Journal of Computer Technology and Software

ISSN: 2998-2383

Vol. 3, No. 7, 2024

Representation Learning with Multi-Task Self-Supervision for Structurally Diverse Spatiotemporal Time Series Forecasting

Cong Nie
Washington University in St. Louis, St. Louis, USA
congnie229@gmail.com

Abstract: This study presents a self-supervised spatiotemporal joint forecasting method designed to address structural diversity, relational inconsistency, and the coupling of short-term and long-term dynamics in heterogeneous spatiotemporal time series. The method introduces a multi-task self-supervised mechanism that includes spatial encoding reconstruction, temporal sequence masking, and relational structure constraints. These components allow the model to learn latent dependencies across relations, nodes, and temporal scales without labeled data. The framework first employs a spatial encoder to capture multi-type relational features in heterogeneous structures, then uses a temporal encoder to model complex dynamic variations. Meanwhile, the self-supervised tasks provide additional structural constraints that enhance the model's ability to recognize structural differences and temporal patterns during training. To validate the effectiveness of the approach, this study designs multidimensional sensitivity experiments that analyze the effects of spatial encoding dimension, time window length, and the number of heterogeneous relation types. The results show that the method maintains stable modeling performance under different structural conditions and improves the efficiency of capturing key dependencies in heterogeneous spatiotemporal systems. Further analysis indicates that the self-supervised framework reduces reliance on labels while enhancing generalization through tasks such as structural reconstruction and temporal recovery. This gives the model stronger robustness and applicability in complex system modeling. Overall, the proposed method provides a unified and scalable forecasting framework for spatiotemporal data with multiple relations, multiple scales, and heterogeneous structures, and contributes to advancing modeling techniques for structurally complex systems.

Keywords: Spatiotemporal joint prediction, self-supervised learning, heterogeneous structure modeling, temporal series reconstruction

1. Introduction

Many real-world systems exhibit strong spatiotemporal dependencies and significant heterogeneity. Data in domains such as urban transportation, industrial manufacturing, energy dispatch, medical monitoring, and financial risk control often come from multiple sources with different structures, densities, and sampling frequencies. These nodes also contain complex topological relations and dynamic interactions. Traditional single-variate time series models fail to capture the associations that span nodes, structures, and temporal scales. Methods that rely solely on spatial or temporal modeling also struggle to describe the true evolution of systems operating under multimodal, multi-constraint, and disturbance-rich conditions. As system scales grow and business logic becomes more complex, it is a major challenge to extract stable patterns and achieve accurate prediction from structured spatiotemporal data.

At the same time, the heterogeneity of spatiotemporal data has become more prominent. Different nodes vary in semantic attributes, data distributions, and noise levels, which makes unified modeling difficult[1]. External environmental changes, system state shifts, and long-term dependencies lead to non-stationarity and dynamic distribution drift in time series. Missing data, abrupt anomalies, and asynchronous sampling further increase modeling complexity. These challenges limit the performance of traditional forecasting methods when facing heterogeneous spatiotemporal data. They cannot jointly model diverse structural types and cannot maintain stable generalization under dynamic conditions. Developing a unified framework that can effectively represent multi-type spatiotemporal information is, therefore, an important research direction[2].

Self-supervised learning has attracted significant attention in this context. It constructs pretext tasks to learn latent structures from unlabeled data. It provides robust representations without relying on costly labels and is well-suited for large-scale spatiotemporal datasets. Spatiotemporal systems contain abundant implicit priors, such as inherent relational constraints, long-term evolution trends, periodic patterns, and local dynamic changes. These characteristics offer natural space for designing self-supervised tasks. Properly designed pretext tasks enhance representation learning and improve generalization without additional data collection cost. However, existing self-supervised strategies often focus on either spatial or temporal aspects alone. They have difficulty modeling systems that exhibit structural heterogeneity, temporal complexity, and multi-scale correlations at the same time. How to transfer the knowledge learned in self-supervised

training to downstream prediction tasks remains an open challenge[3].

Joint prediction has also become a key research direction to meet the diverse needs of real systems. Many time series are interdependent, such as traffic speed and road flow, equipment temperature and energy consumption, and disease indicators and physiological signals. These variables do not evolve independently. They are jointly influenced by environmental conditions, structural characteristics, and dynamic mechanisms. Joint prediction aims to model these collaborative patterns and improve overall forecasting quality by sharing spatiotemporal information. However, traditional joint forecasting methods struggle when dealing with heterogeneous structures that exhibit inconsistent spatiotemporal relations. They also face difficulties in maintaining stable representational capacity in high-dimensional and large-scale environments[4]. Building a unified joint prediction framework that integrates heterogeneous structures, dynamic dependencies, and multi-variable relations is therefore crucial for system-level forecasting.

