

BEAT: Balanced Frequency Adaptive Tuning for Long-Term Time-Series Forecasting

Zhixuan Li¹, Naipeng Chen¹, Seonghwa Choi², Sanghoon Lee², Weisi Lin^{1*}

¹College of Computing and Data Science, Nanyang Technological University, Singapore

²Department of Electrical and Electronic Engineering, Yonsei University, Korea

Abstract

Time-series forecasting is crucial for numerous real-world applications including weather prediction and financial market modeling. While temporal-domain methods remain prevalent, frequency-domain approaches can effectively capture multi-scale periodic patterns, reduce sequence dependencies, and naturally denoise signals. However, existing approaches typically train model components for all frequencies under a unified training objective, often leading to mismatched learning speeds: high-frequency components converge faster and risk overfitting, while low-frequency components underfit due to insufficient training time. To deal with this challenge, we propose BEAT (Balanced frEquency Adaptive Tuning), a novel framework that dynamically monitors the training status for each frequency and adaptively adjusts their gradient updates. By recognizing convergence, overfitting, or underfitting for each frequency, BEAT dynamically reallocates learning priorities, moderating gradients for rapid learners and increasing those for slower ones, alleviating the tension between competing objectives across frequencies and synchronizing the overall learning process. Extensive experiments on seven real-world datasets demonstrate that BEAT consistently outperforms state-of-the-art approaches.

Introduction

Long-term time-series forecasting plays a pivotal role in various real-world applications, such as weather prediction (Zhang et al. 2022), electric load forecasting (Zhou et al. 2021; Gasparin, Lukovic, and Alippi 2022), traffic flow analysis (Jin et al. 2021), and financial market modeling (Lai et al. 2018; Tang et al. 2022). These tasks demand accurate prediction over extended temporal horizons. In recent years, deep learning methods for time-series analysis have achieved significant advancements, characterized by the development of various architectures, including CNN-based (Liu et al. 2022; Wang et al. 2023; Luo and Wang 2024), MLP-based (Oreshkin et al. 2019; Zeng et al. 2023; Wang et al. 2024a), Transformer-based (Wu et al. 2021; Liu et al. 2021, 2024), and LLM-based (Jin et al. 2024; Jia et al. 2024) approaches. Compared with temporal-domain approaches, frequency-domain methods (Wu et al. 2021, 2023; Zhou et al. 2022b) transform time-series data into the frequency domain, enabling the effective capture of large-scale periodic patterns,

reducing sequence dependency lengths, and enhancing robustness through inherent denoising effects.

Despite these advantages, frequency-domain analysis typically decomposes time-series data into multiple sub-series, such as high-frequency, low-frequency, and potentially additional band-pass segments, each representing a distinct temporal scale. High-frequency sub-series often capture short-term fluctuations or noise, while low-frequency sub-series reflect long-term trends or stable periodic patterns. Existing frequency-domain approaches, such as WPMixer (Murad, Aktukmak, and Yilmaz 2025), simply train all frequency sub-series simultaneously under a unified loss function. However, the following challenges are observed: (1) the high-frequency component of the model tends to converge more rapidly, while the low-frequency component of the model requires more time to be adequately learned; (2) early stopping may result in underfitting of the low-frequency components, even if the high-frequency components are well-learned, whereas prolonged training risks overfitting the high-frequency components. This asynchronous learning pace across different frequencies renders a unified loss function insufficient to optimize all sub-series simultaneously, ultimately leading to suboptimal overall performance. Therefore a critical challenge lies in dynamically evaluating the learning progress of each frequency-specific sub-series and introducing timely adjustments to ensure balanced training for both high-frequency and low-frequency components.

To deal with this issue, we propose BEAT (Balanced frEquency Adaptive Tuning), a novel framework that enables real-time monitoring and dynamic adjustment for each frequency sub-series. Specifically, our approach begins by decomposing the ground-truth target series into frequency-specific ones and evaluates whether the model component for a given frequency is converging, overfitting, or underfitting. Consequently, BEAT dynamically modulates the network gradients of the model component for each frequency. For sub-series that converge more quickly and risk overfitting, the framework reduces the back-propagated gradient to decelerate learning. Conversely, for the ones that lag behind, it amplifies the gradients adaptively to accelerate convergence and mitigate underfitting. By overcoming the limitations of a uniform training objective applied across all frequencies, the proposed framework effectively redistributes optimization priorities and synchronizes learning progress across the

*Corresponding author.

frequency spectrum.

Our contributions can be summarized as threefold:

- We perform an in-depth analysis of the optimization conflicts caused by mismatched learning speeds between model components for different frequencies in frequency-based long-range forecasting. We demonstrate how these conflicts negatively influence overall predictive performance with illustrative evidence.
- We propose BEAT, a novel training framework that incorporates real-time monitoring to continuously track learning progress and an adaptive adjustment mechanism to effectively regulate the learning pace for different frequencies, achieving a balanced convergence between high- and low-frequency model components.
- We evaluate BEAT on multiple challenging datasets and experimental results show that BEAT consistently achieves the best long-range forecasting performance by improving the balance between high- and low-frequency sub-series.

Related Work

Temporal-domain Time-Series Forecasting

In recent years, deep learning has emerged as the leading method for time series modeling. This section provides a technical overview of key deep time series models, categorized into three types according to their backbone architecture: CNN-based, MLP-based, and Transformer-based.

CNN-based methods. Convolutional neural networks (He et al. 2016; Gu et al. 2018) have gained prominence for their effectiveness in capturing local features and recognizing patterns for many tasks (Jiang et al. 2024a; Li et al. 2022). SCINet (Liu et al. 2022) employs standard convolutions with a hierarchical downsample-convolve-interact framework to capture dynamic temporal dependencies across varying temporal resolutions of time series data. MICN (Wang et al. 2023) integrates different convolutional kernels to model temporal correlations from both local and global perspectives. ModernTCN (Luo and Wang 2024) enhances traditional TCN (Bai, Kolter, and Koltun 2018) by employing DW-Conv and ConvFFN to capture cross-time and cross-variable dependencies independently.

