EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

A Transformer-Enhanced Multi-Agent Reinforcement Learning Model for Resilience Optimization in Educational Equipment Manufacturing Supply Chains

Yiquan Kong[1,2]

[1]School of Computer Science and Intelligence Education, Lingnan Normal University, Zhanjiang 524048, China
[2]Guangdong Provincial Key Laboratory of Development and Education for Special Needs Children, Zhanjiang 524048, China

Abstract

INTRODUCTION: Educational equipment manufacturing supply chains are vulnerable to demand fluctuations, equipment failures, logistics disruptions, and cross-node risk propagation, while conventional approaches often separate state prediction from recovery decision-making.
OBJECTIVES: This study proposes TMARL-ESCR, a Transformer-enhanced multi-agent reinforcement learning framework for supply chain resilience optimization.
METHODS: The model represents suppliers, manufacturers, logistics providers, and distribution centers as a dynamic network and uses a Transformer to capture long-range temporal dependencies and cross-node interactions. Multitask prediction heads estimate future demand, logistics lead time, available capacity, and disruption risk, and these predictions are fused with current states to support coordinated recovery decisions. Under a centralized training and decentralized execution framework, multiple agents jointly optimize procurement, production, transportation, and inventory reallocation through local-global rewards and explicit operational constraints.
RESULTS:The prediction module achieves a demand WMAPE of 14.38% and a disruption-risk AUC of 0.941. The complete model reaches a 95.2% order fulfillment rate, a 7.1-day recovery time, a 0.5% constraint violation rate, and a resilience score of 0.892. Compared with Transformer-MAPPO, TMARL-ESCR reduces recovery time by 19.3% and improves the resilience score by 5.9%.
CONCLUSION: Experiments integrating M5 Forecasting, AI4I 2020, LaDe, and an SCML-based simulation environment evaluate predictive accuracy and resilience optimization under multiple disruption scenarios, demonstrating improved proactive recovery and coordinated resilience under complex disruptions.

Keywords: Transformer, Multi-agent reinforcement learning, Supply chain resilience, Disruption-risk

Received on 23 July 2026, accepted on 12 August 2026, published on 20 August 2026

doi: 10.4108/eetsis.14151

*Corresponding author. Email: kongyq@lingnan.edu.cn

1. Introduction

The global supply networks are vulnerable to demand volatility, the lack of essential elements, failure of equipment, and breakdowns in logistics. As a result, the goals of manufacturing supply chain management have gone beyond traditional optimization of costs and efficiency to include the maintenance of operational continuity and prompt recovery during disruptions. The concept of supply chain resilience relates to the capability of a system to resist any form of disruption, preserve its most important activities and return to a viable operating level, and has been increasingly studied as a research issue in complex manufacturing systems [1]. Risk exposure, redundancy of resources, and recovery capability depend on the size and structure of supply networks and there are significant variations between them. Supply chain resilience optimization should thus consider the heterogeneity of nodes, network interdependencies and dynamic spread of disruptions [2]. Manufacturing of educational equipment includes a variety of products and parts such as mechanical structural parts, electronic control units, sensors, laboratory consumables, and specific accessories. Multi-node coupling and time variability is also driven by academic calendars, laboratory development programs, and central purchasing schedules, leading to marked multi-node coupling and time variability.

Development of artificial intelligence has opened up fresh possibilities in supply chain state awareness, risk forecasting, and recovery decision-making. The study conducted by Tian et al., based on data on manufacturing companies and structural equation modeling, revealed that data-driven properties could enhance the level of operational performance through the enhancement of supply chain resilience [3]. AI has also been applied to improve risk identification, resource coordination and operational recovery in manufacturing companies [4]. There is a growing focus on ensuring alignment between data, models and business decisions within existing AI implementation frameworks instead of seeing predictive models as isolated tools [5]. In dynamic disruption settings, AI can help detect anomalies, assess their impacts, and formulate adaptive response plans [6]. Its effectiveness, however, depends on supply-chain dynamics and the availability of information. Multistage intelligent decision frameworks are also able to incorporate network configuration, risk assessment and recovery goals into a single optimization process [7].

To identify disruption, Ashraf et al. created a hybrid deep learning framework consisting of an autoencoder, one-class support vector machine and long short-term memory network, which showed that deep learning can be used to detect supply chain disruptions and predict recovery time [8]. Explainable artificial intelligence has also been applied to explain the reasons behind the decisions made under the complex risk environment, thus enhancing the transparency and responsiveness of the supply chain reactions [9]. Generative artificial intelligence has also pushed the boundaries of supply chain information analysis and decision support [10] whereas the combination of AI and big data has increased the ability to detect anomalies and process information in the large-scale supply network [11]. Explainable AI has also been applied in perishable-product supply chains to make associations between customer characteristics and resilience choices [12]. Combination of AI and blockchain has also been investigated to enhance supply chain transparency and disruption-recovery capacity [13]. Fast-moving consumer goods supply chains show that AI-driven risk awareness and operational tuning can be potentially useful in increasing the level of resilience [14]. The other existing research works also indicate that AI can be used in pre-disruption preparedness, during-disruption response, and post-disruption recovery [15]. Nevertheless, its practical effectiveness depends on organizational structures and coordination mechanisms [16]. Organizational capabilities can be a mediator or moderator in the relationship between AI adoption and supply chain resilience [17]. According to the perspective of the organization as information processor, the greatest advantage of AI is enhancing the quality of information sensing and decision-making speed in uncertain contexts [18].

Machine learning has also been used to directly predict the supply chain resilience and complex risk states. In earlier research, machine learning has been employed to investigate the non-linear relationship between trade-network complexity and supply chain resilience [19] and to formulate data-driven resilience prediction models of green supply chains [20]. These solutions have however mostly remained in the form of state forecasting and are unable to come up with direct procurement, production, transportation or inventory adjustment strategies. The use of machine learning and deep learning in demand forecasting has been identified as being significant in supply chain forecasting, but data heterogeneity and long-range dependence continue to be challenging issues [21]. The deep forecasting models can enhance the nonlinear representation of complex demand sequences [22]. The researchers compared the performance of models based on transformers with traditional approaches with the support of the M5 dataset and demonstrated that attention mechanisms can be effectively used to model long-horizon demand forecasting [23]. Transformer models were later applied to probabilistic forecasting problems containing exogenous variables [24], whereas hybrid temporal convolutional network-Transformer architectures have been applied to learn local variation and long-range dependency [25]. However, the current research tends to focus on demand forecasting only, and the simultaneous analysis of demand, lead time, available capacity, and disruption risk is not extensively explored.

Graph-based post-hazard prediction has also been studied at the firm level [26], providing a network-oriented complement to conventional resilience-prediction models.

In contrast to predictive models, reinforcement learning can directly optimize long-term decisions by engaging with an environment. Proximal policy optimization was used by Geevers et al. to multi-echelon inventory optimization and it was proven that deep reinforcement learning can minimize operational costs of complex inventory networks [27]. Multi-agent proximal policy optimization was studied by Mousa et al. to examine the decentralized inventory control and the results revealed that decentralized execution performance could be enhanced if cross-agent information was available to the agents during training [28]. Inventory optimization and dynamic coordination between actors in digital supply chains have since been implemented using multi-agent reinforcement learning [29]. Liu et al. came up with a multi-echelon inventory policy based on heterogeneous-agent proximal policy optimization and proved that multi-agent deep reinforcement learning could minimize system costs and reduce the bullwhip effect [30]. Similar approaches have demonstrated promise to solve distributed decision-making problems in multi-echelon supply chain environments like the Beer Game [31]. Nonetheless, current studies have largely been centered on replenishment and inventory control. The majority of the agents act reactively in response to the present situation, but reactive recovery of supply, manufacturing, logistics, and distribution parties during disruption conditions has got little consideration.

