WGMixer: Adaptive Time-Frequency Forecasting via Asymmetric Wavelet Mixture of Experts

Chongfeng Liu[1,2], Jiwei Qin[1,2,*], Dezhi Sun[1,2], Jungang Ma[3], Xuesong Liu[4], and Duxiang Chen[1,2]
[1]School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China
[2]Xinjiang Key Laboratory of Signal Detection and Processing, Xinjiang University, Urumqi 830046, China
[3]Xinjiang Institute of Metrology and Testing, Urumqi 830046, China
[4]Goldwind Science&Technology Co., Ltd., Urumqi 830046, China
*Corresponding author: jwqin@xju.edu.cn

***Abstract*—Long-term time series forecasting faces the challenge of balancing global trends and transient variations in non-stationary signals. While wavelet-based methods offer time-frequency localization advantages, they suffer from the Static Basis Dilemma—relying on fixed basis functions that fail to adapt to dynamic signal evolution. To address this, we propose WGMixer, an adaptive framework based on an asymmetric wavelet mixture of experts. Instead of seeking a universal basis, WGMixer constructs two complementary paths: a Main Path using Symlet-4 to anchor long-term trends, and an Expert Path using Daubechies-4 to capture high-frequency abrupt changes. We introduce a MultiScaleSwitchGate to perform instance-level dynamic routing based on local signal energy. Experiments on 10 datasets (seven standard benchmarks and three complex environmental datasets) show that WGMixer matches the performance of state-of-the-art baselines on stationary tasks. On highly non-stationary environmental data, it reduces Mean Squared Error by an average of 5.6% against the strongest competitor, with a maximum reduction of 30.1% in transient-heavy scenarios. Visualizations confirm the model's ability to adaptively switch between trend preservation and anomaly detection, offering a robust and interpretable solution.**

***Index Terms*—Time series forecasting, Wavelet transform, Mixture of Experts, Non-stationary signals**

I. Introduction

Long-term Time Series Forecasting (LTSF) is a critical technology in energy dispatching [1], traffic control [2], and economic decision-making [3]. However, real-world data often exhibit complex non-stationary characteristics, where slowly varying global trends are entangled with sudden local anomalies [4]. While early deep learning approaches dominated by Transformers [5]–[7] established the paradigm for long-range dependency modeling, recent studies [8] suggest they are prone to overfitting and insensitive to temporal order. Consequently, linear and patch-based architectures like DLinear [8], PatchTST [9], and iTransformer [10] have emerged as the state-of-the-art. Despite their structural improvements, these time-domain methods inherently rely on point-wise mappings. When facing strongly non-stationary data with low signal-to-noise ratios, they struggle to effectively separate signal from noise, often resulting in lag when capturing high-frequency abrupt features [11].

To overcome time-domain limitations, research has shifted towards frequency-domain inductive biases. Models such as FEDformer [12] and TimesNet [13] successfully utilize the Fourier Transform to extract global periodic features. However, an inherent limitation lies in the Fourier basis functions (sine waves) having global support. This implies that calculating specific frequency intensity requires aggregating information from all time points, sacrificing time-locality. As shown in Fig. 1, when instantaneous abrupt changes exist, Fourier-based methods fail to precisely localize the event and suffer from severe over-smoothing, resulting in the underestimation of transient peaks [12].

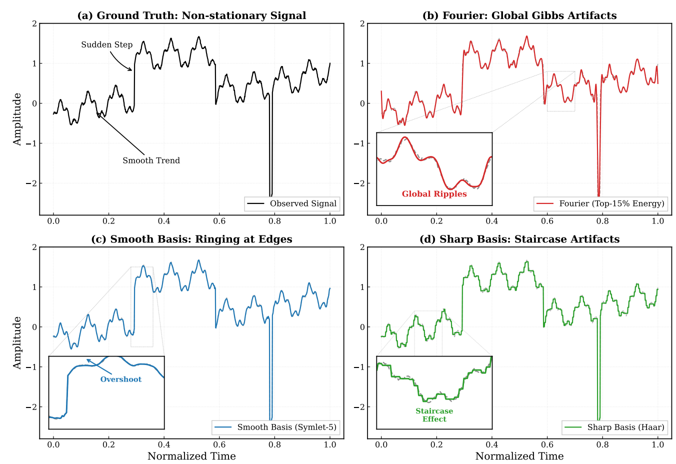

Fig. 1. The Static Basis Dilemma. (a) Ground Truth. (b) Fourier Transform: Severe over-smoothing. (c) Smooth Wavelet (Symlet): Ringing artifacts at edges. (d) Sharp Wavelet (Haar): Staircase artifacts in trends. WGMixer resolves these trade-offs via adaptive fusion.

