Artificial Intelligence and Computing Innovations

Intelligence Science Publishing

https://www.ispublishing.com

ISSN: 3070-4944

Vo.4, No.3, 2024

Towards Intelligent Backend Operations via Unified Temporal Representation and Prediction Modeling

Chen Wang

University of Missouri-Kansas City, Kansas City, USA

wangmingzhitu7@gmail.com

Abstract: This study proposes a unified modeling approach for long- and short-term temporal dependencies to address the challenges of complex metric structures, dynamic distribution changes, and intertwined multidimensional dependencies in backend systems, aiming to achieve high-accuracy and high-robustness system state prediction. The method first performs structured preprocessing and normalization on multi-source monitoring data, extracts multidimensional temporal features through a feature encoding module, and constructs a dual-channel architecture for short-term dynamic modeling and long-term context capture to fully characterize temporal properties of local fluctuations and global trends. On this basis, a multi-scale fusion mechanism is introduced to enable adaptive interaction and integration of features across different temporal granularities, enhancing the model's ability to represent non-stationary signals, sudden load changes, and cross-dimensional dependencies. In addition, residual calibration and dynamic aggregation strategies are designed to mitigate feature shifts in high-noise environments, ensuring stable prediction performance and strong generalization under complex operating conditions. Validation across various dynamic scenarios shows that the proposed method outperforms existing baseline models on key metrics such as MSE, MAE, MAPE, and RMSE, while exhibiting strong adaptability to data disturbances, environmental changes, and sampling granularity variations. The results demonstrate that the proposed approach effectively captures multi-level dependencies in complex time series, providing strong technical support for state modeling, performance prediction, and anomaly detection in backend systems, and laying a methodological foundation for building high-precision intelligent operations and maintenance systems.

Keywords: Time series dependency modeling; multi-scale fusion; feature encoding; system state prediction

1. Introduction

With the rapid evolution of digital infrastructure and the continuous expansion of service scale, backend systems have become the core backbone supporting modern internet services, enterprise applications, and cloud-native platforms. In this context, the stability of system performance, resource scheduling, and service quality relies heavily on the real-time monitoring and accurate analysis of multidimensional operational metrics. These metrics not only reflect the system's health status and resource utilization efficiency but also directly determine user experience, service availability, and business continuity. Compared with traditional monolithic architectures, backend services now exhibit more complex and dynamic characteristics under distributed, microservice-based, and containerized paradigms. Different modules have highly heterogeneous interaction relationships. Runtime workloads show significant volatility and multi-scale variations. Moreover, system states are often influenced by intertwined external and contextual factors. Against this background, extracting temporal dependencies from complex metric data and modeling them effectively has become a key challenge in ensuring stable system operations and enabling intelligent operations and maintenance[1,2].

Time series analysis, as one of the core approaches to backend metric modeling, has long been an essential tool for system operations and performance optimization. However, as service scale grows exponentially and application scenarios diversify, traditional methods have shown clear limitations when facing high-dimensional, non-stationary, and strongly dependent data patterns. On the one hand, backend metrics are often driven by multiple factors simultaneously. Fluctuations and trends vary significantly across different time scales, and relying only on fixed windows or a single time granularity makes it difficult to capture system dynamics comprehensively. On the other hand, the boundary between anomalous events and normal patterns has become increasingly blurred. Short-term transient signals and long-term evolutionary trends are intertwined, forcing models to balance fine-grained change detection with global consistency. In addition, the strong correlations and nonlinear dependencies within metric sequences place higher demands on modeling capabilities. Simple linear or statistical approaches can no longer meet the requirements for accurate prediction and dynamic awareness[3,4].

From an operations perspective, accurately modeling both short-term and long-term temporal dependencies is not only a technical means of improving prediction accuracy but also a fundamental enabler of proactive maintenance and intelligent scheduling. Short-term dependency modeling helps systems respond quickly to sudden load spikes, resource fluctuations, or anomalous events, providing real-time decision support for tasks such as elastic scaling and fault recovery. Long-term dependency modeling, on the other hand, helps identify performance trends, capacity bottlenecks, and potential risks, supporting resource planning, version evolution, and policy optimization. Their collaborative modeling builds a multidimensional understanding of system behavior across different time scales, enabling global awareness and precise control in complex dynamic environments. This is particularly important for the design of future adaptive backend systems[5].

At the same time, in-depth research on long- and short-term dependency modeling provides both theoretical foundations and technical pathways for constructing intelligent operations frameworks. In complex systems, metric sequences often exhibit multi-level characteristics across time spans. Short-term signals reflect immediate state fluctuations, while long-term trends reveal underlying structural evolution. Coordinating these two aspects and achieving semantic complementarity in model design is key to enhancing backend system cognition. This requires models to have hierarchical representation and dynamic fusion capabilities for multi-scale features, as well as the ability to maintain prediction stability and robustness under non-stationary distributions and context shifts. By building a unified temporal representation framework that integrates local detail perception with global trend modeling, systems can gain full-process intelligent support, from current state understanding to future behavior inference[6].