Recent studies are particularly relevant to the positioning of this work. Kotecha and del Rio Chanona [32] combine graph neural networks with multi-agent reinforcement learning for supply-chain inventory control, while Stranieri et al. [33] combine deep reinforcement learning with multistage stochastic programming for inventory management. Ashraf et al. [34] further demonstrate disruption detection within a cognitive digital supply-chain twin. These studies establish graph-aware coordination, hybrid learning-optimization, and digital-twin risk sensing, but they do not jointly connect multi-task forecasts of demand, lead time, available capacity, and disruption risk to heterogeneous supplier, manufacturing, logistics, and distribution agents for disruption recovery. TMARL-ESCR is therefore positioned as an integration and domain-specific extension of these directions rather than as a new Transformer or MARL algorithm.

Taken together, the literature leaves three limitations that are directly relevant to this study. Firstly, demand forecasting, risk of disruption identification, and recovery decision making are usually modeled as independent activities and thus it is unlikely that predictive outputs will be directly used in optimizing subsequent policy. Secondly, available temporal models do not have a common mechanism of depicting the cross-node flow of risk between supplier, manufacturing, logistics and distribution nodes. Finally, multi-agent reinforcement learning may experience conflicts between local goals and working conditions which limit the possibility of policy coordination and action.

In order to overcome these constraints, this paper presents TMARL-ESCR as a resiliency optimization model of educational equipment manufacturing supply chains, incorporating a Transformer with multi-agent reinforcement learning. Initially, the dynamic supply chain state representation is built and a Transformer is applied to predict the future demand, the logistics lead time, the available capacity and disruption risk simultaneously. Secondly, the projections are integrated into the multi-agent decision-making process where supplier, manufacturing, logistics and distribution agents cooperate in recovery in the context of centralized training and decentralized implementation. Lastly, a combined local-global reward structure, operational-constraint penalization, and combined prediction-decision training are presented to transform the AI-based predictions directly into actionable supply chain recovery policies. The predictive performance, overall optimization effectiveness, module contributions, and robustness in the case of complex disruptions are used to assess the proposed model.

2. Transformer–Multi-Agent Reinforcement Learning Model for Resilience Optimization in Educational Equipment Manufacturing Supply Chains

The supply chains that manufacture educational equipment have significant variations in demand, are spread across several geographic locations to source components, have limited production and delivery time frames, and experience cross-node spreading of disruption events. In response to these issues, this paper proposes to develop a supply chain resilience optimization model, TMARL-ESCR, that incorporates both a Transformer and multi-agent reinforcement learning. The initial representation of the model considers suppliers, educational equipment manufacturers, logistics nodes and regional distribution nodes as a dynamic supply chain network. Long-range dependencies on heterogeneous time-series data are extracted using a Transformer, and demand prediction, supply delay, loss of capacity, and disruption risk at nodes are predicted. Then these predictions are incorporated into a multi-agent reinforcement learning framework where agents acting on behalf of various supply chain actors decide about procurement, production, redistribution of inventories, transport and allocation of orders. The model centralizes training and decentralizes implementation to coordinate the behavior of individual agents to enhance the fulfillment of orders, increase the rate of recovery and the continuity of supply chain services as well as reduce the cost of operations.

2.1 Dynamic State Modeling of the Educational Equipment Manufacturing Supply Chain

The educational equipment manufacturing supply chain normally includes suppliers of raw materials and important parts, equipment manufacturers, warehousing and logistics center, regional distribution units and end-demand organizations including schools and laboratory training facilities, as depicted in Figure 1.

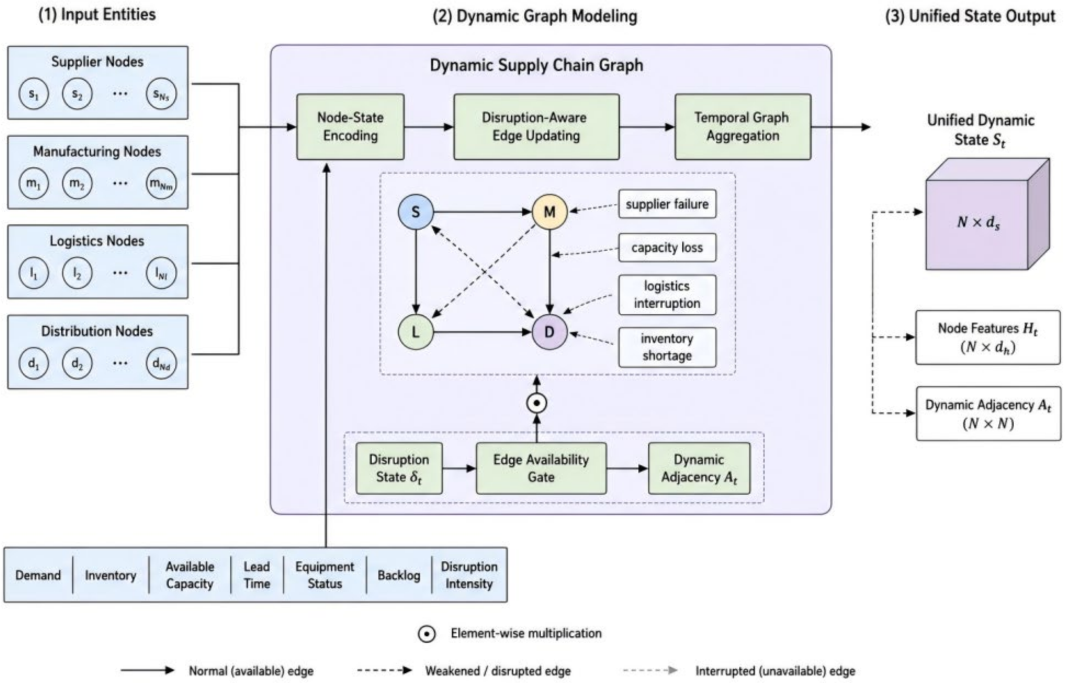

Figure 1. Disruption-aware dynamic supply chain state model

To characterize the time-varying interactions among supply chain entities, the educational equipment manufacturing supply chain is modeled in this study as a time-dependent directed graph that captures the exchange of materials, orders, and state information [35]:

$$G_t=(V,E_t,A_t) \tag{1}$$

where $V=\{v_1,v_2,\ldots,v_N\}$ is the set of supply chain nodes, E_t is the set of active supply relationships at time t, and A_t is the dynamic adjacency matrix. When a supplier suspends production, a transportation route is disrupted, or the inventory of a node falls below the safety-stock level, the corresponding edge weights are updated to reflect the impact of the disruption on supply network availability [36].

Educational equipment products typically comprise multiple classes of materials and components, including structural components, electronic control modules, sensors, laboratory consumables, and specialized accessories [37]. The state of node i at time t is represented as:

$$x_{i,t}=[I_{i,t},D_{i,t},C_{i,t},L_{i,t},P_{i,t},U_{i,t},B_{i,t}] \tag{2}$$

where $I_{i,t}$ represents the stock of raw materials, work-in-progress, and finished products; $D_{i,t}$ refers to the order demand; $C_{i,t}$ represents the available production or transport capacity; $L_{i,t}$ refers to the lead time of procurement and distribution; $P_{i,t}$ is the price of procurement, production, and transportation; $U_{i,t}$ is equipment utilization; and $B_{i,t}$ is the number of back logged orders.

Since demand depends on the start of academic semesters, laboratory development periods as well as centralized procurement schedules, end-market demand has high temporal variation. Demand for product k is characterized by the following equation:

$$D_{k,t+1}=D_{k,t}+\Delta D^{trend}_{k,t}+\Delta D^{season}_{k,t}+\varepsilon_{k,t} \tag{3}$$

where $\Delta D^{trend}_{k,t}$ is the long-term demand trend, $\Delta D^{season}_{k,t}$ is seasonal variations due to academic terms and purchasing cycles, and $\varepsilon_{k,t}$ is a random disturbance term.