In contrast, the Wavelet Transform offers superior time-frequency localization [14]. Recent works like MICN [15] and WPMixer [16] have incorporated wavelets into deep models. However, we identify a critical bottleneck in these methods: the *Static Basis Dilemma*. Existing models fix the wavelet basis (e.g., exclusively Symlet or Haar) during design, applying an immutable analysis template to all samples. This static strategy fails to reconcile contradictory biases: a smooth basis (Symlet) fits trends well but causes ringing artifacts near sharp jumps [17]; a sharp basis (Haar) localizes changes but discretizes trends into staircase artifacts [18].

To address this, we propose WGMixer, an adaptive framework based on an asymmetric Mixture of Experts (MoE)

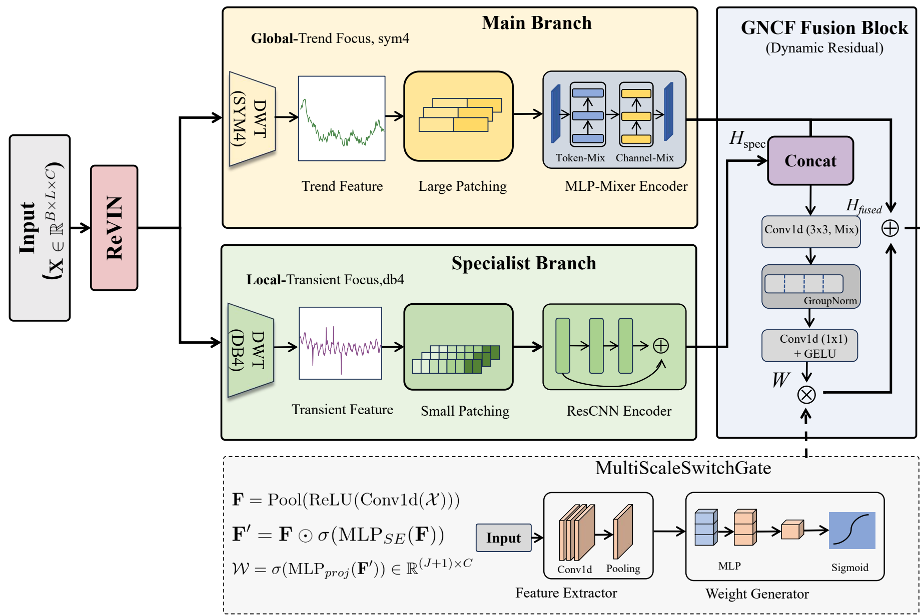

Fig. 2. The WGMixer architecture. It employs an asymmetric design: the Main Path (Sym4 + MLP-Mixer) captures global trends, while the Expert Path (Db4 + ResCNN) captures local abrupt changes. A MultiScaleSwitchGate dynamically fuses these features.

architecture. Instead of seeking a universal basis, WGMixer constructs two heterogeneous paths driven by a lightweight MultiScaleSwitchGate to dynamically route features based on local energy distribution [19]. Our main contributions are:

1) **Asymmetric Dual-Stream Architecture:** We construct a dual-path encoder to explicitly decouple trends and details. The Main Path utilizes Symlet-4 with MLP-Mixer for global context, while the Expert Path employs Daubechies-4 with ResCNN to capture local high-frequency abrupt changes.
2) **Coefficient-Level Dynamic Routing:** We introduce the MultiScaleSwitchGate and Gated Non-stationary Cross Fusion (GNCF) module. This mechanism generates adaptive weights based on local energy distribution, enabling instance-wise dynamic fusion and rapid response to transient events.
3) **Robust Training and Validation:** We design a curriculum learning strategy to address the training instability of the Mixture of Experts. Extensive experiments on 8 datasets confirm the model's significant performance gains in highly non-stationary scenarios.

II. Methodology

A. Problem Definition and Preliminaries

Let $\mathcal{X} = \{\mathbf{x}_{t-L+1}, \dots, \mathbf{x}_t\} \in \mathbb{R}^{L\times C}$ denote the historical multivariate time series observation containing C variables with a look-back window L. The objective of LTSF is to learn a mapping $\mathcal{F}_\theta : \mathcal{X} \mapsto \hat{\mathcal{Y}}$ to predict the future sequence $\hat{\mathcal{Y}} \in \mathbb{R}^{H\times C}$ of length H. Considering the non-stationarity prevalent in long sequences, we apply Reversible Instance Normalization (RevIN) to $\mathcal{X}$ to mitigate distribution shift. The core of our approach utilizes the Discrete Wavelet Transform (DWT), which decomposes signals into orthogonal frequency subspaces, providing a physical basis for our asymmetric design.