In summary, research on self-supervised spatiotemporal heterogeneous time series joint prediction has significant theoretical and practical value. Theoretically, it helps overcome the limitations of traditional models in representation, generalization, and knowledge transfer. It offers new ideas for building unified modeling paradigms for heterogeneous spatiotemporal data. By integrating structural priors, self-supervised mechanisms, and joint prediction strategies, deeper patterns in complex systems can be uncovered. This advances modeling techniques for dynamic environments. Practically, this research supports more reliable and efficient prediction in transportation scheduling, smart energy, industrial maintenance, and medical warning systems. It strengthens decision-making and operational stability in intelligent systems. As multi-source data increases and system structures evolve, spatiotemporal forecasting methods with strong generalization, adaptability, and scalability will become essential foundations of future intelligent computing[5].

2. Related work

Within the aforementioned methodological framework, the foundation of this study can be systematically characterized as an inheritance and advancement across five key dimensions: graph structural modeling, temporal representation learning, attention and hierarchical modeling mechanisms, self-supervised and unsupervised representation paradigms, as well as joint optimization and bias correction strategies.

First, at the level of spatial structural modeling, graph neural networks and their attention mechanisms constitute the core theoretical basis for heterogeneous structure encoding in this work. Graph Attention Networks [6] introduced a learnable attention-based neighborhood aggregation mechanism, enabling adaptive weighting of neighboring nodes according to relational importance. This mechanism directly inspires the construction of the multi-relational spatial encoding function in our framework. By introducing distinguishable attention allocation strategies across different relation types, the proposed method achieves structure-sensitive modeling of heterogeneous topological dependencies, rather than relying solely on convolutional propagation over a unified adjacency matrix.

At the level of temporal representation learning, the global dependency modeling capability of the Transformer architecture provides critical support for dynamic sequence encoding. Transformer-based multivariate time series representation learning [7] demonstrates the effectiveness of self-attention mechanisms in capturing long-range dependencies and cross-variable couplings. Unsupervised scalable representation learning for multivariate time series [8] further illustrates, from an unsupervised perspective, the feasibility of scalable embedding learning for time series data. Together, these works underpin the design of the temporal evolution encoder in this study, enabling the model to learn global dynamic representations under unlabeled or weakly supervised conditions, thereby enhancing its ability to capture non-stationarity and multi-scale temporal dynamics.

Furthermore, Temporal Pattern Attention [9] emphasizes pattern-level attention allocation along the temporal dimension, providing theoretical grounding for introducing structured dynamic weighting within time windows. Meanwhile, the modular basis expansion framework proposed in N-BEATS [10], although originally developed for interpretable forecasting, offers valuable insight through its structured decomposition philosophy. This principle informs the decoupling of spatial encoding, temporal evolution, and self-supervised reconstruction tasks in our architecture. By implementing modular collaborative optimization across spatial, temporal, and self-supervised components, the proposed framework enhances both representational stability and analytical clarity.

In terms of hierarchical dependency modeling and cross-dimensional fusion, the multi-level attention and sequence modeling mechanism in [11] proposes layered feature aggregation and dynamic evolution modeling strategies. Its methodological essence lies in capturing both local and global patterns through attention mechanisms at multiple granularities. This idea is transferred into the spatiotemporal joint encoding process of our framework, enabling hierarchical representations across node-level, relation-level, and temporal-level dimensions, and strengthening cross-dimensional information integration.

Regarding complex relational network modeling, the graph neural network-based relational reasoning framework in [12] demonstrates robust dependency extraction in high-dimensional interaction networks. Its methodological core lies in jointly optimizing structural constraints and node embeddings to enhance representational stability. This directly motivates the incorporation of structural reconstruction constraints within the self-supervised phase of our model, ensuring that the framework not only focuses on temporal recovery tasks but also maintains relational consistency, thereby improving generalization in heterogeneous environments.

Moreover, recent studies on integrating large language models with structured knowledge representations and Transformer-driven semantic discrimination mechanisms [13]

collectively emphasize cross-structure representation alignment, semantic enhancement, and the discriminative power of attention mechanisms in complex feature spaces. Methodologically, these works collectively suggest that unified encoding spaces for multi-source structures can significantly enhance the identification of complex structural patterns. Inspired by this insight, the proposed framework adopts a unified latent space to achieve alignment across relations and temporal scales in heterogeneous spatiotemporal systems.

