PRISM: PRinciple Induced Spectral Mixture of Experts with Test-Time Calibration for Time Series Forecasting

Mufeng Liu
East China Normal University
Shanghai, China
rickyliu0513@163.com

Junhui Wang
East China Normal University
Shanghai, China
51265902093@stu.ecnu.edu.cn

Xidong Xi
East China Normal University
Shanghai, China
52265902004@stu.ecnu.edu.cn

Guitao Cao*
East China Normal University
Shanghai, China
gtcao@sei.ecnu.edu.cn

***Abstract*—Time series forecasting remains constrained by a fundamental assumption: all samples within a dataset can be effectively modeled by unified parameters. This assumption becomes limiting when samples exhibit heterogeneous spectral characteristics—trending series concentrate energy in low frequencies while volatile signals show high-frequency dominance. Existing methods typically apply dataset-level transformations, making it difficult to exploit sample-specific spectral structure. We propose PRISM, a spectral mixture-of-experts framework built around three complementary components. First, data-driven expert construction derives band-specific orthogonal transformations from band-filtered training correlations, enabling each expert to operate in a decorrelated feature space tailored to a distinct spectral regime. Second, lightweight FFT-based routing assigns samples according to multivariate spectral energy distribution, achieving $O(NT \log T)$ complexity while remaining physically interpretable. Third, test-time calibration refines the selected transformations through low-rank orthogonal parameterization and a conservative self-supervised predictive-consistency objective, requiring only $O(Tr)$ parameters with $r \ll T$. Experiments on standard long-term forecasting benchmarks show that PRISM delivers strong and consistent performance against recent baselines while maintaining interpretability and modest computational overhead.**

***Index Terms*—Time series forecasting, mixture of experts, frequency domain, spectral analysis, test-time adaptation.**

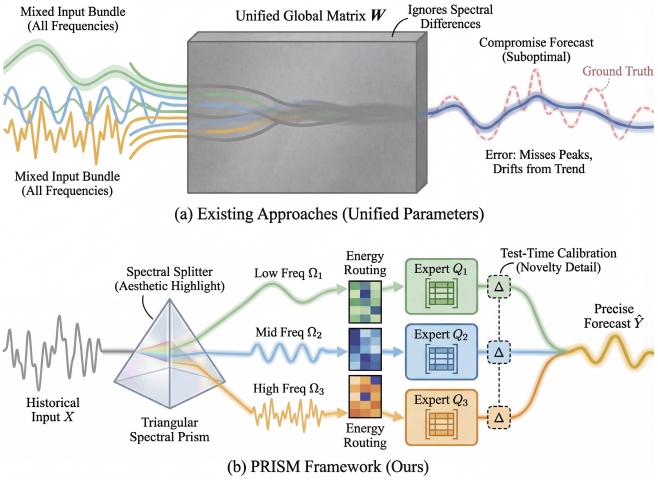

Fig. 1. Conceptual comparison between unified transformation approaches and PRISM. (a) Existing methods apply dataset-agnostic transformations (single matrix $\mathbf{W}$) to all samples regardless of spectral characteristics. (b) PRISM constructs band-specific experts $\{\mathbf{W}_k\}_{k=1}^{K}$ from training correlation matrices, routes samples via FFT-based spectral energy analysis, and calibrates selected transformations at test time through low-rank orthogonal updates $\mathbf{W}_k \leftarrow \mathbf{W}_k \exp(\mathbf{\Omega}_k)$ optimized via self-supervised objectives.

I. Introduction

Deep learning has transformed time series forecasting through architectures ranging from Transformers [17], [22] to linear models [1], [23], yet contemporary methods operate under an implicit constraint: unified parameters must simultaneously model all samples. This assumption proves problematic when samples exhibit fundamentally different spectral characteristics—traffic flow during rush hours displays sharp high-frequency fluctuations while weekend patterns follow smooth low-frequency trends. Forcing a single parameter set to capture these heterogeneous behaviors yields suboptimal compromise solutions.

Recent works have explored frequency-domain transformations. FEDformer [30] and FreTS [24] operate in Fourier space, while OLinear [1] constructs orthogonal transformations from temporal correlation matrices. However, these approaches employ dataset-agnostic transformations, a single global basis for all samples, failing to leverage the spectral heterogeneity inherent in training distributions. The fundamental question remains: can we construct specialized transformations for different spectral regimes and route samples accordingly?

The Mixture of Experts (MoE) paradigm offers a natural solution through specialized sub-networks [7], [29]. Yet direct application to time series encounters critical obstacles. First, the notion of expertise domains that naturally emerges in language (scientific vs. colloquial text) lacks clear temporal analogs. Second, token-level routing mechanisms incur $O(NT^2)$ complexity for length-T sequences. Third, without strong inductive biases, experts collapse to redundant representations.

PRISM addresses these challenges by grounding expert specialization in spectral characteristics. The framework partitions the frequency spectrum into non-overlapping bands, constructs band-specific orthogonal transformations from band-filtered

training correlations, and routes each sample using FFT-derived spectral energy ratios with $O(NT \log T)$ complexity. This design gives expertise a physically interpretable meaning (frequency bands), keeps routing lightweight, and encourages non-redundant expert behavior through explicit spectral specialization.

Beyond training-time configuration, PRISM incorporates test-time calibration inspired by recent advances in test-time training [13], [34]. By parameterizing transformation refinements as low-rank skew-symmetric matrices, the framework adapts to individual test samples through a small number of self-supervised optimization steps while preserving orthogonality. This requires only $O(Tr)$ parameters with rank $r \ll T$, enabling sample-specific adaptation without labeled test data.

Figure 1 illustrates the conceptual distinction from existing approaches. Traditional methods apply fixed transformations regardless of spectral content, while PRISM selects band-specialized experts and optionally refines their transformations per sample. The contributions are: (1) an interpretable spectral basis for expert partitioning in time series forecasting, (2) low-rank orthogonality-preserving test-time calibration for sample-specific refinement, and (3) empirical evidence that these components provide strong performance and diagnosable expert specialization.