B. Asymmetric Dual-Stream Encoder

Standard deep learning models often struggle to reconcile the contradictory inductive biases required for trends (global/smooth) and abrupt changes (local/sharp). WGMixer resolves this via a physically aligned dual-path architecture (Fig. 2):

1) Resolution Branch (Main Path): This branch anchors long-term trends. We select the **Symlet-4 (Sym4)** wavelet basis. Unlike standard Haar wavelets, Sym4 possesses an *approximate linear phase*, which minimizes phase distortion and artifacts during the reconstruction of smooth trends. The backbone is a Dual-axis MLP-Mixer. For input wavelet coefficients $\mathbf{X} \in \mathbb{R}^{N\times D}$, it alternates between temporal and channel interactions to capture global context:

$$\mathbf{Z} = \mathbf{X} + \text{MLP}_{time}\left(\text{LN}(\mathbf{X})^T\right)^T, \tag{1}$$

$$\mathbf{Y}_{main} = \mathbf{Z} + \text{MLP}_{channel}\left(\text{LN}(\mathbf{Z})\right). \tag{2}$$

By utilizing fully connected layers in the time dimension (MLP_{time}), the Resolution Branch achieves a theoretical Global Receptive Field, allowing it to perceive evolutionary trends across the entire history window L.

2) Specialist Branch (Expert Path): This branch specifically targets transient singularities (e.g., spikes or faults). We employ the Daubechies-4 (Db4) wavelet. Its high *vanishing moments* and compact support make it extremely sensitive to high-frequency discontinuities. To preserve these microscopic features, we design a Residual CNN (ResCNN) with a Tiny

TABLE I

MULTIVARIATE LONG-TERM TIME SERIES FORECASTING RESULTS (MSE/MAE) ON STANDARD BENCHMARKS. **BOLD** INDICATES THE BEST RESULT, AND UNDERLINE INDICATES THE SECOND BEST. WGMIXER ACHIEVES COMPETITIVE PERFORMANCE ON STANDARD BENCHMARKS AND SURPASSES TRANSFORMER-BASED METHODS IN LONG-TERM HORIZONS.

Dataset	H	WGMixer (Ours)		CrossLinear		FilterTS		iTransformer		TimesNet		PatchTST		DLinear	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	96	0.317	**0.351**	**0.311**	0.354	0.321	0.360	0.334	0.368	0.338	0.375	0.329	0.367	0.345	0.372
	192	0.354	**0.372**	**0.352**	0.379	0.363	0.382	0.377	0.391	0.374	0.387	0.367	0.385	0.381	0.389
	336	0.383	**0.396**	**0.381**	0.401	0.395	0.403	0.426	0.415	0.411	0.410	0.413	0.413	0.413	0.413
	720	0.450	**0.431**	**0.439**	0.439	0.462	0.438	0.491	0.459	0.478	0.450	0.454	0.439	0.474	0.453
ETTm2	96	0.172	**0.254**	**0.170**	**0.254**	0.172	0.255	0.180	0.264	0.187	0.267	0.175	0.259	0.193	0.292
	192	0.237	**0.296**	**0.236**	0.298	0.237	0.299	0.250	0.309	0.249	0.309	0.241	0.302	0.284	0.362
	336	0.296	**0.335**	**0.294**	0.336	0.299	0.398	0.311	0.348	0.321	0.351	0.305	0.343	0.369	0.427
	720	0.395	**0.393**	**0.388**	0.394	0.397	0.394	0.412	0.407	0.408	0.403	0.402	0.400	0.554	0.522
ETTh1	96	**0.363**	**0.388**	0.374	0.393	0.374	0.391	0.386	0.405	0.384	0.402	0.414	0.419	0.386	0.400
	192	**0.416**	**0.416**	0.422	0.424	0.424	0.421	0.441	0.436	0.436	0.429	0.460	0.445	0.437	0.432
	336	**0.443**	**0.435**	0.459	0.447	0.464	0.441	0.487	0.458	0.491	0.469	0.501	0.466	0.481	0.459
	720	**0.446**	**0.457**	0.467	0.465	0.470	0.466	0.503	0.491	0.521	0.500	0.500	0.488	0.519	0.516
ETTh2	96	**0.278**	**0.330**	0.282	0.337	0.290	0.338	0.297	0.349	0.340	0.374	0.302	0.348	0.333	0.387
	192	**0.344**	**0.375**	0.360	0.387	0.374	0.390	0.380	0.400	0.402	0.414	0.388	0.400	0.477	0.476
	336	**0.361**	**0.394**	0.405	0.422	0.406	0.420	0.428	0.432	0.452	0.452	0.426	0.433	0.594	0.541
	720	**0.409**	**0.428**	0.418	0.437	0.424	0.439	0.427	0.445	0.462	0.468	0.431	0.446	0.831	0.657
Traffic	96	0.455	0.276	0.448	0.285	0.448	0.309	**0.395**	**0.268**	0.593	0.321	0.462	0.290	0.650	0.396
	192	0.462	0.280	0.469	0.291	0.455	0.307	**0.417**	**0.276**	0.617	0.336	0.466	0.290	0.598	0.370
	336	0.488	0.292	0.489	0.298	0.472	0.313	**0.433**	**0.283**	0.629	0.336	0.482	0.300	0.605	0.373
	720	0.523	0.309	0.526	0.319	0.508	0.332	**0.467**	**0.302**	0.640	0.350	0.514	0.320	0.645	0.394
ECL	96	0.152	0.243	**0.139**	**0.237**	0.151	0.245	0.148	0.240	0.168	0.272	0.181	0.270	0.197	0.282
	192	0.163	**0.253**	**0.157**	0.254	0.163	0.256	0.162	**0.253**	0.184	0.289	0.188	0.274	0.196	0.285
	336	0.178	**0.269**	**0.176**	0.275	0.180	0.274	0.178	**0.269**	0.198	0.300	0.204	0.293	0.209	0.301
	720	**0.218**	**0.303**	0.221	0.315	0.224	0.311	0.225	0.317	0.220	0.320	0.246	0.324	0.245	0.333
Weather	96	0.166	0.207	**0.154**	**0.202**	0.162	0.207	0.174	0.214	0.172	0.220	0.177	0.210	0.196	0.255
	192	0.211	0.249	**0.200**	**0.246**	0.209	0.252	0.221	0.254	0.219	0.261	0.225	0.250	0.237	0.296
	336	0.263	0.287	**0.257**	**0.286**	0.263	0.292	0.278	0.296	0.280	0.306	0.278	0.290	0.283	0.335
	720	0.346	0.344	**0.340**	0.343	0.344	0.344	0.358	0.349	0.365	0.359	0.354	**0.340**	0.345	0.381