MLP-based methods. Drawing inspiration that outputs depend linearly on their historical values, Multi-Layer Perceptrons (MLP) have emerged as a favored structure for time series modeling. These approaches handle multivariate data as independent univariate sequences and have shown promising results (Zeng et al. 2023). N-BEATS (Oreshkin et al. 2019), as a purely MLP-driven deep learning model for time series, utilizes deep stacks of linear layers with dual residual branches. TimeMixer (Wang et al. 2024a) proposing an MLP-based multiscale mixing framework posits that time series exhibit varying patterns across different sampling scales.

Transformer-based methods. Motivated by the remarkable success of the Transformer architecture (Vaswani 2017) in sequence-to-sequence tasks (Li et al. 2023; Jiang et al. 2024b), prevailing it to the domain of time series modeling (Zhou et al. 2022b, 2021; Wu et al. 2021). Autoformer (Wu

et al. 2021) pioneers an Auto-Correlation Mechanism. A time delay module is introduced to aggregate similar sub-series from inherent periods instead of focusing on scattered points. Pyraformer (Liu et al. 2021) constructs a multi-resolution C-ary tree and proposes a Pyramidal Attention Mechanism, enabling it to capture both short- and long-temporal dependencies in linear time and space. iTransformer (Liu et al. 2024) further expands the receptive field by tokenizing the entire time series to capture inter-series dependencies.

Frequency-domain Time-Series Forecasting

Due to the presence of both high- and low-frequency components, purely time-domain methods may struggle to manage all frequency bands effectively in long-range scenarios compared with frequency-domain approaches. The ability to decompose signals into sub-series across different frequency ranges makes them suitable for long-horizon forecasting.

As a common practice, TimesNet (Wu et al. 2023) applies FFT (Brigham 1988) to isolate dominant frequencies—those with the highest amplitudes—and rearranges 1D time series into 2D representations based on identified periods. Autoformer (Wu et al. 2021) introduces an Auto-Correlation mechanism alongside efficient FFT to treat data as a real discrete-time signal and capture series-wise correlations. FiLM (Zhou et al. 2022a) employs Frequency Enhanced Layers (FEL), merging Fourier analysis with low-rank approximation to retain key low-frequency components and top eigenspaces, thereby reducing noise and expediting training. FITS (Xu, Zeng, and Xu 2023) uses a low-pass filter (LPF) to remove high-frequency elements beyond a designated cutoff, shrinking model size while maintaining essential information. From the opposite idea, FEDformer (Zhou et al. 2022b) argues that focusing solely on low-frequency components can miss critical variations in time series data. Consequently, to capture a global perspective, the FEDformer randomly selects a fixed number of Fourier components, encompassing both high- and low-frequency ranges.

EV-FGN (Yi et al. 2022) uses a two-dimensional DFT on the spatial-temporal plane of embeddings, combined with graph convolutions to jointly model spatial-temporal dependencies in the frequency domain. FreTS (Yi et al. 2024) leverages the DFT to transform data into frequency-domain spectra and introduces frequency-domain MLPs designed for complex inputs by separately modeling real and imaginary parts. FCVAE (Wang et al. 2024c) incorporates both global and local frequency components into the condition of a Conditional Variational Autoencoder (CVAE). TSLANet (Eldele et al. 2024) proposes a lightweight Adaptive Spectral Block (ASB) that replaces self-attention using Fourier-based global and local filters.

The recent WP-Mixer (Murad, Aktukmak, and Yilmaz 2025) employs multi-level wavelet decomposition, combining patching and patch mixer for local and global information respectively, successfully handling complex data behaviors such as sudden spikes and drops. However, WPMixer faces key issues: it simultaneously trains all frequency sub-series under a unified loss function, with high-frequency components learning quickly while low-frequency components lag, risking underfitting and overfitting due to asynchronous learn-

ing. In contrast, our model BEAT handles this by decomposing the target into frequency-specific sub-series and evaluating their convergence statuses. It adjusts gradients in real time to achieve synchronization of the learning progress across frequencies, overcomes uniform loss limitations, and improves model efficiency.

Methodology

Task Formulation

In time series forecasting with one or multiple observed variates, the goal is to predict the future K time steps $Y = \{y_1, y_2, \dots, y_K\} \in \mathbb{R}^{K \times N}$ with N variates, based on the historical observations $X = \{x_{1-T}, x_{1-T+1}, \dots, x_0\} \in \mathbb{R}^{T \times N}$ with T time steps.

Overall Architecture

As shown in Figure 1, the overall architecture of the proposed BEAT is presented. For the input historical time series sequence X , the Reversible Instance Normalization (RevIN) (Kim et al. 2021) is employed for handling the varied mean and variation in the time-series data. Then the normalized data X_{norm} is transposed and decomposed by the multi-level wavelet decomposition (Mallat 1989; Daubechies 1992; Wang et al. 2018) as approximation X_A and detail coefficients $X_{D_i}, i \in \{1, 2, \dots, f\}$ series corresponding to m frequency levels from low to high.

Then $f + 1$ frequency-specific networks are utilized for capturing the temporal dependencies across the decomposed sub-series within each coefficient X_{D_i} and the approximation X_A .

To demonstrate our proposed BEAT approach, specifically tailored for frequency-domain methods, we utilize WP-Mixer (Murad, Aktukmak, and Yilmaz 2025) as an instantiation. The approximation coefficient series at level f is preserved, ensuring a concise and efficient representation.