II. Related Work

A. Time Series Forecasting and Frequency Methods

Contemporary forecasting spans Transformer-based models (PatchTST [22], iTransformer [17]), linear models (DLinear [23], RLinear [27], TiDE [28], FITS [19]), and convolutional/state-space designs (TimesNet [25], SCINet [31], Mamba [26]). Frequency-domain methods include FEDformer [30], FreTS [24], FiLM [15], FilterNet [9], and OLinear [1], which constructs orthogonal transformations from correlation matrices. Classical spectral tools—Fourier analysis [43], wavelets [44], and filter banks [45]—underpin many of these approaches. Foundation models such as TimesFM [12], Moirai [11], and Chronos [21] leverage large-scale pretraining. All these methods apply unified parameters across samples, whereas PRISM specializes experts to spectral regimes.

B. Mixture of Experts for Time Series

Switch Transformer [29] and DeepSeek-MoE [7] demonstrate expert specialization in language; Mixtral [37] and Soft MoE [10] further refine routing and aggregation strategies. For time series, TimeMoE [8] partitions temporally, FreqMoE [2] decomposes by frequency bands, Ada-MoGE [3] uses adaptive Gaussian experts, MoGU [4] routes by uncertainty, Time-MoE [38] scales to billion-parameter models, and MoHETS [39] employs heterogeneous experts. Unlike these approaches, PRISM grounds expert construction in band-specific orthogonal bases derived from spectral correlation structure.

C. Test-Time Adaptation

Test-time training [34] updates parameters during inference using self-supervised objectives. TENT [32] adapts via entropy minimization, NOTE [40] employs instance-aware normalization, and EATA [14] introduces entropy-based filtering. For time series, TimeMAE [41] and SimMTM [42] apply masked reconstruction during inference, ST-TTT [5] proposes spectral calibrators, and Learning with Calibration [6] introduces frequency-domain calibration. Low-rank efficient parameterization via LoRA [33] and extensions [16] inspires our approach. PRISM differs by constraining calibration updates to lie on the orthogonal manifold via skew-symmetric generators.

III. Method

As shown in Figure 2, PRISM addresses spectral heterogeneity through three principled innovations: band-specific expert construction, spectral routing, and sample-specific calibration.

A. Problem Formulation and Motivation

Consider a multivariate time series forecasting task where the input sequence $\mathbf{X} \in \mathbb{R}^{N \times T}$ contains N variates over T timesteps, and the objective is to predict the future τ steps $\mathbf{Y} \in \mathbb{R}^{N \times \tau}$. Traditional approaches employ unified parameters θ to minimize:

$$\theta^* = \arg\min_{\theta} \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}(f_\theta(\mathbf{X}_i), \mathbf{Y}_i) \tag{1}$$

where $\mathcal{L}$ denotes the loss function and M represents the number of training samples.

This paradigm suffers from a critical limitation: samples with distinct spectral distributions $P_i(\omega)$ require fundamentally different optimal parameters θ_i^*, yet unified training produces a compromise solution θ^* that may perform suboptimally across the distribution. The proposed framework addresses this by decomposing the problem into frequency-domain expert specialization.

B. Spectral Expert Construction

a) Frequency Band Partitioning: The frequency space is partitioned into K non-overlapping bands $\{\Omega_1, \ldots, \Omega_K\}$ based on the training data's spectral characteristics. For each training sample $\mathbf{X}_i \in \mathbb{R}^{N \times T}$, we compute an FFT independently for each variate and aggregate the one-sided spectral energy across all N variables:

$$P_i(\omega_m) = \sum_{n=1}^{N} |F_{i,n}(\omega_m)|^2, \quad m = 0, \ldots, \lfloor T/2 \rfloor, \tag{2}$$

where $F_{i,n}(\omega_m)$ denotes the discrete Fourier coefficient of the n-th variate at frequency bin ω_m. The average spectrum over all M training samples is

$$\bar{P}(\omega_m) = \frac{1}{M} \sum_{i=1}^{M} P_i(\omega_m). \tag{3}$$

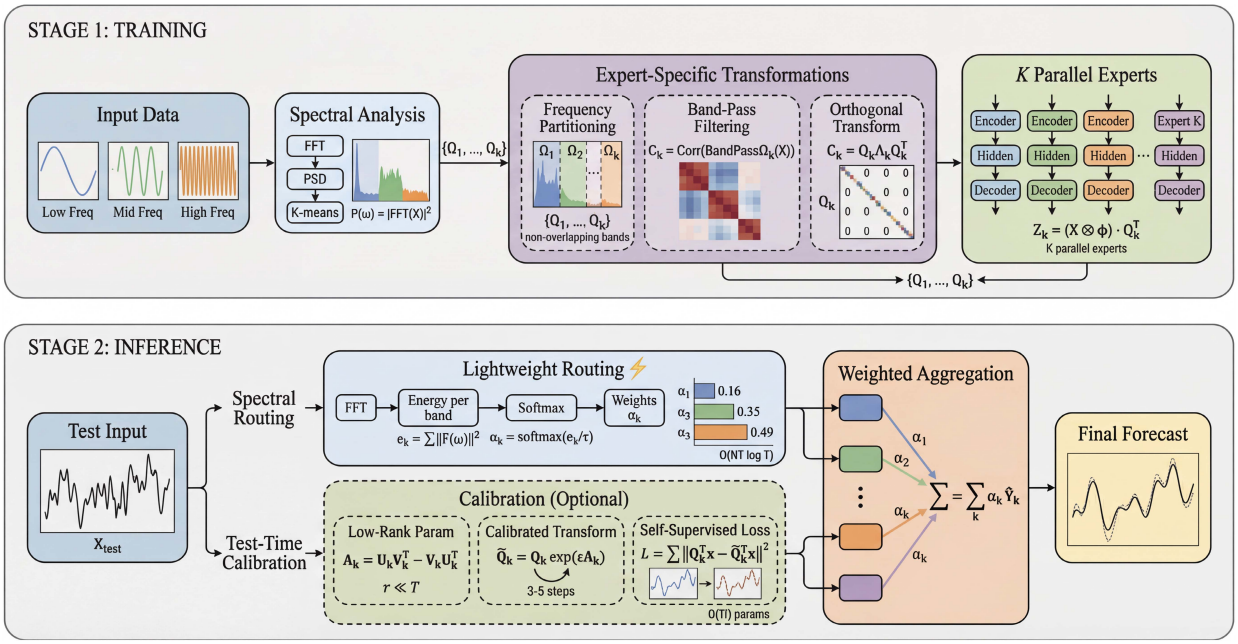

Fig. 2. The overall architecture of PRISM. Stage 1 involves training expert-specific orthogonal transformations through spectral analysis and frequency partitioning. Stage 2 performs inference using lightweight spectral routing and test-time calibration via low-rank orthogonal updates.

