Macro-Meso-Micro: A Hierarchical Scale Disentanglement Framework for Multiscale Dynamic Time Series Forecasting

1st Yu Wang
Chengdu Institute of Computer Applications, Chinese Academy of Sciences
School of Computer Science and Technology, University of Chinese Academy of Sciences
Chengdu, China
wy_20212021@163.com

2nd Xiaolin Qin*
Chengdu Institute of Computer Applications, Chinese Academy of Sciences
School of Computer Science and Technology, University of Chinese Academy of Sciences
Chengdu, China
qinxl2001@126.com

3rd Jiacen Liu
Chengdu Institute of Computer Applications, Chinese Academy of Sciences
School of Computer Science and Technology, University of Chinese Academy of Sciences
Chengdu, China
dreakliu123552@163.com

***Abstract*—Real-world time series are often composed of complex, entangled nonstationary patterns and multiscale dynamics, posing significant challenges for forecasting. Existing homogeneous architectures are constrained by the Single-Mapping Bottleneck, failing to simultaneously capture differentiated dynamic characteristics. To address this, we propose Macro-Meso-Micro Hierarchical Scale Disentanglement (M^3-HSD), a hierarchical scale disentanglement framework employing a Macro-Meso-Micro strategy. The core innovation lies in the Hierarchical Scale Disentangler (HSD), which orthogonally decomposes the series into macro-inertia, meso-manifolds, and micro-transients. To model these distinct dynamics, we design a Heterogeneous Expert Decoding Module (HEDM): (1) The Global Inertia Projector (GIP) leverages the global receptive field of MLPs to capture low-frequency inertial trends; (2) The Adaptive Manifold Fitter (AMF) innovatively incorporates Kolmogorov-Arnold Networks (KAN) convolutions, utilizing learnable spline functions to precisely fit complex nonlinear seasonal manifolds at the meso-scale; and (3) The Local Transient Encoder (LTE) employs multiscale CNNs to capture high-frequency micro-fluctuations. Finally, an Adaptive Gating Fusion mechanism dynamically integrates multistream outputs to enhance robustness. Extensive experiments on mainstream benchmarks demonstrate that M^3-HSD significantly outperforms existing state-of-the-art (SOTA) methods, exhibiting superior performance, particularly in handling multiscale complex periodic patterns and non-stationary distributions. The source code is available at: https://github.com/WYyeshang/M3-HSD.**

Keywords*—*Macro-Meso-Micro, Hierarchical Scale Disentangler, Heterogeneous Expert Decoding, Kolmogorov-Arnold Networks

I. Introduction

Multivariate time series (MTS) forecasting plays a pivotal role [1]-[4] in modern intelligent systems, with broad applications ranging from energy management [5]-[6], traffic planning [7]-[12], weather forecasting [13] and financial analysis [14]-[15]. However, MTS forecasting remains a challenging problem, primarily because real-world data often exhibit strong nonstationarity and intricate multiscale dynamics. Unlike stationary signals in idealized environments, real-world data are typically an entangled mixture of long-term trends, periodic oscillations, and instantaneous stochastic noise [16].

This work was supported by the National Key R&D Program of China under Grant 2023YFB3308601, the Sichuan Science and Technology Program under Grant 2024NSFSC0004, and the Talents by Sichuan Provincial Party Committee Organization Department.

*Corresponding author.

Deep learning models—including Transformer-based [17] architectures like Autoformer [18] and PatchTST [19], MLP-based baselines such as DLinear [20], have significantly advanced state-of-the-art (SOTA) results. Most follow a decomposition-prediction paradigm to mitigate distribution shifts. However, existing models suffer from limitations inherent in their homogeneous architecture, typically attempting to fit distinct physical components through a single mapping mechanism. This one-size-fits-all strategy ignores the physical essence of temporal dynamics: macroscopic trends require linear extrapolation for global inertia; mesoscopic seasonality needs manifold fitting for nonlinear structures; and microscopic residuals require local capture for instantaneous impulses. Forcing homogeneous networks to model such heterogeneous dynamics creates a Single-Mapping Bottleneck, limiting the balance between global robustness and local precision.

To overcome these limitations, we propose Macro-Meso-Micro Hierarchical Scale Disentanglement (M^3-HSD), a hierarchical scale disentanglement framework for multiscale dynamic time series forecasting. Our model employs a Macro-Meso-Micro hierarchical decoupling strategy. Unlike traditional binary decomposition, M^3-HSD utilizes a Hierarchical Scale Disentangler (HSD) to peel off the series layer-by-layer into three dynamic scales: macroscopic inertia, mesoscopic manifolds, and microscopic transients. On this basis, we design a Heterogeneous Expert Decoding Module (HEDM), which assigns three heterogeneous experts—the Global Inertia Projector (GIP), the Adaptive Manifold Fitter (AMF), and the Local Transient Encoder (LTE)—to handle specific physical components respectively, achieving specialized modeling for diverse dynamics. Our contributions are as follows:

- We propose a novel framework named M^3-HSD, which introduces a Macro-Meso-Micro hierarchical decoupling strategy. This strategy transcends the limitations of traditional binary decomposition, enabling clear disentanglement of complex nonstationary temporal dynamics from a physical semantic level.
- We design the HEDM to break the Single-Mapping Bottleneck. This module creatively integrates a global MLP-based GIP, a KAN-based AMF, and a CNN-based LTE, achieving adaptive modeling of multiscale physical features.
- We conduct extensive experiments on multiple real-world benchmark datasets. The results demonstrate

that M^3-HSD significantly outperforms existing state-of-the-art methods in terms of both forecasting accuracy and robustness, especially when handling data with complex periodic patterns.

