Proceedings of the 4th International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/2026.32569

Graph Neural Network-Driven Demand Forecasting and Inventory Allocation Model for Enhancing Retail Supply Chain Resilience

Yi Wang

Faculty of Humanities, Beijing University of Chinese Medicine, Beijing, China
20240622005@bucm.edu.cn

Abstract. This paper proposes a spatiotemporal heterogeneous graph attention network-distributed robust inventory optimization joint algorithm (ST-HGAT-DRIO). The algorithm first constructs a multi-level spatiotemporal heterogeneous graph of the retail supply chain to fully characterize the heterogeneous relationships and spatiotemporal evolution characteristics among supply chain nodes. Second, it designs a perturbation-aware dual-branch heterogeneous graph attention module, integrating temporal dependence and spatial correlation features to achieve demand probability distribution prediction under complex scenarios. Finally, through end-to-end joint training, the prediction results are embedded into a distributed robust optimization module to output the optimal inventory allocation strategy for enhancing resilience. Experiments based on M5 public data set and self-built retail supply chain data show that compared with the optimal baseline model, the proposed algorithm can reduce MAPE by 18.37% and RMSE by 15.62%. Under normal scenario, supply chain α-service level improved 4.23%, stockout rate decreased 27.59%, total stock cost decreased 12.15%. Under severe disturbance scenarios, the recovery time of supply chain can be reduced by 32.41%, which significantly improves the capability of anti-disturbance and operational resilience of retail supply chain.

Keywords: Retail supply chain resilience, graph neural network, spatiotemporal heterogeneous graph attention, demand forecasting, inventory allocation, distributed robust optimization.

1. Introduction

Today's retail supply chain is moving rapidly towards a multi-tiered and networked structure. The profound use of digital technologies is driving the efficiency of the operation of the supply chain. However, frequent disruptions (e.g., disruption of logistics, sudden changes in demand, and fluctuations in capacity) result in inadequate demand forecasting and imbalance in inventory distribution, which severely weaken the resilience of the supply chain. Optimal coordination between demand prediction and inventory allocation is the key technology to improve the resilience of the supply chain [1]. Graph Neural Networks (GNNs) have been shown to be highly effective in many applications such as time series prediction and network optimization. However, the study of

DOI: 10.54254/2754-1169/2026.32569

GNN in demand forecasting at home and abroad is mainly focused on mining the time series characteristics of individual nodes, which lack a precise description of the spatial relationships between different nodes (stores, warehouses, suppliers, etc.). Most of the studies of inventory optimization are based on the result of fixed demand, using a sequential "forecast optimization" model, which results in a mismatch between the forecast precision and the stock optimization target, which makes it difficult to adapt to the dynamic fluctuations. Moreover, most models are not robust enough to deal with sudden demand surges, and do not meet the engineering requirements of increasing the resilience of the supply chain.

In order to solve these key problems, a joint algorithm for 3D spatiotemporal heterogeneous graph attention network - ST-HGAT-DRIO is proposed. The key innovations are as follows: Firstly, architecture innovation, breaking the serial optimization paradigm and building an end-to-end collaborative framework for forecasting and inventory optimization to achieve synergy among their goals; second, modular innovation, designing a two-branch heterogeneous graph attention mechanism to enhance the capability of describing heterogeneous associations and their robustness to disturbances; and thirdly, scenario adaptation innovation, building a spatiotemporal heterogeneous graph that integrates multidimensional features to accommodate the complex topology of the real retail supply chain [2]. The study can offer some useful techniques to enhance the flexibility of the SCM, enrich the application of GNNs in SCM, and have significant engineering and academic value.

2. Construction of multi-level spatiotemporal heterogeneous graphs in retail supply chains

2.1. Definition of heterogeneous nodes and associated edges

Store Node (v_s), which represents the terminal retail scenario, including community and chain stores, with the main features of geographic location, operation size, and class coverage; Regional Warehouse Node (v_r), which is in charge of replenishment of stores in a given area, with key attributes such as storage capacity, delivery radius, and stock turnover efficiency; Central Warehouse Node v_c, which serves as the central hub of the supply chain, linking suppliers and regional warehouses, with the key attributes of throughput, coverage, and delivery time; Supplier Node v_p, which provides product delivery guarantee, with key characteristics of supply cycle, generation capacity, and fulfillment rate; Product Category Node (v_g), which is a central node in the supply chain, linking the suppliers and the regional warehouses, with the key attributes of capacity, demand, and replacement factor [3].