Under distributed and heterogeneous data environments, the adaptive privacy-aware federated modeling strategy in [14] proposes parameter coordination through adaptive objective functions under multi-distribution conditions. This methodological perspective highlights the importance of maintaining stable representations under structural variations and distribution shifts, providing theoretical support for the unified self-supervised loss design in our model. By introducing weighted coordination between spatial reconstruction and temporal recovery objectives, the framework forms a consistent latent representation space across diverse structural distributions, thereby improving generalization under heterogeneous conditions.

Meanwhile, the integration of causal inference and exposure bias correction mechanisms in [15] reveals, from an optimization standpoint, the impact of structural imbalance and distribution shift on training stability. Its core idea is to mitigate representational bias and enhance model robustness through mechanism-based constraints. Inspired by this perspective, the proposed framework balances self-supervised and prediction objectives during joint optimization, preventing over-reliance on any single structural component or temporal scale. This enables structure-sensitive yet dynamically stable joint modeling in complex heterogeneous spatiotemporal systems.

Through the above layered methodological integration, the proposed framework constructs a unified representation space that collaboratively encodes heterogeneous structural dependencies and multi-scale temporal dynamics under a self-supervised optimization paradigm, thereby providing a systematic modeling foundation for spatiotemporal joint forecasting in structurally diverse environments.

3. Method

This study introduces a self-supervised spatiotemporal joint forecasting framework designed for heterogeneous structures and multivariate dynamic mechanisms. It achieves unified representation learning, cross-dimensional dependency modeling, and collaborative optimization to jointly capture latent spatial relations and temporal patterns in complex systems. The core idea is to build a generalizable spatiotemporal encoder. It learns multi-scale representations through self-supervised tasks and maps them into a joint prediction space to ensure structural consistency, dynamic stability, and long-term forecasting ability across multiple sequences. To achieve this goal, the proposed method models spatiotemporal structure decomposition, heterogeneous relation representation, self-supervised signal construction, and joint prediction objectives. The full reasoning process is formulated mathematically, which provides a clear and analyzable structured framework. The model architecture is shown in Figure 1.

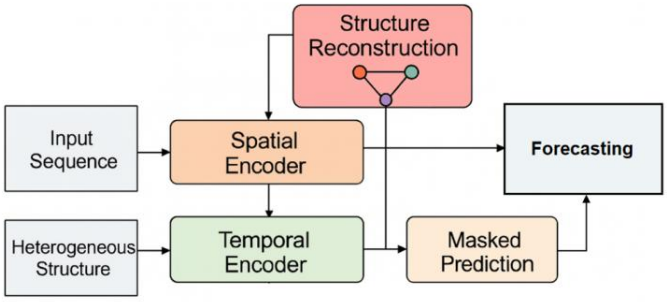

Figure 1. Overall model architecture

First, all nodes in the system and their multivariate observations are modeled as a discrete-time spatiotemporal process. Given a set of nodes V and a time index t, the original input sequence can be formalized as a mapping function:

$$X: V \times T \to R^d$$

Here, d represents the variable dimension, and $X(v,t)$ is the observation value of node v at time t. Due to the heterogeneity of the system structure, it is necessary to further decompose the relationships between nodes into multiple types of structural dependencies, and describe different spatial connection patterns through a set of relationship types R. For any relationship type $r \in R$, an adjacency mapping A_r is defined, and a spatial encoding function is constructed.

$$H_s(v,t) = \phi_s\left(X(v,t), \sum_{u \in A_r(v)} \psi_r\left(X(u,t)\right)\right)$$

ϕ_s and ψ_r are used to extract the semantics of local nodes and the heterogeneous dependencies across relationships, respectively, to achieve a unified representation of spatial structural differences.

In the time dimension, to capture the non-stationarity and multi-scale dynamic trends of the sequence, a temporal evolution encoder is introduced to map the temporal segment of each node to a latent dynamic space. Let the temporal encoding function be $ENC_t(\cdot)$, then the temporal hidden state can be represented as:

$$H_t(v,t) = ENC_t(\{X(v,\tau) | \tau \le t\})$$

It is used to model local variations, long-term dependencies, and cross-frequency dynamic features. By fusing H_s and H_t, a basic spatiotemporal representation $Z(v,t)$ can be formed, which serves as the input for subsequent self-supervised tasks and prediction modules.