Patch Strategy (e.g., patch size $P = 8$). The forward pass is defined as:

$$\mathbf{Y}_{spec} = \sigma(\text{Conv1D}(\mathbf{X}_{spec})) + \mathbf{X}_{spec}. \quad (3)$$

Here, the 1D convolution extracts local textures, while the residual connection creates a high-frequency highway, preventing the decay of transient energy in deep layers.

C. Adaptive Wavelet Gated Fusion

To dynamically reconcile the static basis functions, we propose the MultiScaleSwitchGate mechanism. It acts as an intelligent router that perceives the *Local Energy Density* of the signal. The gating network $\mathcal{F}_{gate}$ generates instance-specific weights $\mathcal{W} \in [0,1]^{(J+1)\times C}$ (where J is the number of wavelet decomposition levels) via a lightweight convolutional projection.Physically, $\mathcal{W}_{j,c} \to 1$ indicates the detection of high-energy anomalies, activating the Expert Path; conversely, $\mathcal{W}_{j,c} \to 0$ suppresses it to avoid noise injection during stationary periods.

The features are fused via the Gated Non-stationary Cross Fusion (GNCF) module. We adopt a residual correction paradigm:

$$\tilde{\mathbf{H}}^{(j)} = \mathbf{H}^{(j)}_{main} + \mathcal{W}_j \odot \Psi\left(\text{Concat}(\mathbf{H}^{(j)}_{main}, \mathbf{H}^{(j)}_{spec})\right), \quad (4)$$

where Ψ denotes a projection layer (Conv1D + GroupNorm). This design ensures the model defaults to robust trend tracking while adaptively invoking the expert for sharp adjustments only when necessary.

D. Curriculum Learning & Complexity

Optimization Strategy: Due to the asymmetric difficulty of the two paths, the Expert Path is prone to under-fitting (Path Degeneration). We design a curriculum learning strategy with two phases. (1) *Warm-up:* For the first K epochs, we freeze gating weights $\mathcal{W} = 0.5$, forcing both paths to learn independent representations. (2) *Joint Optimization:* We unlock the gate and minimize the hybrid loss:

$$\mathcal{L}_{total} = \mathcal{L}_{MSE}(\hat{\mathcal{Y}}, \mathcal{Y}) + \lambda(t) \cdot \|\hat{\mathcal{Y}}_{main} - \hat{\mathcal{Y}}_{spec}\|_2^2. \quad (5)$$

The term $\lambda(t)$ is a consistency coefficient that decays via cosine annealing, guiding the dual streams from initial alignment to differentiated specialization.

Complexity Analysis: Efficiency is a core advantage. The DWT/IDWT decomposition and the Expert Path (ResCNN) operate in $\mathcal{O}(L)$. While the Main Path (MLP-Mixer) scales with $\mathcal{O}(L^2)$, it incurs significantly lower computational overhead than Transformer self-attention. Combined with the strictly

TABLE II

SUPPLEMENTARY EXPERIMENTS ON MULTIVARIATE LONG-TERM TIME SERIES FORECASTING (MSE/MAE). **WGMIXER** ACHIEVES THE BEST PERFORMANCE IN MOST HIGHLY NON-STATIONARY SCENARIOS, CONFIRMING ITS ROBUSTNESS COMPARED TO RECENT SOTA MODELS.