In summary, long- and short-term temporal dependency modeling for backend metrics is not only a technological breakthrough for addressing system complexity and uncertainty but also a key driver in advancing intelligent operations from passive monitoring to proactive decision-making. It bridges the

semantic gap between low-level data collection and high-level policy optimization, providing unified theoretical and practical support for performance prediction, anomaly detection, resource scheduling, and capacity planning. As cloud-native and intelligent infrastructure continue to evolve, research in this direction will push backend systems from being merely observable to becoming predictable, and from being controllable to becoming adaptive, laying a solid foundation for highly reliable, highly elastic, and self-evolving system management paradigms[7].

2. Related Work

Time series modeling of backend metrics has long been an important research topic in intelligent operations and system management. As systems continue to grow in scale and complexity, traditional methods are increasingly inadequate when dealing with large-scale, dynamic, and multidimensional data. Early studies mainly relied on statistical modeling and signal processing techniques. Methods such as autoregression, moving average, and exponential smoothing were used for trend analysis and forecasting of univariate time series. However, these approaches typically assume stationarity and linearity, making them incapable of capturing the nonlinear dependencies and multi-scale interactions that exist in complex systems. As backend systems evolve toward distributed and cloud-native architectures, metric sequences become more heterogeneous and dynamic[8]. Single linear models can no longer meet the demands of multidimensional dependency modeling, which has driven deep learning-based time series analysis to become the mainstream research direction.

Within the deep learning framework, recurrent neural networks and their variants are widely applied to time series forecasting and anomaly detection tasks. These methods can capture temporal dependencies through recursive updates of hidden states, enabling them to model the evolution patterns of backend metrics effectively. As research has progressed, long short-term memory structures and gated recurrent mechanisms have further enhanced the ability to model long sequences, demonstrating strong performance in handling nonlinear dynamics, delayed effects, and multi-scale features. However, traditional recurrent architectures still face limitations when dealing with high-dimensional, multi-source, and asynchronously sampled metric sequences. They often suffer from limited expressive power, gradient decay, and low training efficiency[9]. At the same time, they struggle to balance sensitivity to short-term fluctuations with the ability to capture long-term global trends, leaving room for improvement in multi-scale performance.

To address these limitations, recent research has increasingly focused on attention-based and transformer-based time series modeling frameworks. Attention mechanisms can dynamically allocate feature weights and highlight information at key time points, which helps capture temporal dependencies and global context more effectively[10]. Models built on this principle not only improve the ability to model long-range dependencies but also significantly enhance adaptability under non-stationary and complex patterns. Meanwhile, multi-scale modeling concepts have been introduced into time series analysis. Through hierarchical feature extraction, scale decomposition, and feature fusion, these methods enable the perception and representation of signals at different temporal granularities. Such approaches are highly valuable in backend systems. They can characterize short-term anomalies and fluctuations while also identifying long-term trends and potential structural changes, providing richer temporal information for operational decision-making[11].

Despite significant advances in the expressiveness and application scope of current methods, many challenges remain in real-world backend metric scenarios. First, fusing multi-source heterogeneous data and modeling multidimensional temporal relationships remains difficult. Traditional methods often fail to capture dependencies across different metrics effectively. Second, the balance between short-term transient signals and long-term trend information remains unresolved. Achieving high prediction accuracy and robustness requires preserving fine-grained dynamic features while maintaining global semantic consistency. In addition, the high dynamism of system environments and distribution shifts places further demands on model generalization. Traditional approaches remain fragile when facing distribution changes and anomalous disturbances. Therefore, developing time series analysis frameworks that can jointly model short-term and

long-term dependencies, adapt to complex system dynamics, and achieve high generalization capability is a key direction for future research[12].

3. Method

This study introduces a long- and short-term temporal dependency modeling approach for backend metrics, aiming to capture both local dynamic features and global evolutionary trends of metric data to achieve deep representation and high-precision prediction of complex system behaviors. The core idea is to perform multi-scale analysis of time series through a hierarchical modeling structure. On one hand, a short-term dependency modeling module captures instantaneous fluctuations and sudden changes. On the other hand, a long-term context modeling mechanism characterizes evolutionary trends and global states. These two components are fused at the semantic level to form a unified temporal representation space, providing a solid foundation for subsequent prediction and decision-making. The entire modeling process is centered on the dynamic evolution of time series, emphasizing the interaction among feature extraction, dependency modeling, and semantic fusion to establish a sequence representation framework that combines sensitivity with stability. The model architecture is shown in Figure 1.