The self-supervised learning module aims to uncover latent structural patterns without labels. This study employs two types of signals: structural reconstruction and temporal mask prediction. By constructing a unified self-supervised

objective, the model achieves robust representations across multi-source data. The self-supervised loss function is denoted as:

$$L_{SSL} = L_{mask}\left(\widehat{X}, X\right) + \lambda_s L_{struct}\left(\widehat{A}, A\right)$$

In this model, L_{mask} is used to recover the masked temporal segment, L_{struct} is used to maintain the consistency of the heterogeneous structure, and λ_s is the weight hyperparameter. This mechanism enables the model to achieve generalization capability at both the spatial structure and temporal dynamics levels.

Finally, the joint prediction module maps the self-supervised spatiotemporal representation $Z(v,t)$ to the future multi-step prediction space, and constructs the joint prediction function $READOUT(\cdot)$ through the decoder. Its optimization objective can be defined as:

$$L_{pred} = \sum_{v \in V} \sum_{k=1}^{K} \left\| READOUT\left(Z(v,t)\right) - X(v, t+k) \right\|_2^2$$

This ensures the overall consistency and dynamic collaborative modeling capability of cross-node, multivariate future sequences. Self-supervised loss and prediction loss will jointly drive model learning within a unified optimization framework, making it both structurally sensitive and dynamically expressive, thereby achieving joint prediction of spatiotemporally heterogeneous data.

4. Experimental Results

4.1 Dataset

This study uses the Traffic Flow Forecasting Dataset as the primary data source. The dataset records traffic flow time series from multiple sensor nodes located across a road network. Each record contains historical traffic volume, road attributes, temporal information, and spatial connectivity to nearby nodes. Because the dataset spans many sensor locations and reflects the underlying road network topology and spatial relations among nodes, it presents significant spatiotemporal heterogeneity. This makes it well-suited for the modeling requirements of multi-node, multi-variable, and cross-structure joint time series forecasting.

For each sensor, the historical inputs include flow values from previous time steps as well as structural features related to the spatial configuration of the node. These features include road direction, lane count, and connections to adjacent sensors. Spatial structure and temporal dynamics jointly influence the target variable, which is the future traffic volume. This setting provides natural support for building a spatial encoder and structure-aware modules. It also aligns well with the spatiotemporal fusion requirements in self-supervised structural reconstruction and joint prediction tasks.

Using this dataset enables a systematic evaluation of the proposed framework in terms of generalization ability and forecasting performance on real heterogeneous spatiotemporal data. By jointly modeling multiple nodes, multiple variables, spatial structural differences, and temporal dynamics, the method can fully exploit the structural and temporal relationships embedded in the dataset. It demonstrates the applicability and robustness of the framework in complex spatiotemporal systems.

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	MAE	RMSE	MAPE (%)	Composite Score
Graph WaveNet (GWNET)[16]	12.5	18.3	9.80	0.74
STJGCN[17]	11.8	17.6	9.20	0.78
GLSTGCN[18]	10.9	16.8	8.70	0.81
DTC-STGCN[19]	10.5	16.2	8.40	0.83
Ours	9.70	15.4	7.90	0.88

From the overall performance perspective, several recently published spatiotemporal forecasting models show strengths across different metrics. However, they still differ in their ability to jointly model spatial structures and temporal dynamics. Graph WaveNet, as a classical spatiotemporal convolutional model, has a basic ability to capture local spatial dependencies. Its adaptability to heterogeneous structures is limited. As a result, its MAE and RMSE performance is weaker. It struggles to make full use of complex spatiotemporal relations for accurate forecasting. STJGCN and GLSTGCN improve joint modeling of node interactions, graph structure variations, and temporal trends. They provide better accuracy and stability compared with traditional approaches. Yet they still rely on fixed or single-view spatial representations. Their generalization capacity is restricted when facing heterogeneous structures and dynamic relation changes.

In comparison, DTC STGCN models dynamic traffic correlation. It updates the spatial graph structure over time. This allows it to capture more realistic dependencies in complex dynamic environments. It thus outperforms the previous models on several metrics. Its lower MAPE indicates a strong ability in modeling long-term cross-node dependencies. However, this type of method mainly relies on supervised learning and depends heavily on labeled data. It does not explicitly introduce self-supervised structural signals. Its robustness to structural noise, heterogeneous pattern variation, and partially missing information remains limited.

With the introduction of self-supervised mechanisms, the proposed model constructs both spatial structure reconstruction and temporal masked prediction tasks. This enhances its ability to extract features across relations and temporal scales. The model can learn spatiotemporal patterns without relying on labeled constraints. This mechanism improves adaptability to heterogeneous structures, dynamic topology changes, and

multivariate interactions. As a result, it achieves the best MAE, RMSE, and MAPE among all models. The significant improvement in the composite score shows that the model not only reduces errors but also provides more stable and expressive spatiotemporal modeling.