Next $f + 1$ Frequency-Specific Monitors (FSM) (as shown in Figure 1 (c)) are designed and deployed for each frequency-specific network to track their optimization statuses during training. Besides, the Dynamical Gradient Balancer (DGB) (presented in Figure 1 (d)) is presented and equipped for each frequency-specific network to modulate their back-propagated gradients during the backward stage for balancing their learning processes to avoid overfitting or underfitting. If one network for a frequency converges over-quick or over-slow, FSM can recognize the inappropriate optimization status and transfer it to DGB for timely adjustment.

The Wavelet Reconstruction module takes all the predicted approximation \hat{Y}_A and detail wavelet coefficient series \hat{Y}_{D_i} from all frequency-specific networks for transforming the predicted future time series into the temporal domain as \hat{Y}_{norm} . Finally, the transposition is applied and the RevIN is used for de-normalization, generating the final output \hat{Y} .

Frequency-Specific Monitor

The Frequency-Specific Monitor (FSM) is designed to track the training status of the network for each frequency, supporting the dynamical adjustment of the learning pace for

these networks inspired by (Peng et al. 2022). To monitor the learning status for each frequency, it is necessary to compare their prediction with the corresponding targets. To this end, we take advantage of the ground-truth sequence Y and utilize the wavelet decomposition to transform Y in the temporal domain as the approximation coefficient series Y_A and detail coefficient series Y_{D_i} where $i \in \{1, 2, \dots, f\}$ and f is the wavelet decomposition level. The wavelet decomposition process can be formulated as:

$$Y_A, Y_{D_f}, Y_{D_{f-1}}, \dots, Y_{D_1} = DWT(Y, f, \psi), \quad (1)$$

where the DWT represents the Discrete Wavelet Transform, and ψ is the wavelet type. Here only the approximation coefficient series at level f is reserved for efficient representation. The wavelet families Coiflets, Daubechies, Biorthogonal, and Symlets are evaluated as potential wavelet types.

Then we calculate the discrepancy δ between the predicted and decomposed ground-truth targets for all frequencies, formulated as below:

$$\delta_A = MSE(Y_A, \hat{Y}_A), \quad (2)$$

$$\delta_{D_i} = MSE(Y_{D_i}, \hat{Y}_{D_i}), \quad i \in \{1, 2, \dots, f\}, \quad (3)$$

where MSE is the Mean Squared Error to measure the discrepancy. Then the mean value of the discrepancies for all detail coefficient series is calculated by:

$$\mu = \frac{\sum_{i=1}^f \delta_{D_i}}{f}. \quad (4)$$

And the discrepancy ratios r^v , where $v \in \{1, 2, \dots, f+1\}$, for all frequencies can be calculated with:

$$r_A = \frac{\delta_A}{\mu}, \quad r_{D_i} = \frac{\delta_{D_i}}{\mu}, \quad i \in \{1, 2, \dots, f\}. \quad (5)$$

The discrepancy ratio r^v quantitatively measures the discrepancy between the predicted and ground-truth coefficient series for each frequency, representing the relative learning speed with the help of normalization using the mean of the discrepancies regarding all detail coefficient series. The mean discrepancy is calculated only for the detail coefficients because they capture high-frequency information, which is more sensitive to prediction errors and crucial for analyzing signal details. The approximation coefficient, representing low-frequency information, is treated separately to avoid dominating the mean value.

Dynamical Gradient Balancer

Since the learning processes of different frequency-specific networks are not synchronized, it is essential to adjust the learning speed of these networks dynamically to balance their learning processes in time. To this end, the Dynamical Gradient Balancer (DGB) is presented to modulate the back-propagated gradients based on the calculated discrepancy ratios of all frequencies.