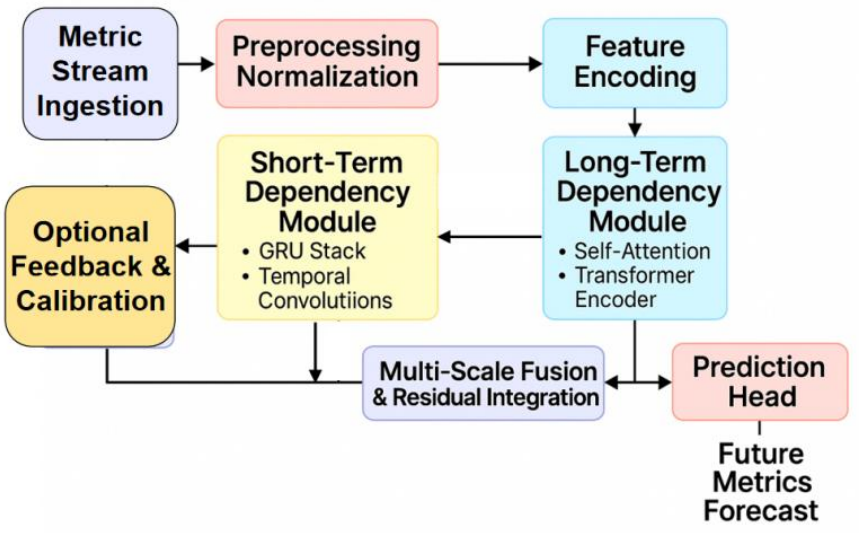

Figure 1. Overall framework model

First, let the indicator sequence of the backend system be a multidimensional vector sequence $\{x_t\}_{t=1}^{T}$ with time step T, where $x_t \in \mathbb{R}^d$ represents the d-dimensional indicator observation at time step t. To model the sequence, the original observation is first projected into a latent space representation through a nonlinear feature mapping function $Enc(\cdot)$, so that complex dynamic dependencies can be captured later:

$$h_t = Enc(x_t) \tag{1}$$

Where $h_t \in \mathbb{R}^m$ represents the hidden layer representation of time step t, and m is the dimension of the latent space.

In short-term dependency modeling, this study characterizes the local dynamics of the sequence through a gated recursive mechanism to fully capture the nonlinear correlation between adjacent time steps. The state update of the recursive unit can be expressed as:

$$s_t = GRU(h_t, s_{t-1}) \tag{2}$$

Where s_t is the short-term state vector and $GRU(\cdot)$ represents the gated recurrent unit, which is used to integrate the current input and historical information, emphasizing the dynamic response within a short time window.

To further capture the long-term evolution trend in the sequence, this study introduces the self-attention mechanism to model global dependencies. Given a sequence representation $\{h_t\}$, the attention distribution can be defined as follows:

$$\alpha_{t,i} = \frac{\exp\left((h_t W_q)(h_i W_k)^T / \sqrt{d_k}\right)}{\sum_{j=1}^{T} \exp\left((h_t W_q)(h_j W_k)^T / \sqrt{d_k}\right)} \tag{3}$$

Where W_q and W_k are the learnable query and key matrices, respectively, and $\alpha_{t,i}$ represents the attention weight of time step t on historical position i, which is used to measure the contribution of different time points to the current state. Based on the attention distribution, the long-term dependency representation can be calculated as:

$$c_t = \sum_{i=1}^{T} \alpha_{t,i} (h_i W_v) \tag{4}$$

Where W_v is the value matrix and c_t represents the long-term representation vector that incorporates global context information.

Subsequently, the short-term state s_t and the long-term context c_t are fused at the semantic level to form a unified temporal representation that captures dynamic features at all time scales. The fusion process can be expressed as:

$$z_t = \sigma(W_s s_t + W_c c_t + b) \tag{5}$$

Where W_s and W_c are linear transformation matrices, b is the bias term, and $\sigma(\cdot)$ is the nonlinear activation function. z_t is the final temporal feature vector, which combines local response sensitivity with global semantic consistency.

Finally, to achieve predictions for future time steps, the model performs nonlinear regression mapping based on the fused representation:

$$\widehat{x}_{t+1} = Dec(z_t) \tag{6}$$

Where $Dec(\cdot)$ represents the decoding function, which is used to map the high-dimensional time series representation back to the original indicator space, and $\widehat{x}_{t+1}$ is the indicator prediction value for the future moment.

In summary, this method achieves collaborative representation of long- and short-term dependencies in backend metric time series by constructing a hierarchical modeling mechanism that spans from local to global levels. The recurrent structure ensures the model's sensitivity to local dynamic features, the self-attention mechanism enables the capture of global semantics, and feature fusion establishes a complementary relationship between the two. Together, they theoretically form a time series analysis framework capable of adapting to complex dynamic environments with multi-scale perception capabilities. This design not only

provides a structured representation foundation for understanding backend system behaviors but also lays the groundwork for higher-level intelligent decision-making and resource scheduling in the future.