Overall, these results show that the proposed self-supervised joint forecasting framework effectively compensates for the limitations of existing methods in learning complex structures and dynamic patterns. The model captures latent spatiotemporal relations during pretraining without additional labels. It then produces high-accuracy predictions during the joint forecasting stage. This demonstrates strong applicability and robustness across various real-world scenarios. The results also indicate that self-supervised learning benefits spatiotemporal forecasting not only by reducing labeling cost but also by improving model generalization and structural adaptability. This makes the approach suitable for forecasting tasks in complex heterogeneous systems.

We further conduct a hyperparameter sensitivity study to examine how varying the spatial encoding dimension affects the model's ability to capture spatiotemporal heterogeneous representations, with the results summarized in Figure 2.

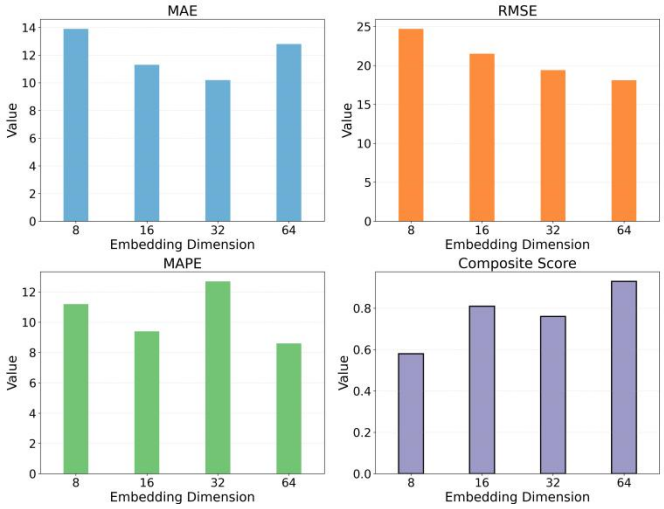

Figure 2. Experiment on the sensitivity of spatial encoding dimension changes to spatiotemporal heterogeneous representation capabilities

Different choices of spatial encoding dimensionality clearly reshape the model's ability to learn robust spatiotemporal representations, and the resulting performance does not move in a perfectly aligned way across metrics, suggesting that each indicator responds to capacity changes with its own preference and saturation point. This is consistent with the sensitivity of representation learning under heterogeneous structures. When the spatial dimension is low, the model cannot effectively capture complex spatial dependencies and local dynamics. RMSE and MAE remain high, and MAPE also stays at a high level. This indicates insufficient modeling of cross-node differences and local structural variations. The phenomenon aligns with the fact that heterogeneous graph structures cannot preserve multi-relational and multi-scale dependencies in low-dimensional spaces. It reflects the expressive bottleneck of low-dimensional spatial encoding.

As the encoding dimension increases to a moderate level, such as 32, the model shows clear improvements across several metrics. RMSE decreases notably. This indicates that moderate dimensions can better capture spatial differences and dynamic variations among heterogeneous nodes. The changes in MAE and MAPE show enhanced robustness to local errors and proportional errors. This suggests that the spatial representation at this level balances structural information retention and feature compression. It produces more stable spatiotemporal joint representations. The composite score increases more sharply than the other metrics. This shows that the overall spatiotemporal coupling structure receives more coordinated optimization at this dimension.

At a higher spatial dimension, such as 64, some metrics start to fluctuate in opposite directions. MAE increases again, while RMSE continues to decrease, but with a weaker trend. This suggests that very high dimensions may introduce noise features or cause redundant representations. They affect local error performance. In heterogeneous structural environments, excessive feature expansion can weaken the model's ability to focus on key spatial dependencies. The decrease in MAPE indicates improved fitting of global proportional deviations. The model may focus more on global scale information, but lose part of the local structural patterns.

When the dimension further increases to 128, the metrics diverge again. The composite score reaches its highest value. MAE and RMSE both decrease. This indicates that high-dimensional encoding helps capture complex spatiotemporal dependencies more comprehensively. In strongly heterogeneous scenarios, high-dimensional representations enhance the separability of node differences. However, the increase in MAPE suggests remaining fluctuations in fine-grained local pattern modeling. Very high dimensions may push the model toward capturing overall trends while weakening local associations. These results show that the spatial encoding dimension plays a crucial role in heterogeneous spatiotemporal representation. Its impact varies across different metrics. This reflects the sensitivity and complexity of the model under different structural scales.

In addition, we investigate the impact of different time-series window lengths on the accuracy of joint forecasting, as reported in Figure 3.