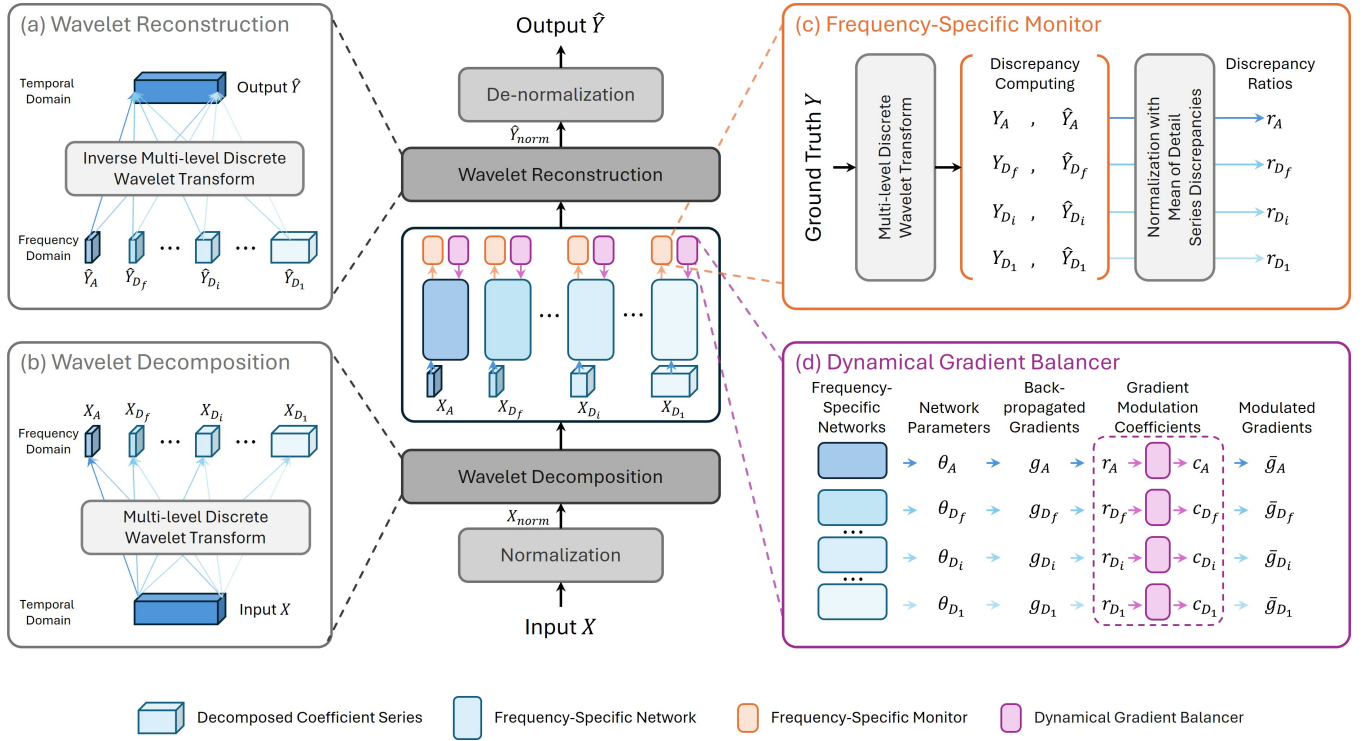


Figure 1: Overall architecture of the proposed BEAT. (a) and (b) are the illustrations of the wavelet decomposition and reconstruction, utilizing the multi-level discrete wavelet transform and the inverse operation. (c) demonstrates the process of monitoring the discrepancy ratio for each frequency. (d) presents the process of calculating the gradient modulation coefficients, which are applied to adjust the gradients being back-propagated. Best viewed in color and zoom in.

Specifically, for a frequency-specific network ϕ^v , where $v \in \{1, 2, \dots, f+1\}$, its network parameters θ^v are updated as follows when applying the Gradient Decent approach:

$$\theta_{u+1}^v = \theta_u^v - \alpha \nabla_{\theta^v} L(\theta_u^v), \quad (6)$$

where α is the learning rate, u is the optimization step, and L is the loss of the v -th network. The ∇_{θ^v} is the back-propagated gradient, which can be denoted as g^v .

Next, we calculate the gradient modulation coefficient c^v for each g^v according to the discrepancy ratio r^v which reflects the learning status quantitatively of each frequency-specific network. To be specific, the gradient modulation coefficient c^v is calculated following:

$$c^v = \begin{cases} \frac{1}{1+e^{-0.5(r^v-1)}} + 0.5, & \text{if } r^v > 1, \\ \frac{1}{r^v}, & \text{if } r^v \leq 1. \end{cases} \quad (7)$$

Then the back-propagated gradient of each frequency-specific network's parameters can be updated as:

$$\theta_{u+1}^v = \theta_u^v - \alpha \bar{g}^v L(\theta_u^v), \quad (8)$$

where $\bar{g}^v = c^v g^v = c^v \nabla_{\theta^v}$, representing the modulated gradients, which are appropriate for balancing the learning process among different frequency-specific networks and achieving harmonious convergence. For a network with a

Algorithm 1: The proposed BEAT framework

Input: Historical time series data X , wavelet decomposition level f , wavelet type ψ , future time steps K , initialized frequency-specific network's parameters θ^v , $v \in \{1, 2, \dots, f+1\}$, learning rate α .

for $u = 1$ **to** U **do**

 Sample a mini-batch B_u from X ;

 Feed-forward B_u to the model;

 Calculate the loss \mathcal{L}_u of the model;

 Use back-propagation to calculate g^v for θ^v ;

 Calculate discrepancy ratio r^v with Eqn. 1, 2, 3, 4, 5;

 Calculate modulation coefficient c^v with Eqn. 7;

 Update the gradient of θ^v with Eqn. 8.

end for

faster learning process, where $r^v > 0$, its back-propagated gradients will be decreased to slow down its learning speed to avoid overfitting and being dominant over the learning. On the contrary, regarding those networks with slower learning processes, where $r^v \leq 1$, their back-propagated gradients will be increased to accelerate their learning speed to avoid underfitting.

Dataset	# VAR	Size	FRQ
Weather	21	(36792, 5271, 10540)	10 min
ETTh1/h2	7	(8545, 2881, 2881)	1 hour
ETTm1/m2	7	(34465, 11521, 11521)	15 min
Traffic	862	(12185, 1757, 3509)	1 hour
ECL	321	(18317, 2633, 5261)	1 hour

¹ # VAR denotes the number of variates.

² FRQ means the frequency of data sampling.

Table 1: Descriptions of datasets used for evaluation.

Training and Inference

During training, the proposed BEAT framework is taking effect to balance the learning progress of different frequency-specific networks to avoid overfitting or underfitting. The overall pipeline is shown in Algorithm 1. It is worth noticing that after calculating the loss \mathcal{L}_u of the entire model, the gradient g^v for the parameters θ^v of each frequency-specific networks are only computed but not updated. The gradient of θ^v is updated using Equation 8 for applying the gradient modulation coefficient c^v on the gradient g^v . The loss function used for training is SmoothL1Loss with the default threshold value set as 1. The proposed BEAT framework is only deployed during training and not used during inference, keeping the model efficient.

Experiments

Experimental Settings

Datasets. We benchmark our method over seven challenging datasets in long-term series forecasting. (1) The weather dataset is collected every 10 minutes throughout 2020, containing 21 meteorological indicators such as air temperature and humidity. (2) ETT dataset (Zhou et al. 2021) contains data collected from electricity transformers, including load and oil temperature measurements recorded every 15 minutes from July 2016 to July 2018. There are four sub-datasets in ETT, including ETTh1, ETTh2, ETTm1, and ETTm2. (3) The traffic dataset comprises hourly data from the California Department of Transportation, detailing road occupancy rates measured by various sensors on San Francisco Bay Area freeways. (4) ECL dataset (Li et al. 2019) records hourly electricity consumption data from 321 clients. Table 1 provides data statistics of these datasets. Size denotes the number of time points in the (Train, Validation, Test) sets, respectively.