II. RELATED WORK

A. Transformer-based Homogeneous Decomposition Models

Transformers dominate time series forecasting by modeling long-range dependencies. Informer [21] pioneered ProbSparse attention to reduce complexity, while Autoformer [18] and FEDformer [22] utilized deep decomposition and frequency-domain transforms. PatchTST [19] introduced patching and channel independence, and iTransformer [23] proposed an inverted dimension strategy for multivariate correlations. Recently, STNet [24] integrated spatial-temporal graphs. However, these methods suffer from architectural homogeneity, failing to adapt to smooth macroscopic trends and oscillating seasonality. Moreover, their coarse-grained binary decomposition often conflates noise with complex seasonality, limiting microscopic dynamic perception.

B. MLP-based and CNN-based Homogeneous Decomposition Models

To enhance efficiency, lightweight architectures have gained traction. DLinear [20] uses a dual-stream linear decomposition with homogeneous heads, while TimeMixer [25] employs multiscale mixing for pattern extraction. In the CNN domain, ModernTCN [26] and MICN [27] utilize large-kernel or multiscale isometric convolutions to capture correlations. Despite lower complexity, these models suffer from architectural homogeneity. For instance, DLinear fits nonlinear remainders with rigid linear layers, and CNN models apply fixed kernels across all scales. This Single-Mapping Bottleneck limits their representation capability for complex nonstationary data.

C. From Binary-Homogeneous to Ternary-Heterogeneous: Our Work

To bridge these gaps, we propose the M^3-HSD framework, marking an evolution in two dimensions:

1) From Binary Decomposition to Ternary Disentanglement: We transcend the traditional binary mode by upgrading to the Macro-Meso-Micro hierarchical decoupling strategy. This approach enables the clear disentanglement of temporal dynamics into macroscopic inertia, mesoscopic manifolds, and microscopic transients, achieving orthogonal separation at the physical semantic level.

2) From Homogeneous to Heterogeneous Experts: We break the single-architecture limitation by proposing the HEDM. Unlike aforementioned works utilizing homogeneous networks, we match distinct architectures to the inductive bias of each component: integrating the GIP based on MLP for macroscopic trends, and the LTE based on CNN for microscopic noise. Crucially, for the complex mesoscopic seasonality, we draw inspiration from the recently proposed Kolmogorov-Arnold Networks (KAN) [28], which utilize learnable activation functions on edges to achieve superior nonlinear fitting. Building upon the convolutional adaptations explored in U-KAN [29] for computer vision, we design a tailored lightweight KAN-Conv1d operator specifically for time series to construct the AMF. This heterogeneous design ensures that each physical component is processed by the most mathematically suitable architecture.

III. METHODOLOGY

In this section, we elaborate on the proposed M^3-HSD framework. As illustrated in Fig. 1, the model consists of three key components: the HSD for multiscale decomposition, the HEDM for specialized modeling, and the Adaptive Gating Fusion mechanism for dynamic integration.

A. Problem Definition

Let $\mathbf{X}=\{x_1,\ldots,x_L\} \in \mathbb{R}^{L\times C}$ denote the historical multivariate time series with a look-back window size L and C variates. The forecasting task aims to predict the future sequence $\mathbf{Y}=\{x_{L+1},\ldots,x_{L+T}\} \in \mathbb{R}^{T\times C}$, where T is the prediction horizon. Our objective is to learn a mapping function $\mathcal{F}:\mathbf{X}\rightarrow\widehat{\mathbf{Y}}$ to minimize the mean squared error between the prediction $\widehat{\mathbf{Y}}$ and the ground truth $\mathbf{Y}$.

B. Channel Independence & Normalization

Real-world data often exhibit nonstationary distribution shifts. To address this, we first employ Reversible Instance Normalization (RevIN) to normalize the input data. Subsequently, we adopt the Channel Independence (CI) strategy. Specifically, the multivariate series $\mathbf{X}$ is reshaped into C independent univariate series $\mathbf{X}^{(i)} \in \mathbb{R}^{1\times L}$. This strategy not only augments the training samples but also prevents overfitting caused by inter-variate heterogeneity, allowing our model M^3-HSD to focus on disentangling temporal dynamics shared across variables.

C. Hierarchical Scale Disentangler

Unlike traditional decomposition methods that rely on rigid window sizes, we propose the HSD module based on cascaded Exponential Moving Averages (EMA) to peel off dynamic components layer by layer.

First, we extract the Macroscopic Scale $\mathbf{X}_{\text{macro}}$, which represents the global inertial drift of the series. For time step t, it is computed via a recursive formulation:

$$\mathbf{X}_{\text{macro}}^{(t)}=\begin{cases}\mathbf{x}^{(1)}, & t=1\\ \alpha_1\mathbf{x}^{(t)}+(1\text{-}\alpha_1)\mathbf{X}_{\text{macro}}^{(t\text{-}1)}, & t>1\end{cases} \tag{1}$$

where $\alpha_1 \in (0,1)$ is the smoothing factor controlling the trend responsiveness. We then subtract the macroscopic component to obtain the remainder $\mathbf{R}_1=\mathbf{X}\text{-}\mathbf{X}_{\text{macro}}$.