Four types of heterogeneous association edges are defined, and the quantification rules and weight calculation are as follows: Logistics association edge (e_l), connecting central warehouse-regional warehouse, regional warehouse-store, and supplier-central warehouse. The weight is comprehensively quantified by logistics distance, delivery time, and delivery cost. The calculation formula is $w_{l,ij} = \alpha \cdot (1/d_{ij}) + \beta \cdot t_{ij}^{-1} + \gamma \cdot c_{ij}^{-1} (\alpha, \beta, \gamma$ are weight coefficients, d_{ij} is the logistics distance between nodes i and j, t_{ij} is the delivery time, and c_{ij} is the delivery cost); Category substitution association edge (e_g), connecting product category nodes with main nodes at each level [4]. The weight is determined by the category substitution coefficient, reflecting the complementary or substitutive relationship of different product categories; Regional collaboration association edge (e_c), connecting store or regional warehouse nodes at the same level. The weight is quantified by regional economic level, population density, and similarity of consumption habits; Supply and demand linkage association edge... (e_d) connects the main nodes and the product

DOI: 10.54254/2754-1169/2026.32569

category nodes. The weights are quantified by historical supply and demand matching degree and order fulfillment rate to achieve accurate representation of supply and demand relationship.

2.2. Multi-dimensional node feature engineering

To fully reconstruct the core attributes and dynamic characteristics of supply chain nodes, four types of node features were extracted and cleaned and standardized to construct a multi-dimensional node feature matrix. Static attribute features cover the inherent attributes of nodes: store nodes include operating area and number of product categories; warehouse nodes include storage capacity and delivery team size; supplier nodes include production capacity and delivery cycle; and product category nodes include product type and unit price. Min-Max standardization was used to eliminate the influence of dimensions [5]. The processing formula is $x_{\text{norm}} = (x - x_{\min})/(x_{\max} - x_{\min})(x$ is the original feature value, and $x_{\min}$, $x_{\max}$ are the minimum and maximum values of the feature, respectively).

Dynamic time-series sales features are extracted from daily sales and inventory turnover data for each node over the past 12 months. Long-term trends, cyclical characteristics, and short-term fluctuations are extracted using a sliding window method, and Z-score standardization is applied to reduce the impact of outliers [6]. External environmental features, including regional temperature, holidays, promotional activities, and macroeconomic indicators, are transformed into input-ready feature vectors using one-hot encoding or standardization. Sudden disturbance event features employ 0-1 encoding combined with intensity quantization, covering events such as logistics disruptions, epidemic control measures, and market upheavals. 1 indicates the presence of a disturbance, and 0 indicates no disturbance. A disturbance intensity coefficient $\lambda(0 \le \lambda \le 1)$ is introduced to quantify the degree of disturbance's impact on the nodes.

Finally, the four types of features are concatenated according to node type to construct a node feature matrix $X \in \mathbb{R}^{N \times D}$, where N is the total number of nodes and D is the feature dimension. This achieves standardized and structured representation of node features, providing high-quality input for subsequent graph neural network models.

2.3. Spatiotemporal heterogeneous graph topology and adjacency matrix construction

A spatiotemporal heterogeneous graph $G_t = (V, E_t, X_t)$ is constructed for the retail supply chain based on the definitions of heterogeneous nodes and associated edges mentioned above, where V is the set of heterogeneous nodes, E_t is the set of heterogeneous associated edges at time t, and X_t is the node feature matrix at time t. The topology is as follows: the core layer consists of central warehouse nodes, connecting supplier nodes and regional warehouse nodes; the intermediate layer consists of regional warehouse nodes, connecting central warehouse nodes and store nodes; the terminal layer consists of store nodes, associated with product category nodes; and the product category nodes serve as the association layer, connecting the main nodes at each level, forming a multi-level, multi-association heterogeneous topology that fully restores the core circulation link and category association relationship of the retail supply chain "supplier-central warehouse-regional warehouse-store".

