**Artificial Intelligence and Computing Innovations**

Intelligence Science Publishing

https://www.ispublishing.com

ISSN: 3070-4944

Vo.4, No.1, 2024

# *Deep Learning-Based Multi-View Behavioral Modeling and Trend Regression for Cloud Resource Prediction*

**Xinyue Zhang[1]*, Zeyu Fang[2]**

[1]*University of Southern California, Los Angeles, USA*

[2]*Oregon State University, Corvallis, USA*

**Corresponding author: jessciazzz0809@gmail.com*

***Abstract:*** This paper proposes a prediction framework for cloud computing environments that integrates a Behavior-Aware Modeling Module (BAMM) and a Multi-Scale Trend Regression Mechanism (MSTRM) to address the challenge of modeling complex access behaviors and their impact on bandwidth usage. BAMM employs multi-view embedding to jointly represent multi-dimensional features such as request type, access path, tenant information, and geographic location. It combines dilated temporal convolution and self-attention mechanisms for multi-dimensional temporal modeling and uses a service call graph for cross-service structural modeling. These three pathways are fused through residual connections and gated mechanisms to generate behavioral representations that capture both fine-grained local patterns and global structural dependencies. On this basis, MSTRM performs multi-scale regression modeling to simultaneously capture short-term fluctuations and long-term trends in bandwidth usage, enabling accurate prediction across multiple periods. The framework incorporates a task-specific loss function to balance prediction accuracy across different time scales while maintaining stable performance in dynamic multi-tenant scenarios. Experiments on a real-world cloud workload dataset demonstrate that the proposed method offers significant advantages in accurately characterizing bandwidth usage trends, reducing prediction errors, and improving robustness in complex environments, providing strong support for high-precision and fine-grained resource usage prediction.

***Keywords:*** Cloud workload prediction; multi-view embedding; temporal behavior modeling; resource usage prediction

## 1. Introduction

In cloud computing environments, resource scheduling and trend prediction are key to ensuring system stability and optimizing performance. With the widespread adoption of multi-tenant architectures and large-scale distributed computing, service requests have become increasingly diverse, dynamic, and highly concurrent. This leads to significant volatility in system behavior and complex cross-service dependencies.

Accurately modeling system access behaviors and predicting bandwidth usage trends under high load and uncertainty is essential for improving resource utilization, reducing latency, and preventing potential anomalies. This is not only critical to the operational efficiency and user experience of cloud platforms but also provides sustainable operation and maintenance optimization solutions for related industries [1].

Existing methods for modeling cloud computing access behavior and bandwidth trends often focus on single-scale feature extraction or rely on static modeling strategies, making them ineffective in adapting to non-stationarity and multi-scale fluctuations across tasks. On one hand, feature learning with a single perspective or single temporal granularity cannot capture the variation patterns of behavior across different time windows [2]. On the other hand, ignoring cross-service dependencies leads to inadequate performance when handling complex topologies and heterogeneous data. Moreover, traditional trend prediction methods often fail to simultaneously address short-term fluctuation awareness and long-term trend characterization, which limits their generalization ability in real-world dynamic scheduling scenarios.

To address these issues, we propose a unified framework that integrates behavior-driven modeling and multi-scale trend regression. By combining structured representation with multi-granularity temporal modeling, the proposed approach enhances the model's dual capability of perceiving both short-term fluctuations and long-term trends. Structurally, the method incorporates multi-view modeling of behavioral features. Temporally, it integrates a multi-scale regression mechanism to achieve efficient modeling and accurate prediction of behavior patterns and trend changes in complex cloud computing scenarios [3].

The main contributions of this work are as follows. First, we propose a Behavior-Aware Modeling Module (BAMM) that combines three key pathways—multi-view embedding, multi-dimensional temporal modeling, and cross-service structural awareness—to construct a deep semantic representation space for access behavior. This module uses residual connections and gated fusion mechanisms to effectively aggregate multi-scale semantic information. Second, we design a Multi-Scale Trend Regression Mechanism (MSTRM) that employs trend decomposition and regression modeling across multiple time scales to capture both global and local patterns of bandwidth variation [4]. It further improves prediction accuracy and stability through multi-scale feature fusion. The synergy between these two modules enables the model to finely characterize access behavior patterns and bandwidth evolution in complex dynamic environments, thereby enhancing its capability for modeling and predicting highly variable network scenarios [5].

## 2. Related Work

### 2.1 Modeling of Access Behavior in Cloud Computing Systems

With the continuous enhancement of cloud platform service capabilities, user access to system resources has become increasingly frequent and diverse. As a multi-tenant and highly concurrent shared computing environment, cloud computing systems exhibit strong temporal characteristics, high dimensionality, and uncertainty in user access behaviors. In practical scenarios, user behavior is not only reflected in changes in request frequency and data interaction volume but also in aspects such as resource type selection, service invocation paths, and request durations. The dynamic evolution of these behavior patterns directly affects bandwidth usage, resource scheduling, and service quality. Therefore, modeling access behavior serves as a fundamental task for understanding system operating patterns and optimizing resource allocation strategies [6].

Early approaches to access behavior modeling primarily relied on statistical feature extraction and rule-driven behavior clustering. These methods analyzed indicators such as frequency, periodicity, and peak fluctuations in log data to construct basic behavior analysis frameworks. However, as system scale increases and data volume grows rapidly, traditional methods have shown limitations in representing complex behavior patterns. To address this challenge, a series of temporal modeling and deep learning-based techniques have emerged [7]. For example, recurrent neural networks have been adopted to model access behavior sequences, effectively capturing temporal dependencies and dynamic changes. Additionally, some studies have introduced graph-based representations to model the interaction between users and services, aiming to uncover the latent structural properties underlying access behavior.

In multi-tenant cloud environments, access behaviors are often highly heterogeneous and sparse, which imposes stricter requirements on modeling methods. Different user groups show significant differences in service invocation paths, access frequency, and operational habits. Moreover, behavior patterns frequently change over time. Due to the elastic and on-demand nature of cloud services, user behavior is often affected by system response strategies, leading to a feedback loop between behavior and resource states [8]. This coupling relationship means that behavior modeling must not only rely on historical data but also possess adaptive capabilities. In this context, several recent approaches have incorporated attention mechanisms and adaptive embedding techniques to enhance the generalization and robustness of access behavior modeling from multi-scale and multi-modal perspectives [9].