The inventory of material or product k at node i is updated according to:

$$I_{i,k,t+1}=I_{i,k,t}+Q^{in}_{i,k,t}+Y_{i,k,t}-Q^{out}_{i,k,t}-S_{i,k,t} \tag{4}$$

$Q^{in}_{i,k,t}$ and $Q^{out}_{i,k,t}$ represent the inbound and outbound quantities, respectively; $Y_{i,k,t}$ represents the local production quantity; $S_{i,k,t}$ and represents the quantity delivered in response to actual orders. The equation (4) can be used to calculate the supplier inventories, work-in-process inventories in manufacturing plants and finished goods inventories in distribution center.

To characterize capacity reductions caused by raw-material shortages, equipment failures, natural disasters, or transportation disruptions, the capacity availability ratio $\rho_{i,t}$ is introduced:

$$C_{i,t}^{ava}=C_{i,t}^{nom}\rho_{i,t},\rho_{i,t}=1-\delta_{i,t} \tag{5}$$

where $C_{i,t}^{nom}$ is the nominal capacity of node i, and $\delta_{i,t}\in[0,1]$ is the measure of a disruption. The value $\delta_{i,t}=0$ denotes regular functioning, and $\delta_{i,t}=1$ implies full node breakdown.

The order fulfillment rate of the supply chain at time t is defined as:

$$\eta_t=\frac{\sum_{k=1}^{K} S_{k,t}}{\sum_{k=1}^{K} D_{k,t}+\epsilon} \tag{6}$$

where $S_{k,t}$ is the actual delivery amount of product k, and ϵ is a small constant that is used to avoid dividing by zero.

Supply chain resilience metric is defined by simultaneously looking at order fulfillment, recovery speed, and operating cost

$$R=\omega_1 \frac{1}{T}\sum_{t=1}^{T}\eta_t+\omega_2\left(1-\frac{T_{rec}}{T_{max}}\right)-\omega_3\frac{C_{tot}}{C_{max}} \tag{7}$$

where T_{rec} is the recovery time of the supply chain to achieve the desired level of service, C_{tot} is the overall expenditure caused by the purchase, manufacturing, stockpiling, shipment, and expedient changes, ω_1 , ω_2 and ω_3 are weighting factors. This measure does not assess a policy on the basis of cost or fulfillment rate but measures the supply chain resiliency to deal with disruptions, service continuity, and restoration of business activities. Because the composite score is weight-dependent, the same score definition is applied to all methods and scenarios, and the composite value is interpreted together with fulfillment rate, recovery time, and operating cost rather than as a standalone performance claim.

2.2 Transformer-Based Demand and Disruption-Risk Prediction

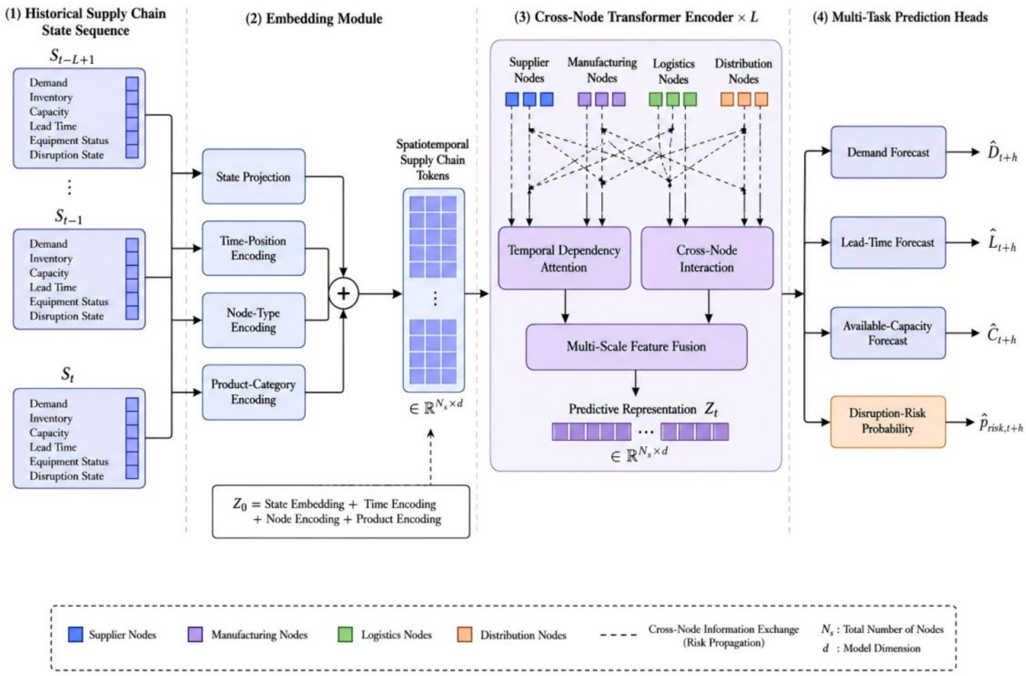

Figure 2. Cross-node Transformer for multitask state and risk prediction

The figure above (Figure 2) demonstrates that traditional recurrent neural networks tend to fail to maintain long-range dependencies in the context of long-horizon demands changes and spread of risks in the supply chain network nodes. The proposed approach thus uses a Transformer to simultaneously model heterogeneous data sources such as past orders, inventory positions, capacities, equipment status and supplier delivery history and logistics lead time, and offers estimates of future states on which multiple agents can make decisions later on [38].

Let the historical supply chain state sequence of length L be:

$$X_t=[s_{t-L+1},s_{t-L+2},\ldots,s_t] \tag{8}$$

where s_t represents the joint state of all the supply chain nodes at time t. The projected input sequence is then mixed with time-position, node-type, and product-category embeddings:

$$H^{(0)}=X_tW_e+E_{pos}+E_{node}+E_{prod} \tag{9}$$

where W_e is the projection matrix of the input, E_{pos} is the temporal order, E_{node} is used to identify the supplier, manufacturing, logistics, and distribution nodes and E_{prod} refers to various products and category of educational equipment.

The query, key and value matrices of the l-th Transformer encoder layer are calculated as:

$$Q^{(l)}=H^{(l-1)}W_Q^{(l)},K^{(l)}=H^{(l-1)}W_K^{(l)},V^{(l)}=H^{(l-1)}W_V^{(l)} \tag{10}$$

The scaled dot-product attention is given by:

$$Attn\left(Q^{(l)},K^{(l)},V^{(l)}\right)=softmax\left(\frac{Q^{(l)}K^{(l)\top}}{\sqrt{d_k}}\right)V^{(l)} \tag{11}$$

d_k with being the dimensionality of the key vectors. Attention weights are features that indicate dependencies between various time steps, supply chain nodes and material conditions [39]. As an example, as the duration of delivery delays by a critical electronic-component supplier keeps growing, the model may correlate such data with further changes in the production schedule and the lack of available end-products.

The multi-head attention output is expressed as:

$$MHA\left(H^{(l-1)}\right)=Concat(h_{ead1},...,head_H)W_O \tag{12}$$

with H being the number of attention heads. Various heads may focus on different patterns related to demand cycles, supply delays, node failures, and congestion of logistics.

The encoder output is achieved after the residual connections, layer normalization and feed-forward network:

$$H^{(l)}=LN\left[\tilde{H}^{(l)}+FFN\left(\tilde{H}^{(l)}\right)\right] \tag{13}$$

where:

$$\tilde{H}^{(l)}=LN\left[H^{(l-1)}+MHA\left(H^{(l-1)}\right)\right] \tag{14}$$

and L_e indicates the size of Transformer encoder layers. In order to detect possible disruptions at a single node, a classification head of disruption risks is added:

$$\hat{p}_{i,t+\mathrm{h}}^{risk}=\sigma\left(W_r\mathrm{h}_{i,t+\mathrm{h}}+b_r\right),\mathrm{h}=1,\ldots,H_f \tag{15}$$

and where $\hat{p}_{i,t+\mathrm{h}}^{risk}$ is the probability of a supply disruption, capacity loss or serious delivery delay at node i in the h-th future period.