Baselines. There are nine state-of-the-art baseline methods for long-term time series forecasting are compared, including WPMixer (Murad, Aktukmak, and Yilmaz 2025), TimeMixer (Wang et al. 2024a), iTransformer (Liu et al. 2024), TSMixer (Ekambaram et al. 2023), PatchTST (Nie et al. 2023), Crossformer (Zhang and Yan 2023), TiDE (Das et al. 2023), and DLinear (Zeng et al. 2023).

Implementation Details. We develop BEAT using Pytorch (Paszke et al. 2019) based on the TSLib (Wang et al. 2024b) codebase. Following the previous method (Ekambaram et al. 2023), we set the historical time step T as 96

Index	FSM	DGB	Weather		ETTh1	
			MSE	MAE	MSE	MAE
(1)			0.243	0.269	0.422	0.423
(2)	✓	✓	0.239	0.263	0.415	0.419

Table 2: Ablation study of designed components in BEAT. FSM and DGB represent the proposed Frequency-Specific Monitor and Dynamical Gradient Balancer. The average performance over four prediction lengths including {96, 192, 336, 720} are reported for each dataset’s test set.

and follow the unified setting for a fair comparison for all experiments. The prediction length K varies within the range of {96, 192, 336, 720}. For each dataset, all experiments are conducted on the same training set and evaluated on the same test set. A single NVIDIA RTX 4090 with 24GB memory is used for all experiments.

Metrics. The Mean Squared Error (MSE) and Mean Absolute Error (MAE) are employed for evaluation. MSE and MAE are defined as:

$$\text{MSE} = \frac{1}{K} \sum_{k=1}^K (y_k - \hat{y}_k)^2, \quad (9)$$

$$\text{MAE} = \frac{1}{K} \sum_{k=1}^K |y_k - \hat{y}_k|, \quad (10)$$

where for the k -th sample in a time series, y_k is the ground-truth data and the \hat{y}_k is the predicted one. The lower MSE and MAE indicate better performance.

Long-Term Time Series Forecasting Results

Table 3 shows the performance of BEAT and compared methods on the test set of seven challenging datasets sampled from the real world. All experiments are under the same setting of historical time step length 96 for fair comparison. As shown in the table, the proposed method BEAT achieves the highest number of first-place rankings across all prediction lengths and datasets.

Specifically, our designed frequency-specific learning process monitoring and dynamic gradient adjusting for frequency-decomposition-based approaches outperforms the performance of the existing state-of-the-art frequency-based approach WPMixer, which treats the different decomposed frequencies equally and suffers from learning speed conflict, prone to be over-fitting and under-fitting for frequency-specific networks with different learning processes. Compared with the state-of-the-art Linear-based approach TimeMixer which decomposes time series as different scales and mixes different seasonal and trend series, BEAT archives better performance. Moreover, compared to the Transformer-based model iTransformer, BEAT achieves more first-ranking counts. It is worth noticing that BEAT achieves well performance against counterpart methods on challenging datasets like Weather and ETT that have no clear periodic patterns,