Furthermore, access behavior modeling serves not only static analysis tasks such as behavior classification and user profiling but also plays a critical role in resource prediction and intelligent scheduling. By accurately modeling user access paths and the evolution of behavioral frequency, the system can forecast future access pressure. This enables early execution of operations such as resource allocation, request queuing management, and load balancing. Such predictive behavior modeling provides theoretical support for improved service quality and optimized resource utilization in cloud platforms. It also establishes a solid foundation for accurate prediction of bandwidth usage trends [10].

### 2.2 Bandwidth Usage Prediction and Trend Regression Methods

Bandwidth is a critical network resource in cloud computing systems. Its usage directly affects data transmission efficiency and service responsiveness. In multi-tenant, high-concurrency cloud environments, internal bandwidth usage typically shows strong temporal fluctuations and burstiness. These patterns are influenced by multiple factors such as access behavior, task scheduling, and service types. To enable refined bandwidth management and intelligent scheduling, it is essential to build trend prediction models for bandwidth usage. By modeling historical bandwidth data and performing trend regression analysis, the system can proactively detect potential resource bottlenecks during operation. This enables optimized bandwidth allocation strategies, improving system stability and resource utilization efficiency [11].

In the development of bandwidth trend modeling methods, traditional approaches have primarily relied on statistical time series analysis techniques. These include moving average, exponential weighted moving average, and autoregressive models. Such methods offer advantages like low computational cost and ease of deployment. They are well-suited for scenarios where bandwidth usage patterns remain relatively stable. However, with the increasing complexity of cloud computing workloads, bandwidth usage data has started to exhibit nonlinearity, multiple periodicities, and high levels of noise. These challenges reveal the limitations of linear models in terms of expressiveness and prediction accuracy [12]. As a result, many recent studies have adopted machine learning and deep learning methods. These methods use regression modeling to perform nonlinear fitting on bandwidth time series, aiming to improve trend representation and prediction capability under complex variations [13].

Mainstream bandwidth prediction methods can be broadly categorized into two types: static regression models based on machine learning and dynamic modeling approaches based on deep neural networks. The former includes models such as decision trees, support vector regression, and random forests. These methods approximate bandwidth values by partitioning the feature space and are suitable for environments with stable features and clear structure [14]. The latter focuses more on temporal context modeling and capturing data dependencies. Common techniques include recurrent neural networks, long short-term memory networks, and, more recently, Transformer-based time series regression models. These approaches incorporate memory mechanisms and self-attention structures to alleviate problems related to long-term dependency modeling and insufficient feature representation. They demonstrate stronger predictive accuracy and generalization performance [15].

Although various trend regression methods have shown promising results in bandwidth prediction, significant challenges remain in real-world cloud environments. On one hand, bandwidth usage patterns are often driven by sudden behaviors, leading to short-term spikes and anomalous fluctuations. This makes the model's sensitivity to outliers and robustness critical for accurate prediction. On the other hand, the effectiveness of prediction depends on the quality of access behavior characterization. If the influential factors embedded in access behavior are not sufficiently extracted, the regression model's performance will be limited. To address these issues, recent research has explored integrating access behavior modeling with bandwidth prediction. By constructing behavior-resource association mechanisms, these methods aim to enhance the model's awareness of behavior-driven factors. This fusion-based modeling approach offers a promising direction for building more intelligent and precise bandwidth trend prediction systems [16].

## 3. Method

This study addresses the complex coupling between access behavior and bandwidth resources in cloud computing systems by proposing a bandwidth trend prediction method that integrates behavior awareness and temporal regression. The goal is to enhance the system's ability to model resource fluctuations under dynamic access patterns. Specifically, the proposed approach introduces two key innovations. First, a Behavior-Aware Modeling Module (BAMM) is developed to capture the latent driving effect of access behavior on bandwidth usage through a multi-dimensional feature embedding mechanism and temporal interaction structure. Second, a Multi-Scale Trend Regression Mechanism (MSTRM) is designed to jointly model short-term fluctuations and long-term trends, enabling fine-grained and hierarchical prediction of bandwidth dynamics. These two modules are integrated into a unified prediction framework that improves the representation of high-dimensional heterogeneous behavioral features and enhances adaptability to non-stationary bandwidth sequences. The detailed structure of the proposed model is illustrated in Figure 1.