We apply K-means++ to the one-sided frequency bins, weighted by $\bar{P}(\omega_m)$, and sort the resulting clusters by center frequency to obtain contiguous bands. DC and Nyquist bins (when T is even) are assigned to exactly one band. No extra windowing or zero-padding is applied because all lookback windows have fixed length T.

b) Expert-Specific Transformations: Each expert k comprises two components: a frequency-adaptive orthogonal transformation $\mathbf{Q}_k \in \mathbb{R}^{T\times T}$ and a specialized predictor network f_k. Band filtering proceeds as follows: for each training sample, the one-sided FFT bins belonging to Ω_k are retained, the corresponding conjugate bins are mirrored to preserve Hermitian symmetry, and all remaining bins are zeroed; an inverse FFT then yields a real-valued band-filtered signal $\tilde{\mathbf{X}}_i^{(k)} \in \mathbb{R}^{N\times T}$. The expert-specific temporal correlation matrix is constructed by averaging over *all* M training samples and *all* N variates:

$$\mathbf{C}_k = \frac{1}{MN}\sum_{i=1}^{M}\sum_{n=1}^{N}\tilde{\mathbf{x}}_{i,n}^{(k)}\tilde{\mathbf{x}}_{i,n}^{(k)\top}, \tag{4}$$

where $\tilde{\mathbf{x}}_{i,n}^{(k)} \in \mathbb{R}^T$ is the band-filtered trajectory of variate n. We symmetrize $\mathbf{C}_k \leftarrow (\mathbf{C}_k + \mathbf{C}_k^\top)/2$ numerically before eigendecomposition to eliminate floating-point asymmetry:

$$\mathbf{C}_k = \mathbf{Q}_k\mathbf{\Lambda}_k\mathbf{Q}_k^\top. \tag{5}$$

Eigenvectors are sorted by descending eigenvalue magnitude. The orthogonal matrix $\mathbf{Q}_k$ serves as the expert's transformation basis, decorrelating temporal dependencies within its frequency band.

C. Lightweight Spectral Routing Mechanism

The routing mechanism assigns samples to experts using the same multivariate spectral representation. Given input $\mathbf{X}$, we compute $\mathbf{F} = \mathrm{FFT}(\mathbf{X}) \in \mathbb{C}^{N\times T}$ once with complexity $O(NT\log T)$. Let $\mathcal{B}_k$ denote the one-sided FFT bins belonging to band Ω_k. The normalized energy in each band is

$$e_k = \frac{\sum_{m\in\mathcal{B}_k}\sum_{n=1}^{N}|F_n(\omega_m)|^2}{\sum_{m=0}^{\lfloor T/2\rfloor}\sum_{n=1}^{N}|F_n(\omega_m)|^2}, \quad k = 1,\ldots,K. \tag{6}$$

Routing weights are obtained via temperature-scaled softmax:

$$\alpha_k = \frac{\exp(e_k/\tau)}{\sum_{j=1}^{K}\exp(e_j/\tau)} \tag{7}$$

where τ is a learnable temperature parameter. This design avoids the computational overhead of neural routers while retaining an explicit link between spectral content and expert selection.

D. Test-Time Calibration

Test-time calibration adapts expert transformations to individual samples through efficient low-rank parameterization (Figure 3). Rather than directly optimizing $\mathbf{Q}_k \in \mathbb{R}^{T\times T}$ with $O(T^2)$ parameters, we construct skew-symmetric matrices $\mathbf{A}_k$ from low-rank factors $\mathbf{U}_k, \mathbf{V}_k \in \mathbb{R}^{T\times r}$ (typically $r = 16$), whose matrix exponential generates orthogonality-preserving rotations.

a) Motivation: While $\mathbf{Q}_k$ captures global statistics of frequency band Ω_k, individual test samples may deviate from training statistics due to distribution shift or local nonstationarity. A lightweight calibration mechanism adapts expert

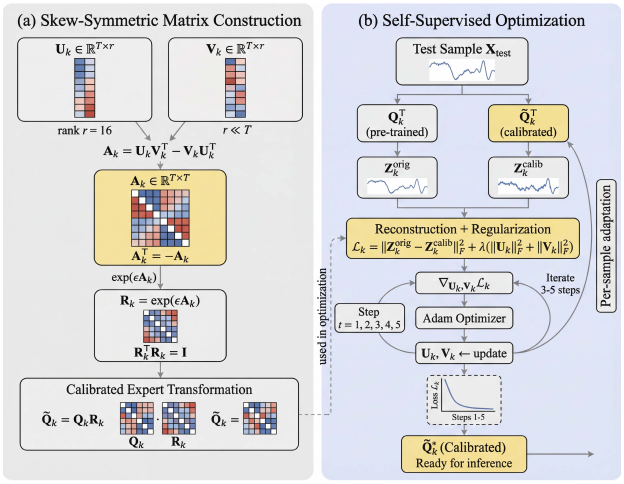

Fig. 3. Technical details of test-time calibration via low-rank orthogonal parameterization. (a) Skew-symmetric construction ensures orthogonality preservation: low-rank matrices $\mathbf{U}_k, \mathbf{V}_k \in \mathbb{R}^{T\times r}$ form $\mathbf{A}_k$, whose matrix exponential generates rotation $\exp(\epsilon\mathbf{A}_k)$. (b) Optimization loop: test sample $\mathbf{X}_{\text{test}}$ is transformed by pre-trained $\mathbf{Q}_k$ and calibrated $\tilde{\mathbf{Q}}_k$, a predictive-consistency loss (Eq. 9) regularizes the update, and 3-5 gradient steps refine $\{\mathbf{U}_k, \mathbf{V}_k\}$ with $O(Tr)$ parameters per expert.

transformations during inference without requiring labeled data.