The Transformer training objective jointly accounts for continuous-state prediction errors and disruption-risk classification loss:

$$\mathrm{L}_{TF}=\lambda_d\mathrm{L}_{demand}+\lambda_l\mathrm{L}_{lead}+\lambda_c\mathrm{L}_{capacity}+\lambda_r\mathrm{L}_{risk} \tag{16}$$

The optimization of demand, lead-time, and capacity forecasting is based on mean squared error, and the training of disruption-risk prediction is done with weighted cross-entropy. The problem of class imbalance due to significantly higher number of normal operating samples than the number of disruption samples is resolved by assigning more weight to high-risk samples.

2.3 Multi-Agent Collaborative Recovery and Response Decision-Making

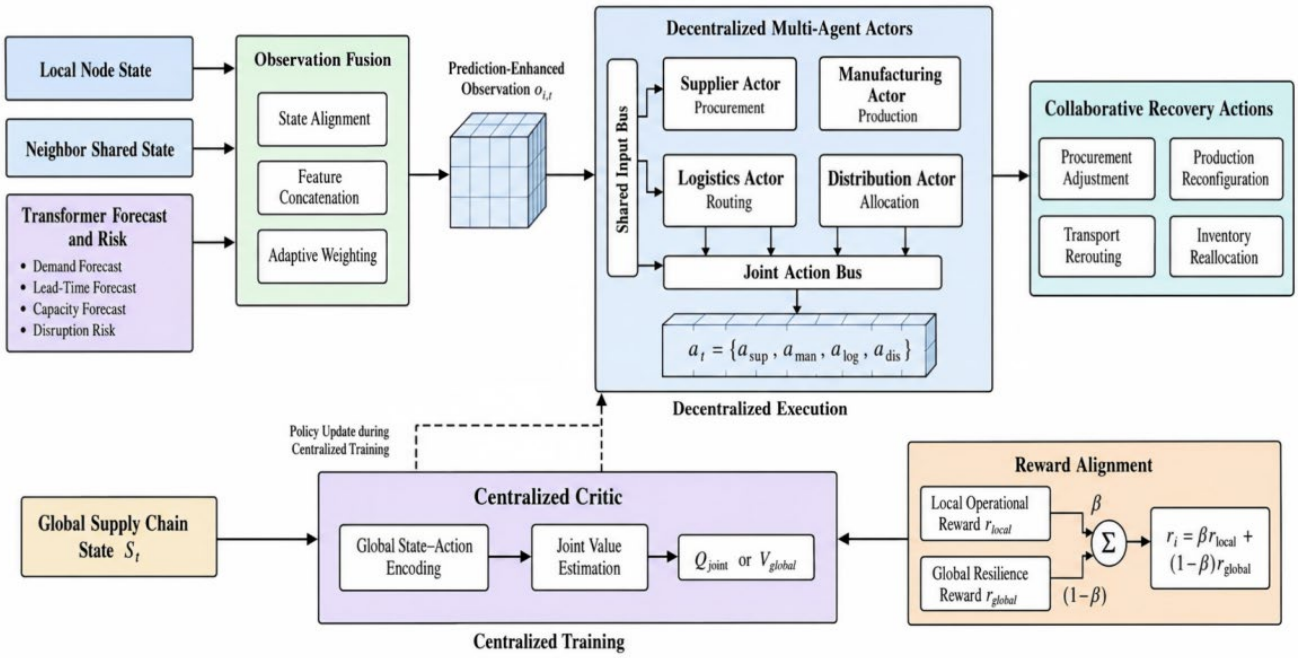

Figure 3. Prediction-enhanced multi-agent collaborative recovery

Figure 3 shows that when future demand and disruption-risk forecasts are derived, the educational equipment manufacturing supply chain optimization problem can be formulated as a partially observable multi-agent Markov decision process:

$$\mathrm{M}=\langle N,S,\{O_i\}_{i=1}^{N_a},\{A_i\}_{i=1}^{N_a},P,\mathrm{R},\gamma \tag{17}$$

where N is the set of agents, S is the global state space, and O_i and A_i are the local observation space and action space of agent i, respectively. P denotes the state-transition function, R is the reward function, and γ is the discount factor.

Based on the functional roles of the actors in the educational equipment supply chain, the agents are divided into suppliers, manufacturing, logistics, and distributors agents. Local observation of agent i is defined by:

$$o_{i,t}=\left[x_{i,t},x_{N_i,t},\hat{Y}_{i,t+1:t+H_f},\hat{p}^{risk}_{i,t+1:t+H_f}\right] \tag{18}$$

in which $x_{N_i,t}$ represents the shareable states of neighboring collaborative nodes, whereas the last two terms are the future-state forecasts and disruption-risk probabilities that are produced by the Transformer [40]. The reinforcement learning policies do not then simply react to the current inventory and order states but can actively stockpile, source from different sources or reserve transportation capacity.

The joint action of all agents is represented as:

$$a_t=\left[a_{1,t},a_{2,t},\ldots,a_{N_a,t}\right] \tag{19}$$

The supplier agent decides on supplier selection, procurement volume as well as the percentage of emergency purchase. The manufacturing agent defines production batch size, switching of production lines and overtime capacity. The logistics agent chooses transportation mode, routing choice and rush shipping ratios. The distribution agent makes a decision about the reallocation of regional inventory, order priority, and deferral delivery plans [41].

In order to avoid the minimization of the cost incurred by an individual actor to the detriment of supply chain performance across the board, a reward system that incorporates both local and global rewards is built [42]. Local reward of agent i is defined as:

$$r^{loc}_{i,t}=-\alpha_1 C^{op}_{i,t}-\alpha_2 C^{inv}_{i,t}-\alpha_3 C^{delay}_{i,t}-\alpha_4 C^{switch}_{i,t} \tag{20}$$

where $C^{op}_{i,t}$ represents procurement, production or transportation costs; $C^{inv}_{i,t}$ is inventory holding cost; $C^{delay}_{i,t}$ is delayed-delivery penalty; and $C^{switch}_{i,t}$ means the cost of supplier switching, production line reconfiguration or adjustment of transportation route.

The global reward evaluates the overall operating condition of the supply chain:

$$r^{glo}_t=\beta_1\eta_t+\beta_2\Delta R_t-\beta_3 B_t-\beta_4 V_t \tag{21}$$

where η_t is the order fulfillment rate, ΔR_t is the change in the resilience level, B_t is the system-wide order backlog, and V_t is the extent of breach of procurement, capacity, inventory or transportation constraints.

The final reward received by agent i is

$$r_{i,t}=\mu r^{loc}_{i,t}+(1-\mu)r^{glo}_t \tag{22}$$

where $\mu_{i,t}\in[0,1]$ balances individual and system-level objectives. Under severe disruptions, μ can be reduced to increase the emphasis placed by individual actors on overall supply chain recovery. The effect of this local-global mixing is examined in Figure 11(d), where intermediate reward weights provide the best overall resilience and extreme weighting reduces either coordination or local-policy flexibility.

During decentralized execution, each agent independently generates an action based on its local observation:

$$a_{i,t}\sim\pi_{\theta_i}\left(a_{i,t}|o_{i,t}\right) \tag{23}$$

where π_{θ_i} is the policy network whose parameters are determined by θ_i. In the centralized training, the value network gets the global state and joint action:

$$Q_\varphi\left(s_t,a_{1,t},\ldots,a_{N_a,t}\right)=E\left[\sum_{\tau=t}^{T}\gamma^{\tau-t}r_\tau\right] \tag{24}$$

where φ represents the parameters of the centralized critic. This model allows the critic to account on the coupling of the actions of various individuals and at the same time permits each enterprise or node to implement its policies independently through the use of locally accessible information in deployment [43].