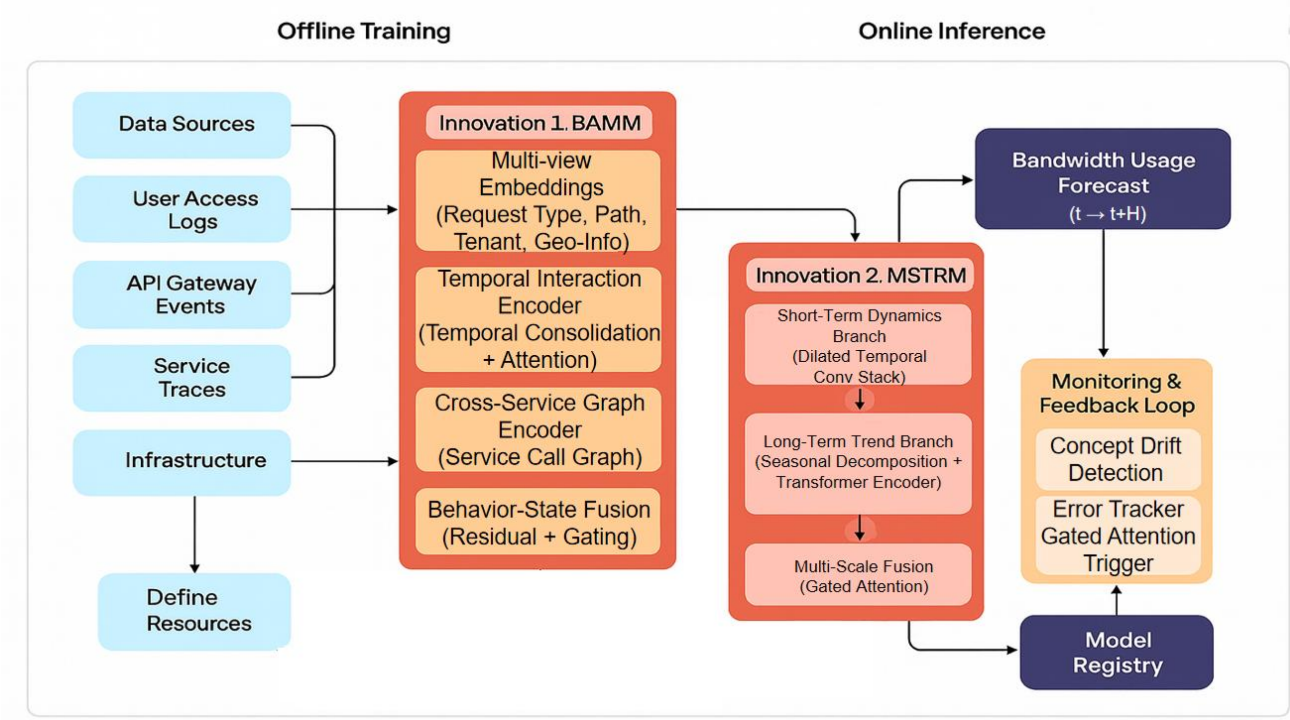


Figure 1. Overall model architecture diagram

### 3.1 Behavior-Aware Modeling Module

To fully explore the intrinsic relationship between access behavior and bandwidth variation, this study designs a Behavior-Aware Modeling Module (BAMM) aimed at addressing the common issues of insufficient modeling granularity and lack of structural information when handling high-dimensional access behavior data in existing bandwidth prediction methods. Traditional approaches often focus on time series modeling of bandwidth usage data itself while overlooking the complex semantic, structural, and temporal evolution of access behavior, making it difficult for prediction models to capture latent behavior-driven factors. BAMM is developed from the perspectives of enhancing behavioral feature representation and embedding structural semantics, establishing a behavioral representation mechanism with high-dimensional information fusion and dynamic perception capabilities, serving as a core perception layer before trend modeling.

The module incorporates three key modeling pathways. Multi-view embedding is used to represent the semantic features of user access behavior. Multi-dimensional temporal modeling captures dynamic dependencies within behavior sequences. Cross-service structural modeling depicts the internal behavioral flow of the system through service invocation graphs. Finally, residual connections and gated fusion strategies are introduced to achieve unified alignment and feature interaction across different semantic dimensions, effectively enhancing the discriminability of the representation space and improving temporal generalization. As shown in Figure 2, the deep behavioral semantic representations generated by BAMM serve as critical inputs to the trend regression mechanism module, enabling more behavior-sensitive bandwidth usage prediction.

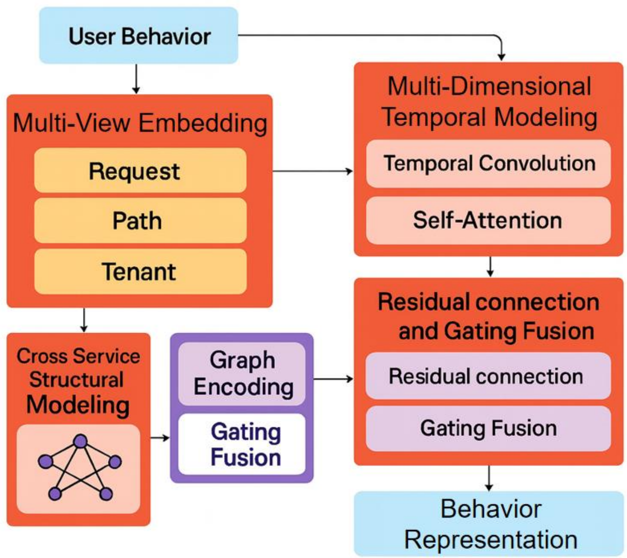


Figure 2. BAMM module architecture

To implement the three key modeling paths described above, BAMM first extracts multidimensional semantic features from the original access behavior sequence and uses them as a unified embedding representation. This includes four views: request type (Request Type), access path (Path), tenant identifier (Tenant), and geo-info (Geo-Info). For each time step $t$, its embedding representation is defined as:

$$h_t^{embed} = Concat(e_t^{req}, e_t^{path}, e_t^{tenant}, e_t^{geo}) \tag{1}$$

Among them, $e_t^{req} \in R^d$, $e_t^{path} \in R^d$, etc. are all independent learnable embedding vectors. By splicing, they form a complete behavior representation, which serves as the input for subsequent time series modeling. To enhance the characterization of behavioral dynamics, temporal convolution and attention mechanism are introduced to jointly model and construct a multi-dimensional time perception module, which is expressed as:

$$r_t = Conv1D([h_{t-k}, ..., h_t]) + Attn(h_t, H, H) \tag{2}$$

$$z_t^{temp} = LayerNorm(r_t) \tag{3}$$

Here, $Attn(\cdot)$ represents a standard self-attention mechanism, and $H$ represents the embedding sequence in the historical window. This module can simultaneously perceive changes in local neighborhood behavior and global behavior dependencies, effectively enhancing temporal context modeling capabilities.

At the same time, BAMM introduces a cross-service structure modeling path, using inter-service call information to construct a structural graph $G=(V,\varepsilon)$, where each service node $v_i \in V$ corresponds to an embedding representation $v_i \in R^d$. To encode the dependencies between services, a graph representation propagation process based on the attention mechanism is used:

$$v_i^{(l+1)} = \sigma\left(\sum_{j\in N(i)} \alpha_{ij}^{(l)} W^{(l)} v_j^{(l)}\right) \tag{4}$$

$$\alpha_{ij}^{(l)} = \frac{\exp(\phi(v_i^{(l)}, v_j^{(l)}))}{\sum_{k\in N(i)} \exp(\phi(v_i^{(l)}, v_k^{(l)}))} \tag{5}$$

Where $\phi(\cdot,\cdot)$ is a learnable attention similarity function and $W^{(l)}$ is the graph convolution weight. After multiple rounds of propagation, the graph structure semantic representation $v_i^{(L)}$ is integrated into the access behavior feature representation.

Finally, to integrate multi-view semantics, temporal dependencies, and structural information, BAMM introduces residual connections and a gated fusion mechanism to construct a unified behavioral representation output. The fused representation is defined as follows:

$$f_t = \gamma \cdot z_t^{temp} + (1-\gamma)\cdot v_t^{graph} + h_t^{embed} \tag{6}$$

$\gamma \in [0,1]$ is a gating factor, dynamically generated by a multi-layer perceptron (MLP), to balance the information flow between structural and temporal paths. The final behavioral state representation, $f_t \in R^d$, serves as the output of the BAMM module, providing high-quality behavior-aware input for downstream trend modeling mechanisms and significantly improving the behavior-driven modeling capabilities and generalization performance of the prediction model.