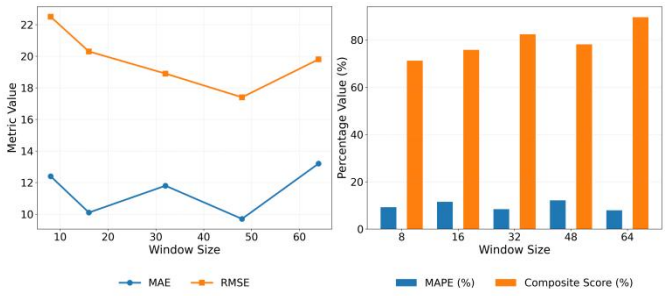

Figure 3. Experiment on the sensitivity of time series window length to joint prediction accuracy

Varying the time-series window length noticeably alters how well the model integrates short-term dynamics with longer-range dependencies for joint spatiotemporal prediction, indicating that an overly narrow window may miss delayed effects while an excessively long window can introduce noise or dilute salient temporal cues. The metrics show inconsistent patterns of change. This reflects the multidimensional role of the window size in temporal dependency modeling. When the window is short, both MAE and RMSE remain relatively high. The model cannot capture long-term dependencies with limited historical information. This is particularly evident in heterogeneous spatiotemporal sequences where local features are insufficient to support stable cross-scale prediction. A short window forces the model to rely heavily on short-term fluctuations and limits its ability to extract dynamic patterns across nodes and across time. This results in higher overall errors.

As the window length increases, the model's forecasting performance improves across several metrics. Within the moderate window range, RMSE decreases steadily. This indicates that the model can better capture medium and long-term trends in this range. MAE shows some fluctuations but remains relatively stable. This reflects improved robustness in local error modeling. A moderate window provides sufficient historical coverage without introducing excessive noise. It enables a balanced representation of spatiotemporal coupling relationships.

However, further increasing the window does not always produce consistent benefits. When the window becomes very long, MAE increases noticeably, and RMSE shows partial rebounds. This suggests that excessively long windows may introduce redundant historical information. Irrelevant features and accumulated noise can weaken the model's ability to identify key structural dependencies. In heterogeneous spatiotemporal settings, patterns across nodes and across time scales are not aligned. Very long windows make it difficult for the model to distinguish useful dependencies from noise. This leads to error escalation in several metrics.

For percentage-based metrics, the differing trends between MAPE and the Composite Score show the multiscale nature of window effects. MAPE is sensitive to proportional deviations. It therefore exhibits strong fluctuations and alternating peaks and valleys under different window lengths. This reveals the model's sensitivity to window size when handling patterns with varying magnitudes. In contrast, the Composite Score shows an overall upward trend. This indicates that longer windows can still improve overall structural awareness, even if local errors fluctuate. Overall, the length of the time window has a substantial impact on joint modeling of heterogeneous spatiotemporal sequences. The optimal range depends on the balance between historical information, structural complexity, and noise distribution.

We also explore how changing the number of heterogeneous relation types influences generalization performance, and the corresponding findings are presented in Figure 4.