4. Experimental Results

4.1 Dataset

This study uses a large-scale cluster monitoring and workload trace dataset as the data source for method validation, namely the Alibaba Cluster Trace (ClusterData 2018). The dataset is collected from production-level cloud infrastructure and provides long-span time-series records describing cluster resource usage and workload dynamics. It contains multi-dimensional signals that are highly representative of backend system operation, such as CPU utilization, memory usage, disk and network statistics, and workload intensity indicators derived from task and instance behaviors. As a publicly accessible dataset, it offers strong reproducibility and allows fair comparison with existing time-series forecasting baselines.

The metric sequences in this dataset exhibit typical backend characteristics, including multi-scale temporal patterns, strong non-stationarity, periodic fluctuations, and sudden burst events caused by workload surges or scheduling changes. Moreover, the indicators are coupled across different dimensions and system components, which naturally forms complex cross-variable dependencies and propagation effects. Such properties align well with the core challenges targeted in this paper, namely capturing short-term transient dynamics while preserving long-term evolution trends, and modeling intertwined dependencies among heterogeneous metrics.

To construct the multivariate forecasting benchmark, we perform structured preprocessing on the raw trace records by aligning sampling intervals, filtering incomplete segments, and normalizing each metric to a consistent scale. The resulting multivariate time series enables systematic evaluation of the proposed dual-channel long-short dependency modeling and multi-scale fusion mechanisms under realistic cloud-backend dynamics. This setting provides a reliable testbed for verifying the model's capability in trend prediction, burst response, and robust representation learning under distribution shifts.

4.2 Experimental Results

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

Table1. Comparative experimental results

Model	MSE	MAE	MAPE (%)	RMSE	Model	MSE	MAE	MAPE (%)
Crossformer[13]	0.324	0.145	4.87	0.569	**Crossformer[13]**	0.324	0.145	4.87
SOFTS[14]	0.298	0.138	4.52	0.546	**SOFTS[14]**	0.298	0.138	4.52
TimePro[15]	0.312	0.142	4.75	0.558	**TimePro[15]**	0.312	0.142	4.75
MMFNet[16]	0.287	0.133	4.30	0.536	**MMFNet[16]**	0.287	0.133	4.30
Ours	0.275	0.128	4.12	0.524	**Ours**	0.275	0.128	4.12
Model	**MSE**	**MAE**	**MAPE (%)**	**RMSE**	**Model**	**MSE**	**MAE**	**MAPE (%)**
Crossformer[13]	0.324	0.145	4.87	0.569	**Crossformer[13]**	0.324	0.145	4.87

SOFTS[14]	0.298	0.138	4.52	0.546	**SOFTS[14]**	0.298	0.138	4.52

From the perspective of overall error levels, the proposed long- and short-term dependency modeling achieves the best results across all four metrics (MSE = 0.275, RMSE = 0.524, MAE = 0.128, MAPE = 4.12%). Compared with representative methods centered on channel interaction and cross-dimensional attention, such as Crossformer, SOFTS, and TimePro, the consistent performance improvements indicate that the model achieves not only lower global fitting errors but also better bias control at the point level and lower relative errors after scale normalization. Combined with the architecture of "feature encoding - parallel short-term/long-term modeling - multi-scale fusion - prediction head," the results show that multi-scale semantic aggregation and residual integration effectively alleviate error amplification caused by non-stationarity and multi-source coupling.

From the perspective of collaborative modeling of short-term fluctuations and long-term trends, MMFNet performs relatively well on RMSE and MAE (0.536 and 0.133, respectively) due to its frequency-domain decomposition, but it still lags behind our fusion-based representation. This demonstrates that relying solely on frequency segmentation cannot fully capture the compound patterns of backend metrics, such as sudden load spikes, slow drifts, and periodic disturbances. In contrast, parallel modeling of short-term dynamic channels (such as GRU or TCN) and long-term contextual channels (such as self-attention or Transformer), followed by multi-scale fusion within a unified representation space, can suppress high-frequency noise while preserving low-frequency structural information. This enables a more robust balance between overall error reduction and peak sensitivity.

From the perspective of metric interpretation, the continuous decrease in MSE and RMSE reflects simultaneous improvements in global fitting quality and sensitivity to extreme values, demonstrating the model's stronger ability to absorb and buffer anomalous spikes and long-tail disturbances. The reduction in MAE indicates more convergent point-level deviations during normal periods, which helps reduce false alarms near service quality thresholds. The improvement in MAPE suggests that the relative error remains more controllable when facing heterogeneous service and metric scales, making the proposed method well-suited for unified threshold management and policy evaluation in cross-service and cross-module backend metric scenarios.