2.4 Joint Prediction–Decision Optimization and Resilience Constraints

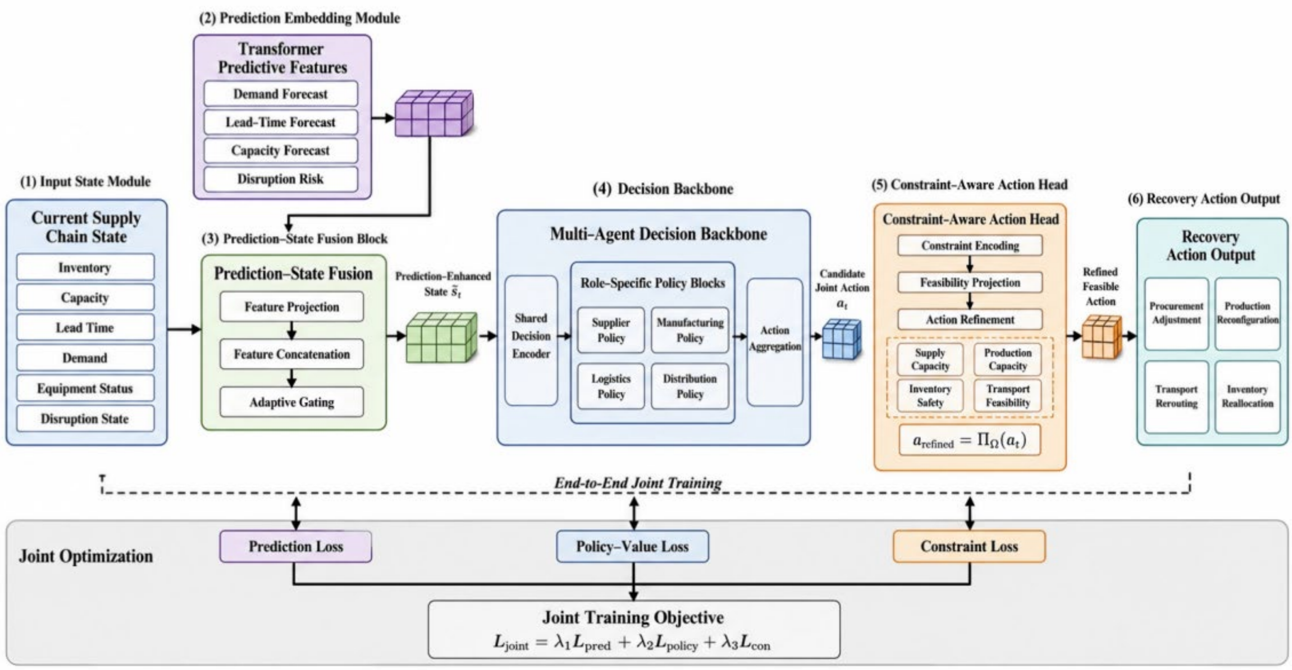

Figure 4. Prediction-enhanced constraint-aware decision network

According to Figure 4, to make sure that the predictions made by the Transformer directly contribute to recovery decisions, the present state of the supply chain, future state predictions, disruption risks representations, and predictive time properties are merged into a state improved by prediction:

$$\tilde{s}_t=\left[s_t,\hat{Y}_{t+1:t+H_f},\hat{p}^{risk}_{t+1:t+H_f},z_t^{TF}\right] \tag{25}$$

where z_t^{TF} is the temporal representation that is obtained by the last Transformer layer. In this form, there are higher-order dependencies between the historical demand, supply delays, and the spread of risks, which can be used to supplement the restricted representational ability of state variables built by hand.

The policy optimization follows the MAPPO family of PPO-based centralized-training/decentralized-execution methods [44]. MAT [45] provides a related Transformer-based MARL architecture, but it formulates multi-agent control itself as sequence modeling; TMARL-ESCR instead uses the Transformer primarily to augment role-specific agents with predictive state information.

To update the policy of each agent, multi-agent proximal policy optimization is used. The clipped objective of the policy of the agent i is:

$$\mathrm{L}_i^{clip}=E_t\left[min\left(r_{i,t}(\theta_i)\hat{\mathrm{A}}_{i,t},clip\left(r_{i,t}(\theta_i),1\text{-}\varepsilon,1+\varepsilon\right)\hat{\mathrm{A}}_{i,t}\right)\right] \tag{26}$$

where:

$$r_{i,t}(\theta_i)=\frac{\pi_{\theta_i}\left(a_{i,t}|o_{i,t}\right)}{\pi_{\theta_i^{old}}\left(a_{i,t}|o_{i,t}\right)} \tag{27}$$

$\hat{A}_{i,t}$ is the estimation of the advantage function, and ε is policy-update clipping parameter. Clipping mechanism constrains the size of policy changes between successive training iterations, thus enhancing training stability in the presence of complex disruption conditions.

The loss function of the centralized value network is

$$\mathrm{L}_V=E_t\left[\left(V_\varphi(s_t)\text{-}\widehat{G}_t\right)^2\right]) \tag{28}$$

where, $\widehat{G}_t$ is the observed cumulative return. In order to maintain exploration of policies, the concept of policy entropy is introduced as

$$\mathrm{H}_i=\text{-}E_{a_{i,t}}\left[log\pi_{\theta_i}\left(a_{i,t}|o_{i,t}\right)\right] \tag{29}$$

The process of decision-making in educational equipment supply chains should meet the constraints of supply capacity, production capacity, balance between inventories, and order fulfillment. The severity of the constraint violation can hence be expressed as a penalty term:

$$\mathrm{L}_{con}=\sum_{t=1}^{T}\left[\left(Q_{i,t}\text{-}Q_{i,t}^{max}\right)_+^2+\left(Y_{i,t}\text{-}C_{i,t}^{ava}\right)_+^2+\left(I_{i,t}^{min}\text{-}I_{i,t}\right)_+^2\right] \tag{30}$$

where $(x)_+=max(0,x)$ and $Q_{i,t}^{max}$ is the maximum supply or transportation capacity and $I_{i,t}^{min}$ is the minimum safety-stock level. This penalty mitigates the infeasible procurement, production and inventory reallocation decisions caused by the policy. It has been found in previous research that integrating explicit constraints with continuous-action reinforcement learning enhances the feasibility of supply chain inventory policies [46]. More broadly, Constrained Policy Optimization treats feasibility as an explicit policy-search requirement rather than only as reward shaping; in TMARL-ESCR, the constraint term serves the narrower purpose of suppressing infeasible supply-chain actions.

To ensure that the coordination of data-driven policy learning and supply chain operational optimization takes place through reinforcement learning, the reinforcement learning objective is simultaneously modeled with the prediction and constraint objectives. The sum total of the training goals is given by

$$L_{total}=\lambda_{TF}L_{TF}-\lambda_{\pi}\sum_{i=1}^{N_a}L_i^{clip}+\lambda_V L_V-\lambda_H\sum_{i=1}^{N_a}H_i+\lambda_{con}L_{con} \tag{31}$$

where λ_{TF}, λ_{π}, λ_V, λ_H, and λ_{con} are the weighting coefficients for the prediction loss, policy objective, value-estimation loss, entropy regularization, and constraint penalty, respectively.

In the process of model training, the Transformer is initially trained on historical normal-operation data and simulated disruption data to acquire the temporal patterns of demand variation, delivery delays and node-level risks. It is then allowed to freeze its parameters as multi-agent policies are learned in the presence of demand shocks, supplier failure, equipment failure and logistics disruption. The stage enhances the responsiveness of the decision model to changes in the distribution and operational uncertainty. Lastly, joint fine-tuning is performed with a lower learning rate such that the representations of risk generated by the prediction module are more closely matched with reinforcement learning goals.

Mechanisms of reinforcement learning based on the nature of supply chain issues may enhance the stability and generalization of policies in stochastic settings. When deployed, each agent develops procurement and production, transportation and inventory adjustment strategies with respect to its local state and the Transformer predictions. It allows taking active measures on resource allocation prior to any disruption and quick switching to alternative sourcing, restructuring capacities and redistribution of inventories across regions in case of a disruption.