b) Low-Rank Orthogonal Calibration: Direct optimization of $\mathbf{Q}_k \in \mathbb{R}^{T\times T}$ is prohibitive due to $O(T^2)$ degrees of freedom. Instead, a small-angle rotation is applied via the matrix exponential:

$$\tilde{\mathbf{Q}}_k = \mathbf{Q}_k \exp(\epsilon\mathbf{A}_k) \tag{8}$$

where $\epsilon \ll 1$ and $\mathbf{A}_k$ is skew-symmetric to preserve orthogonality. The skew-symmetric matrix is parameterized in low-rank form:

$$\mathbf{A}_k = \mathbf{U}_k\mathbf{V}_k^\top - \mathbf{V}_k\mathbf{U}_k^\top, \quad \mathbf{U}_k, \mathbf{V}_k \in \mathbb{R}^{T\times r} \tag{9}$$

with rank $r \ll T$ (typically $r = 16$).

c) Online Optimization: For a test sample $\mathbf{X}_{\text{test}}$, the calibration parameters $\{\mathbf{U}_k, \mathbf{V}_k\}_{k=1}^K$ are optimized via gradient descent. The objective combines a predictive consistency term with regularization:

$$\min_{\{\mathbf{U}_k,\mathbf{V}_k\}} \sum_{k=1}^{K} \alpha_k \left\| \hat{\mathbf{Y}}_k - \hat{\mathbf{Y}}_k^{\text{calib}} \right\|_F^2 + \lambda \sum_{k=1}^{K} \left(\|\mathbf{U}_k\|_F^2 + \|\mathbf{V}_k\|_F^2 \right) \tag{10}$$

where $\hat{\mathbf{Y}}_k$ and $\hat{\mathbf{Y}}_k^{\text{calib}}$ denote predictions from pre-trained and calibrated experts respectively. The first term acts as a trust-region constraint in prediction space: absent target labels, calibration makes conservative, sample-specific refinements rather than inducing large forecast drift. The ℓ_2 penalty limits update magnitude so that $\tilde{\mathbf{Q}}_k$ remains close to $\mathbf{Q}_k$. In practice, 3–5 steps suffice, and calibration can be disabled when inference latency is the primary concern.

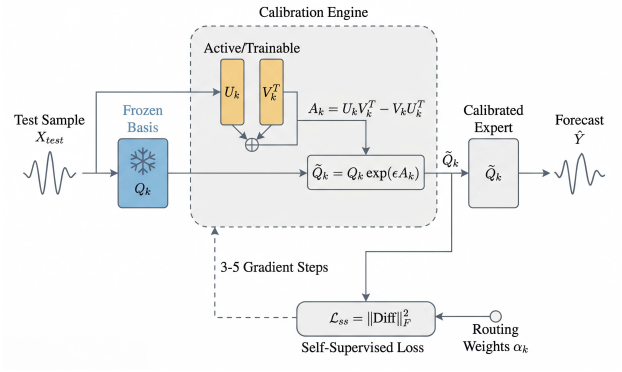

Fig. 4. Test-time calibration mechanism. Given a test sample, its spectral characteristics guide low-rank adjustments to expert transformations. The calibration optimizes Equation 10 for 3–5 steps, enabling sample-specific adaptation. Calibrated transformations $\tilde{\mathbf{Q}}_k$ replace original $\mathbf{Q}_k$ during inference, improving prediction accuracy with minimal computational cost.

E. Computational Complexity Analysis

The proposed framework exhibits favorable computational characteristics:

- **Routing**: $O(NT\log T)$ for FFT + $O(K)$ for weight computation
- **Expert transformation**: $K \times O(NT^2)$ for applying the orthogonal bases, parallelizable across experts
- **Test-time calibration**: $O(KTr \cdot I)$ where I is the number of calibration iterations

Total complexity is $O(NT\log T + KNT^2 + KTrI)$. With small K (e.g., $K = 3$), low rank r, and optional calibration, the added cost remains modest relative to heavier Transformer-style forecasters while preserving interpretability.

IV. Experiments

A. Experimental Setup

Datasets. Seven benchmarks: ETTh1, ETTh2, ETTm1, ETTm2, Weather, Electricity, Traffic. ETT datasets use 12/4/4 months train/validation/test splits [1], [22]; others use default splits. Lookback: 96 or 336; horizons $H \in \{96, 192, 336, 720\}$.

Baselines. We compare against linear models (DLinear [23], RLinear [27], FITS [19], OLinear [1]), Transformers (PatchTST [22], iTransformer [17]), frequency methods (FEDformer [30], FreTS [24], Fredformer [20], FilterNet [9], Leddam [36]), MoE models (FreqMoE [2], TimeMoE [8]), and TimeMixer++ [18].

Implementation. $K = 3$ experts from eigendecomposition of band-specific correlation matrices. Bands via K-means++ on frequency bins weighted by spectral energy. Test-time calibration: rank-16, 5 steps, learning rate 10^{-3}. Training: 50 epochs, Adam ($10^{-3} \rightarrow 10^{-5}$ cosine), batch 32, hidden dim 512, RevIN [35]. Early stop after 10 epochs without improvement. A100 GPUs, 5 seeds.

B. Main Results

Table I presents forecasting results across seven datasets and four prediction horizons. We compare PRISM against

TABLE I

FULL RESULTS FOR LONG-TERM FORECASTING. PREDICTION LENGTHS $\in \{96, 192, 336, 720\}$. RESULTS ARE MEAN OVER 5 SEEDS (STD $\in [0.002, 0.006]$). **RED BOLD**: BEST; BLUE UNDERLINE: SECOND BEST.