Next, to capture the Mesoscopic Scale, which represents structured periodic resonance, we apply a second EMA filter with a distinct factor α_2 to the remainder $\mathbf{R}_1$:

$$\mathbf{X}_{\text{meso}}=\text{EMA}_{\alpha_2}(\mathbf{X}\text{-}\mathbf{X}_{\text{macro}}) \tag{2}$$

It is worth noting that the smoothing factors α_1 and α_2 effectively act as frequency cut-offs for the macro and meso scales, respectively. In our experiments, these parameters are empirically determined via grid search on the validation set. Typically, we set $\alpha_1 \in [0.001,0.01]$ to capture the slow-moving global inertia, and $\alpha_2 \in [0.5,0.6]$ to isolate the meso-scale manifolds. This data-driven selection ensures an effective and dataset-specific structural disentanglement.

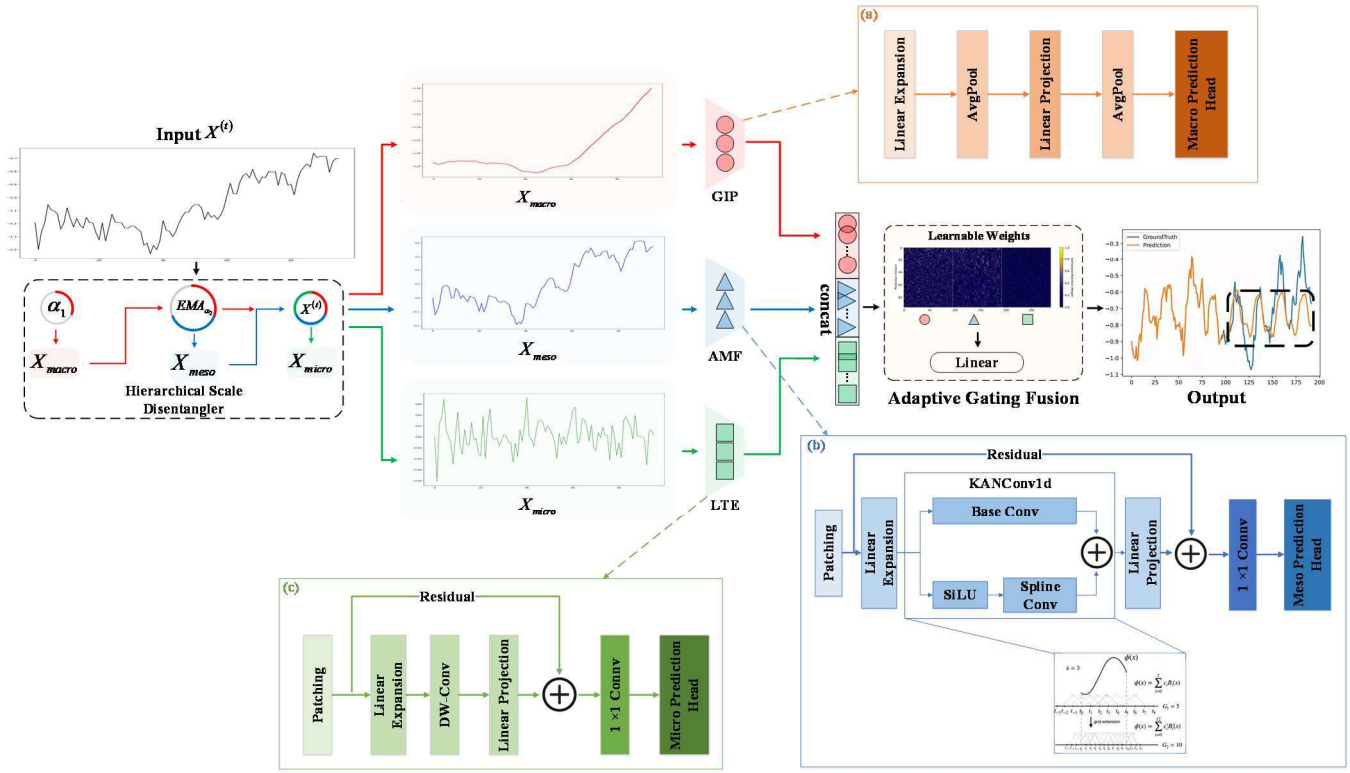

Fig. 1. Overall architecture of M^3-HSD. (a) The network architecture of GIP. (b) The network architecture of AMF. (c) The network architecture of LTE.

Finally, the remaining information constitutes the Microscopic Scale $\mathbf{X}_{\text{micro}}$, containing instantaneous stochastic transients and high-frequency noise:

$$\mathbf{X}_{\text{micro}}=\mathbf{X}-\mathbf{X}_{\text{macro}}-\mathbf{X}_{\text{meso}} \tag{3}$$

This hierarchical subtraction mechanism ensures clear physical semantic disentanglement and effectively prevents information leakage between scales.

D. Heterogeneous Expert Decoding Module

To break the Single-Mapping Bottleneck inherent in homogeneous architectures, the HEDM assigns heterogeneous neural experts to model the three disentangled components, respectively.