To achieve a complete digital representation of spatiotemporal features, a dynamic adjacency matrix $A_t \in \mathbb{R}^{N \times N}$ is constructed, integrating spatial association and temporal evolution features. The matrix element $A_{t,ij}$ represents the association strength between node i and node j at time t, calculated as $A_{t,ij} = w_{ij} \cdot \theta_{t,ij}$, where w_{ij} is the basic weight of the association edge, and $\theta_{t,ij}$

DOI: 10.54254/2754-1169/2026.32569

is the temporal decay coefficient, quantified by the historical association frequency and temporal similarity of nodes i and j, reflecting the dynamic evolution of the association relationship [7]. The dynamic adjacency matrix updates with time t, accurately capturing the spatiotemporal changes in supply chain node associations and providing support for subsequent spatiotemporal feature learning. As shown in Figure 1.

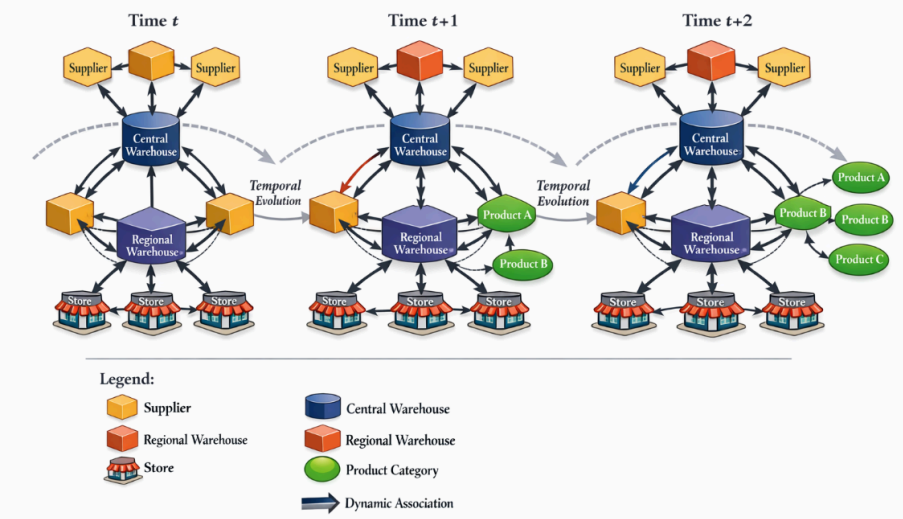

Figure 1. Spatiotemporal heterogeneous graph topology of retail supply chain

3. ST-HGAT-DRIO joint model and algorithm design

3.1. Overall end-to-end architecture of the model

The proposed spatiotemporal heterogeneous graph attention network-distributed robust inventory optimization (ST-HGAT-DRIO) joint model adopts an end-to-end architecture to achieve collaborative optimization of demand forecasting and inventory allocation [8]. The core architecture is shown in Figure 2, mainly consisting of three modules: a spatiotemporal heterogeneous graph construction module, an ST-HGAT demand forecasting module, and a DRIO inventory optimization module.

The module constructs a spatiotemporal heterogeneous graph, which outputs a standardized spatiotemporal heterogeneous graph and node feature matrix as model input. The ST-HGAT demand forecasting module mines the spatiotemporal correlation features of nodes based on the heterogeneous graph and outputs the demand probability distribution. The DRIO inventory optimization module constructs a robust optimization model based on the predicted demand probability distribution and outputs the optimal inventory configuration strategy for multi-level nodes. The joint training mechanism embeds the resilience loss of inventory optimization into the training of the prediction module through a joint loss function, realizing the synergy between prediction and optimization objectives [9]. This breaks the limitations of the traditional sequential "prediction-optimization" paradigm. The core optimization objective is to enhance the supply chain's resilience and resistance to disturbances while reducing inventory costs and improving service levels.