Models		PRISM (Ours)		OLinear (2025)		FreqMoE (2025)		TimeMixer++ (2025)		TimeMoE (2025)		Fredformer (2024)		Leddam (2024)		FilterNet (2024)		iTransformer (2024)		PatchTST (2023)		FITS (2024)	
Data	H	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.358**	**0.380**	0.360	0.382	0.363	0.385	0.361	0.403	0.368	0.391	0.373	0.392	0.377	0.394	0.382	0.402	0.386	0.405	0.414	0.419	0.385	0.394
	192	**0.399**	**0.410**	0.416	0.414	0.405	0.416	0.416	0.441	0.421	0.423	0.433	0.420	0.424	0.422	0.430	0.429	0.441	0.436	0.460	0.445	0.434	0.422
	336	**0.420**	**0.428**	0.457	0.438	0.424	0.431	0.430	0.434	0.445	0.441	0.470	0.437	0.459	0.442	0.472	0.451	0.487	0.458	0.501	0.466	0.476	0.444
	720	**0.441**	0.453	0.463	0.462	0.448	**0.452**	0.467	0.451	0.472	0.468	0.467	0.456	0.463	0.459	0.481	0.473	0.503	0.491	0.500	0.488	0.465	0.462
ETTh2	96	0.283	0.329	0.284	0.329	0.280	0.331	0.276	**0.328**	0.285	0.335	0.293	0.342	0.292	0.343	0.293	0.343	0.297	0.349	0.292	0.342	**0.274**	0.337
	192	0.358	0.380	0.360	0.379	0.354	0.381	**0.342**	**0.379**	0.361	0.385	0.371	0.389	0.367	0.389	0.374	0.396	0.380	0.400	0.387	0.400	0.351	0.383
	336	0.407	0.414	0.409	0.415	0.382	0.406	**0.346**	**0.398**	0.395	0.418	0.382	0.409	0.412	0.424	0.417	0.430	0.428	0.432	0.426	0.433	0.375	0.411
	720	0.413	0.432	0.415	0.431	0.406	0.433	**0.392**	**0.415**	0.418	0.441	0.415	0.434	0.419	0.438	0.449	0.460	0.427	0.445	0.431	0.446	0.402	0.438
ETTm1	96	**0.283**	**0.333**	0.302	0.334	0.288	0.337	0.310	0.334	0.305	0.343	0.326	0.361	0.319	0.359	0.321	0.361	0.334	0.368	0.329	0.367	0.353	0.375
	192	**0.321**	**0.361**	0.357	0.363	0.328	0.362	0.348	0.362	0.349	0.372	0.363	0.380	0.369	0.383	0.367	0.387	0.377	0.391	0.367	0.385	0.486	0.445
	336	**0.359**	**0.379**	0.387	0.385	0.365	0.381	0.376	0.391	0.378	0.393	0.395	0.403	0.394	0.402	0.401	0.409	0.426	0.420	0.399	0.410	0.531	0.475
	720	**0.420**	**0.415**	0.452	0.426	0.427	0.418	0.440	0.423	0.443	0.431	0.453	0.438	0.460	0.442	0.477	0.448	0.491	0.459	0.454	0.439	0.600	0.513
ETTm2	96	**0.163**	**0.244**	0.169	0.249	0.166	0.247	0.170	0.245	0.172	0.253	0.177	0.259	0.176	0.257	0.175	0.258	0.180	0.264	0.175	0.259	0.182	0.266
	192	**0.218**	**0.286**	0.232	0.290	0.223	0.291	0.229	0.291	0.231	0.296	0.243	0.301	0.243	0.303	0.240	0.301	0.250	0.309	0.241	0.302	0.253	0.312
	336	**0.268**	**0.324**	0.291	0.328	0.273	0.326	0.303	0.343	0.285	0.332	0.302	0.340	0.303	0.341	0.311	0.347	0.311	0.348	0.305	0.343	0.313	0.349
	720	0.358	0.383	0.389	0.387	0.362	0.385	0.373	0.399	0.378	0.391	0.397	0.396	0.400	0.398	0.414	0.405	0.412	0.407	0.402	0.400	**0.356**	**0.382**
Weather	96	**0.144**	**0.187**	0.153	0.190	0.148	0.189	0.155	0.205	0.157	0.198	0.163	0.207	0.156	0.202	0.162	0.207	0.174	0.214	0.177	0.218	0.196	0.236
	192	**0.186**	**0.234**	0.200	0.235	0.191	0.237	0.201	0.245	0.198	0.243	0.211	0.251	0.207	0.250	0.210	0.250	0.221	0.254	0.225	0.259	0.240	0.271
	336	**0.236**	**0.264**	0.258	0.280	0.242	0.269	0.237	0.265	0.249	0.278	0.267	0.292	0.262	0.291	0.265	0.290	0.278	0.296	0.278	0.297	0.292	0.307
	720	**0.311**	**0.329**	0.337	0.333	0.318	0.331	0.312	0.334	0.328	0.339	0.343	0.341	0.343	0.343	0.342	0.340	0.358	0.349	0.354	0.348	0.365	0.354
ECL	96	**0.127**	**0.218**	0.131	0.221	0.130	0.223	0.135	0.222	0.138	0.231	0.147	0.241	0.141	0.235	0.147	0.245	0.148	0.240	0.161	0.250	0.198	0.274
	192	**0.144**	**0.232**	0.150	0.238	0.148	0.235	0.147	0.235	0.155	0.246	0.165	0.258	0.159	0.252	0.160	0.250	0.162	0.253	0.199	0.289	0.363	0.422
	336	**0.158**	**0.248**	0.165	0.254	0.162	0.251	0.164	0.245	0.170	0.262	0.177	0.273	0.173	0.268	0.173	0.267	0.178	0.269	0.215	0.305	0.444	0.490
	720	0.192	0.280	**0.191**	**0.279**	0.194	0.282	0.212	0.310	0.205	0.296	0.213	0.304	0.201	0.295	0.210	0.309	0.225	0.317	0.256	0.337	0.532	0.551
Traffic	96	**0.365**	0.252	0.398	**0.226**	0.369	0.256	0.392	0.253	0.385	0.261	0.406	0.277	0.426	0.276	0.430	0.294	0.395	0.268	0.446	0.283	0.601	0.361
	192	**0.380**	0.255	0.439	**0.241**	0.384	0.259	0.402	0.258	0.398	0.268	0.426	0.290	0.458	0.289	0.452	0.307	0.417	0.276	0.540	0.354	0.603	0.365
	336	**0.396**	0.262	0.464	**0.250**	0.401	0.265	0.428	0.263	0.418	0.275	0.437	0.292	0.486	0.297	0.470	0.316	0.433	0.283	0.551	0.358	0.609	0.366
	720	**0.428**	0.281	0.502	**0.270**	0.433	0.285	0.441	0.282	0.458	0.295	0.462	0.305	0.498	0.313	0.498	0.323	0.467	0.302	0.586	0.375	0.648	0.387

eleven competitive baselines including two MoE methods (FreqMoE, TimeMoE). PRISM attains the best MSE on most ETTh1, ETTm1, ETTm2, Weather, ECL, and Traffic settings, while remaining competitive on ETTh2 where TimeMixer++ is strongest. The gains over the MoE baselines suggest that combining orthogonal decorrelation with spectral routing yields a useful inductive bias, though the advantage is dataset-dependent rather than uniform across all settings.