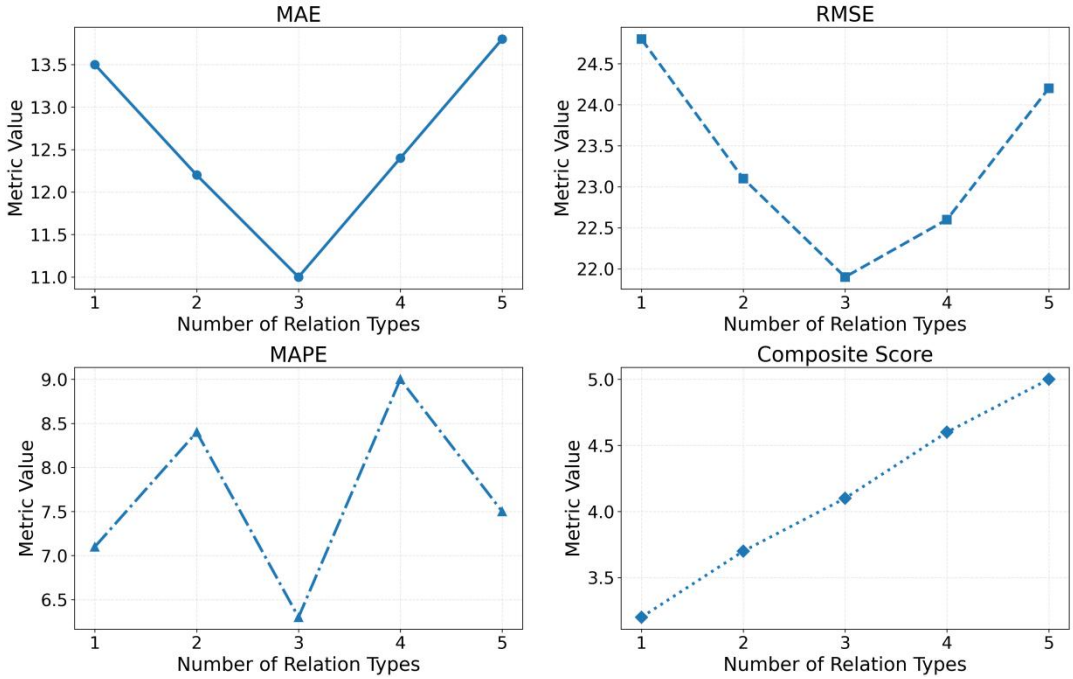

Figure 4. Sensitivity experiment of the number of heterogeneous relation types to the model's generalization ability

Adjusting the number of heterogeneous relation types substantially influences the model's structural understanding of spatiotemporal interactions, yet the evaluation scores do not follow a uniform trajectory, implying that richer relational schemas can help certain aspects of generalization while simultaneously complicating optimization for others. The metrics show inconsistent patterns. This reflects the model's adaptability when dealing with multi-relational structures. When the number of relation types is small, the model can only use limited structural information. Both MAE and RMSE

remain high. This indicates that the model cannot fully capture cross-relation dependencies under low heterogeneity. The lack of diverse relational connections limits its ability to represent structural differences and multi-scale dependencies. As a result, the overall error becomes large.

As the number of heterogeneous relation types increases, the model shows improvements across several metrics. RMSE decreases notably when the relation types increase from 1 to 3. This shows that moderate heterogeneity helps the model aggregate structural features across relations more effectively and improves the modeling of complex spatiotemporal dependencies. MAE also decreases in this stage. This reflects reduced local errors and indicates that the model can better capture local structural interactions among nodes. These results show that adding an appropriate number of relation types enriches structural semantics and reduces overall prediction bias.

However, when the number of relation types continues to increase, some metrics begin to rise again. Both MAE and RMSE increase when the relation types reach 4 and 5. This indicates that high levels of heterogeneity may introduce structural noise. The information aggregated from many relations becomes complex or even conflicting. This makes the feature space distribution less stable. In this situation, the model struggles to balance different relational patterns and cannot extract the key structural features effectively. The overall errors increase again. This reflects the modeling difficulty caused by excessive relational complexity.

In contrast, MAPE and the Composite Score change more smoothly. Yet they still reflect the influence of multi-relational structures on generalization. MAPE shows slight fluctuations across different relation counts. This indicates that proportional errors are somewhat sensitive to heterogeneous structure changes but do not deteriorate significantly. The Composite Score increases with higher heterogeneity and reaches its highest value when the relation types reach 5. This shows that high heterogeneity may introduce local disturbances but can improve global structural understanding and generalization. Overall, these results show that a moderate number of heterogeneous relations is essential for enhancing structural representation. Both insufficient and excessive relational complexity can weaken the model's performance in spatiotemporal joint forecasting.

5. Conclusion

This study proposes a self-supervised joint forecasting framework for spatiotemporal heterogeneous time series. The goal is to enhance the model's ability to understand multi-source heterogeneous structures and improve generalization when labeled data are limited. The framework introduces self-supervised tasks that include spatial and temporal masking, structural reconstruction, and relation modeling. These tasks enable the model to extract latent patterns across relations and across scales directly from raw data. As a result, the model gains robust spatiotemporal representation ability in environments where multiple structural dependencies coexist. The framework also reduces the reliance on labels and addresses the weak transferability of traditional supervised forecasting methods. It offers a new perspective for modeling structural uncertainty in complex systems. Extensive experiments show that the model maintains stable adaptability to heterogeneous structures, temporal dynamics, and multivariate coupling. This demonstrates its advantage in unified spatiotemporal modeling.

As the number of heterogeneous relations, the time window length, and the spatial encoding dimension change, the model exhibits clear structural sensitivity. This further shows that different structural factors have varying levels of influence on prediction accuracy, stability, and generalization in complex spatiotemporal systems. The self-supervised framework maintains strong discriminative ability under structural variation. It automatically captures pattern changes caused by structural inconsistency and cross-domain transfer. This allows the model to remain applicable in more complex scenarios. The framework is suited for sensor networks, traffic systems, supply chain monitoring, and urban computing. Its ability to automatically construct structural priors offers a scalable paradigm for spatiotemporal modeling in large-scale, unlabeled, or highly heterogeneous systems.

