. Furthermore, Table V shows that our Extended DFT (TFA) consistently outperforms Standard DFT across all horizons on Traffic, validating that our frequency alignment effectively resolves the spectral mismatch issue.

Finally, Fig. 3 visualizes the learned adaptive weights. For **Sequence (a)**, distinct high-weight bands appear during a transient periodic event, indicating dynamic frequency-domain reliance. In contrast, **Sequence (c)** shows lower weights, prompting reliance on the time-domain backbone due to weaker periodicity. This instance-wise adaptability confirms our gating mechanism prevents the "one-size-fits-all" limitation.

V. Conclusion

In this paper, we proposed DDM-Mixer, a framework synergizing adaptive time-domain decomposition with frequency-domain global modeling. By utilizing Extended DFT for spectral interpolation and dimensional alignment, we effectively resolved the frequency mismatch issue. Extensive experiments validate the robustness of this dual-domain approach.

Acknowledgment

This work is supported by the Joint Innovation Laboratory for Future Observation, Zhejiang University, focusing on the research of cloud-native observability technology based on GuanceCloud(H22010).

References

[1] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[2] C. C. Holt, "Forecasting seasonals and trends by exponentially weighted moving averages," *International journal of forecasting*, vol. 20, no. 1, pp. 5–10, 2004.

[3] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "Deepar: Probabilistic forecasting with autoregressive recurrent networks," *International journal of forecasting*, vol. 36, no. 3, pp. 1181–1191, 2020.

[4] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," *arXiv preprint arXiv:2001.04451*, 2020.

[5] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 11 106–11 115.

[6] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," *Advances in neural information processing systems*, vol. 34, pp. 22 419–22 430, 2021.

[7] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting," in *International conference on machine learning*. PMLR, 2022, pp. 27 268–27 286.

[8] S. Liu, H. Yu, C. Liao, J. Li, W. Lin, A. X. Liu, and S. Dustdar, "Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting," in *# PLACEHOLDER_PARENT_METADATA_VALUE#*, 2022.

[9] Y. Zhang and J. Yan, "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting," in *The eleventh international conference on learning representations*, 2023.

[10] Y. Nie, "A time series is worth 64words: Long-term forecasting with transformers," *arXiv preprint arXiv:2211.14730*, 2022.

[11] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "itransformer: Inverted transformers are effective for time series forecasting," *arXiv preprint arXiv:2310.06625*, 2023.

[12] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "N-beats: Neural basis expansion analysis for interpretable time series forecasting," *arXiv preprint arXiv:1905.10437*, 2019.

[13] C. Challu, K. G. Olivares, B. N. Oreshkin, F. G. Ramirez, M. M. Canseco, and A. Dubrawski, "Nhits: Neural hierarchical interpolation for time series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 6, 2023, pp. 6989–6997.

[14] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 9, 2023, pp. 11 121–11 128.

[15] A. Das, W. Kong, A. Leach, S. Mathur, R. Sen, and R. Yu, "Long-term forecasting with tide: Time-series dense encoder," *arXiv preprint arXiv:2304.08424*, 2023.

[16] Z. Li, Z. Rao, L. Pan, and Z. Xu, "Mts-mixers: Multivariate time series forecasting via factorized temporal and channel mixing," *arXiv preprint arXiv:2302.04501*, 2023.

[17] M. Liu, A. Zeng, M. Chen, Z. Xu, Q. Lai, L. Ma, and Q. Xu, "Scinet: Time series modeling and forecasting with sample convolution and interaction," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5816–5828, 2022.

[18] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "Timesnet: Temporal 2d-variation modeling for general time series analysis," *arXiv preprint arXiv:2210.02186*, 2022.

[19] H. Wang, J. Peng, F. Huang, J. Wang, J. Chen, and Y. Xiao, "Micn: Multi-scale local and global context modeling for long-term series forecasting," in *The eleventh international conference on learning representations*, 2023.

[20] Y. Hu, P. Liu, P. Zhu, D. Cheng, and T. Dai, "Adaptive multi-scale decomposition framework for time series forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 16, 2025, pp. 17 359–17 367.

[21] S. Wang, H. Wu, X. Shi, T. Hu, H. Luo, L. Ma, J. Y. Zhang, and J. Zhou, "Timemixer: Decomposable multiscale mixing for time series forecasting," *arXiv preprint arXiv:2405.14616*, 2024.

[22] D. Cao, Y. Wang, J. Duan, C. Zhang, X. Zhu, C. Huang, Y. Tong, B. Xu, J. Bai, J. Tong *et al.*, "Spectral temporal graph neural network for multivariate time-series forecasting," *Advances in neural information processing systems*, vol. 33, pp. 17 766–17 778, 2020.

[23] T. Zhou, Z. Ma, Q. Wen, L. Sun, T. Yao, W. Yin, R. Jin *et al.*, "Film: Frequency improved legendre memory model for long-term time series forecasting," *Advances in neural information processing systems*, vol. 35, pp. 12 677–12 690, 2022.

[24] K. Yi, Q. Zhang, W. Fan, S. Wang, P. Wang, H. He, N. An, D. Lian, L. Cao, and Z. Niu, "Frequency-domain mlps are more effective learners in time series forecasting," *Advances in Neural Information Processing Systems*, vol. 36, pp. 76 656–76 679, 2023.

[25] Z. Xu, A. Zeng, and Q. Xu, "Fits: Modeling time series with $10k$ parameters," *arXiv preprint arXiv:2307.03756*, 2023.

[26] H. Ye, J. Chen, S. Gong, F. Jiang, T. Zhang, J. Chen, and X. Gao, "Atfnet: Adaptive time-frequency ensembled network for long-term time series forecasting," *arXiv preprint arXiv:2404.05192*, 2024.

[27] A. Shabani, A. Abdi, L. Meng, and T. Sylvain, "Scaleformer: Iterative multi-scale refining transformers for time series forecasting," *arXiv preprint arXiv:2206.04038*, 2022.

[28] Z. Ni, H. Yu, S. Liu, J. Li, and W. Lin, "Basisformer: Attention-based time series forecasting with learnable and interpretable basis," *Advances in Neural Information Processing Systems*, vol. 36, pp. 71 222–71 241, 2023.

[29] Y. Liu, C. Li, J. Wang, and M. Long, "Koopa: Learning non-stationary time series dynamics with koopman predictors," *Advances in neural information processing systems*, vol. 36, pp. 12 271–12 290, 2023.

[30] K. Yi, Q. Zhang, W. Fan, H. He, L. Hu, P. Wang, N. An, L. Cao, and Z. Niu, "Fouriergnn: Rethinking multivariate time series forecasting from a pure graph perspective," *Advances in neural information processing systems*, vol. 36, pp. 69 638–69 660, 2023.

[31] S. Lin, W. Lin, W. Wu, H. Chen, and J. Yang, "Sparsetsf: Modeling long-term time series forecasting with 1k parameters," *arXiv preprint arXiv:2405.00946*, 2024.

[32] P. Chen, Y. Zhang, Y. Cheng, Y. Shu, Y. Wang, Q. Wen, B. Yang, and C. Guo, "Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting," *arXiv preprint arXiv:2402.05956*, 2024.

[33] D. Luo and X. Wang, "Moderntcn: A modern pure convolution structure for general time series analysis," in *The twelfth international conference on learning representations*, 2024, pp. 1–43.