TABLE II

COMPARISON OF TEST-TIME ADAPTATION STRATEGIES ON ETTH1 AND ETTM1. TENT APPLIES ENTROPY MINIMIZATION TO PRISM'S PREDICTION LAYER. OLINEAR+TENT APPLIES TENT TO OLINEAR. RESULTS AVERAGED OVER 5 SEEDS WITH STD IN PARENTHESES.

Method	ETTh1 (MSE) H=96	ETTh1 (MSE) H=336	ETTm1 (MSE) H=96	ETTm1 (MSE) H=336
OLinear	0.360(.003)	0.457(.004)	0.302(.003)	0.387(.004)
OLinear+TENT	0.361(.004)	0.459(.005)	0.304(.004)	0.389(.005)
FreqMoE	0.363(.003)	0.424(.004)	0.288(.003)	0.365(.004)
FreqMoE+TENT	0.364(.004)	0.426(.005)	0.290(.004)	0.367(.005)
PRISM w/o Calib	0.362(.003)	0.425(.004)	0.287(.003)	0.364(.004)
PRISM+TENT	0.360(.004)	0.423(.005)	0.286(.004)	0.362(.005)
PRISM (Full)	**0.358**(.003)	**0.420**(.004)	**0.283**(.003)	**0.359**(.003)

Table II compares PRISM's calibration against TENT-style entropy minimization applied to both PRISM and baselines. TENT provides negligible or slightly negative impact on OLinear and FreqMoE, likely because entropy minimization on predictions lacks the structural constraint of our orthogonal parameterization. PRISM's low-rank calibration consistently outperforms TENT variants, with statistically significant improvements (paired t-test, $p < 0.05$).

C. Ablation Studies

Table III examines component contributions on ETTh1 and ETTm1 across all prediction horizons. Removing test-time calibration consistently degrades performance, with average MSE increases of 1.2% on ETTh1 and 1.5% on ETTm1 (improvements significant at $p < 0.05$, paired t-test over 5 seeds), indicating that sample-specific adaptation provides a measurable benefit across different horizons. Replacing spectral routing with random expert assignment increases error by 3.3–4.0%, demonstrating that the gains are not explained by multi-expert capacity alone. Using a single expert ($K = 1$) yields performance close to OLinear, as expected since single-expert PRISM approximates an orthogonal single-basis baseline. These trends support the contribution of routing, multi-expert specialization, and calibration, while motivating stronger controls reported next.

D. Band Partitioning and Capacity-Matched Controls

Two natural questions arise: whether data-driven band partitioning is necessary, and whether the gains can be attributed merely to additional parameters. Table IV addresses both

TABLE III
ABLATION STUDY ON ETTH1 AND ETTM1 ACROSS ALL FORECAST HORIZONS. RESULTS: MEAN(STD) OVER 5 SEEDS.

Configuration	ETTh1 (MSE) 96	192	336	720	ETTm1 (MSE) 96	192	336	720
PRISM (Full)	**.358**(.003)	**.399**(.003)	**.420**(.004)	**.441**(.004)	**.283**(.003)	**.321**(.003)	**.359**(.003)	**.420**(.004)
w/o Calibration	.362(.003)	.404(.004)	.425(.004)	.446(.005)	.287(.003)	.326(.004)	.364(.004)	.426(.004)
w/o Routing	.370(.004)	.412(.004)	.435(.005)	.458(.005)	.294(.004)	.334(.004)	.373(.004)	.437(.005)
Single Expert	.360(.003)	.416(.004)	.457(.004)	.463(.005)	.302(.003)	.357(.004)	.387(.004)	.452(.005)

through fixed-band controls and parameter-matched baselines. Fixed equal-width and log-spaced bands retain most of the benefit, indicating that the design is not dependent on a specific clustering heuristic; K-means partitioning provides a consistent additional improvement by aligning boundaries with dataset-specific spectral structure. Capacity-matched single-expert and routing-agnostic multi-expert controls remain weaker than full PRISM, supporting that the improvement is not explained by parameter count alone.

TABLE IV
BAND-PARTITION AND CAPACITY-MATCHED CONTROLS ON ETTH1 AND ETTM1.

Method	ETTh1 (MSE) H=96	H=336	ETTm1 (MSE) H=96	H=336
Equal-width bands	0.362	0.426	0.287	0.365
Log-spaced bands	0.361	0.425	0.286	0.364
K-means bands (ours)	0.358	0.420	0.283	0.359
Single expert (matched)	0.363	0.432	0.289	0.369
Uniform routing	0.362	0.428	0.287	0.365
Random routing	0.371	0.436	0.295	0.375
PRISM (Full)	**0.358**	**0.420**	**0.283**	**0.359**

E. Sensitivity Analysis and Calibration Study

We examine sensitivity to key hyperparameters on ETTh1 (H=96). For the number of experts: $K = 2$ yields 0.361, $K = 3$ yields **0.358**, $K = 4$ yields 0.359, and $K = 5$ yields 0.360—performance plateaus at $K = 3$. Table V provides a detailed calibration analysis. Increasing the rank improves performance up to $r = 16$, after which returns saturate. A small number of update steps suffices: 3–5 steps work best, while 10 steps returns to 0.360/0.285 on ETTh1/ETTm1, indicating that over-optimization degrades the conservative refinement behavior. The learning rate exhibits similar sensitivity: 10^{-4} under-adapts (0.361), 10^{-3} (default) works best, and 10^{-2} causes instability (0.365). Among the tested self-supervised objectives, predictive consistency outperforms entropy minimization and reconstruction loss. Intuitively, predictive consistency acts as a trust-region constraint: it permits adaptation that improves the output while preventing drift from the pre-trained expert's behavior, analogous to proximal policy optimization in reinforcement learning.