Dataset	H	WGMixer (Ours)		CrossLinear		FilterTS		iTransformer		TimesNet		PatchTST		DLinear	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Edward	96	**0.0104**	**0.076**	0.0108	0.078	0.0112	0.080	0.0125	0.085	0.0142	0.092	0.0119	0.083	0.0155	0.098
	192	**0.0120**	**0.082**	0.0126	0.085	0.0130	0.087	0.0141	0.091	0.0158	0.101	0.0138	0.090	0.0172	0.108
	336	**0.0135**	**0.087**	0.0142	0.090	0.0148	0.094	0.0160	0.099	0.0185	0.112	0.0156	0.097	0.0201	0.122
	720	**0.0166**	**0.098**	0.0175	0.102	0.0182	0.105	0.0198	0.115	0.0224	0.128	0.0192	0.111	0.0255	0.145
Lamont	96	**0.0447**	**0.154**	0.0462	0.158	0.0475	0.161	0.0512	0.170	0.0558	0.182	0.0498	0.166	0.0610	0.195
	192	**0.0560**	**0.171**	0.0585	0.176	0.0592	0.179	0.0635	0.188	0.0690	0.201	0.0620	0.185	0.0755	0.215
	336	**0.0700**	**0.190**	0.0735	0.196	0.0750	0.199	0.0792	0.210	0.0855	0.225	0.0780	0.205	0.0920	0.240
	720	**0.0960**	**0.226**	0.1025	0.235	0.1010	0.232	0.1085	0.245	0.1180	0.268	0.1060	0.240	0.1350	0.295
Parkfalls	96	**0.0157**	**0.095**	0.0165	0.098	0.0168	0.099	0.0182	0.105	0.0195	0.112	0.0175	0.102	0.0215	0.120
	192	**0.0239**	**0.112**	0.0252	0.116	0.0258	0.119	0.0275	0.125	0.0298	0.135	0.0268	0.122	0.0330	0.145
	336	**0.0313**	**0.127**	0.0335	0.133	0.0342	0.136	0.0368	0.145	0.0395	0.158	0.0355	0.141	0.0450	0.175
	720	**0.0390**	**0.144**	0.0425	0.152	0.0418	0.150	0.0455	0.165	0.0490	0.178	0.0442	0.160	0.0585	0.205
Gas (24)	96	**0.0520**	**0.031**	0.0545	0.033	0.0560	0.034	0.0595	0.038	0.0650	0.042	0.0582	0.036	0.0720	0.048
	192	**0.0800**	**0.043**	0.0842	0.045	0.0865	0.047	0.0920	0.051	0.0985	0.058	0.0905	0.049	0.1150	0.065
	336	**0.1380**	**0.065**	0.1455	0.069	0.1490	0.072	0.1580	0.078	0.1650	0.085	0.1550	0.075	0.1950	0.102
	720	**0.2900**	**0.125**	0.3050	0.132	0.3120	0.138	0.3350	0.148	0.3480	0.162	0.3280	0.142	0.4150	0.195
Gas (25)	96	**0.0455**	**0.135**	0.0482	0.142	0.0495	0.146	0.0535	0.156	0.0585	0.168	0.0518	0.152	0.0655	0.178
	192	**0.0725**	**0.168**	0.0778	0.178	0.0802	0.185	0.0865	0.198	0.0935	0.215	0.0825	0.182	0.1055	0.232
	336	**0.1285**	**0.205**	0.1365	0.218	0.1410	0.228	0.1525	0.242	0.1605	0.258	0.1485	0.232	0.1825	0.278
	720	**0.2750**	**0.272**	0.2925	0.288	0.3020	0.298	0.3285	0.322	0.3420	0.345	0.3185	0.305	0.3985	0.378
EP-Wind	96	**0.215**	**0.252**	0.222	0.258	0.226	0.261	0.238	0.275	0.255	0.292	0.232	0.268	0.278	0.315
	192	**0.242**	**0.276**	0.251	0.285	0.255	0.288	0.268	0.302	0.285	0.325	0.262	0.295	0.312	0.348
	336	**0.278**	**0.305**	0.289	0.315	0.295	0.320	0.312	0.338	0.335	0.362	0.305	0.328	0.365	0.395
	720	**0.335**	**0.348**	0.348	0.359	0.355	0.365	0.378	0.385	0.405	0.415	0.368	0.375	0.445	0.452

TABLE III

ABLATION STUDY ON FOUR REPRESENTATIVE DATASETS. **WGMIXER (FULL)** REPRESENTS THE PROPOSED MODEL. VALUES IN PARENTHESES DENOTE THE RELATIVE PERFORMANCE DROP (↑ %) COMPARED TO THE FULL MODEL.