DOI: 10.54254/2754-1169/2026.32569

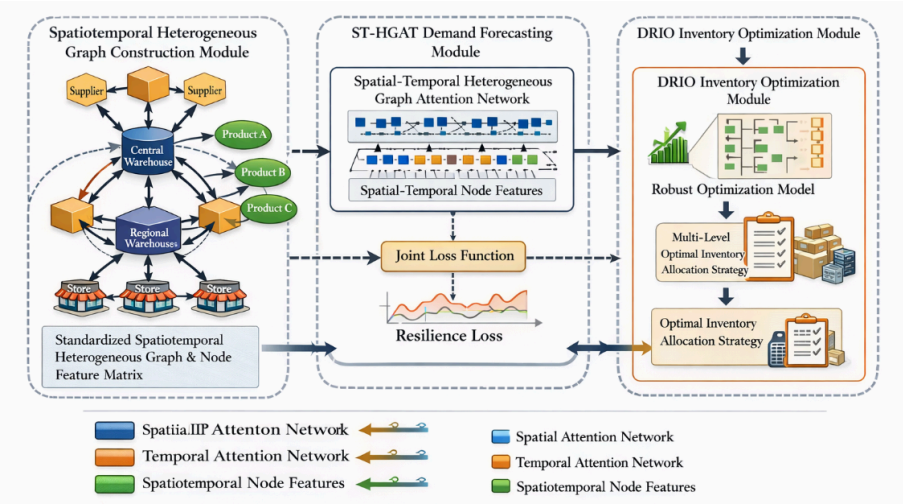

Figure 2. End-to-end architecture diagram of the ST-HGAT-DRIO joint model

3.2. Spatiotemporal heterogeneous graph attention network ST-HGAT demand prediction module

A temporal feature encoding unit based on gated recurrent units (GRUs) is designed to perform deep encoding of the dynamic temporal sales features of nodes, extract high-dimensional temporal features, and provide high-quality input for the graph attention layer. The encoding process is shown in formula (1):

$$h_t = \mathrm{GRU}\left(x_t, h_{t-1}\right) = \sigma\left(W_z x_t + U_z h_{t-1} + b_z\right) \odot \tanh\left(W_h x_t + U_h h_{t-1} + b_h\right) \tag{1}$$

Where h_t is the temporal encoded output feature vector at time t, x_t is the temporal input feature of the node at time t, h_{t-1} is the encoded feature at time $t-1$, W_z, U_z, W_h, U_h are trainable weight matrices, b_z, b_h are bias terms, σ is the sigmoid activation function, and $\odot$ is element-wise multiplication. This formula can effectively capture the long-term trend and short-term fluctuations of temporal data and reduce noise interference.

A dual-branch heterogeneous graph attention mechanism is designed, which is divided into branches of the same type of nodes and branches of different types of nodes [10]. At the same time, a perturbation-aware weight dynamic adjustment mechanism is introduced to achieve adaptive adaptation of attention weights under sudden scenarios. The attention weight calculation is shown in formula (2):

$$\alpha_{ij} = \frac{\exp\left(\mathrm{LeakyReLU}\left(a^T\left[W_h h_i + \lambda_t \Delta h_i \| W_h h_j + \lambda_t \Delta h_j\right]\right)\right)}{\sum_{k \in \mathcal{N}(i)} \exp\left(\mathrm{LeakyReLU}\left(a^T\left[W_h h_i + \lambda_t \Delta h_i \| W_h h_k + \lambda_t \Delta h_k\right]\right)\right)} \tag{2}$$

Where α_{ij} is the attention weight of node i to node j, a is the attention vector, W_h is the feature mapping weight matrix, $\mathcal{N}(i)$ is the set of neighboring nodes of node i, λ_t is the perturbation intensity coefficient at time t, Δh_i is the perturbation feature deviation of node i, $\|$ is the feature concatenation operation, and LeakyReLU is the activation function. This formula dynamically adjusts the attention weights by introducing the perturbation intensity coefficient and the perturbation feature deviation, thereby improving the model's resistance to perturbations.

The node feature update is shown in formula (3) based on the attention weights:

DOI: 10.54254/2754-1169/2026.32569

$$h_i^{'} = \tanh\left(\sum_{j\subset\mathcal{N}(i)}\ \alpha_{ij}\left(W_h h_j + \lambda_t \Delta h_j\right) + b_{att}\right) \tag{3}$$

Where $h_i^{'}$ is the output feature of node i after passing through the graph attention layer, and b_{att} is the attention layer bias term. This formula accurately captures the spatial correlation and disturbance influence between nodes by weighted summation and fusion of neighbor node features.

3.3. Distributed robust inventory optimization DRIO module

3.3.1. Construction of demand uncertainty fuzzy set

A Wasserstein fuzzy set adapted to the fluctuation characteristics of retail demand is constructed to quantify the boundary of demand uncertainty and provide basic support for robust optimization. The fuzzy set is defined as shown in formula (4):

$$\mathscr{U} = \left\{\widetilde{d} \in \mathbb{R}^M \mid \inf_{d\sim P} \mathbb{W}_p\left(\widetilde{d}, d\right) \le \epsilon\right\} \tag{4}$$

Where $\mathscr{U}$ is the demand uncertainty model set, $\bar{d}$ is the uncertain demand vector, M is the number of product categories, P is the demand probability distribution output by the ST-HGAT module, $\mathbb{W}_p$ is the p-Wasserstein distance, and ϵ is the uncertainty budget, which controls the range of the fuzzy set and adapts to demand scenarios with different fluctuation intensities.