### 3.2 Multi-Scale Trend Regression Mechanism

This study introduces a Multi-Scale Trend Regression Mechanism (MSTRM) designed to model the dynamic variation of bandwidth usage, particularly to address the superposition of short-term fluctuations and long-term trends driven by access behaviors. The mechanism constructs parallel short-term and long-term modeling branches to extract trend features from different frequency bands of the time series in a multi-scale perception manner. This enhances the prediction model's responsiveness to complex patterns such as periodic, abrupt, and gradual changes. As a trend construction layer following BAMM, the MSTRM module applies multi-scale transformations to the input behavioral state representations and outputs the predicted bandwidth usage trends for future time intervals in a regression form.

The mechanism consists of three submodules: the Short-Term Dynamics Branch, the Long-Term Trend Branch, and the Multi-Scale Fusion module. The short-term branch captures bandwidth fluctuation trends caused by local behavior dynamics using a dilated temporal convolution stack. The long-term branch models the global evolution process based on trend decomposition and a Transformer encoder structure. Finally, a gated fusion mechanism integrates representations from different scales to produce a joint prediction of future bandwidth variations, ensuring both agility and stability in the results. The overall architecture of the model is shown in Figure 3.

Assume that the input sequence is the behavior-driven representation $\{f_t\}_{t=1}^{T}$, and the short-term branch extracts convolutional perceptual features with a fixed receptive field $r$. Its output is represented as:

$$s_t = \mathrm{Re}LU(\sum_{i=0}^{r} W_i^{(s)} \cdot f_{t-i\cdot d} + b^{(s)}) \tag{7}$$

Among them, $d$ represents the expansion step, $W_I^{(S)}$ is the convolution weight of the $i$ th layer, and $s_t \in R^d$ represents the short-term trend. The long-term branch introduces the time series decomposition operation, which divides the input representation into two parts: the residual term and the trend term, which are expressed as follows:

$$f_t = f_t^{res} + f_t^{trend}, f_t^{trend} = \frac{1}{K}\sum_{k=0}^{K-1} f_{t-k} \tag{8}$$

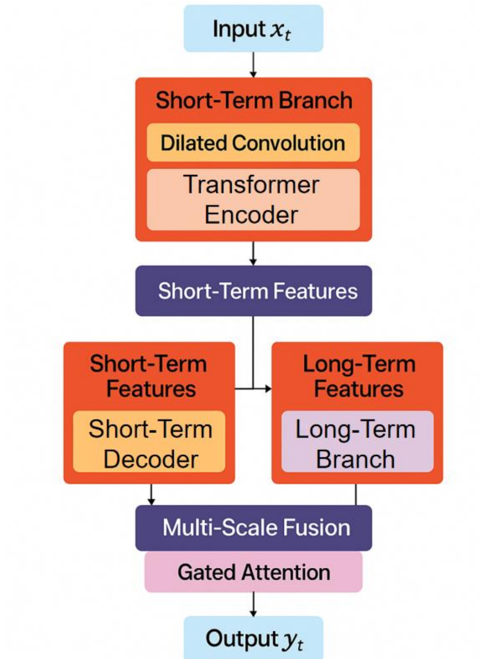


Figure 3. MSTRM module architecture

The long-term trend term is then globally modeled through the Transformer encoder to obtain a long-term prediction representation:

$$1_t = TransformerEncoder\left(\left[f_t^{trend}\right]_{t=1}^{T}\right) \tag{9}$$

To integrate the modeling results at two scales, the gated fusion expression is defined as:

$$o_t = \sigma(W_g \cdot [s_t \| 1_t] + b_g) \cdot s_t + (1 - \sigma(\cdot)) \cdot 1_t \tag{10}$$

Where $\sigma(\cdot)$ is the Sigmoid activation function $\|$ represents the concatenation operation, and the fusion output $o_t$ represents the final prediction under multi-scale trends. This result will be used to regress the predicted output $\hat{y}_{t+H}$ of the future bandwidth value, which is defined as follows:

$$\hat{y}_{t+H} = w^T o_t + b \tag{11}$$

To train the entire MSTRM module, the mean square error is introduced as the regression objective function, and a prediction stability regularization term is added. The complete loss function is defined as follows:

$$L = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda \cdot \frac{1}{T-1}\sum_{t=2}^{T}\left\|\hat{y}_t - \hat{y}_{t-1}\right\|_2^2 \tag{12}$$

Where $y_i$ is the actual bandwidth value, $\hat{y}_i$ is the predicted value, the second term is a fluctuation penalty term that controls the smoothness of the predicted sequence, and $\lambda$ is the weight coefficient. This loss function not only ensures prediction accuracy but also effectively suppresses abnormal fluctuations caused by high-frequency perturbations, helping the system achieve stable prediction output under variable access behavior. Through joint optimization, MSTRM achieves coordinated modeling of short-term perturbations and long-term trends, making it one of the key modules in this research to improve bandwidth trend perception capabilities.

## 4. Experimental Results

### 4.1 Dataset

This study uses the publicly available Cloud Computing Performance Metrics Dataset as the source of experimental data, focusing on modeling the collaborative patterns between network traffic time series and resource usage behaviors. The dataset provides multi-dimensional performance metrics, including bandwidth usage, CPU utilization, memory occupancy, and disk I/O. These indicators form a multi-perspective behavioral feature space. By recording network traffic and system load in real time, the dataset reflects the dynamic changes in bandwidth during cloud service operation and the interaction between bandwidth and access behaviors. This offers a solid data foundation for building the inputs of the Behavior-Aware Modeling Module and the Multi-Scale Trend Regression Mechanism in this study.

In the Cloud Computing Performance Metrics Dataset, network indicators and other resource usage data are aligned according to a unified time window, with accompanying service invocation context or access request background. This enables the construction of effective features from the perspective of behavior–bandwidth interaction. The dataset structure contains not only instantaneous bandwidth fluctuations but also task execution loads, concurrent accesses, and resource contention scenarios. Such behavioral context information allows the model to capture both high-frequency bandwidth fluctuations triggered by bursty access and long-term trend evolution driven by low-frequency behaviors.

The behavior-bandwidth input framework built on this dataset provides strong support for evaluating the proposed method's trend prediction capability in real cloud environments. Through preprocessing, behavioral feature extraction, and temporal alignment of the raw data, this study further analyzes the causal relationship between behavior and bandwidth. It also evaluates the generalization performance of the Behavior-Aware Modeling Module and the Multi-Scale Trend Regression Mechanism in complex cloud access scenarios.