The main purpose of the suggested model is not simply to forecast the occurrence of supply chain disruptions, but rather, to convert the risk predictions into actionable and integrated multi-actor decisions. The Transformer gives details on future demand and changing risks whereas multi-agent reinforcement learning distributes recovery work between suppliers, manufacturers, logistics nodes, and distribution nodes. The resilience constraints guarantee that the optimized policies together take order fulfillment, recovery time, operating cost, and realistic implementation requirements.

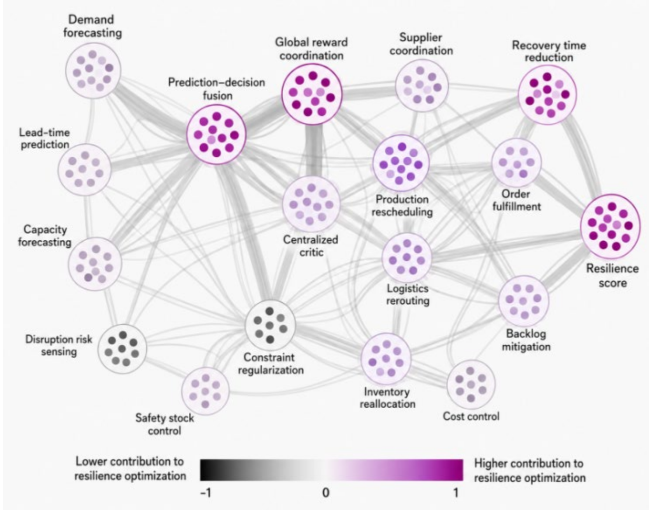

Figure 5. Mechanism association network of predictive characteristics, co-related decisions, and resilience-related results in the TMARL-ESCR

Using Figure 5, the mechanism association network of TMARL-ESCR is shown to demonstrate the relationships between predictive information and coordinated decision mechanisms and supply chain resilience outcomes. The left side nodes are related to information on demand, lead-time, available-capacity and disruption-risk produced by the Transformer-based multitask prediction module. Central nodes are associated with core mechanisms, namely the fusion of prediction-decision-making, the centralized critic, global coordination of rewards, and operational constraints. Nodes on the right side are associated with recovery actions, such as coordination with suppliers, rescheduling of production, re-routing of the logistics, and reallocation of inventory, along with measures of their outputs (order fulfillment, backlog reduction, cost control, recovery time) and a composite resilience score.

However, the network shows that the prediction module cannot directly affect the end results of resilience. It merely affects the recovery behaviors of each actor by increasing their knowledge of the future operating environment and possible threats. The alignment of local strategies with the general recovery goal through centralized estimation of values and coordinated global rewards, plus the existence of operational constraints will guarantee that the created actions meet the needs of supply, production, inventory, and transport. TMARL-ESCR thus forms a full decision chain starting with state forecasting and risk awareness and ending up with coordinated recovery and resilience optimization between multiple actors.

3. Experiments and Results

3.1 Datasets and Experimental Settings

In order to determine the usefulness of TMARL-ESCR in demand forecasting, disruption-risk identification and coordinated supply chain recovery, the experiments integrate the use of public data-driven modeling and multi-agent supply chain simulation. In the demand domain, M5 Forecasting dataset is applied to derive multi-product and multi-region demand trends. The AI4I 2020 dataset is used on the manufacturing side to create the operating states of equipment and production-capacity disturbances. To represent delivery lead times and abnormal delays, the LaDe dataset is used on the logistics side. Decisions concerning procurement, production, transportation and inventory are simultaneously modeled in a multi-agent supply chain using SCML.

Time series of product-store in M5 are connected to educational equipment SKU-regional demand sequence. Information about equipment conditions and failure along with the use of AI4I is employed to dynamically change the capacity of manufacturing nodes, whereas information about delivery-time distributions (LaDe) is used to create the normal transport lead times and regional logistic delays. Simulated supply chain consists of primary suppliers, backup suppliers, manufacturing plants, logistics centers, regional distribution centers and end-demand nodes. Five disruption scenarios have been identified namely: supply disruptions, production disruptions, logistics disruptions, demand shocks, and compound disruptions.

The three public datasets are used as empirical drivers rather than as a literal record of a single educational-equipment enterprise. M5 product-store series provide the temporal demand patterns assigned to SKU-region demand nodes; AI4I equipment-condition and failure records determine manufacturing-node equipment states and capacity-loss events; and LaDe delivery-time observations define baseline lead-time distributions and abnormal logistics delays. These mapped exogenous signals are then injected into the common SCML supplier-plant-logistics-distribution environment, so all decision methods are evaluated under the same material-flow and network structure while retaining the temporal and disruption characteristics of the source data.

Training set is mostly composed of normal operation conditions, individual disruptions, and some specific double-disruptions. Test set also comprises various types of disruptions including triple and quadruple combined disruptions never seen in training to assess robustness and generalization across situations. Every reinforcement learning algorithm assumes the same state space, action set, training budget, and stochastic scenario-generation tool.The datasets, simulated supply chain configuration, disruption scenarios, sequence settings, and major model parameters used in the experiments are summarized in Table 1.

Table 1. Datasets, experimental environment, and major parameter settings

Category	Item	Setting
Public data	Demand data	M5 Forecasting
	Manufacturing data	AI4I 2020
	Logistics data	LaDe
Environment	Multi-agent environment	SCML-based
	Educational equipment SKUs	240
	Suppliers / plants	30 / 4
	Logistics / distribution centers	6 / 10
Scenario	Disruption types	Supply, production, logistics, demand, compound
Sequence	Historical window / forecast horizon	30 / 14 days
Transformer	Layers / attention heads	4 / 8
	Hidden / FFN dimension	128 / 256
MARL	Discount factor / PPO clip	0.99 / 0.20
Evaluation	Independent random seeds	10

The learning rates of the Transformer, actor and critic networks are respectively 1x10 -4, 3x10 -4 and 1x10-4. The dropout rate is 0.1 and the generalized advantage estimation coefficient is 0.95. Every reinforcement learning experiment is repeated with 10 independent random seeds.

Prediction module is determined based on weighted mean absolute percentage error and root mean square error in case of demand forecasting, mean absolute error in case of logistic lead-time and available capacity forecasting, and F1-score and area under the receiver operating characteristic curve in case of disruption risk identification. The supply chain decisions are assessed in terms of order fulfillment rate, recovery time, overall operating cost, average backlog, constraint violation rate, and aggregate resiliency score.

The total operating cost is provided in the form of a normalized cost index that compares with ROP-SS, which has an index of 100. Recovery time refers to the amount of time that it takes the system service level to be restored to 95 percent of what it was before the disruption. All reinforcement-learning results are reported as means over 10 independent random seeds.

3.2 Comparative Experiments and Overall Performance Analysis

The various time-series models are initially compared in terms of demand forecasting, lead-time prediction, estimation of available capacity and disruption-risk analysis. Baselines consist of LSTM, GRU, TCN, Informer, Temporal Fusion Transformer and a conventional transformer.

Table 2. Overall performance comparison of different prediction models

Method	Demand WMAPE/%↓	Demand RMSE↓	Lead-time MAE/h↓	Capacity MAE/%↓	Risk F1 ↑	AUC ↑
LSTM	19.42	8.36	3.48	7.92	0.782	0.846
GRU	18.77	8.01	3.31	7.55	0.794	0.859
TCN	17.93	7.54	3.08	7.11	0.807	0.872
Informer	16.88	7.02	2.84	6.68	0.829	0.894
TFT	16.21	6.76	2.73	6.43	0.841	0.906
Transformer	15.96	6.61	2.69	6.29	0.849	0.914
Proposed Predictor	14.38	5.94	2.31	5.62	0.882	0.941

Conventional recurrent neural networks do not perform well in long term demand variation and abrupt disruption state changes as depicted in Table 2. Informer, TFT, and Transformer can produce lower prediction errors by using attention-based modeling approach. Demand WMAPE (standard Transformer) is 15.96% with the demand WMAPE of proposed predictor being 14.38% which is a relative decrease of about 9.9%. Logistics lead-time MAE is reduced by 2.69 - 2.31 hours and the available-capacity MAE is also reduced by 6.29 - 5.62 %.