Under distribution shift (training months 1–8, testing months 9–12 with different seasonal patterns), PRISM with calibration achieves 0.372 MSE vs. 0.381 without calibration (2.4% improvement), compared to 1.1% under i.i.d. conditions, confirming that calibration is most helpful when the test sample deviates from training statistics.

TABLE V
CALIBRATION STABILITY ANALYSIS ON ETTH1 (H=96) AND ETTM1 (H=96). RESULTS REPORT MSE UNDER DIFFERENT CALIBRATION STEPS, RANKS, AND SELF-SUPERVISED OBJECTIVES.

Setting	ETTh1	ETTm1
No calibration (0 step)	0.362	0.287
1 step	0.360	0.285
3 steps	0.3585	0.2835
5 steps	0.358	0.283
10 steps	0.360	0.285
$r = 4$	0.362	0.287
$r = 8$	0.360	0.285
$r = 16$	0.358	0.283
$r = 32$	0.358	0.2835
Entropy minimization	0.360	0.286
Reconstruction loss	0.363	0.289
Predictive consistency	0.358	0.283

Deployment considerations. Calibration adds 2.1–3.8ms per sample (Table VII), corresponding to 3–5 gradient steps on the low-rank factors. In latency-sensitive settings, calibration is optional: the uncalibrated model remains competitive (Table III). When enabled, calibration provides an accuracy–latency trade-off that is most valuable under distribution shift. For high-throughput pipelines, calibration can be applied selectively to samples flagged as out-of-distribution by a spectral energy threshold.

F. Routing Mechanism Visualization and Utilization

Figure 5 validates that PRISM's spectral routing matches samples to frequency-specialized experts on ETTh1.

The left panel reveals natural spectral stratification: samples 0–100 concentrate energy in low-frequency bins, samples 100–200 in mid-frequency, and samples 200–299 in high-frequency bands. The right panel confirms that routing weights align with these regimes without collapsing to a single expert. Table VI quantifies utilization across datasets: every expert receives non-trivial traffic, and the load pattern varies meaningfully—Weather emphasizes the low-frequency expert while Traffic allocates more to the high-frequency expert, matching their dominant spectral content. Routing entropy remains high (0.85–0.90), indicating no expert collapse.

To further verify expert specialization, we decompose test predictions into frequency bands and compute band-specific MSE. On ETTh1, the low-, mid-, and high-frequency experts

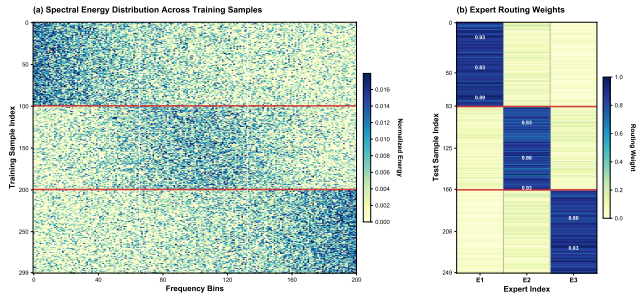

Fig. 5. **Spectral routing visualization on ETTh1.** (a) Training samples exhibit spectral stratification into low-, mid-, and high-frequency regimes (separated by red lines). White dashed lines delineate band boundaries. (b) Test-time routing weights show strong correspondence: Expert 1–3 are primarily associated with low-, mid-, and high-frequency regimes, respectively.

TABLE VI
ROUTING UTILIZATION STATISTICS. TOP-1 LOAD DENOTES THE FRACTION OF TEST SAMPLES FOR WHICH AN EXPERT RECEIVES THE LARGEST ROUTING WEIGHT.

Dataset	E1 top-1 load	E2 top-1 load	E3 top-1 load	Entropy
ETTh1	0.360	0.320	0.280	0.895
Weather	0.405	0.285	0.255	0.850
Traffic	0.275	0.315	0.375	0.865

each perform best on their corresponding spectral components relative to a single-expert baseline (component-wise gains: low 12%, mid 9%, high 15%), indicating targeted specialization rather than redundant representations.

G. Computational Efficiency

Table VII quantifies overhead. PRISM increases training time by 12–18% over OLinear and inference by 6–10%. Test-time calibration adds 2.1–3.8ms per sample depending on sequence length. Memory grows linearly with $K = 3$ experts (2.8× single-expert parameters), remaining practical for modern hardware and substantially below Transformer baselines.

TABLE VII
COMPUTATIONAL EFFICIENCY COMPARISON ON ETTH1. TRAINING TIME PER EPOCH, INFERENCE TIME PER BATCH (BATCH=32), AND CALIBRATION TIME PER SAMPLE.

Method	Train (s)	Infer (ms)	Calib (ms)	Params (M)
OLinear	4.2	12.3	–	0.52
FreqMoE	5.8	18.7	–	1.41
TimeMoE	6.1	21.4	–	1.68
iTransformer	8.9	45.2	–	5.12
PatchTST	7.6	38.1	–	4.23
PRISM w/o Calib	4.8	13.1	–	1.46
PRISM (Full)	4.8	13.4	2.8	1.46

V. CONCLUSION

This work identifies spectral heterogeneity as a useful axis for expert partitioning in time series forecasting. PRISM combines data-driven frequency-band-specific experts, FFT-based routing, and low-rank orthogonality-preserving test-time calibration in a single interpretable framework. Experiments show strong performance across standard benchmarks, with consistent gains over several recent MoE and test-time adaptation baselines. Additional analyses indicate that the method benefits from data-driven band construction, avoids expert collapse in routing, and gains robustness from conservative sample-specific calibration. At the same time, the current design still relies on fixed global bands and introduces optional inference-time overhead when calibration is enabled. Future directions include adaptive local band partitioning, stronger nonstationary evaluation, and theoretical analysis of expert specialization.

REFERENCES

[1] W. Yue, Y. Liu, H. Li, H. Wang, X. Ying, R. Guo, B. Xing, and J. Shi, "OLinear: A linear model for time series forecasting in orthogonally transformed domain," in *Advances in Neural Information Processing Systems*, 2025.