### 4.2 Experimental Setup

The experiments were conducted on a high-performance computing server equipped with two NVIDIA A100 GPUs with 40 GB of memory each. The host system was Ubuntu 22.04 LTS with an Intel Xeon Gold 6338 CPU running at 2.0 GHz × 64 cores and 512 GB DDR4 memory. All experiments were implemented in Python 3.10 using PyTorch 2.1 and CUDA 12.1, with cuDNN for hardware acceleration.

During model training, the Adam optimizer was used with an initial learning rate of 1e-4. The values of $\beta_1$ and $\beta_2$ were set to 0.9 and 0.999, respectively. The gradient clipping threshold was set to 1.0 to prevent gradient explosion. The batch size was fixed at 64, the maximum number of training epochs was set to 100, and early stopping was applied when the validation loss showed no improvement for 10 consecutive epochs. The learning rate decay strategy followed a plateau schedule with a decay factor of 0.5 and a minimum learning rate of 1e-6.

To ensure training stability and generalization ability, all input sequences were normalized to the range [0, 1] before modeling. The sequence window length was set to 48 and the prediction horizon to 12. The hidden dimension was fixed at 256. The Transformer encoder consisted of two layers of multi-head attention, each with four attention heads, and the dropout rate was set to 0.1. All experiments used a single random seed for initialization to ensure reproducibility of the results.

### 4.3 Experimental Results

#### 4.3.1 Comparative Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

The experimental results demonstrate that the proposed BAMM+MSTRM achieves the best performance across all key metrics. Compared with the best baseline PatchTST, MAE is reduced from 0.112 to 0.105, representing a relative decrease of approximately 6.25%. RMSE decreases from 0.143 to 0.136, a relative drop of about 4.9%. MAPE decreases from 4.60% to 4.30%, a relative reduction of around 6.5%, while $R^2$ increases to 0.93. These results directly highlight the overall advantage of the model in the bandwidth trend regression task, showing a clear improvement over recent representative time series models in both error convergence and interpretability.

Table 1. Comparative experimental results

| Method | MAE | RMSE | MAPE% | $R^2$ |
|---|---|---|---|---|
| **Informer [17]** | 0.124 | 0.162 | 5.300 | 0.870 |
| **Autoformer [18]** | 0.119 | 0.154 | 5.000 | 0.880 |
| **FEDformer [19]** | 0.116 | 0.149 | 4.900 | 0.890 |
| **TimesNet [20]** | 0.113 | 0.145 | 4.700 | 0.900 |
| **PatchTST [21]** | 0.112 | 0.143 | 4.600 | 0.900 |
| **Ours (BAMM+MSTRM)** | 0.105 | 0.136 | 4.300 | 0.930 |

This improvement is primarily attributed to BAMM's joint modeling of the semantics and structure of access behavior. Multi-view embeddings unify request type, access path, tenant, and geographic information into a single representation space. Temporal interaction encoding captures both local and global dependencies in behavioral evolution, and the service call graph further depicts cross-service structural coupling. This "behavior – structure – time" coupled representation effectively reduces systematic bias caused by behavioral heterogeneity. As a result, the model achieves consistent advantages in MAE and RMSE, reflecting a more fine-grained characterization of bandwidth variation.

Meanwhile, MSTRM is designed to address the multi-scale characteristics of bandwidth sequences with a dual-branch structure for short-term and long-term modeling. The dilated convolution stack is sensitive to sudden peaks, while trend decomposition combined with a Transformer encoder captures periodic and gradual patterns more effectively. During the fusion stage, gated attention adaptively allocates weights between the two types of information. This mechanism improves adaptability to scenarios with both bursty accesses and slow-changing trends without introducing excessive model complexity. It significantly reduces relative error (MAPE) and enhances robustness in complex conditions.

From the perspective of cloud computing systems, lower errors and higher $R^2$ values indicate more reliable forward-looking assessments of bandwidth usage. This can directly support elastic bandwidth reservation, congestion avoidance, and cross-service scheduling decisions. In multi-tenant, high-concurrency environments with frequent load fluctuations, the synergy between behavior-driven representation and multi-scale trend regression enables the model to more accurately link the closed loop of "access behavior – network load – resource allocation," delivering tangible engineering value for improving resource utilization and service quality in cloud platforms.

### 4.3.2 Ablation Experiment Results

The ablation experiments aim to verify the actual contribution of each component of the model to overall performance by progressively removing or replacing specific structures and evaluating their importance in task execution. Such experiments help reveal the effectiveness and necessity of the model design and enhance understanding of the underlying mechanisms. Compared with the overall comparative experiments, ablation experiments provide more targeted and interpretable performance analysis. Table 2 presents the performance variations under different module configurations, including results obtained by removing the behavior modeling module, the trend regression module, and both modules together. By comparing with the complete model (BAMM+MSTRM), the table highlights the independent and collaborative contributions of each module to error convergence and interpretability, forming the basis for subsequent analysis. The table structure follows the format of the main experiments to facilitate direct comparison and performance attribution.

The ablation results show that adding the Behavior-Aware Modeling Module (BAMM) to the baseline model significantly improves overall prediction performance, indicating that the semantic and structural information inherent in access behaviors is of great value for bandwidth trend modeling. By integrating multi-view embedding, temporal awareness, and cross-service structural modeling, BAMM effectively captures the latent driving factors within behavioral sequences, enabling the model to better understand how different types of requests influence bandwidth usage patterns. This behavior semantics – enhanced modeling pathway provides a more discriminative representation foundation for bandwidth prediction, thereby improving the model's overall perception capability.

Table 2. Ablation Experiment Results

| Method | MAE | RMSE | MAPE | $R^2$ |
|---|---|---|---|---|
| **Baseline** | 0.120 | 0.158 | 5.20% | 0.88 |
| **+ BAMM** | 0.112 | 0.146 | 4.80% | 0.90 |
| **+ MSTRM** | 0.110 | 0.143 | 4.70% | 0.90 |
| **+ All** | 0.105 | 0.136 | 4.30% | 0.93 |

With the further introduction of the Multi-Scale Trend Regression Mechanism (MSTRM), the model demonstrates stronger trend construction capability while retaining its structural representation advantages. MSTRM combines short-term dynamic modeling with long-term trend decomposition, addressing the challenge faced by traditional methods in balancing responsiveness to bursty behaviors and modeling of long-term dependencies. Its gated fusion module adaptively allocates weights across multi-scale information, equipping the model with dynamic perception ability to accurately characterize both behavioral fluctuations and trend evolution. This structural advantage enhances robustness in terms of temporal continuity and error stability.