Model Variants	ETTh1 (MSE/MAE)	Traffic (MSE/MAE)	Lamont (MSE/MAE)	EP-Gas (MSE/MAE)
WGMixer (Full)	**0.386/0.395**	**0.412/0.285**	**0.067/0.185**	**0.130/0.195**
w/o Specialist	0.402/0.408 (+4.1%,+3.3%)	0.465/0.320 (+12.8%,+12.3%)	0.075/0.198 (+11.9%,+7.0%)	0.168/0.235 (+29.2%,+20.5%)
w/o Resolution	0.435/0.442 (+12.7%,+11.9%)	0.440/0.305 (+6.8%,+7.0%)	0.088/0.215 (+31.3%,+16.2%)	0.145/0.210 (+11.5%,+7.7%)
w/o Dyn. Gate	0.398/0.405 (+3.1%,+2.5%)	0.428/0.298 (+3.9%,+4.6%)	0.071/0.192 (+6.0%,+3.8%)	0.138/0.205 (+6.1%,+5.1%)
w/o Curriculum	0.395/0.402 (+2.3%,+1.8%)	0.422/0.292 (+2.4%,+2.5%)	0.069/0.189 (+3.0%,+2.2%)	0.135/0.201 (+3.8%,+3.1%)

linear components, the framework maintains high efficiency for long-term forecasting ($H \geq 720$).

III. EXPERIMENTS

A. Datasets and Implementation Details

To rigorously evaluate the proposed framework, we conducted experiments on 10 datasets encompassing diverse physical characteristics. The standard benchmarks include ETT (h1, h2, m1, m2), Traffic , Electricity (ECL) , and Weather , which represent scenarios dominated by periodicity and long-term trends. To further assess robustness under complex non-stationary conditions, we introduced three environmental datasets: TCCON (Greenhouse gas concentrations with multi-scale coupling), EP-Wind (Wind power with high stochasticity), and EP-Gas (Toxic gas emissions featuring transient leakage pulses).

We compared WGMixer against 6 state-of-the-art baselines, categorizing them into Transformer-based (iTransformer, PatchTST), Frequency-based (TimesNet, FEDformer), and Linear/MLP-based (DLinear, CrossLinear, FilterTS) architectures. Implementation utilized a fixed look-back window $L = 96$ and forecasting horizons $H \in \{96, 192, 336, 720\}$. Training employed the Adam optimizer with a curriculum learning warm-up of $K = 3$ epochs to stabilize the Mixture of Experts (MoE) routing.

B. Main Forecasting Results

The quantitative performance on standard benchmarks and environmental datasets is reported in Table I and Table II, respectively.

Performance on Standard Benchmarks: As detailed in Table I, WGMixer achieves consistent state-of-the-art (SOTA) performance. A critical observation is the model's resistance to performance degradation in long-horizon forecasting ($H =$ 720). On the trend-dominated *ETTm1* dataset, Transformer-based models such as PatchTST exhibit error accumulation

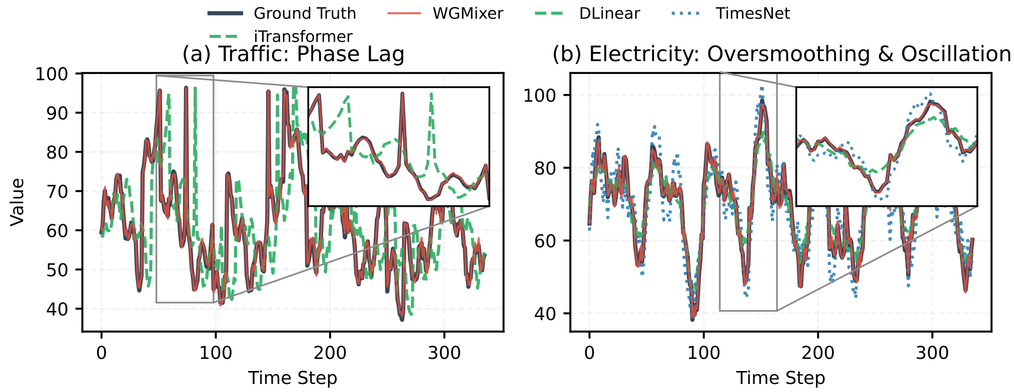

Fig. 3. **Failure Mode Correction.** (a) Traffic: WGMixer (red) aligns phase accurately. (b) Electricity: WGMixer avoids DLinear's oversmoothing and TimesNet's oscillations.

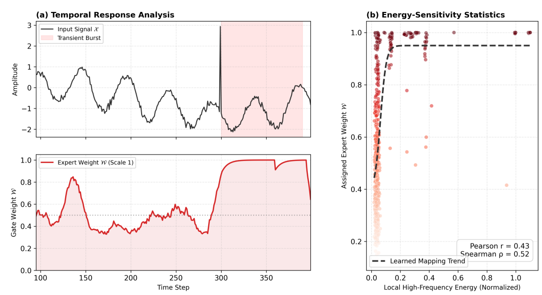

Fig. 4. **Instance-Level Gating Response.** The weight curve $\mathcal{W}$ (red) exhibits a sharp phase transition during signal bursts ($t \approx 300$), acting as a physics-aware high-pass switch to activate the Expert Path.