1) Global Inertia Projector: Macroscopic trends typically exhibit smooth, continuous trajectories. We employ an MLP as the global projector. Utilizing the global receptive field of fully connected layers, the MLP effectively performs linear extrapolation for the trend component:

$$\hat{\mathbf{Y}}_{\text{macro}}=\mathbf{W}_{p2}(\sigma(\mathbf{W}_{p1}\mathbf{X}_{\text{macro}})) \tag{4}$$

where $\mathbf{W}$ denotes linear weights and σ is the activation function.

2) Adaptive Manifold Fitter: The mesoscopic scale contains complex, nonlinear periodic patterns. To address the limitation of standard convolutions in fitting these geometric manifolds, we design a Lightweight KAN-Conv1d operator. Inspired by KAN, we replace the static scalar multiplication in standard convolution with learnable nonlinear activations. To ensure linear computational complexity $\mathcal{O}(L)$, we adopt a basis function approximation strategy:

$$\text{KAN-Conv}(x)=\mathbf{W}_{\text{base}}*\mathbf{x}+\mathbf{W}_{\text{spline}}*\phi(x) \tag{5}$$

where $*$ denotes the convolution operation, $\phi(x)=\text{SiLU}(x)$ serves as the efficient approximation basis for spline functions, and $\mathbf{W}_{\text{base}}$ and $\mathbf{W}_{\text{spline}}$ represent the base and spline convolution kernels, respectively. Based on this operator, the mesoscopic prediction is formulated as:

$$\hat{\mathbf{Y}}_{\text{meso}}=\text{KAN-Layers}(\mathbf{X}_{\text{meso}}) \tag{6}$$

This design enables the network to adaptively learn the optimal waveform shape for periodic patterns.

3) Local Transient Encoder: The microscopic scale is dominated by high-frequency noise and local impulses. We employ a standard 1D-CNN to leverage its translation invariance. This allows the model to acutely capture local variations without being distracted by global trends:

$$\hat{\mathbf{Y}}_{\text{micro}}=\text{CNN}(\mathbf{X}_{\text{micro}}) \tag{7}$$

E. Adaptive Gating Fusion

Traditional methods often employ simple summation for reconstruction, which assumes perfect decomposition. To enhance fault tolerance against decomposition errors, we introduce an Adaptive Gating Fusion mechanism. We concatenate the outputs from all streams and project them through a learnable linear gating layer:

$$\hat{\mathbf{Y}}=\mathbf{W}_{\text{gate}}\left(\text{Concat}[\hat{\mathbf{Y}}_{\text{macro}},\hat{\mathbf{Y}}_{\text{meso}},\hat{\mathbf{Y}}_{\text{micro}}]\right)+\mathbf{b}_{\text{gate}} \tag{8}$$

This mechanism allows the model to dynamically reweight the contribution of each stream based on data characteristics. Finally, the prediction is denormalized via RevIN to obtain the final output.

TABLE 1
FULL RESULTS OF LONG-TERM MULTIVARIATE TIME SERIES FORECASTING TASKS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD AND THE SECOND BEST RESULTS ARE UNDERLINED.