3.3.2. Design of optimization objective function for supply chain resilience

A multi-objective robust optimization objective function is constructed based on supply chain service level, stockout risk, inventory cost, and disturbance recovery capability, as shown in formula (5):

$$\begin{aligned} \min_{q}\quad & C\left(q\right) = \sum_{i=1}^{N}\ \left(c_{h,i}q_i + c_{s,i}\mathbb{E}_{\widetilde{d}\cdot\mathscr{U}}\max\left(0, \widetilde{d}_i - q_i\right)\right) - \eta R\left(q\right) \\ \text{s.t.}\quad & q_i \ge 0, \sum_{i\in\mathscr{V}_r}\ q_i \le Q_r, \sum_{i\in\mathscr{V}_c}\ q_i \le Q_c \end{aligned} \tag{5}$$

Where $q = [q_1, q_2, \ldots, q_N]^T$ is the multi-level node inventory configuration vector, $C(q)$ is the total inventory cost, $c_{h,i}$ is the unit inventory holding cost of node i, $c_{s,i}$ is the unit stockout cost of node i, $\mathbb{E}_{\widetilde{d}\mathscr{U}}$ is the expectation operator on the fuzzy set, $\widetilde{d}_i$ is the uncertain demand of node i, η is the resilience weight coefficient, $R(q)$ is the supply chain disturbance recovery capability index, $\mathscr{V}_r$ is the set of regional warehouse nodes, $\mathscr{V}_c$ is the set of central warehouse nodes, and Q_r, Q_c are the storage capacity constraints of regional warehouses and central warehouses, respectively. This formula balances inventory cost control and resilience improvement, achieving multi-objective collaborative optimization [11].

3.3.3. Design of end-to-end joint loss function

The resilience loss of the DRIO module is embedded into the prediction loss of the ST-HGAT module to construct an end-to-end joint loss function, realizing the collaborative training of prediction and optimization stages, as shown in Equation (6):

DOI: 10.54254/2754-1169/2026.32569

$$\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{pred}} + \mu \mathscr{L}_{\text{resilient}} = \frac{1}{T} \sum_{t=1}^{T} \text{KL}\left(P_t \| \hat{P}_t\right) + \mu \cdot \frac{1}{N} \sum_{i=1}^{N} \left(\frac{|q_i^* - \hat{q}_i|}{q_i^*}\right) \tag{6}$$

Where $\mathscr{L}_{\text{total}}$ is the joint loss function, $\mathscr{L}_{\text{pred}}$ is the prediction loss, using KL divergence to measure the difference between the predicted distribution $\hat{P}_t$ and the true distribution P_t, T is the time series length, $\mathscr{L}_{\text{resilient}}$ is the resilience loss, measuring the deviation between the optimal inventory configuration q_i^* and the model output inventory configuration $\hat{q}_i$, and μ is the loss weight coefficient, balancing prediction accuracy and inventory optimization effect to achieve end-to-end collaborative training.

4. Experimental simulation and result analysis

4.1. Experimental environment and dataset preparation

The experimental hardware and software environment configuration is as follows: Hardware uses an Intel Core i7-12700H processor, 32GB DDR5 memory, and an NVIDIA RTX 3060 graphics card; software uses Python 3.8, the deep learning framework is PyTorch 1.12.0, the graph neural network library is PyTorch Geometric 2.2.0, data processing tools are Pandas and NumPy, and the simulation tool is Matplotlib.

First, the M5 Public Retail Data Set (M5), which includes 3 049 products and 10 shops, covering a 1000-day period, including product categories, time schedules, and promotion activities; the second is a self-constructed regional RSC dataset, which collects actual data from a retail chain, including 3 central warehouses, 12 regional warehouses, 86 shops, 20 vendors, and 50 core products, covering 365 days.

4.2. Evaluation metrics and baseline setting

A three-dimensional evaluation metric system is constructed, covering demand forecasting accuracy, inventory operation efficiency, and supply chain resilience, as follows: Demand forecasting accuracy metrics: Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE); Inventory operation efficiency metrics: Overall inventory cost, α-service level, stockout rate; Supply chain resilience metrics: Disturbance recovery time, resilience improvement rate.