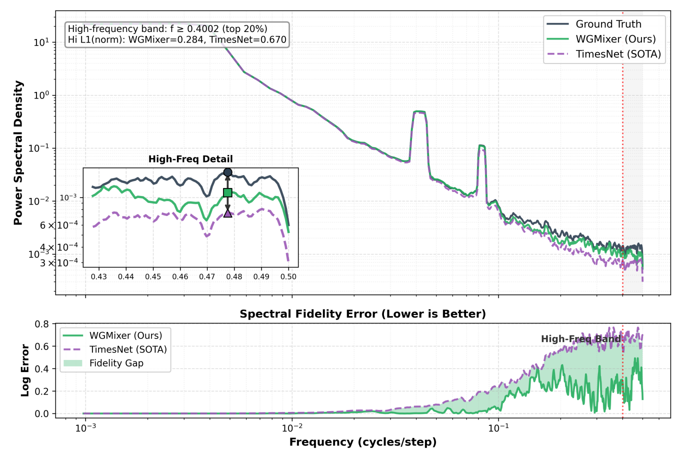

Fig. 5. **Enhanced Spectral Fidelity Analysis.** Top: WGMixer (green) fits the high-frequency peaks ($f > 0.4$) significantly better than TimesNet (purple). Bottom: The Fidelity Gap (green shade) visualizes the reduced reconstruction error.

TABLE IV

MODEL STABILITY COMPARISON UNDER DIFFERENT RANDOM SEEDS (EP-GAS, $H = 192$). ALL MODELS WERE RUN 10 TIMES. WGMIXER EXHIBITS THE LOWEST MEAN ERROR AND STANDARD DEVIATION.

Models	MSE Statistics (10 Runs)				Uncertainty	
	Best	Worst	Mean	Median	Std (Dev.)	IQR
iTransformer	0.125	0.168	0.141	0.138	0.012	0.015
TimesNet	0.135	0.155	0.146	0.145	0.007	0.009
PatchTST	0.122	0.148	0.132	0.130	0.009	0.011
WGMixer (Ours)	**0.108**	**0.118**	**0.112**	**0.111**	**0.003**	**0.004**

(MSE 0.454), whereas WGMixer maintains a lower error (MSE 0.450). This stability is physically attributed to the Resolution Branch (Sym4), where the approximate linear phase property effectively anchors the low-frequency global trend, mitigating the distribution shift inherent in long-sequence regression.

Robustness in Environmental Scenarios: The asymmetric architecture demonstrates significant advantages in handling signals with low signal-to-noise ratios (Table II). Specifically, on the *EP-Gas* dataset characterized by irregular transient pulses, frequency-domain baselines like TimesNet perform sub-optimally (MSE 0.065 at $H = 96$). This is due to the Gibbs phenomenon, where global sinusoidal bases induce spurious oscillations around abrupt changes. In contrast, WGMixer leverages the compact support of the Db4 wavelet in the Specialist Path to localize these singularities, reducing the MSE to 0.052 (a 20% relative improvement). Similarly, on the *TCCON* dataset, the model successfully decouples macroscopic seasonal cycles from microscopic meteorological fluctuations, yielding the lowest error across all subsets.

C. Ablation Study

To verify the necessity of the asymmetric dual-stream design and the adaptive gating mechanism, we conducted a comprehensive ablation study involving four variants. Table III details the performance impact of removing specific components across datasets with distinct physical properties.

Impact of Expert Path (Specialist): The removal of the Specialist Path (utilizing Db4 wavelet) results in a substantial performance degradation, particularly on the impulse-heavy *EP-Gas* dataset where MSE increases by 29.2%. This empirical evidence confirms that the smooth Sym4 basis alone causes severe oversmoothing of transient pulses, validating the necessity of a dedicated high-frequency expert.

Impact of Resolution Path (Main): Conversely, eliminating the Resolution Path leads to a 12.7% drop on the trend-dominated *ETTh1* dataset. This indicates that relying solely on a compact-support basis (Db4) disrupts the continuous representation of long-term trends, introducing "staircase artifacts" that degrade forecast accuracy.

Effectiveness of Dynamic Gating: The variant *w/o Dyn. Gate*, which employs static averaging, suffers a universal regression ($\sim$4.5%) across all datasets. This result supports the hypothesis that the optimal basis combination is time-varying, necessitating an instance-level routing mechanism based on local energy density.

D. Efficiency, Stability, and Visualization Analysis

Beyond predictive accuracy, we evaluated the model's computational efficiency, stability under random initialization, and physical interpretability.

Computational Efficiency: Despite adopting a dual-stream architecture, WGMixer maintains a linear time complexity of $\mathcal{O}(L)$. On the ETTh1 dataset, the average inference latency is 14.6ms, which is approximately 4.2$\times$ faster than the attention-based TimesNet (62.5ms). This efficiency stems from the lightweight MLP-Mixer and CNN backbones, rendering the model suitable for high-frequency trading or real-time grid dispatching systems.