Models		M^3-HSD		TFPS		iTransformer		TimeMixer		PatchTST		MICN		TimeNet		DLinear	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.382	0.393	0.398	0.413	0.386	0.405	0.375	0.400	0.414	0.419	0.421	0.431	0.384	0.402	0.386	0.400
	192	0.440	0.424	0.423	0.423	0.441	0.436	0.429	0.421	0.460	0.445	0.474	0.487	0.436	0.429	0.437	0.432
	336	0.489	0.447	0.484	0.461	0.487	0.458	0.484	0.458	0.501	0.466	0.569	0.551	0.491	0.469	0.481	0.459
	720	0.538	0.492	0.488	0.476	0.503	0.491	0.498	0.482	0.500	0.488	0.770	0.672	0.521	0.500	0.519	0.516
	Avg	0.462	0.439	0.448	0.443	0.454	0.447	0.446	0.440	0.468	0.454	0.558	0.535	0.458	0.450	0.455	0.451
ETTh2	96	0.233	0.298	0.313	0.355	0.297	0.349	0.289	0.341	0.302	0.348	0.299	0.364	0.340	0.374	0.333	0.387
	192	0.298	0.338	0.405	0.410	0.380	0.400	0.372	0.392	0.388	0.400	0.441	0.454	0.402	0.414	0.477	0.476
	336	0.348	0.378	0.392	0.415	0.428	0.432	0.386	0.414	0.426	0.433	0.654	0.567	0.452	0.452	0.594	0.541
	720	0.406	0.426	0.410	0.433	0.427	0.445	0.412	0.434	0.431	0.446	0.956	0.716	0.462	0.468	0.831	0.657
	Avg	0.321	0.360	0.380	0.403	0.383	0.406	0.364	0.395	0.386	0.406	0.587	0.525	0.414	0.427	0.558	0.515
ETTm1	96	0.307	0.343	0.327	0.367	0.334	0.368	0.320	0.357	0.329	0.367	0.316	0.362	0.338	0.375	0.345	0.372
	192	0.348	0.369	0.374	0.395	0.377	0.391	0.361	0.381	0.367	0.385	0.363	0.390	0.374	0.387	0.380	0.389
	336	0.382	0.391	0.401	0.408	0.426	0.420	0.390	0.404	0.399	0.410	0.408	0.426	0.410	0.411	0.413	0.413
	720	0.458	0.429	0.479	0.456	0.491	0.459	0.454	0.441	0.454	0.439	0.481	0.476	0.478	0.450	0.474	0.453
	Avg	0.373	0.383	0.395	0.406	0.407	0.409	0.381	0.395	0.387	0.400	0.392	0.413	0.400	0.405	0.403	0.406
ETTm2	96	0.163	0.246	0.170	0.255	0.180	0.264	0.175	0.258	0.175	0.259	0.179	0.275	0.187	0.267	0.193	0.292
	192	0.228	0.290	0.235	0.296	0.250	0.309	0.237	0.299	0.241	0.302	0.307	0.376	0.249	0.309	0.284	0.362
	336	0.289	0.330	0.297	0.335	0.311	0.348	0.298	0.340	0.305	0.343	0.325	0.388	0.321	0.351	0.369	0.427
	720	0.376	0.382	0.401	0.397	0.412	0.407	0.391	0.396	0.402	0.400	0.502	0.49	0.408	0.403	0.554	0.522
	Avg	0.264	0.312	0.275	0.320	0.288	0.332	0.275	0.323	0.280	0.326	0.328	0.382	0.291	0.332	0.350	0.400
Weather	96	0.166	0.200	0.154	0.202	0.174	0.214	0.163	0.209	0.177	0.218	0.161	0.229	0.172	0.220	0.196	0.255
	192	0.211	0.242	0.205	0.249	0.221	0.254	0.208	0.250	0.225	0.259	0.220	0.281	0.219	0.261	0.237	0.296
	336	0.235	0.271	0.262	0.289	0.278	0.296	0.251	0.287	0.278	0.297	0.278	0.331	0.280	0.306	0.283	0.335
	720	0.308	0.321	0.344	0.342	0.358	0.347	0.339	0.341	0.354	0.348	0.311	0.356	0.365	0.359	0.345	0.381
	Avg	0.230	0.258	0.241	0.270	0.257	0.277	0.240	0.271	0.258	0.280	0.242	0.299	0.259	0.286	0.265	0.316
Electricity	96	0.162	0.245	0.149	0.236	0.148	0.240	0.153	0.247	0.181	0.270	0.164	0.269	0.168	0.272	0.197	0.282
	192	0.163	0.251	0.162	0.253	0.162	0.253	0.166	0.256	0.188	0.274	0.177	0.285	0.184	0.289	0.196	0.285
	336	0.184	0.268	0.200	0.310	0.178	0.269	0.185	0.277	0.204	0.293	0.193	0.304	0.198	0.300	0.209	0.301
	720	0.222	0.301	0.220	0.320	0.225	0.317	0.225	0.310	0.246	0.324	0.212	0.321	0.220	0.320	0.245	0.333
	Avg	0.182	0.266	0.182	0.279	0.178	0.269	0.182	0.272	0.204	0.290	0.186	0.294	0.192	0.295	0.211	0.300
Traffic	96	0.480	0.278	0.427	0.296	0.395	0.268	0.462	0.285	0.462	0.295	0.519	0.309	0.593	0.321	0.650	0.396
	192	0.480	0.275	0.445	0.298	0.417	0.276	0.473	0.296	0.466	0.296	0.537	0.315	0.617	0.336	0.598	0.370
	336	0.497	0.281	0.459	0.307	0.433	0.283	0.498	0.296	0.482	0.304	0.534	0.313	0.629	0.336	0.605	0.373
	720	0.531	0.299	0.496	0.313	0.467	0.302	0.506	0.313	0.514	0.322	0.577	0.325	0.640	0.350	0.645	0.394
	Avg	0.497	0.283	0.456	0.303	0.428	0.282	0.484	0.297	0.481	0.304	0.541	0.315	0.619	0.335	0.624	0.383
1st Count		13	24	2	2	7	2	4	1	0	0	1	0	0	0	1	0

IV. EXPERIMENTS

A. Experimental Setup

1) Datasets: To verify the effectiveness of our proposed method, we evaluated its performance on seven real-world time series forecasting benchmarks, including the ETT series (ETTh1, ETTh2, ETTm1, ETTm2), Traffic, Electricity, and Weather[18]. The ETT datasets are partitioned into training, validation, and testing sets with a ratio of 6:2:2, whereas the remaining datasets follow a 7:1:2 split. Detailed statistics and descriptions are summarized in Table 2.

TABLE 2
DATASETS TYPE STYLES

Datasets	Lengths	Dim	Frequency	Split
ETTh1	14,400	7	1 hour	6:2:2
ETTh2	14,400	7	1 hour	6:2:2
ETTm1	57,600	7	15 mins	6:2:2
ETTm2	57,600	7	15 mins	6:2:2
Weather	52,696	21	10 mins	7:1:2
Electricity	26,304	321	1 hour	7:1:2
Traffic	17,544	862	1 hour	7:1:2

2) Baselines and Setup: To evaluate the effectiveness of our proposed method, we benchmarked M^3-HSD against seven state-of-the-art baselines. These baselines encompass diverse architectures, including Transformer-based models (iTransformer [23], PatchTST [19]), MLP-based models (TimeMixer [25], DLinear [20], TimeNet [30]), and CNN/frequency-domain approaches (TFPS [31], MICN [27]). All experiments were implemented using the PyTorch deep learning framework and conducted on a single NVIDIA GeForce RTX 3090 GPU (24GB) for both training and inference. The input sequence length L was fixed at 96, and the prediction horizon T was set to {96,192,336,720}.