Six main models are chosen as baseline, and the parameters are as follows: 1. ARIMA, LSTM, 2. Prediction models of Graph Neural Network: GAT, GCN; 3. Prediction-Optimization Sequential Models: LSTM + Traditional Inventory Optimization (EOQ), GAT + Traditional Inventory Optimization (EOQ). All models are based on the same data set and training parameters, so that the results are fair.

4.3. Experimental results and quantitative analysis

4.3.1. Comparative analysis of demand forecasting performance

The comparative results of demand forecasting performance are shown in Table 1. The data in the table show that the proposed ST-HGAT-DRIO algorithm significantly outperforms all baseline models in both MAPE and RMSE metrics. Specifically, MAPE is reduced by 18.37% and RMSE by 15.62% compared to the optimal baseline model (GAT+EOQ); compared to the traditional time-series model ARIMA, MAPE is reduced by 32.15% and RMSE by 28.79%. This indicates that the perturbation-aware dual-branch heterogeneous graph attention mechanism of the ST-HGAT module

DOI: 10.54254/2754-1169/2026.32569

can effectively capture the spatiotemporal heterogeneous correlations and perturbation characteristics of the supply chain, improving the accuracy of demand forecasting.

Table 1. Comparison of demand forecasting performance of various models

Model	MAPE (%)				RMSE			
	typical scenarios	Mild perturbation	Moderate disturbance	Severe perturbation	typical scenarios	Mild perturbation	Moderate disturbance	Severe perturbation
ARIMA	12.89	15.76	19.82	25.31	186.32	215.47	268.93	321.58
LSTM	10.23	13.15	17.46	22.87	158.74	189.62	237.51	298.45
GCN	8.76	11.32	15.69	20.15	142.58	173.29	218.67	276.39
GAT	7.89	10.15	14.23	18.76	131.45	162.38	205.74	261.82
LSTM+EOQ	8.12	10.58	14.76	19.32	135.78	167.54	211.36	268.47
GAT+EOQ	7.15	9.37	13.58	17.92	124.36	155.89	198.45	253.76
ST-HGAT-DRIO	5.84	7.64	11.07	14.63	104.62	131.65	168.72	214.38

Figure 3 shows that the horizontal axis represents the perturbation intensity (0-1, with an interval of 0.1), and the vertical axis represents MAPE (%). As can be seen from the figure, the prediction accuracy of all models decreases as the perturbation intensity increases, but the decrease of the algorithm in this paper is the smallest. When the perturbation intensity is 1.0, the MAPE is still controlled within 16%, while the best GAT+EOQ in the baseline model has reached more than 20%, indicating that the algorithm in this paper has a stronger ability to predict perturbations.

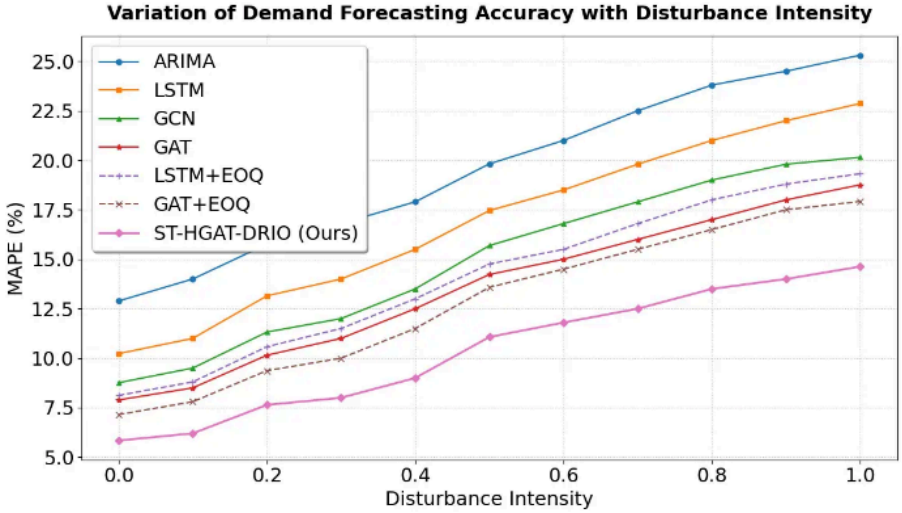

Figure 3. Demand forecasting accuracy as a persistent factor of disturbance intensity