The proposed model has an F1-score of 0.882 and an AUC of 0.941 on the task of disruption-risk identification (see Figure 6), as opposed to the standard Transformer which has an F1-score of 0.849 and an AUC of 0.914. The findings show that node-type encoding, cross-node dependency modeling, and multitask prediction enhance not only continuous-state predictions but also the recognition of abnormal supply chain conditions.

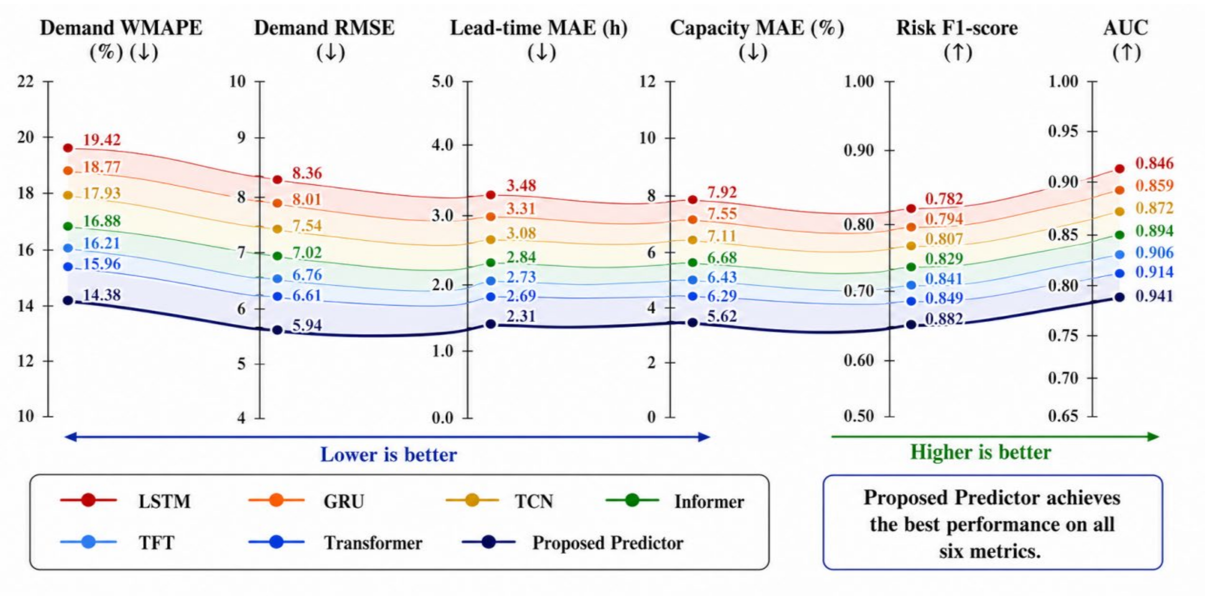

Figure 6. Comparative performance of different prediction models

Table 3. Overall performance of different supply chain resilience optimization methods

Method	Fulfillment Rate/%↑	Recovery Time/day↓	Cost Index↓	Average Backlog↓	Violation Rate/% ↓	Resilience Score ↑
ROP-SS	82.7	18.9	100.0	18.6	0.0	0.642
Rolling-MILP	89.1	12.8	104.6	11.7	0.0	0.748
IPPO	87.4	15.1	97.5	13.9	4.8	0.711
MADDPG	89.8	12.6	98.9	10.9	3.2	0.759
MAPPO	91.5	10.9	98.1	9.1	2.0	0.801
Transformer-MAPPO	93.4	8.8	98.7	7.1	1.5	0.842
TMARL-ESCR	95.2	7.1	96.8	5.4	0.5	**0.892**

As shown in Table 3, the conventional ROP-SS policy maintains high operational feasibility but cannot proactively adjust procurement, production, and inventory decisions according to future disruption risks. It is thus able to achieve just an order fulfillment rate of 82.7 percent and take 18.9 days to recover during a disruption. The fulfillment rate is improved in Rolling-MILP by centralized optimization, however, the cost index goes up to 104.6, which means that more effort would be made to use inventory and emergency supplies when disturbances occur.

MAPPO is one of the reinforcement learning methods that can be used to reduce the decision conflicts between the supply chain actors through centralized training. The forecast with Transformer-based predictions raises the fulfillment rate by 93.4 per cent and minimizes recovery time to 8.8 days when Transformer-MAPPO is applied, which shows that future-state forecasting will help react proactively before any disruption happens.

The full TMARL-ESCR model gives a fulfillment rate of 95.2, recovery time of 7.1 days, and composite resilience score of 0.892. The comparison between the strongest baseline, Transformer-MAPPO, and TMARL-ESCR shows that the fulfillment rate is increased by 1.8 percentage points, the recovery time is decreased by 19.3 percent, and the resilience score is increased by 5.9 per cent.

In particular, the TMARL-ESCR provides a cost index of 96.8 which means that their performance improvement is not at the expense of significant increases in stock or emergency purchasing. The outcome shows a better joint-decision system based on the use of predictive information, coordination across actors, and restrictions based on resilience.

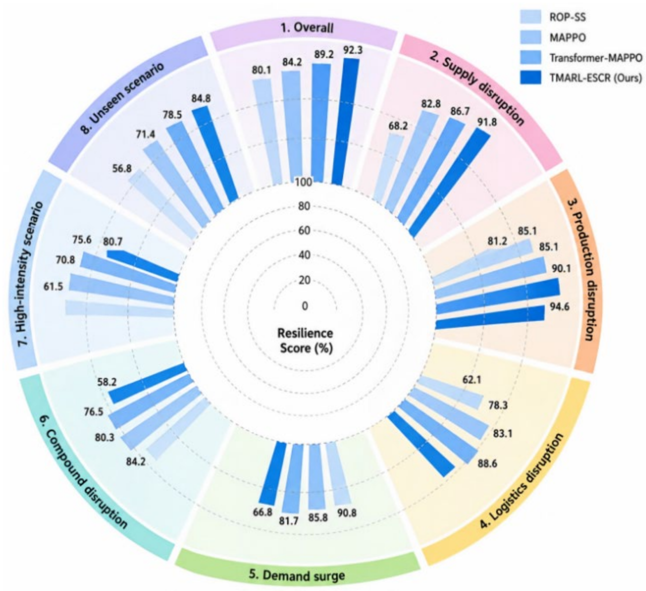

Figure 7. Scenario-wise resilience scores of different supply chain optimization methods

The composite resilience scores of four representative methods in the overall test setting and seven disruption scenarios are compared in Figure 7. In order to enhance the legibility of the radial chart, resilience scores between 0-1 are multiplied by 100 and shown on a percentage basis. TMARL-ESCR obtains the best score in all situations and Transformer-MAPPO usually performs better than MAPPO. The implication of this finding is that predictions of the future demand, lead time and capacity are useful in making proactive allocations of resources prior to the occurrence of any disruption.

Each of the methods demonstrates a reduction in performance during compound disruptions, high-intensity disruptions and new combinations of disruptions. However, TMARL-ESCR still has quite high scores of resilience meaning that there is improvement in the stability of recovery when there are complex conditions with regard to global reward coordination, centralization of value estimation and operational constraints.

3.3 Ablation Study and Mechanism Analysis

In order to determine how much each core mechanism contributes, seven model variants are created through the individual elimination of the Transformer prediction module, risk prediction, cross-node interaction, multi-scale fusion, centralized critic, global reward, and constraint penalty.