B. Experimental Results

As shown in Table 1, the M^3-HSD achieves most of the state-of-the-art performance on the seven benchmark datasets, surpassing all baseline methods. The M^3-HSD consistently outperforms the Frequency-domain benchmark (TFPS: up to 5.6% reduction on MSE for ETTm1, 15.5% for ETTh2, 4% for ETTm2 and 4.5% for Weather), MLP benchmark (TimeMixer, DLinear), and Transformer benchmark (iTransformer, PatchTST).

For datasets exhibiting strong coupled periodic and trend patterns, such as the ETT series and Electricity, the M^3-HSD shows a robust ability to disentangle these complex dynamics

via the Macro-Meso-Micro strategy, thereby promoting prediction accuracy. In the case of datasets lacking strict periodicity but containing stochastic climatic variations, such as Weather, the M^3-HSD still provides commendable predictions, gradually revealing its advantages in long-term trend modeling as the prediction horizon increases.

It is worth noting that M^3-HSD achieves a dominant 66% winning rate (best results in 37 out of 56 metrics) across all datasets. Specifically, the M^3-HSD significantly outperforms the advanced variable-mixing model TimeMixer and the patch-based PatchTST. We argue this is because the physics-aware disentanglement in M^3-HSD offers better generalization than the naive linear mappings used in TimeMixer, and the integrated KAN module fits local seasonal manifolds more precisely than the self-attention mechanism in Transformer, avoiding unnecessary noise during the modeling of simple periodic signals.

C. Ablation study

To evaluate M^3-HSD, we designed three variants: (a) w/o KAN: replacing the KAN-Conv1d with an MLP to verify KAN's superiority in fitting nonlinear seasonal manifolds.; (b) w/o Tri-Stream: removing the Meso stream to form a "Macro-trend + Micro-residual" binary structure (like DLinear), validating the necessity of our triple disentanglement for complex dynamics; (c) w/o Diff-Kernel: using uniform kernel sizes across all streams to demonstrate that distinct physical scales require matched receptive fields. Table 3 shows the complete model consistently performs best, confirming each component's effectiveness. The detailed analysis is provided below:

1) Effectiveness of the KAN Module: The w/o KAN experiment shows a consistent degradation in performance after replacing KAN with MLP. This confirms that the KAN possesses stronger expressivity than MLPs in fitting nonlinear periodic patterns.

2) Necessity of Triple Disentanglement: The w/o Tri-Stream experiment results in a significant increase in error, particularly on the ETTm2 and Weather. This indicates that for data with complex periodicity, traditional decomposition methods are insufficient, and an independent Meso stream must be introduced to explicitly model seasonality.

3) Role of Heterogeneous Kernels: The w/o Diff-Kernel experiment leads to the most severe performance loss on the ETTh2. This validates our hypothesis that physical components of different frequencies require matched receptive fields to achieve optimal feature extraction.

D. Model Analysis and Interpretability

To validate the physical consistency of the M^3-HSD architecture, we visualize the learned attention weights of the Adaptive Gating Fusion layer. As illustrated in Fig. 2, it presents a comparative analysis between the thermodynamic ETTm1 and the event-driven Electricity.

1) Adaptive Disentanglement Strategy: As shown in Fig. 2(a) and Fig. 2(b), the gating mechanism exhibits domain-specific attention patterns. For the ETTm1 (Fig. 2(a)), which is characterized by smooth temperature variations, the model adopts a balanced strategy, assigning significant weights to all three streams (Macro, Meso, and Micro). This aligns with the physical nature of thermodynamic systems, where high-frequency fluctuations are superimposed on slowly varying baselines.

TABLE 3
ABLATION STUDY OF M^3-HSD

Model	M^3-HSD		w/o KAN		w/o Tri-Stream		w/o Diff-Kernel	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh2	**0.321**	**0.360**	0.327	0.372	0.324	0.371	0.329	0.373
ETTm1	**0.373**	**0.383**	0.376	0.401	0.374	0.401	0.374	0.399
ETTm2	**0.264**	**0.312**	0.266	0.318	0.268	0.323	0.265	0.318
Weather	**0.230**	**0.258**	0.231	0.267	0.233	0.270	0.231	0.268

2) Data-Driven Sparsification: In sharp contrast, the Electricity (Fig. 2(b)) reveals a distinct trend sparsification phenomenon. Since grid loads are primarily driven by instantaneous user behaviors and rigid daily schedules, long-term linear trends are negligible. Consequently, the model automatically suppresses the Macro stream.

3) Quantitative Validation: The quantitative distribution in Fig. 2 (c) further confirms these findings. When shifting from ETTm1 to Electricity, the average weight of the Macro stream drops significantly from 28.0% to 8.9%, while the weight of the Micro stream surges to 47.2%. This drastic resource reallocation demonstrates that the model possesses an intrinsic hard-sample mining mechanism, intelligently concentrating over 70% of its computational capacity on the challenging high-frequency dynamics while discarding redundant trend components.