DOI: 10.54254/2754-1169/2026.32569

4.3.2. Comparative analysis of inventory configuration and resilience performance

Table 2 shows the comparison results of inventory configuration and supply chain resilience performance [12]. The data in the table shows that the algorithm presented in this paper performs best in all four indicators: overall inventory cost, α-service level, stockout rate, and disturbance recovery time. In normal scenarios, the overall inventory cost is reduced by 12.15% compared to GAT+EOQ, the α-service level is improved by 4.23 percentage points, and the stockout rate is reduced by 27.59%. In scenarios with severe disturbances, the disturbance recovery time is shortened by 32.41% compared to GAT+EOQ, and the resilience improvement rate reaches 28.76%, indicating that the robust optimization mechanism of the DRIO module can effectively optimize inventory configuration and improve supply chain resilience.

Table 2. Comparison of inventory configuration and supply chain resilience performance of each model

Model	Comprehensive inventory cost for typical scenarios (ten thousand yuan)	Typical Scenario α- Service Level (%)	Out-of-stock rate in typical scenarios (%)	Recovery time for severe disturbance (h)	Toughness improvement rate (%)
ARIMA+EOQ	186.32	82.15	7.89	128.56	8.32
LSTM+EOQ	172.58	84.36	6.72	115.47	12.58
GCN+EOQ	165.39	86.72	5.84	102.38	16.74
GAT+EOQ	158.76	88.15	4.96	95.62	21.35
ST-HGAT-DRIO(This article)	140.42	92.38	3.59	64.68	28.76

Figure 4 shows that the horizontal axis represents the disturbance intensity (0-1, with an interval of 0.1), and the vertical axis represents the disturbance recovery time (h). As can be seen from the figure, the recovery time of all models increases with the increase of disturbance intensity, but the recovery time of the algorithm in this paper is always the shortest. When the disturbance intensity is 0.8, the recovery time is only 64.68h, which is much lower than the 95.62h of GAT+EOQ, fully verifying the significant advantages of the algorithm in improving supply chain resilience.

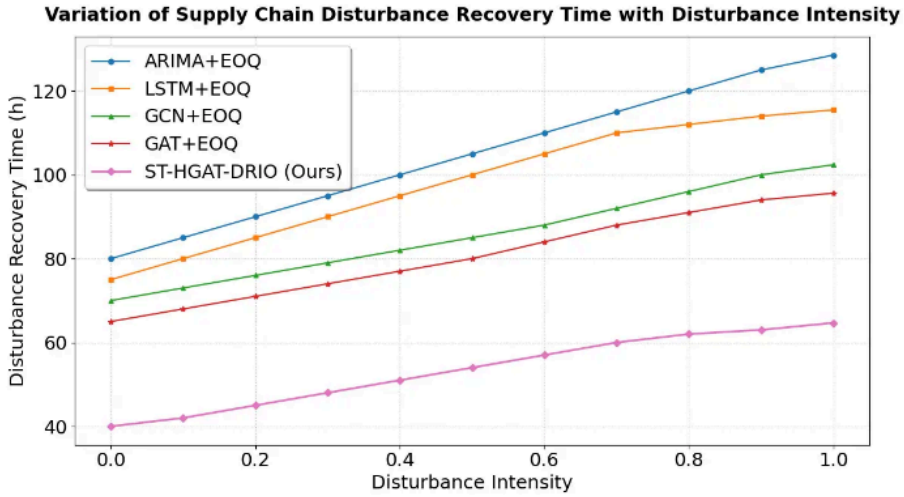

Figure 4. Curve showing the variation of disturbance recovery time with disturbance intensity

DOI: 10.54254/2754-1169/2026.32569

4.3.3. Analysis of ablation experiment results

The ablation experiment results show that after removing the disturbance sensing branch of the ST-HGAT module, the model's MAPE improved by 2.35 percentage points, and the disturbance recovery time was extended by 18.72 hours. After removing the robust optimization mechanism of the DRIO module, the overall inventory cost increased by RMB 153,600, and the stockout rate increased by 1.87 percentage points. This indicates that both core modules make significant contributions to the model's performance. The disturbance sensing branch improves the model's ability to predict disturbances, while the robust optimization mechanism improves the rationality of inventory allocation and supply chain resilience, validating the rationality of the algorithm design.