A horizontal comparison shows that Crossformer and SOFTS, which emphasize cross-variable dependencies, demonstrate certain advantages in multivariate scenarios. However, when metrics exhibit complex characteristics such as cross-service propagation and multi-timescale coupling, a single attention channel or aggregation strategy cannot simultaneously capture both short- and long-term dependencies. TimePro is more specialized in modeling long-sequence dependencies but is less sensitive to short-term bursts. In contrast, the proposed method employs explicit dual-channel parallel modeling combined with subsequent multi-scale fusion and residual calibration, enabling structured absorption and alignment of both global trends and local perturbations. As a result, it achieves comprehensive superiority across all four key metrics and aligns more closely with the objectives of long- and short-term temporal dependency modeling for backend metrics.

This paper also conducts comparative experiments on the hyperparameter sensitivity of short-term channel depth and hidden dimension to error convergence. The experimental results are shown in Figure 2.

From the overall trend, as the depth and hidden dimensions of the short-term channel gradually increase, the model exhibits a "decrease followed by stabilization or slight fluctuation" pattern across all four error metrics. MSE decreases from 0.315 to a minimum of 0.283, and MAE converges from 0.148 to 0.131. This indicates that a deeper sequential modeling structure enhances the representation of local dynamic features, significantly improving the model's accuracy in capturing rapid fluctuations and short-

term dependencies. However, when the depth increases further, the rate of error reduction diminishes or even slightly rebounds, suggesting that excessive model complexity may lead to representational redundancy and gradient dissipation, causing performance gains to become marginal.

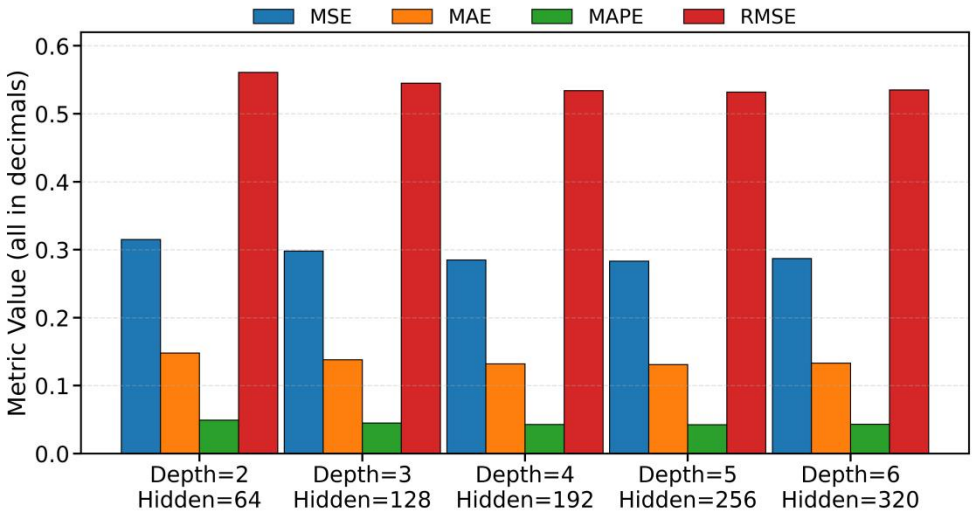

Figure 2. Hyperparameter Sensitivity Assessment of Short-Term Channel Depth and Hidden Dimension on Error Convergence

From the perspective of relative error (MAPE), as parameter capacity grows, the prediction accuracy for metrics of different scales continues to improve, decreasing from 0.0490 to 0.0425. This shows that greater feature diversity helps mitigate the impact of metric distribution heterogeneity. This trend aligns particularly well with the characteristics of backend metric sequences. In multidimensional and non-stationary scenarios, finer-grained analysis of signal patterns by the short-term dependency structure contributes to maintaining stable relative errors across metrics of varying scales. However, MAPE increases slightly under the maximum configuration, indicating that excessive parameter redundancy may introduce noise fitting and weaken the model's generalization ability in complex environments.

The change in RMSE further demonstrates that, as depth and dimensionality increase, the model's ability to capture extreme fluctuations and anomalous peaks improves significantly, with the error decreasing from 0.561 to 0.532. This shows that the parallel short-term modeling channel has higher sensitivity to typical backend metric behaviors such as "sudden load spikes" and "short-term jitter," helping to reduce prediction bias under large fluctuations. This capability is particularly important for maintaining service quality, as most false positives and false negatives originate from unstable predictions in extreme regions.

Overall, this hyperparameter sensitivity experiment reveals a nonlinear relationship between parameter configuration and performance in the short-term dependency modeling module. Moderate increases in depth and dimensionality can significantly enhance short-term dynamic modeling capability and error convergence efficiency. However, excessive expansion may lead to overfitting and noise accumulation. By finding a balance between model complexity and representational power, the proposed structure can better capture sudden patterns and short-term dynamics in backend metric sequences, providing a solid foundation for subsequent multi-scale fusion and collaborative long- and short-term modeling.