Methods		BEAT	WPMixer	TimeMixer	iTransformer	TSMixer	PatchTST	Crossformer	TiDE	Dlinear
		Ours	2025	2024	2024	2023	2023	2023	2023	2023
Metric		MSE MAE	MSE MAE	MSE MAE	MSE MAE	MSE MAE	MSE MAE	MSE MAE	MSE MAE	MSE MAE
Weather	96	0.160 0.198	0.162 0.204	0.163 0.209	0.174 0.214	0.175 0.247	0.186 0.227	0.195 0.271	0.202 0.261	0.195 0.252
	192	0.207 0.241	0.209 0.246	0.208 0.250	0.221 0.254	0.224 0.294	0.234 0.265	0.209 0.277	0.242 0.298	0.237 0.295
	336	<u>0.258</u> 0.279	0.263 0.287	0.251 0.287	0.278 0.296	0.262 0.326	0.284 0.301	0.273 0.332	0.287 0.335	0.282 0.331
	720	0.329 0.332	0.339 0.338	0.339 0.341	0.358 0.347	0.349 0.348	0.356 0.349	0.379 0.401	0.351 0.386	0.345 0.382
	Avg	0.239 0.263	0.243 0.269	0.240 0.271	0.258 0.278	0.253 0.304	0.265 0.285	0.264 0.320	0.271 0.320	0.265 0.315
ETTh1	96	0.360 0.392	0.368 0.394	0.375 0.400	0.386 0.405	0.387 0.411	0.460 0.447	0.423 0.448	0.479 0.464	0.397 0.412
	192	0.417 0.414	0.420 0.418	0.429 0.421	0.441 0.436	0.441 0.437	0.512 0.477	0.471 0.474	0.525 0.492	0.446 0.441
	336	0.443 0.427	0.452 0.433	0.484 0.458	0.487 0.458	0.507 0.467	0.546 0.496	0.570 0.546	0.565 0.515	0.489 0.467
	720	0.439 0.443	0.449 0.449	0.498 0.482	0.503 0.491	0.527 0.548	0.544 0.517	0.653 0.621	0.594 0.558	0.513 0.510
	Avg	0.415 0.419	0.422 0.423	0.447 0.440	0.454 0.447	0.466 0.467	0.516 0.484	0.529 0.522	0.541 0.507	0.461 0.457
ETTh2	96	0.273 0.329	0.282 0.334	0.289 0.341	0.297 0.349	0.308 0.357	0.308 0.355	0.745 0.584	0.400 0.440	0.340 0.394
	192	0.355 0.378	0.359 0.385	0.372 0.392	0.380 0.400	0.395 0.404	0.393 0.405	0.877 0.656	0.528 0.509	0.482 0.479
	336	0.368 0.401	0.374 0.404	0.386 0.414	0.428 0.432	0.428 0.434	0.427 0.436	1.043 0.731	0.643 0.571	0.591 0.541
	720	0.399 0.416	0.405 0.427	0.412 0.434	0.427 0.445	0.443 0.451	0.436 0.450	1.104 0.763	0.874 0.679	0.839 0.661
	Avg	0.349 0.381	0.355 0.387	0.364 0.395	0.383 0.407	0.394 0.412	0.391 0.411	0.942 0.684	0.611 0.550	0.563 0.519
ETTm1	96	0.307 0.342	0.314 0.350	0.320 0.357	0.334 0.368	0.331 0.378	0.352 0.374	0.404 0.426	0.364 0.387	0.346 0.374
	192	0.355 0.369	0.358 0.375	0.361 0.381	0.377 0.391	0.386 0.399	0.390 0.393	0.450 0.451	0.398 0.404	0.382 0.391
	336	0.378 0.390	0.384 0.395	0.390 0.404	0.426 0.420	0.426 0.421	0.421 0.414	0.532 0.515	0.428 0.425	0.415 0.415
	720	0.445 0.427	0.448 0.432	0.454 0.441	0.491 0.459	0.489 0.465	0.462 0.449	0.666 0.589	0.487 0.461	0.473 0.451
	Avg	0.371 0.382	0.376 0.388	0.381 0.395	0.407 0.410	0.408 0.416	0.406 0.407	0.513 0.495	0.419 0.419	0.404 0.408
ETTm2	96	0.164 0.245	0.171 0.253	0.175 0.258	0.180 0.264	0.179 0.282	0.183 0.270	0.287 0.366	0.207 0.305	0.193 0.293
	192	0.228 0.290	0.234 0.294	0.237 0.299	0.250 0.309	0.244 0.305	0.255 0.314	0.414 0.492	0.290 0.364	0.284 0.361
	336	0.285 0.327	0.292 0.333	0.298 0.340	0.311 0.348	0.320 0.357	0.309 0.347	0.597 0.542	0.377 0.422	0.382 0.429
	720	0.384 0.386	0.387 0.390	0.391 0.396	0.412 0.407	0.419 0.432	0.412 0.404	1.730 1.042	0.558 0.524	0.558 0.525
	Avg	0.265 0.312	0.271 0.317	0.275 0.323	0.288 0.332	0.290 0.344	0.290 0.334	0.757 0.610	0.358 0.404	0.354 0.402
ECL	96	0.149 0.240	0.150 0.241	0.153 0.247	0.148 0.240	0.190 0.299	0.190 0.296	0.219 0.314	0.237 0.329	0.210 0.302
	192	0.160 0.251	0.162 0.252	0.166 0.256	0.162 0.253	0.216 0.323	0.199 0.304	0.231 0.322	0.236 0.330	0.210 0.305
	336	0.178 0.268	0.179 0.270	0.185 0.277	0.178 0.269	0.226 0.334	0.217 0.319	0.246 0.337	0.249 0.344	0.223 0.319
	720	0.215 0.299	0.217 0.304	0.225 0.310	0.225 0.317	0.250 0.353	0.258 0.352	0.280 0.363	0.284 0.373	0.258 0.350
	Avg	0.176 0.265	0.177 0.267	0.182 0.272	0.178 0.270	0.220 0.327	0.216 0.318	0.244 0.334	0.251 0.344	0.225 0.319
Traffic	96	0.459 0.278	0.465 0.286	0.462 0.285	0.395 0.268	0.499 0.344	0.526 0.347	0.644 0.429	0.805 0.493	0.650 0.396
	192	0.470 0.282	0.475 0.290	0.473 0.296	0.417 0.276	0.540 0.370	0.522 0.332	0.665 0.431	0.756 0.474	0.598 0.370
	336	0.487 0.294	0.489 0.296	0.498 0.296	0.433 0.283	0.557 0.378	0.517 0.334	0.674 0.420	0.762 0.477	0.605 0.373
	720	0.510 0.312	0.527 0.318	0.506 0.313	0.467 0.302	0.586 0.397	0.552 0.352	0.683 0.424	0.719 0.449	0.645 0.394
	Avg	0.482 0.292	0.489 0.297	0.484 0.297	0.428 0.282	0.546 0.372	0.529 0.341	0.667 0.426	0.760 0.473	0.625 0.383
1st Count:		28 30	0 0	1 0	7 6	0 0	0 0	0 0	0 0	0 0

Table 3: Comparison with state-of-the-art methods on the test sets of seven challenging datasets for long-term time-series forecasting. The historical time step is set as 96 for fairness. The prediction length varies in the range of {96, 192, 336, 720}. Counts of ranking first for each metric across all methods are reported. For each metric, the best performance is in bold, while the second is underlined.

resulting in reduced predictability due to their inherent irregularities.

Index	Wavelet Type	Weather		ETTh1	
		MSE	MAE	MSE	MAE
(1)	Coiflets	0.242	0.268	0.415	0.419
(2)	Daubechies	0.239	0.263	0.418	0.420
(3)	Biorthogonal	0.240	0.265	0.422	0.419
(4)	Symlets	0.244	0.269	0.419	0.421

Table 4: Ablation study of the wavelet type’s design choice.