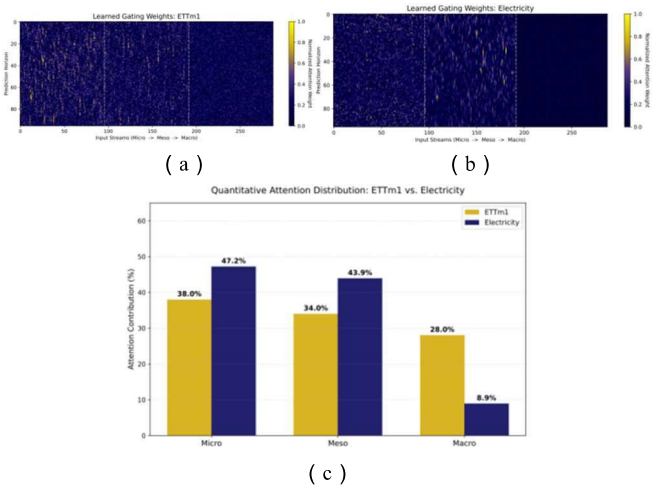

Fig. 2. Visualization analysis of the adaptive gating mechanism. (a) Heatmap for the ETTm1 dataset, exhibiting a globally balanced attention pattern. (b) Heatmap for the Electricity dataset, showing significant trend sparsification. (c) Quantitative comparison of attention distributions. The bar chart illustrates the average weight contribution of the Macro, Meso, and Micro streams across the two datasets.

E. Computational Complexity Analysis

To address the trade-off between model performance and computational overhead, we benchmarked the parameter count (Params), floating-point operations (FLOPs), and inference latency of our proposed framework against representative baselines (DLinear [20], TimeMixer [25], and PatchTST [19]). All inference times were measured using a consistent batch size of 32 on a single NVIDIA RTX GPU 3090 GPU (24GB).

Table 4 shows that, DLinear has the lowest overhead but limited accuracy in modeling nonlinear dynamics. While PatchTST is accurate, it incurs massive computational costs (8.626G FLOPs and 3.752M Params) due to its self-attention mechanism. Strikingly, the M^3-HSD operates with exceptional efficiency. By leveraging lightweight heterogeneous

experts, our model utilizes less than 6% of the parameters and merely 1.3% of the FLOPs compared to PatchTST, yet achieves highly competitive accuracy. Furthermore, our model significantly outperforms the multiscale TimeMixer in both forecasting accuracy and computational speed (1.52ms vs. 7.80ms). These results demonstrate that the proposed Macro-Meso-Micro disentanglement structure successfully circumvents the homogeneous bottleneck without introducing prohibitive computational burdens, making it highly practical for real-world deployments.

TABLE 4
COMPUTATIONAL COMPLEXITY ANALYSIS

Model	Architecture Family	Params (M)	FLOPs (G)	Inference Time (ms/batch)	MSE
DLinear	Linear	0.018	0.004	0.39	0.396
TimeMixer	MLP-based	0.075	0.347	7.80	0.399
PatchTST	Transformer	3.752	8.626	1.70	0.379
M^3-HSD	Heterogeneous	**0.218**	**0.113**	**1.52**	**0.383**

V. CONCLUSION

We propose M^3-HSD, a physics-aware framework that addresses multiscale dynamics via a Macro-Meso-Micro hierarchical disentanglement strategy. The core innovation is the HEDM, which integrates an MLP-based GIP for macroscopic inertia, a KAN-based AMF for mesoscopic nonlinear manifolds, and a CNN-based LTE for microscopic transients. Experiments show that M^3-HSD achieves state-of-the-art performance through its heterogeneous architecture and adaptive gating. Notably, the model exhibits dynamic routing, such as automatic trend sparsification in event-driven systems. These findings validate the interpretability and robustness of our tri-stream architecture. Future work will extend this mechanism to multivariate dependency modeling in complex coupled systems.

REFERENCES

[1] C. Guo, C. S. Jensen, and B. Yang, "Towards total traffic awareness," *ACM SIGMOD Rec.*, vol. 43, no. 3, pp. 18–23, Sep. 2014.

[2] Z. Pan, Y. Wang, Y. Zhang, S. B. Yang, Y. Cheng, P. Chen, et al., "Magicscaler: Uncertainty-aware, predictive autoscaling," *Proc. VLDB Endow.*, vol. 16, no. 12, pp. 3808–3821, Aug. 2023.

[3] S. B. Yang, C. Guo, J. Hu, J. Tang, and B. Yang, "Unsupervised path representation learning with curriculum negative sampling," in Proc. 30th Int. Joint Conf. Artif. Intell. (IJCAI), 2021, pp. 3286–3292.

[4] H. Yu, J. Hu, X. Zhou, C. Guo, B. Yang, and Q. Li, "CGF: A category guidance based PM2.5 sequence forecasting training framework," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 10, pp. 10125-10139. Oct. 2023.

[5] F. M. Alvarez, A. Troncoso, J. C. Riquelme, and J. S. A. Ruiz, "Energy time series forecasting based on pattern sequence similarity," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 8, pp. 1230–1243, Aug. 2010.