This paper also evaluates the environmental sensitivity to changes in load burstiness and request arrival distribution. The experimental results are shown in Figure 3.

From the overall trend, as load burst intensity and the uncertainty of request arrival distributions increase, the model exhibits significant divergence across the four error metrics. MSE and RMSE show a continuous and steep upward trajectory, rising from 0.180 and 0.300 to 0.760 and 1.050, respectively. This indicates that under intensified load shocks, challenges in global fitting accuracy and sensitivity to extreme values increase sharply. These results reveal the destructive nature of burst traffic patterns on temporal structures. Short-term dependencies struggle to capture rapid changes in anomalous fluctuations, while the predictive capability of long-term context is suppressed by severe non-stationarity, causing errors to become highly sensitive to peak events.

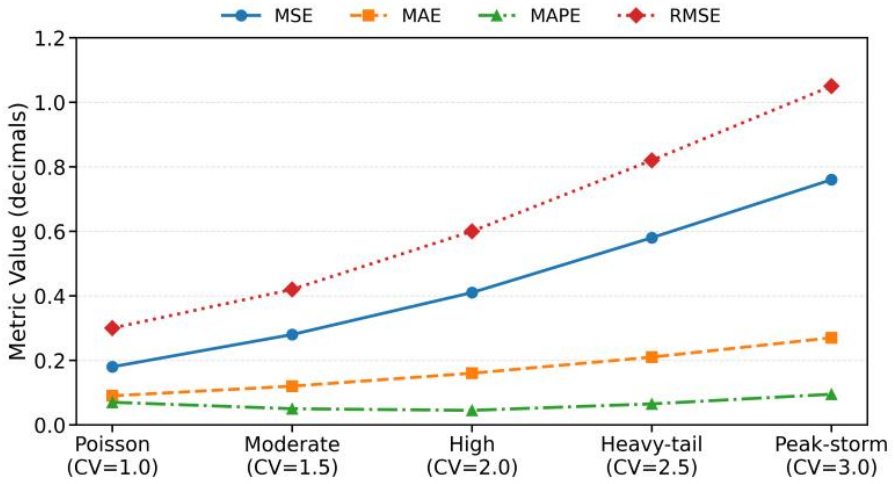

Figure 3. Environmental sensitivity assessment under changes in load burstiness and request arrival distribution

The increase in MAE is relatively moderate, rising from 0.090 to 0.270. This reflects the model's ability to maintain relatively stable point-level prediction performance during regular periods and steady-state intervals. It suggests that the short-term modeling module proposed in this study remains robust against low to moderate disturbances and can still capture the fundamental trends of key metrics. However, as burstiness intensifies, the accumulation of point-to-point errors also becomes more pronounced. This indicates that the short-term dependency window cannot fully adapt to rapid shifts in input distributions, highlighting the need for stronger contextual memory mechanisms to balance steady-state accuracy and dynamic adaptability.

MAPE exhibits a non-monotonic "U-shaped" curve, first decreasing from 0.070 to 0.045 and then rising to 0.095. This trend reflects the model's relative error adaptability in moderate burst scenarios. When fluctuations remain within the modelable range, feature normalization and dynamic adjustment mechanisms effectively suppress relative errors across metrics of different scales. However, as the distribution deviates further from the steady state, amplified effects from metric scale disparities and cross-service coupling cause proportional prediction errors to rise again, highlighting the complexity of multi-variable alignment and scale modeling under highly dynamic conditions.

Considering the evolution of all four metrics, load burstiness and request distribution shifts not only raise the overall error levels but also exacerbate the separation between different error dimensions. In particular, the sharp increase in RMSE indicates that extreme-value responses increasingly dominate overall prediction performance, which aligns closely with the "spike-steady" alternating behavior characteristic of backend systems. The proposed long- and short-term dependency fusion structure demonstrates strong robustness and adaptability under low to moderate disturbances. However, under highly bursty load conditions, reducing extreme-value response errors through enhanced contextual modeling and uncertainty regulation mechanisms will be a key direction for future model optimization.

Finally, this study evaluated the hyperparameter sensitivity of sequence window length and stride settings to long-term and short-term information coverage. The experimental results are shown in Figure 4.

MSE and RMSE show a continuous downward trend as the window length increases from 32 to 96, decreasing from 0.340 and 0.583 to 0.290 and 0.538, respectively. This indicates that a moderate window length and stride can better balance local fluctuations and long-term dependency structures in the time series. When the input sequence coverage is appropriate, the short-term channel can fully capture the dynamic characteristics of local temporal patterns, while the long-term channel can leverage contextual information across time segments for deeper semantic modeling. As a result, prediction accuracy and error convergence are significantly improved.