Index	Levels of Decomposition f	Weather		ETTh1	
		MSE	MAE	MSE	MAE
(1)	1	0.239	0.264	0.416	0.419
(2)	2	0.239	0.263	0.417	0.420
(3)	3	0.240	0.266	0.415	0.419
(4)	4	0.243	0.265	0.416	0.421
(5)	5	0.244	0.266	0.418	0.422

Table 5: Ablation study of levels of decomposition in the wavelet transform.

Ablation Study

We conduct extensive ablation studies on the test sets of the Weather and ETTh1 datasets to evaluate the effectiveness of the proposed BEAT framework. To be specific, we examine BEAT from the aspect of technical designs, wavelet type, discrepancy calculation design, and gradient modulation design in the following. For all ablation studies, the average performance over four prediction lengths, including {96, 192, 336, 720}, is reported for each dataset’s test set.

Effect of BEAT. The effect of the designed Frequency-Specific Monitor (FSM) and Dynamical Gradient Balancer (DGB) is evaluated and shown in Table ?? . When both FSM and DGB are not employed, the performance is not well since the learning process of networks for different frequencies could conflict. The performance is improved with the proposed FSM and DGB for monitoring and modulating the learning process dynamically, achieving a balanced training between frequencies.

Wavelet Type. To evaluate the effect of the wavelet type, four wavelet families are considered, including Coiflets, Daubechies, Biorthogonal, and Symlets. As shown in Table 4, the performance achieves the best on the Weather and ETTh1 datasets when employing Daubechies, and Coiflets wavelet types, respectively, showing specific types should be chosen for different datasets with particular data patterns.

Levels of Decomposition. As shown in Table 5, the effect of levels f for the decomposition in the wavelet transform is evaluated. The best performance of BEAT on Weather and ETTh1 datasets is achieved when the levels f are set as 2 and

Index	Discrepancy Calculation	Weather		ETTh1	
		MSE	MAE	MSE	MAE
(1)	RMSE	0.244	0.268	0.421	0.425
(2)	R-Squared	0.241	0.265	0.418	0.423
(3)	MSE	0.239	0.263	0.415	0.419
(4)	MAE	0.240	0.264	0.416	0.419

Table 6: Ablation study of the design for discrepancy calculation.

3, respectively. Setting the level f as 4 and 5 doesn’t bring better performance. This is because higher decomposition levels can over-smooth the data or remove essential details, preventing further performance improvements.

Index	Modulation	Weather		ETTh1	
		MSE	MAE	MSE	MAE
(1)	Loss	0.242	0.268	0.420	0.421
(2)	Gradient	0.239	0.263	0.415	0.419

Table 7: Ablation study of the design for learning process modulation.

Discrepancy Calculation Design. To evaluate the effect of the discrepancy calculation metric used in Equations 2 and 3, we conduct experiments on the metric for measuring the discrepancy between the predicted coefficients and the ground-truth ones, reflecting the learning process of networks for different frequencies. As shown in Table 6, using MSE for calculating the discrepancy can achieve the best performance against other metrics, including RMSE, R-Squared, and MAE. MSE outperforms other metrics because by squaring the errors, it disproportionately penalizes larger discrepancies, leading to a more accurate measurement.

Learning Process Modulation Design. We conduct experiments to evaluate the performance when applying modulation on the total loss with the average of coefficient c^v , where $v \in \{1, 2, \dots, f\}$, calculated by Equation 7. As shown in Table 7, compared with our proposed DGB, which applies modulation to the back-propagated gradients of networks for corresponding frequencies, modulating the loss is less effective. This is because the modulation affects the entire model and cannot adjust each network individually considering its learning process.

Conclusion

In this paper, we explore a challenging problem in frequency-based time-series forecasting: the asynchronous convergence of high- and low-frequency sub-series. Although frequency-domain methods offer advantages for long-range forecasting, existing approaches overlook the mismatched learning dynamics across frequencies, leading to optimization conflicts and degraded performance. To tackle this, we propose BEAT

(Balanced frEQUENCY Adaptive Tuning), a framework that dynamically monitors the convergence of each frequency component and adaptively adjusts their gradient flow. By amplifying gradients for underfitting parts and suppressing overfitting ones, BEAT synchronizes learning across frequencies. Experiments on multiple benchmarks show that BEAT consistently improves long-range forecasting by resolving optimization conflicts caused by imbalanced learning speeds.