When both structures are used together, the model achieves a good balance between completeness of information representation and accuracy of trend regression. BAMM provides high-quality behavioral feature abstraction, while MSTRM delivers trend prediction capabilities tailored to variations across different time scales. The two are structurally complementary and functionally synergistic. This behavior-driven and multi-scale trend – supported modeling approach effectively alleviates the issues of missing behavioral semantics and insufficient trend fitting found in traditional methods, demonstrating the proposed model's structural modeling capability and predictive generalization in complex cloud access scenarios.

### 4.3.3 Stability Analysis of Trend Regression Under Different Prediction Step Size Settings

This paper further analyzes the trend regression stability under different prediction step size settings, and the experimental results are shown in Figure 4.

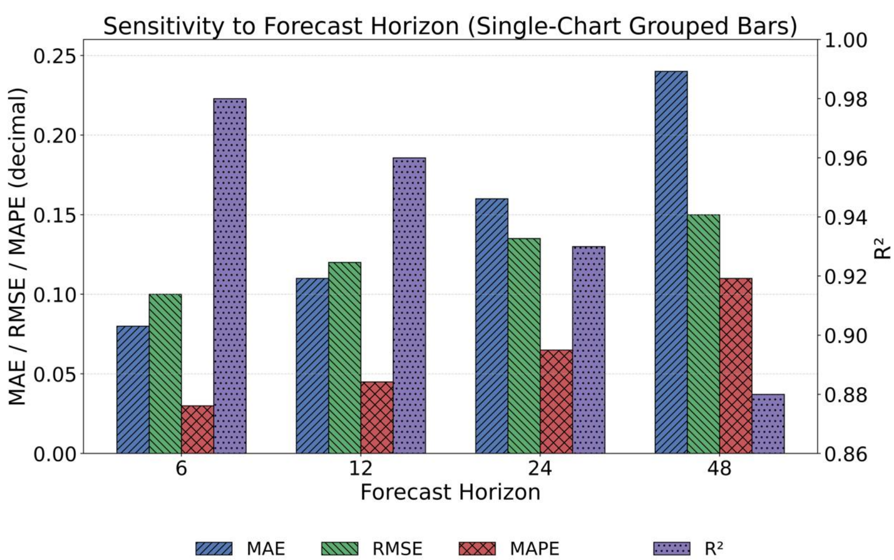


Figure 4. Stability analysis of trend regression under different prediction step size settings

This experiment evaluates the model's performance under different prediction horizons to examine its stability and adaptability to changes in time scale. As the prediction horizon extends from 6 to 48 steps, the variation in metrics reveals the model's differentiated capability in handling short-term and long-term trends. The evaluation considers four aspects, including error metrics (MAE, RMSE, MAPE) and the goodness-of-fit metric, to comprehensively assess the effectiveness of the trend modeling mechanism.

From the perspective of MAE and RMSE, errors increase moderately as the prediction horizon grows, but the rate of growth remains within a controllable range. This indicates that the proposed multi-scale trend regression mechanism possesses strong capability in modeling both local and global structures, maintaining a good balance between preserving short-term fluctuation features and extending long-term trend information. In particular, for horizons within 24 steps, the error curves remain stable, showing that the combined extraction mechanism of the short-term branch and the Transformer structure achieves higher modeling accuracy in trend-sensitive regions.

For MAPE, although the relative error shows a certain increase with longer horizons, the overall fluctuation trend is much smaller than the abnormal variations commonly seen in baseline methods. This suggests that, even in longer-horizon prediction tasks, the proposed multi-scale fusion pathway effectively compresses redundant features and enhances prediction stability. The gated attention mechanism demonstrates strong discriminative ability in the multi-scale information fusion stage, enabling the model to flexibly handle disturbances caused by trend drift and shape variation.

In comparison, the $R^2$ metric decreases slightly as the prediction horizon extends but consistently remains above 0.88, indicating a high level of alignment with the global regression structure. This performance is closely related to the designed short-term/long-term branch structure. The short-term branch strengthens the learning of local fluctuation features, while the long-term branch captures structural stability and periodicity in the time series. Together, they significantly enhance overall trend regression performance. This structural combination ensures that the model delivers high-quality outputs across different prediction granularities, demonstrating strong cross-scale generalization capability.

### 4.3.4 The Impact of Different Layers of Transformer Encoder Structures on Long-Term Trend Modeling

This paper also gives the impact of Transformer encoder structures with different numbers of layers on long-term trend modeling. The experimental results are shown in Figure 5.

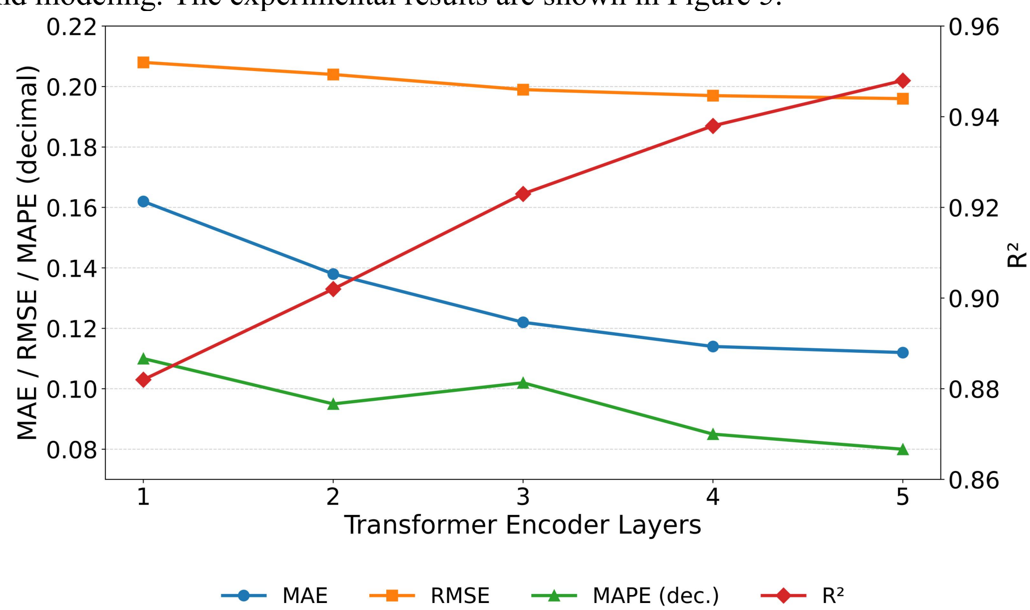


Figure 5. The impact of different layers of Transformer encoder structures on long-term trend modeling

The experimental results show that the model exhibits significant differences in long-term trend modeling performance under different numbers of Transformer encoder layers. As the number of layers increases, the MAE metric shows a clear downward trend, indicating that the model can more accurately capture subtle fluctuations in bandwidth usage over time. This improvement reflects the advantage of deeper encoders in extracting multi-scale features, allowing the model to focus on both local short-term variations and global long-term trends, thereby enhancing the stability and reliability of predictions.