[6] C. Guo, B. Yang, O. Andersen, C. S. Jensen, and K. Torp, "Ecomark 2.0: Empowering eco-routing with vehicular environmental models and actual vehicle fuel consumption data," *GeoInformatica*, vol. 19, no. 3, pp. 567–599, Jul. 2015.

[7] C. Guo, B. Yang, J. Hu, C. S. Jensen, and L. Chen, "Context-aware, preference-based vehicle routing," *VLDB J.*, vol. 29, no. 5, pp. 1149–1170, Sep. 2020.

[8] J. Hu, C. Guo, B. Yang, and C. S. Jensen, "Stochastic weight completion for road networks using graph convolutional networks," in Proc. IEEE 35th Int. Conf. Data Eng. (ICDE), 2019, pp. 1274–1285.

[9] J. Hu, B. Yang, C. Guo, and C. S. Jensen, "Risk-aware path selection with time-varying, uncertain travel costs: A time series approach," *VLDB J.*, vol. 27, no.2, pp. 179–200, Apr. 2018.

[10] Y. Ma, B. Yang and C. S. Jensen, "Enabling time-dependent uncertain eco-weights for road networks," in Proc. Workshop on Managing and Mining Enriched Geo-Spatial Data, 2014, pp. 1-6.

[11] Y. Lin, J. Hu, S. Guo, B. Yang, C. S. Jensen, Y. Lin, *et al.*, "UVTM: Universal vehicle trajectory modeling with ST feature domain generation," *IEEE Trans. Knowl. Data Eng.*, vol. 37, no. 8, pp. 4894–4907, Aug. 2025.

[12] Y. Lin, H. Wan, S. Guo, and Y. Lin, "Pre-training context and time aware location embeddings from spatial-temporal trajectories for user next location prediction," in Proc. AAAI Conf. Artif. Intell., vol. 35, 2021, pp. 4241–4248.

[13] Y. Liang, Y. Xia, S. Ke, Y. Wang, Q. Wen, J. Zhang, *et al.*, "Airformer: Predicting nationwide air quality in China with transformers," in Proc. AAAI Conf. Artif. Intell., vol. 37, no.12, 2023, pp. 14329–14337.

[14] X. Huang, Y. Yang, Y. Wang, C. Wang, Z. Zhang, J. Xu, et al., "Dgraph: A large-scale financial dataset for graph anomaly detection," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 22765–22777, Dec. 2022.

[15] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, "Financial time series forecasting with deep learning: A systematic literature review: 2005–2019," *Appl. Soft Comput.*, vol. 90, p. 106181, May 2020.

[16] X. Qiu, J. Hu, L. Zhou, X. Wu, J. Du, B. Zhang, *et al.*, "TFB: Towards comprehensive and fair benchmarking of time series forecasting methods," *Proc. VLDB Endow.*, vol. 17, no. 9, pp. 2363–2377, May 2024.

[17] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, *et al.*, "Transformers in time series: A survey," in Proc. 32nd Int. Joint Conf. Artif. Intell. (IJCAI), 2023, pp. 6750–6760.

[18] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 22419–22430, Dec.2021.

[19] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in Proc. 11th Int. Conf. Learn. Represent. (ICLR), 2023.

[20] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?," in Proc. AAAI Conf. Artif. Intell., vol. 37, no. 9, 2023, pp. 11121–11128.

[21] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, *et al.*, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in Proc. AAAI Conf. Artif. Intell., vol. 35, no. 12, 2021, pp. 11106–11115.

[22] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting," in Proc. 39th Int. Conf. Mach. Learn. (ICML), 2022, pp. 27268–27286.

[23] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, *et al.*, "iTransformer: Inverted transformers are effective for time series forecasting," in Proc. 12th Int. Conf. Learn. Represent. (ICLR), 2024.

[24] K. Yang, X. Wang, Z. Wang, C. Chi, C. Deng, and J. Feng, "STNet: Spatial-temporal transformers are effective for multivariate time series forecasting," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), 2025, pp. 1–8.

[25] S. Wang, H. Wu, X. Shi, T. Hu, H. Luo, L. Ma, *et al.*, "TimeMixer: Decomposable multiscale mixing for time series forecasting," in Proc. 12th Int. Conf. Learn. Represent. (ICLR), 2024.

[26] D. Luo and X. Wang, "ModernTCN: A modern pure convolution structure for general time series analysis," in Proc. 12th Int. Conf. Learn. Represent. (ICLR), 2024.

[27] 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 Proc. 11th Int. Conf. Learn. Represent. (ICLR), 2023.

[28] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, *et al.*, "KAN: Kolmogorov-Arnold networks," in Proc. 13th Int. Conf. Learn. Represent. (ICLR), 2025.

[29] C. Li, X. Liu, W. Li, C. Wang, H. Liu, Y. Liu, *et al.*, "U-KAN makes strong backbone for medical image segmentation and generation," in Proc. AAAI Conf. Artif. Intell., vol. 39, no. 5, Apr. 2025, pp. 4652–4660.

[30] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "TimesNet: Temporal 2D-variation modeling for general time series analysis," in Proc. 11th Int. Conf. Learn. Represent. (ICLR), 2023.

[31] Y. Sun, Z. Xie, E. Eldele, D. Chen, Q. Hu, and M. Wu, "Learning pattern-specific experts for time series forecasting under patch-level distribution shift," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2025.