References

- Bai, S.; Kolter, J. Z.; and Koltun, V. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.
- Brigham, E. 1988. The fast Fourier transform and its applications.
- Das, A.; Kong, W.; Leach, A.; Mathur, S.; Sen, R.; and Yu, R. 2023. Long-term forecasting with TiDE: Time-series dense encoder. In *arXiv preprint arXiv:2304.08424*.
- Daubechies, I. 1992. Ten lectures on wavelets. *Society for industrial and applied mathematics*.
- Ekambaram, V.; Jati, A.; Nguyen, N.; Sinthong, P.; and Kalagnanam, J. 2023. TSMixer: Lightweight mlp-mixer model for multivariate time series forecasting. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 459–469.
- Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; and Li, X. 2024. Tslanet: Rethinking transformers for time series representation learning. *arXiv preprint arXiv:2404.08472*.
- Gasparin, A.; Lukovic, S.; and Alippi, C. 2022. Deep learning for time series forecasting: The electric load case. *CAAI Transactions on Intelligence Technology*, 7(1): 1–25.
- Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. 2018. Recent advances in convolutional neural networks. *Pattern recognition*, 77: 354–377.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Jia, F.; Wang, K.; Zheng, Y.; Cao, D.; and Liu, Y. 2024. GPT4MTS: Prompt-based Large Language Model for Multimodal Time-series Forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23343–23351.
- Jiang, X.; Jin, S.; Lu, L.; Zhang, X.; and Lu, S. 2024a. Weakly Supervised Monocular 3D Detection with a Single-View Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10508–10518.
- Jiang, X.; Jin, S.; Zhang, X.; Shao, L.; and Lu, S. 2024b. MonoMAE: Enhancing Monocular 3D Detection through Depth-Aware Masked Autoencoders. In *Advances in Neural Information Processing Systems*.
- Jin, K.; Wi, J.; Lee, E.; Kang, S.; Kim, S.; and Kim, Y. 2021. TrafficBERT: Pre-trained model with large-scale data for long-range traffic flow forecasting. *Expert Systems with Applications*, 186: 115738.
- Jin, M.; Wang, S.; Ma, L.; Chu, Z.; Zhang, J. Y.; Shi, X.; Chen, P.-Y.; Liang, Y.; Li, Y.-F.; Pan, S.; and Wen, Q. 2024. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. In *International Conference on Learning Representations*.
- Kim, T.; Kim, J.; Tae, Y.; Park, C.; Choi, J.-H.; and Choo, J. 2021. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*.
- Lai, G.; Chang, W.-C.; Yang, Y.; and Liu, H. 2018. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, 95–104.
- Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.-X.; and Yan, X. 2019. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in Neural Information Processing Systems*, 32.
- Li, Z.; Ye, W.; Jiang, T.; and Huang, T. 2022. 2D Amodal Instance Segmentation Guided by 3D Shape Prior. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, 165–181.
- Li, Z.; Ye, W.; Terven, J.; Bennett, Z.; Zheng, Y.; Jiang, T.; and Huang, T. 2023. MUVA: A New Large-Scale Benchmark for Multi-view Amodal Instance Segmentation in the Shopping Scenario. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23504–23513.
- Liu, M.; Zeng, A.; Chen, M.; Xu, Z.; Lai, Q.; Ma, L.; and Xu, Q. 2022. SCINet: Time series modeling and forecasting with sample convolution and interaction. *Advances in Neural Information Processing Systems*, 35: 5816–5828.
- Liu, S.; Yu, H.; Liao, C.; Li, J.; Lin, W.; Liu, A. X.; and Dustdar, S. 2021. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International Conference on Learning Representations*.
- Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2024. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*.
- Luo, D.; and Wang, X. 2024. ModernTCN: A modern pure convolution structure for general time series analysis. In *International Conference on Learning Representations*.
- Mallat, S. G. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7): 674–693.
- Murad, M. M. N.; Aktukmak, M.; and Yilmaz, Y. 2025. WP-Mixer: Efficient Multi-Resolution Mixing for Long-Term Time Series Forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*.
- Oreshkin, B. N.; Carpov, D.; Chapados, N.; and Bengio, Y. 2019. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*.

- Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- Peng, X.; Wei, Y.; Deng, A.; Wang, D.; and Hu, D. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8238–8247.
- Tang, Y.; Song, Z.; Zhu, Y.; Yuan, H.; Hou, M.; Ji, J.; Tang, C.; and Li, J. 2022. A survey on machine learning models for financial time series forecasting. *Neurocomputing*, 512: 363–380.
- Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.
- Wang, H.; Peng, J.; Huang, F.; Wang, J.; Chen, J.; and Xiao, Y. 2023. MICN: Multi-scale local and global context modeling for long-term series forecasting. In *International Conference on Learning Representations*.
- Wang, J.; Wang, Z.; Li, J.; and Wu, J. 2018. Multilevel wavelet decomposition network for interpretable time series analysis. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2437–2446.
- Wang, S.; Wu, H.; Shi, X.; Hu, T.; Luo, H.; Ma, L.; Zhang, J. Y.; and ZHOU, J. 2024a. TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting. In *International Conference on Learning Representations*.
- Wang, Y.; Wu, H.; Dong, J.; Liu, Y.; Long, M.; and Wang, J. 2024b. Deep Time Series Models: A Comprehensive Survey and Benchmark.
- Wang, Z.; Pei, C.; Ma, M.; Wang, X.; Li, Z.; Pei, D.; Rajmohan, S.; Zhang, D.; Lin, Q.; Zhang, H.; et al. 2024c. Revisiting VAE for Unsupervised Time Series Anomaly Detection: A Frequency Perspective. In *Proceedings of the ACM on Web Conference 2024*, 3096–3105.
- Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *International Conference on Learning Representations*.
- Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, 22419–22430.
- Xu, Z.; Zeng, A.; and Xu, Q. 2023. FITS: Modeling time series with 10k parameters. *arXiv preprint arXiv:2307.03756*.
- Yi, K.; Zhang, Q.; Fan, W.; Wang, S.; Wang, P.; He, H.; An, N.; Lian, D.; Cao, L.; and Niu, Z. 2024. Frequency-domain MLPs are more effective learners in time series forecasting. *Advances in Neural Information Processing Systems*, 36.
- Yi, K.; Zhang, Q.; Hu, L.; He, H.; An, N.; Cao, L.; and Niu, Z. 2022. Edge-Varying Fourier Graph Networks for Multivariate Time Series Forecasting. *arXiv preprint arXiv:2210.03093*.
- Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 11121–11128.
- Zhang, G.; Yang, D.; Galanis, G.; and Androulakis, E. 2022. Solar forecasting with hourly updated numerical weather prediction. *Renewable and Sustainable Energy Reviews*, 154: 111768.
- Zhang, Y.; and Yan, J. 2023. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*.
- Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 11106–11115.
- Zhou, T.; Ma, Z.; Wen, Q.; Sun, L.; Yao, T.; Yin, W.; Jin, R.; et al. 2022a. Film: Frequency improved legendre memory model for long-term time series forecasting. *Advances in neural information processing systems*, 35: 12677–12690.
- Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; and Jin, R. 2022b. FEDFormer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, 27268–27286.