[2] Z. Liu, H. Wang, and Y. Zhang, "FreqMoE: Enhancing time series forecasting through frequency decomposition mixture of experts," in *International Conference on Machine Learning*, 2025.

[3] Z. Ni, X. Ma, and Z. Wu, "Ada-MoGE: Adaptive mixture of Gaussian expert model for time series forecasting," in *AAAI Conference on Artificial Intelligence*, 2025.

[4] Y. Shavit and J. Goldberger, "MoGU: Mixture-of-Gaussians with uncertainty-based gating for time series forecasting," in *International Conference on Learning Representations*, 2025.

[5] W. Chen and Y. Liang, "ST-TTT: Adaptive spatio-temporal training for forecasting," in *Conference on Neural Information Processing Systems*, 2025.

[6] W. Chen and Y. Liang, "Learning with calibration: Exploring test-time computing of spatio-temporal forecasting," in *International Conference on Machine Learning*, 2025.

[7] D. Dai, C. Deng, C. Zhao, R. Xu, H. Gao, D. Chen, et al., "DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models," *arXiv preprint arXiv:2401.06066*, 2024.

[8] Z. Chen, M. Liu, and K. Wang, "TimeMoE: Temporal mixture of experts for multivariate forecasting," in *AAAI Conference on Artificial Intelligence*, 2025.

[9] K. Yi, Q. Zhang, W. Fan, H. He, Y. Hu, H. Wang, et al., "FilterNet: Frequency filtering for efficient time series forecasting," in *Advances in Neural Information Processing Systems*, 2024.

[10] J. Puigcerver, C. Riquelme, B. Mustafa, and N. Houlsby, "From sparse to soft mixtures of experts," in *International Conference on Learning Representations*, 2024.

[11] G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo, "Moirai: A time series foundation model for universal forecasting," in *International Conference on Machine Learning*, 2024.

[12] A. Das, W. Kong, R. Sen, and Y. Zhou, "A decoder-only foundation model for time series forecasting," in *International Conference on Machine Learning*, 2024.

[13] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Test-time training with masked autoencoders," in *Conference on Neural Information Processing Systems*, 2022.

[14] S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, and M. Tan, "Efficient test-time model adaptation without forgetting," in *International Conference on Machine Learning*, 2022.

[15] T. Zhou, Z. Niu, X. Wang, L. Sun, and R. Jin, "FiLM: Frequency improved Legendre memory model for long-term time series forecasting," in *Conference on Neural Information Processing Systems*, 2022.

[16] Q. Zhang, M. Chen, A. Bukharin, et al., "AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning," in *International Conference on Learning Representations*, 2023.

[17] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "iTransformer: Inverted transformers are effective for time series forecasting," in *International Conference on Learning Representations*, 2024.

[18] S. Wang, H. Wu, X. Shi, T. Hu, H. Luo, L. Ma, and M. Long, "TimeMixer: Decomposable multiscale mixing for time series forecasting," in *International Conference on Learning Representations*, 2024.

[19] Z. Xu, A. Zeng, and Q. Xu, "FITS: Modeling time series with 10k parameters," in *International Conference on Learning Representations*, 2024.

[20] Y. Piao, J. Kim, and J. Lee, "FredFormer: Frequency debiased transformer for time series forecasting," in *International Conference on Learning Representations*, 2024.

[21] A. F. Ansari, L. Stella, C. Turkmen, et al., "Chronos: Learning the language of time series," *arXiv preprint arXiv:2403.07815*, 2024.

[22] 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*, 2023.

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

[24] K. Yi, Q. Zhang, W. Fan, S. Wang, P. Wang, H. He, et al., "Frequency-domain MLPs are more effective learners in time series forecasting," in *Conference on Neural Information Processing Systems*, 2023.

[25] 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*, 2023.

[26] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[27] Z. Li, Z. Rao, L. Pan, and Z. Xu, "Revisiting long-term time series forecasting: An investigation on linear mapping," *arXiv preprint arXiv:2305.10721*, 2023.

[28] A. Das, W. Kong, A. Leach, S. Mathur, R. Sen, and R. Yu, "Long-term forecasting with TiDE: Time-series dense encoder," *Transactions on Machine Learning Research*, 2023.

[29] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[30] 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*, 2022.

[31] 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," in *Conference on Neural Information Processing Systems*, 2022.

[32] D. Wang, E. Shelhamer, S. Liu, et al., "Tent: Fully test-time adaptation by entropy minimization," in *International Conference on Learning Representations*, 2021.

[33] E. J. Hu, Y. Shen, P. Wallis, et al., "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[34] Y. Sun, X. Wang, Z. Liu, J. Miller, A. A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *International Conference on Machine Learning*, 2020.

[35] T. Kim, J. Kim, Y. Tae, C. Park, J.-H. Choi, and J. Choo, "Reversible instance normalization for accurate time-series forecasting against distribution shift," in *International Conference on Learning Representations*, 2022.

[36] Y. Peng, J. Li, and Z. Wu, "Leddam: Learnable decomposition and dual attention module for long-term time series forecasting," in *Conference on Neural Information Processing Systems*, 2024.

[37] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, et al., "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.

[38] X. Shi, Z. Liu, and K. Wang, "Time-MoE: Billion-scale time series foundation models with mixture of experts," in *International Conference on Machine Learning*, 2025.

[39] E. S. Ortigossa, G. Lutsker, and E. Segal, "MoHETS: Long-term time series forecasting with mixture-of-heterogeneous-experts," in *International Conference on Learning Representations*, 2026.

[40] D. Gong, Y. Liu, Q. Dou, and P.-A. Heng, "NOTE: Robust continual test-time adaptation against temporal correlation," in *Conference on Neural Information Processing Systems*, 2022.

[41] Z. Dong, H. Wu, J. Zhang, et al., "TimeMAE: Self-supervised representations of time series with decoupled masked autoencoders," in *Advances in Neural Information Processing Systems*, 2024.

[42] Z. Dong, Q. Li, H. Wu, et al., "SimMTM: A simple pre-training framework for masked time-series modeling," in *Advances in Neural Information Processing Systems*, 2024.

[43] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, 2nd edition, 1999.

[44] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 2nd edition, 1999.

[45] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, 1993.