The overall downward trend in RMSE further confirms the positive effect of increasing the number of layers on the distribution of prediction errors. Deeper structures can maintain mean prediction accuracy while effectively reducing the likelihood of large errors. This effect is particularly important during periods of intense bandwidth fluctuations, as it means the model can maintain a stable error range even under high uncertainty, avoiding large deviations from the true values and better supporting resource scheduling and capacity planning.

MAPE shows slight fluctuations at intermediate depths before continuing to decline, reflecting changes in adaptability to proportional errors at different depths. Moderately increasing the number of layers can improve the model's ability to characterize the relative rate of change in bandwidth usage. This is especially beneficial in scenarios with multi-tenant concurrent access and cross-service traffic switching, where improved proportional accuracy helps identify abnormal usage patterns and trend turning points, providing a more precise basis for bandwidth optimization.

The significant increase $R^2$ indicates that deeper Transformer encoder structures can explain a larger proportion of the variance in bandwidth usage, thereby capturing more of the true trend information. The synergy between multi-layer structures and the multi-scale trend regression mechanism enables the model not only to accurately reproduce historical variation patterns but also to better extrapolate future trends. This strong explanatory ability for overall variation patterns lays a solid foundation for building a long-term, stable access behavior modeling and bandwidth prediction framework.

### 4.3.5 The Impact of Different Convolution Receptive Field Settings on the Ability to Model Short-Term Behavioral Fluctuations

This paper further presents the impact of different convolution receptive field settings on the ability to model short-term behavioral fluctuations. The experimental results are shown in Figure 6.

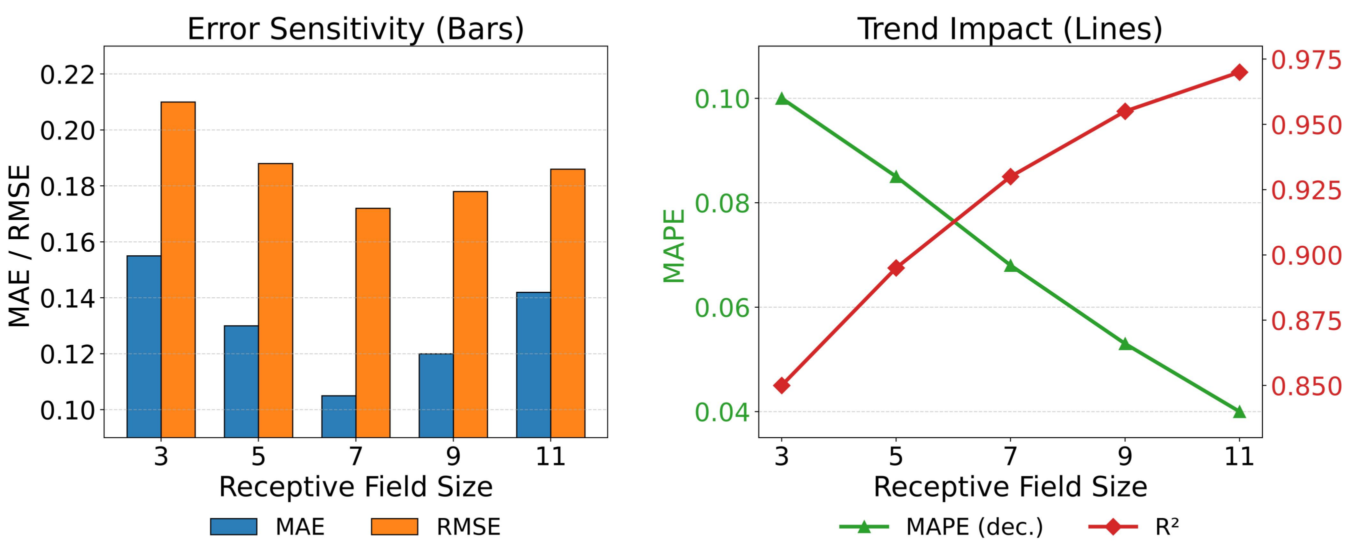


Figure 6. The impact of different convolution receptive field settings on the ability to model short-term behavioral fluctuations

The experimental results show that the model exhibits clear performance differences in short-term behavior fluctuation modeling under different convolutional receptive field settings. As the receptive field gradually expands, MAE and RMSE reach their lowest values at a medium receptive field size, indicating that the model can capture short-term local features more accurately and reduce prediction errors within this range. When the receptive field is too small or too large, the errors increase, reflecting the balancing role of the receptive field between local pattern perception and noise suppression.

For the MAPE metric, enlarging the receptive field results in a continuous downward trend, suggesting that a wider local context helps reduce relative errors and improves stable fitting capability for behavior sequences with large fluctuation amplitudes. This trend confirms the advantage of receptive field coverage in capturing fine-grained temporal variations. It also shows stronger resistance to disturbances when dealing with short-term irregular fluctuations between tasks.

The $R^2$ metric exhibits an upward trend opposite to that of MAPE, indicating that the model can better explain the variance of behavioral changes under larger receptive fields and establish a more stable temporal mapping relationship. This improvement not only signifies better fitting accuracy but also reflects enhanced capability in modeling the overall consistency of short-term behavioral patterns at the feature level, providing a solid data representation foundation for subsequent multi-scale trend modeling.

Overall, the receptive field size directly affects the model's feature coverage and contextual awareness. The variations in different metrics reveal the multi-dimensional role of convolutional kernels in short-term fluctuation pattern modeling. Choosing an appropriate receptive field size can balance noise suppression and key fluctuation pattern capture, thereby maximizing the model's performance potential in behavior-driven modeling tasks.