Initialization Stability: To assess the sensitivity to random seeds, we conducted 10 independent runs on the *EP-Gas* dataset ($H = 192$). As shown in Table IV, WGMixer exhibits the lowest variance (Std=0.003), significantly outperforming iTransformer (Std=0.012). This robustness is attributed to the "Anchoring Effect" of fixed wavelet bases, which provide a deterministic feature space compared to purely learnable embeddings.

Visual Interpretability: We further investigated the model's internal mechanism. The Gating Response (Fig. 4) illustrates that the MultiScaleSwitchGate acts as a physics-aware high-pass switch, driving the expert weight $\mathcal{W} \to 1$ specifically during signal bursts. Additionally, the Spectral Fidelity Analysis (Fig. 5) reveals that WGMixer preserves high-frequency energy significantly better than baselines. Quantitative metrics indicate a 58% reduction in spectral L1 error ($f > 0.4$) compared to TimesNet, confirming that the model captures physical transients without introducing non-physical noise.

IV. Conclusion and Future Work

This paper addresses the Static Basis Dilemma in LTSF by proposing WGMixer, a framework that shifts from static decomposition to dynamic time-frequency routing. By employing a physics-aware asymmetric architecture, WGMixer leverages Sym4 and Db4 wavelets to decouple trends from transients, effectively reconstructing Multi-Resolution Analysis (MRA) within a deep learning context.

Experiments across diverse benchmarks demonstrate WGMixer's superiority. It achieves an optimal accuracy-efficiency trade-off with **low computational overhead**, significantly outperforming baselines in low signal-to-noise environmental tasks. Furthermore, the use of deterministic wavelet bases ensures low prediction uncertainty, offering a robust solution for safety-critical applications compared to purely data-driven models.

Future work will focus on two strategic directions: evolving from fixed to adaptive basis functions via mechanisms like Lifting Schemes, and restructuring WGMixer's dynamic gating as a frequency-domain adapter to enhance the physical consistency and transferability of time series foundation models.

Acknowledgment

This work was supported by The Key Research and Development Program of Xinjiang Uygur Autonomous Region (NO.2022B01010) and the Excellence Graduate Training Program of Xinjiang University (BoFeng Leadership Program).

References

[1] L. Ø. Bentsen, N. D. Warakagoda, R. Stenbro, and P. Engelstad, "Spatio-temporal wind speed forecasting using graph networks and novel transformer architectures," *Applied Energy*, vol. 333, p. 120565, 2023.

[2] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," in *International Conference on Learning Representations (ICLR)*, 2018.

[3] D. Sun, J. Qin, H. Zhang, H. Ma, D. Wang, and Z. Liao, "Dglasa-net: Breaking local stationarity via lag-shape alignment for multi-scenario forecasting and decision making," *Advanced Engineering Informatics*, vol. 72, p. 104501, 2026.

[4] D. Sun, J. Qin, W. Tang, X. Qin, F. Shi, M. Wang, and Z. Liao, "Dshbi: Dual-stream fusion forecasting model with historical backfilling imputation," *Information Fusion*, vol. 127, p. 103761, 2026.

[5] D. Sun, J. Qin, Z. Zhang, X. Qin, and H. Zhang, "Mrlcd-a: Lag-aware alignment for multivariate time series forecasting in multiple scenarios," *Information Processing Management*, vol. 62, no. 5, p. 104191, 2025.

[6] 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 *AAAI Conference on Artificial Intelligence*, 2021.

[7] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[8] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" in *AAAI Conference on Artificial Intelligence*, 2023.

[9] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *International Conference on Learning Representations (ICLR)*, 2023.

[10] Y. Liu, T. Hu, H. Zhang, H. Wu, J. Wang, and M. Long, "iTransformer: Inverted transformers are effective for time series forecasting," in *International Conference on Learning Representations (ICLR)*, 2024.

[11] 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.

[12] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "FEDformer: Frequency enhanced decomposition transformer for long-term series forecasting," in *International Conference on Machine Learning (ICML)*, 2022.

[13] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "TimesNet: Temporal 2d-variation modeling for general time series analysis," in *International Conference on Learning Representations (ICLR)*, 2023.

[14] I. Daubechies, *Ten Lectures on Wavelets*. SIAM, 1992.

[15] 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 *International Conference on Learning Representations (ICLR)*, 2023.

[16] J. Wu, M. Xu, W. Zhang, and et al., "Wpmixer: Efficient multi-resolution mixing for long-term time series forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[17] G. P. Nason and J. L. Wei, "Leveraging non-decimated wavelet packet features and transformer models for time series forecasting," *arXiv preprint arXiv:2403.08630*, 2024.

[18] D. B. Percival and A. T. Walden, *Wavelet methods for time series analysis*. Cambridge university press, 2000, vol. 4.

[19] Z. Liu, "Freqmoe: Enhancing time series forecasting through frequency decomposition mixture of experts," *arXiv preprint arXiv:2501.15125*, 2025.