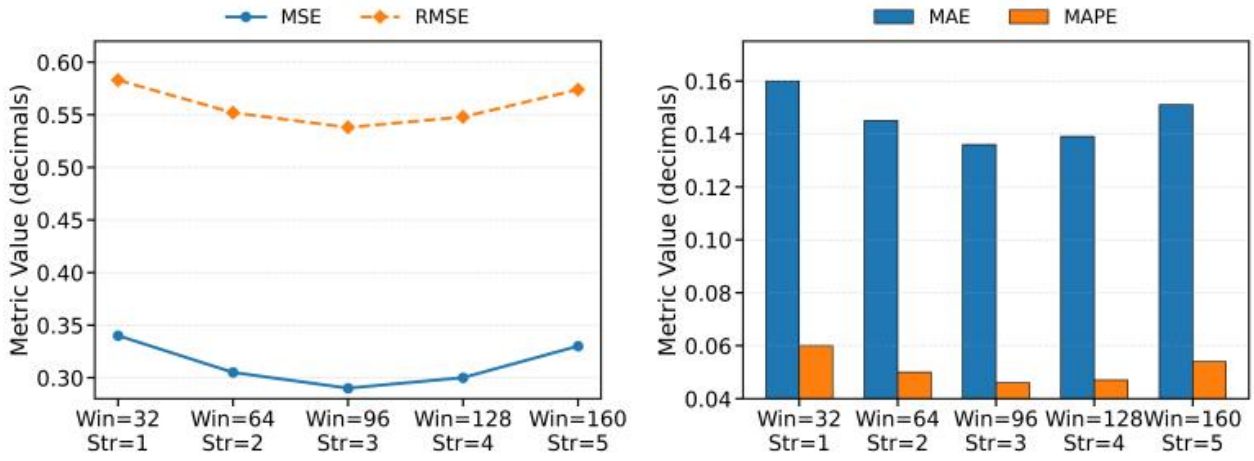

Figure 4. Hyperparameter sensitivity experiment of sequence window length and stride setting on long-term and short-term information coverage

When the window length is further increased to 128 and 160, both error metrics rebound, with MSE rising to 0.330 and RMSE to 0.574. This reflects the fact that overly long time segments introduce more non-stationary factors and contextual noise, reducing the model's ability to extract temporal features. At the same time, a large stride reduces the sampling granularity, causing some fine-grained dynamic features to be missed and weakening the sensitivity of short-term dependency modeling. This phenomenon reveals the "too large-too small" performance boundary that exists for window length and stride in capturing temporal structures.

The changes in MAE and MAPE further reveal the model's behavior in terms of point-to-point error and relative proportional error. MAE decreases to 0.136 under moderate settings and then slightly increases, indicating that the model can maintain stable local accuracy under normal conditions. MAPE shows a more distinct V-shaped pattern, dropping to 0.046 when the window length and stride are optimally matched, which demonstrates optimal control of proportional error on a global scale. However, when the coverage range is too wide or the stride too large, normalized errors are easily amplified by scale shifts, leading to significant deterioration in proportional metrics.

Overall, the tuning of window length and stride not only directly determines feature coverage but also affects the complementarity and fusion of long- and short-term information. Smaller configurations tend to capture high-frequency disturbances but may lose trend structures, while larger configurations increase the risk of noise and undersampling. The configuration of Win = 96 and Str = 3 achieves the best balance between coverage density and temporal span, providing sufficient information support for collaborative modeling of long- and short-term dependencies and demonstrating globally optimal modeling performance across all error metrics.

5. Conclusion

This study proposes a unified modeling framework for long- and short-term temporal dependencies to address the multi-scale dependency structures, non-stationary distributions, and dynamic interaction complexities present in backend system metrics. The proposed method significantly improves prediction accuracy, robustness, and generalization capability. By introducing a collaborative mechanism for short-term dynamic modeling and long-term context capture, the model achieves deep representation and adaptive fusion of multi-granularity temporal features. It demonstrates stable error convergence and high predictive reliability when dealing with typical scenarios such as sudden load spikes, periodic disturbances, and multidimensional signal coupling. In addition, the design of the feature encoding layer and fusion structure enables efficient extraction of global trends and local variations from high-dimensional and heterogeneous monitoring data. This provides structured and interpretable feature representations for backend metric prediction, advancing the development of time series modeling methods and offering theoretical support for system-level state awareness.

The results show that the proposed method significantly improves prediction performance and anomaly response efficiency in complex backend environments, maintaining lower error and stronger adaptability under highly dynamic and non-stationary conditions. This capability is crucial for cloud computing platforms, intelligent operations systems, and industrial IoT infrastructures. In automated monitoring and resource orchestration scenarios, the method can predict performance fluctuations and potential bottlenecks earlier and more accurately, providing more reliable data-driven decision support for service quality assurance, resource scheduling optimization, and anomaly defense strategies. In enterprise applications and critical infrastructure management, the proposed framework can also serve as a core component of the intelligent perception layer, laying the technical foundation for building adaptive and predictive service management systems.