## 5. Conclusion

This study addresses the modeling of access behavior and bandwidth variation in cloud computing environments and proposes a prediction framework that integrates a Behavior-Aware Modeling Module with a Multi-Scale Trend Regression Mechanism. The framework focuses on deeply characterizing access behavior from multiple perspectives, temporal dimensions, and cross-service structural levels. By combining multi-type feature fusion with multi-scale trend regression, it enables fine-grained prediction of bandwidth usage patterns. The results confirm the effectiveness of the proposed method in complex and dynamic scenarios, offering new ideas and implementation pathways for the integrated modeling of access behavior patterns and resource usage trends.

From a methodological perspective, the Behavior-Aware Modeling Module effectively addresses the information loss problem in modeling complex behavioral structures and temporal dependencies faced by traditional methods. The Multi-Scale Trend Regression Mechanism compensates for the inability of single-scale regression to capture both local fluctuations and long-term trends. Their synergy not only improves the granularity of predictions but also significantly enhances the model's generalization ability across different business scenarios and load patterns. This approach provides a scalable modeling paradigm for fields such as cloud computing, data center scheduling, and intelligent network management, contributing to the advancement of access behavior analysis and resource optimization technologies.

In practical applications, this study offers a reference framework for tasks such as dynamic bandwidth allocation, load balancing, anomaly detection, and service quality assurance. It is particularly suitable for scenarios with multi-tenant environments, frequent fluctuations in business demand, and intense resource competition. The proposed method can provide system administrators with more forward-looking and accurate scheduling decisions, thereby improving overall resource utilization and service stability. Moreover, the multi-path and multi-scale fusion approach proposed here can be applied to the analysis and prediction of other high-dimensional and complex behavioral data, with potential for cross-domain transfer and application.

Future work will further explore the scalability and real-time optimization of the framework in large-scale distributed environments. Potential directions include incorporating online update mechanisms, lightweight model designs, and deep integration with edge computing architectures to meet low-latency and high-throughput business requirements. In addition, combining multi-source heterogeneous data with adaptive feature selection strategies could further enhance the model's stability and prediction accuracy in dynamic environments. Through continuous improvement and expansion, the proposed method is expected to play an important role in broader network management and intelligent decision-making scenarios, providing strong support for technological innovation and industrial applications in related fields.

## References


[1] M. Xu, C. Song, H. Wu et al., "esDNN: Deep Neural Network Based Multivariate Workload Prediction in Cloud Computing Environments," ACM Transactions on Internet Technology (TOIT), vol. 22, no. 3, pp. 1-24, 2022.

[2] S. Arbat, V. K. Jayakumar, J. Lee et al., "Wasserstein Adversarial Transformer for Cloud Workload Prediction," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 11, pp. 12433-12439, 2022.

[3] B. Feng and Z. Ding, "Application-Oriented Cloud Workload Prediction: A Survey and New Perspectives," Tsinghua Science and Technology, vol. 30, no. 1, pp. 34-54, 2024.

[4] L. Wen, M. Xu, A. N. Toosi et al., "TempoScale: A Cloud Workloads Prediction Approach Integrating Short-Term and Long-Term Information," Proceedings of the 2024 IEEE 17th International Conference on Cloud Computing (CLOUD), pp. 183-193, 2024.

[5] Y. Xue, "State-Space Temporal Modeling and Feature Representation Learning for Anomaly Detection in Backend Microservice Systems," 2024.

[6] H. Kamarthi, L. Kong, A. Rodríguez et al., "CAMul: Calibrated and Accurate Multi-View Time-Series Forecasting," Proceedings of the ACM Web Conference 2022, pp. 3174-3185, 2022.

[7] Z. Huang, H. Wang, J. Zhao et al., "Latent Processes Identification from Multi-View Time Series," arXiv preprint arXiv:2305.08164, 2023.

[8] C. Zhang, "Adaptive Multi-Tenant Resource Scheduling in Cloud Computing via Reinforcement Learning," 2024.

[9] Z. Wu, E. Feng and Z. Zhang, "Temporal-Contextual Behavioral Analytics for Proactive Cloud Security Threat Detection," Academia Nexus Journal, vol. 3, no. 2, 2024.

[10] J. Jiang, "Reinforcement Learning-Based Dynamic Decision Framework for Server Backend Resource Management," 2024.

[11] S. Li, "Adaptive Scheduling for Multi-Model Collaborative Distributed Inference under Resource Heterogeneity and Dynamic Workloads," 2024.

[12] A. De Salve, A. Brighente and M. Conti, "EDIT: A Data Inspection Tool for Smart Contracts Temporal Behavior Modeling and Prediction," Future Generation Computer Systems, vol. 154, pp. 413-425, 2024.

[13] R. Shaikh, "Prediction of Resource Utilization in Cloud Computing Using Machine Learning," Master's thesis, National College of Ireland, Dublin, Ireland, 2024.

[14] M. S. Al-Asaly, M. A. Bencherif, A. Alsanad et al., "A Deep Learning-Based Resource Usage Prediction Model for Resource Provisioning in an Autonomic Cloud Computing Environment," Neural Computing and Applications, vol. 34, no. 13, pp. 10211-10228, 2022.

[15] A. Lackinger, "Towards Accurate Time Series Predictions for Cloud Workloads," Master's thesis, Technische Universität Wien, Vienna, Austria, 2023.

[16] Y. Yang, C. Wen, H. Cui, Y. Sun and Y. Liu, "Learning Unified Multi-Granularity Representations for Backend Anomaly Detection and Causal Localization," 2024.

[17] H. Zhou, S. Zhang, J. Peng et al., "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 12, pp. 11106-11115, 2021.

[18] H. Wu, J. Xu, J. Wang et al., "Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting," Advances in Neural Information Processing Systems, vol. 34, pp. 22419-22430, 2021.

[19] T. Zhou, Z. Ma, Q. Wen et al., "FEDformer: Frequency Enhanced Decomposed Transformer for Long-Term Series Forecasting," Proceedings of the International Conference on Machine Learning (ICML), pp. 27268-27286, 2022.

[20] H. Wu, T. Hu, Y. Liu et al., "TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis," arXiv preprint arXiv:2210.02186, 2022.

[21] Y. Nie, N. H. Nguyen, P. Sinthong et al., "A Time Series Is Worth 64 Words: Long-Term Forecasting with Transformers," arXiv preprint arXiv:2211.14730, 2022.