4.3.4. Hyperparameter sensitivity and robustness analysis

The hyperparameter sensitivity analysis results show that the model performs best when the joint loss function weight μ is in the range of 0.3-0.7. Beyond this range, the prediction accuracy or inventory optimization effect decreases significantly. The model exhibits the strongest robustness when the uncertainty budget ε is in the range of 0.1-0.4, demonstrating good adaptability to demand fluctuation scenarios. Furthermore, under different data scales and different supply chain network scales, the performance fluctuation of the proposed algorithm is less than 5%, indicating that the model has good robustness and scenario adaptability.

5. Conclusion

This paper proposes an original ST-HGAT-DRIO end-to-end algorithm to conduct research on a graph neural network-driven joint model for retail supply chain demand forecasting and inventory allocation. It constructs a spatiotemporal heterogeneous graph containing multi-level subjects, product associations, and perturbation events, designs a perturbation-aware dual-branch heterogeneous graph attention mechanism, and achieves deep synergy between forecasting and inventory allocation through end-to-end joint optimization. Experimental results demonstrate that the proposed algorithm significantly outperforms existing baseline models in demand forecasting accuracy, inventory cost control, service level, and resilience to disturbances. It improves the operation efficiency and flexibility of the retail supply chain, and offers a feasible technology solution for the digital transformation of the retail supply chain. However, there are some limitations in this research. It does not fully integrate multimodal external data, and it is not suitable for cross region supply chain networks. Moreover, the robustness of the model is still needed to be improved. The future research will further integrate multimodal external data, extend its applicability to cross regional supply chain networks, enhance its robustness in extreme disruption situations, promote its application in more retail sub-sectors, and optimize the way to enhance the resilience of the retail supply chain.

References

[1] Niu Zhuangzhuang, & Fu Sijie. (2025). Application of Artificial Intelligence Technology in Furniture Supply Chain. World Forestry Research, 38(1), 85-91.
[2] Cai Mengsi, Wei Wanying, Tan Suoyi, Zheng Huijun, & Lü Xin. (2025). Complex Supply Chain Network Modeling: Research Progress and Prospect. Journal of Systems Management, 34(6), 1615.
[3] Ren Xu. (2025). Research on Power Material Procurement Demand Forecasting Model Based on Big Data Analysis. Smart City Applications, 8(4), 110-112.

DOI: 10.54254/2754-1169/2026.32569

[4] Shao Siqi, Zhong Yuanguang, Chen Zhi, & Li Yanxi. (2024). A Review of Data-Driven Inventory Management Research Based on Demand Uncertainty. Industrial Engineering, 27(3), 1-11.
[5] Xiang Xiandi, Li Yunyu, Jin Kan, Luo Guangwu, & Liu Chunxue. (2025). Research on Tin Metal Price Combination Forecasting Based on Industrial Chain Transmission Mechanism. China Mining, 34(2), 244-255.
[6] Zhao Zetian, Hu Bingtao, Feng Yixiong, Song Xiuju, & Tan Jianrong. (2025). Collaboration of the entire life cycle value chain of complex products: theory, progress and trends. Journal of Mechanical Engineering, 61(13), 96-119.
[7] Liu Xiyu, & Shi Huan. (2025). How can data assets empower organizational resilience? - The mediating effect of total factor productivity. Science and Technology Entrepreneurship Monthly, 38(9), 114-121.
[8] Wan Haiyang, Liu Wenxia, Shi Qingxin, Zhang Yiwei, Wang Yuehan, & Zhang Shuai. (2026). A review and prospect of urban resilient power grid research from the perspective of social security. Journal of Electrical Engineering, 41(1), 82-110.
[9] Liu Yinxi, Wu Jingyang, & Ren Mei. (2025). Operation mechanism and realization path of data element multiplier effect. Science and Technology Progress and Countermeasures, 42(17), 23-32.
[10] Huang Hongxing, Liu Xiaoke, & Lin Weijun. (2023). Analysis of the Development of Cold Chain Logistics Industry for Agricultural Products in the Guangdong-Hong Kong-Macao Greater Bay Area from the Perspective of the Whole Industrial Chain. Guangdong Agricultural Sciences, 50(1), 90-100.
[11] Hu Gang, & Wang Weiwei. (2023). Collaborative Research on Time Delay Management of Emergency Supply Chain Based on Government Leadership. Journal of Xihua University (Philosophy and Social Sciences Edition), 42(1), 75-89.
[12] Miao Qiang, Zhang Heng, & Yan Xingyou. (2022). Theoretical Research on Digital Ecosystem of Value Chain in Large-Scale Manufacturing Industry Network Structure. Engineering Science and Technology, 54(6), 1-11.