Beyond its direct applications in backend monitoring and prediction, the proposed approach also has transferability to broader time series analysis and decision optimization tasks. Mechanisms such as multi-scale fusion, context-aware modeling, and dynamic feature alignment can provide general solutions for sequence modeling problems in fields such as financial risk control, energy scheduling, traffic flow management, and smart manufacturing. These capabilities can further drive predictive analytics frameworks to evolve from static modeling toward adaptive and intelligent approaches. Furthermore, the proposed method offers new perspectives for research on data-driven system interpretability, model stability analysis, and the construction of automated decision-making pipelines, providing important theoretical and methodological foundations for the design of future intelligent systems.

Looking ahead, future research can be further expanded and deepened in several directions. One direction is to incorporate dynamically adjustable windows, adaptive strides, and multi-level attention mechanisms into the model structure to enhance its ability to capture complex temporal dependencies. Another is to combine the current method with federated learning, graph-based modeling, or reinforcement learning strategies to explore a unified modeling framework under conditions of distribution shift, data silos, and multi-task collaboration. Moreover, online deployment and real-time feedback optimization of the model in large-scale real-world systems will be an important area of future work. This will not only improve the engineering usability and self-evolution capability of the algorithm but also advance intelligent monitoring, predictive operations, and autonomous scheduling systems toward higher levels of intelligence.

References

[1] Liu M, Zeng A, Chen M, et al. Scinet: Time series modeling and forecasting with sample convolution and interaction[J]. Advances in Neural Information Processing Systems, 2022, 35: 5816-5828.

[2] Alkilane K, He Y, Lee D H. MixMamba: Time series modeling with adaptive expertise[J]. Information Fusion, 2024, 112: 102589.

[3] Zhang W, Yin C, Liu H, et al. Irregular multivariate time series forecasting: A transformable patching graph neural networks approach[C]//Forty-first International Conference on Machine Learning. 2024.

[4] I. Kaufman and O. Azencot, "Analyzing deep transformer models for time series forecasting via manifold learning,"arXiv preprint arXiv:2410.13792, 2024.

[5] Y. Wang, H. Wu, J. Dong, G. Qin, H. Zhang, Y. Liu, et al., "Timexer: Empowering transformers for time series forecasting with exogenous variables,"Advances in Neural Information Processing Systems, vol. 37, pp. 469-498, 2024.

[6] X. Yang, "Trend-Fluctuation Decomposition with Deep Residual Networks for System Forecasting," 2024.

[7] Z. Qiu, "A Multi-Scale Deep Learning and Uncertainty Estimation Framework for Comprehensive Anomaly Detection in Cloud Environments," 2023.

[8] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?,"in Proc. AAAI Conf. Artificial Intelligence, vol. 37, no. 9, pp. 11121-11128, Jun. 2023.

[9] F. Chen, "AI-Augmented Anomaly Detection via Generative Distribution Modeling and Uncertainty Quantification in Cloud Systems," 2024.

[10] X. Sun, Y. Yao, X. Wang, P. Li and X. Li, "AI-driven health monitoring of distributed computing architecture: Insights from XGBoost and SHAP," 2024 4th International Conference on Communication Technology and Information Technology (ICCTIT), pp. 480-484, Dec. 2024.

[11] Y. Wang, "Semantic-Driven Large Model Scheduling for Distributed Systems via Unified Representation and Policy Generation," 2024.

[12] R. Ilbert, A. Odonnat, V. Feofanov, A. Virmaux, G. Paolo, T. Palpanas, and I. Redko, "Samformer: Unlocking the potential of transformers in time series forecasting with sharpness-aware minimization and channel-wise attention,"arXiv preprint arXiv:2402.10198, 2024.

[13] Zhang Y, Yan J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting[C]//The eleventh international conference on learning representations. 2023.

[14] Han L, Chen X Y, Ye H J, et al. Softs: Efficient multivariate time series forecasting with series-core fusion[J]. Advances in Neural Information Processing Systems, 2024, 37: 64145-64175.

[15] R. Cristian, P. Harsha, C. Ocejo, G. Perakis, B. Quanz, I. Spantidakis, and H. Zerhouni, "Inter-series transformer: Attending to products in time series forecasting,"arXiv preprint arXiv:2408.03872, 2024.

[16] Ma A, Luo D, Sha M. MMFNet: Multi-Scale Frequency Masking Neural Network for Multivariate Time Series Forecasting[J]. arXiv preprint arXiv:2410.02070, 2024. .