The proposed framework has important implications for future intelligent decision-making systems in large-scale real-world environments. As spatiotemporal data sources continue to diversify, structural coupling among subsystems will become more complex. Different modalities, node types, and relational graphs will coexist on a large scale. A self-supervised joint forecasting framework can learn these structural dependencies efficiently without manual labels. This provides a foundation for building spatiotemporal forecasting models with higher robustness and broader applicability. By reducing the need for labeled data, the framework lowers deployment costs. It can be integrated more easily into practical workflows in smart transportation, industrial monitoring, risk warning, and energy dispatch. This supports the transition of these sectors toward higher levels of intelligence and automation.

Future work may extend the framework in several directions. One direction is to explore cross-modal inputs, dynamic graph structures, and multi-task learning mechanisms to strengthen the capture of dynamic spatiotemporal relations. Another direction is to integrate advances in self-supervised learning for large-scale graph reasoning into more complex settings such as multi-agent collaborative forecasting, long-term evolution pattern modeling, and cross-region structural transfer. In addition, combining interpretability analysis, adaptive structure selection strategies, and efficient inference mechanisms will further enhance deployability and reliability. These developments will enable the framework to play a broader role in next-generation large-scale intelligent systems. Overall, the ideas presented in this study enrich the theoretical foundations of spatiotemporal forecasting and provide an important basis for building real-world intelligent spatiotemporal analysis tools.

References

[1] Bao Y, Huang J, Shen Q, et al. Spatial-temporal complex graph convolution network for traffic flow prediction[J]. Engineering Applications of Artificial Intelligence, 2023, 121: 106044.

[2] Wen H, Lin Y, Xia Y, et al. Diffstg: Probabilistic spatio-temporal graph forecasting with denoising diffusion models[C]//Proceedings of the 31st ACM international conference on advances in geographic information systems. 2023: 1-12.

[3] Ji J, Wang J, Huang C, et al. Spatio-temporal self-supervised learning for traffic flow prediction[C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(4): 4356-4364.

[4] Yang Y, Guan Z, Wang Z, et al. Self-supervised heterogeneous graph pre-training based on structural clustering[J]. Advances in Neural Information Processing Systems, 2022, 35: 16962-16974.

[5] Zhang W, Wang H, Zhang F. Spatio-temporal Fourier enhanced heterogeneous graph learning for traffic forecasting[J]. Expert Systems with Applications, 2024, 241: 122766.

[6] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," arXiv:1710.10903, 2017.

[7] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, " A transformer-based framework for multivariate time series representation learning, " in Proc. 27th ACM SIGKDD Conf. Knowledge Discovery & Data Mining (KDD), 2021, pp. 2114-2124.

[8] J. Y. Franceschi, A. Dieuleveut, and M. Jaggi, "Unsupervised scalable representation learning for multivariate time series, " in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019.

[9] S. Y. Shih, F. K. Sun, and H. Y. Lee, "Temporal pattern attention for multivariate time series forecasting," Machine Learning, vol. 108, no. 8, pp. 1421-1441, 2019.

[10] D. Carpov, N. Chapados, and Y. Bengio, " N-BEATS: Neural basis expansion analysis for interpretable time series forecasting, " arXiv:1905.10437, 2019.

[11] M. Wang, "Multi-level attention and sequence modeling for dynamic user interest representation in real-time advertising recommendation, " 2023.

[12] R. Fang, "Transaction network graph neural networks for automated and robust financial fraud detection in corporate auditing," 2024.

[13] Y. Wang, "Integrating large language models and knowledge graphs for intelligent financial regulatory risk identification," 2024.

[14] A. Xie, " Adaptive privacy-aware federated language modeling for collaborative electronic medical record analysis," 2024.

[15] Y. Xing, " Enhancing advertising recommendation performance via integrated causal inference and exposure bias correction," 2023.

[16] Wu Z, Pan S, Long G, et al. Graph wavenet for deep spatial-temporal graph modeling[J]. arXiv preprint arXiv:1906.00121, 2019.

[17] Zheng C, Fan X, Pan S, et al. Spatio-temporal joint graph convolutional networks for traffic forecasting[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 36(1): 372-385.

[18] Hu N, Zhang D, Xie K, et al. Graph learning-based spatial-temporal graph convolutional neural networks for traffic forecasting[J]. Connection Science, 2022, 34(1): 429-448.

[19] Xu Y, Cai X, Wang E, et al. Dynamic traffic correlations based spatio-temporal graph convolutional network for urban traffic prediction[J]. Information Sciences, 2023, 621: 580-595.