Table 4. Ablation results of TMARL-ESCR

Variant	Fulfillment Rate/% ↑	Recovery Time/day ↓	Violation Rate/% ↓	Resilience Score ↑
w/o Transformer	88.9	8.6	0.6	0.809
w/o Risk Prediction	91.1	8.3	0.3	0.834
w/o Cross-node Interaction	90.4	8.9	0.4	0.826
w/o Multi-scale Fusion	91.5	8.1	0.4	0.843
w/o Centralized Critic	88.0	9.6	0.7	0.791
w/o Global Reward	86.6	10.0	0.6	0.776
w/o Constraint Penalty	93.6	7.5	6.5	0.831
Full TMARL-ESCR	95.2	7.1	0.5	0.892

Table 4 findings indicate that the core modules influence supply chain resilience in three complementary areas: predictive awareness, actor coordination, and operational feasibility. Deletion of Transformer prediction module decreases the fulfillment rate by 95.2 to 88.9 per cent and raises the recovery time by 7.1 to 8.6 days which proves that a future-state forecasting is a better way to allocate resources proactively before disruptions occur. Elimination of cross-node interaction decreased the resilience score to 0.826, which means that the model of temporal changes at each node will not be enough to reflect the correlated spreading of risk between supply, manufacturing, and logistics nodes.

Centralized critic and global reward affect the overall recovery performance most significantly. The removal of these elements will lower the resilience score to 0.791 and 0.776 respectively and reduce the fulfillment rate to 88.0 and 86.6 respectively. These findings show that centralization of values estimates and global objective alignment are important in order to prevent local optima and ensure coordination between multiple agents. With the elimination of the constraint penalty the fulfillment rate is quite high at 93.6% and the rate of violation of constraints rises by 0.5 percentage points to 6.5 percent. It implies that the primary task of the constraint mechanism is not to enhance reward-related metrics directly, but to minimize infeasible procurement, production, inventory, and transportation behavior.

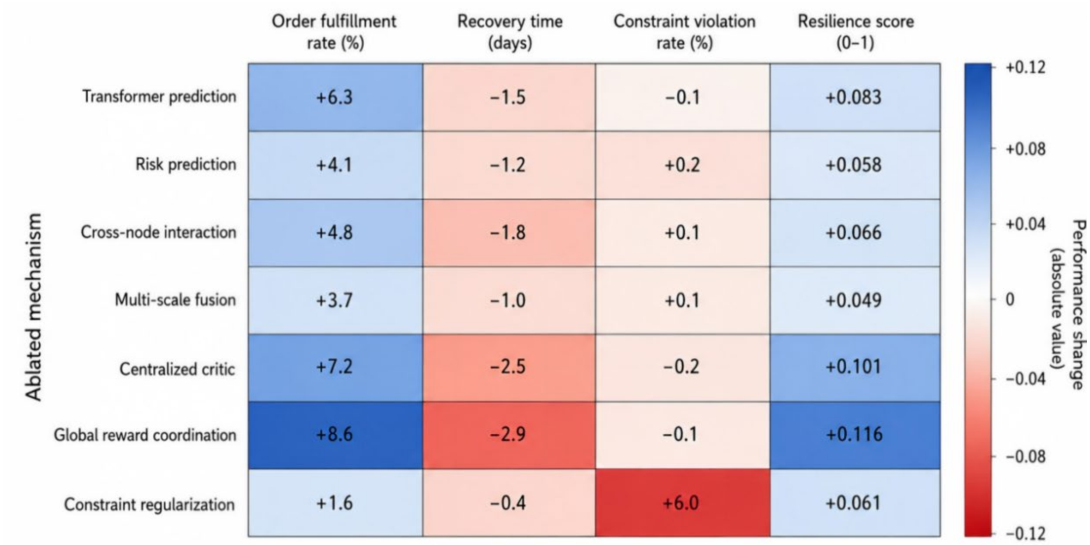

Figure 8. The image shows heat map of the impact of major TMARL-ESCR mechanisms on the supply chain resilience indicators

To explore the impact of core mechanisms on resilience optimization in more detail, Figure 8 shows a heatmap of the difference between absolute performances of the complete TMARL-ESCR model and its ablated versions in terms of fulfillment rate, recovery time, constraint violation rate and composite resilience score. The global reward coordination and the centralized critic have the biggest influence on overall performance. Elimination of these mechanisms lowers the fulfillment rate by 8.6 and 7.2 percentage points respectively, raises recovery time by 2.9 and 2.5 days and lowers the composite resilience score by 0.116 and 0.101 respectively. Such findings support the hypothesis that cross-actor objective alignment and centralized value estimation are essential to multi-agent recovery coordinated.

Prediction based on transformer, interaction between nodes and predicting risks are equally significant factors in recovery performance and it demonstrates that the awareness of future state and risk propagation modeling helps to make preventive decisions prior to the occurrence of any disrupts. Multi-scale fusion is less effective although the hierarchical representation of predictive features is enhanced. Constraint regularization has a small direct impact on fulfillment rate and recovery time, but elimination of constraint regularization leads to an increase in the constraint violation rate by 6.0 percentage points. As a result, the mechanism mainly guarantees that procurement, production, inventory, and transportation-related decisions can be operational. On the whole, the modules help to enhance TMARL-ESCR overall with predictive awareness, cross-actor coordination, and feasibility control.

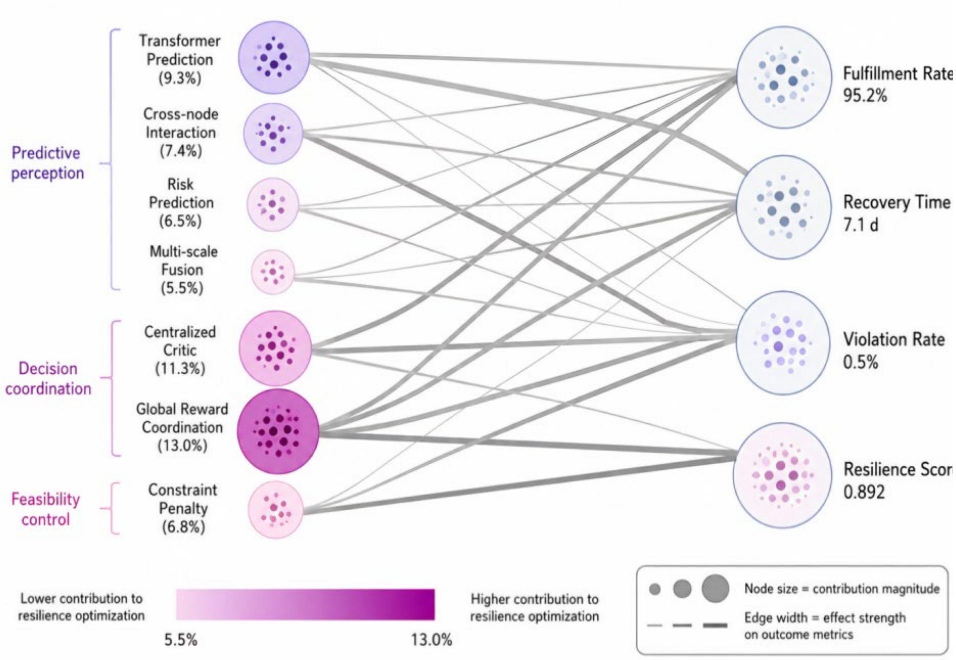

Figure 9. Ablation-derived mechanism contribution network of TMARL-ESCR

In order to provide more quantification in the role played by every core mechanism in supply chain resilience, a mechanism contribution network is built based on the results of the ablations, as it is depicted in Figure 9. Left side nodes stand for three groups of mechanisms: predictive awareness, decision coordination and feasibility control. The size and color intensity of nodes are defined by proportionate decrease in the composite resilience score when the given module is removed.

The global reward coordination and centralized critic have contribution rates of 13.0 percent and 11.3 percent respectively which are higher than the contribution rates of other components. It implies that it is necessary to match the goals of local decisions and the goals of global recovery as well as the values estimation across actors in the process of centralized training to ensure a coordinated multi-agent recovery. The prediction of transformers, cross-node interactions, prediction of risks and multi-scale fusion account at 9.3, 7.4, 6.5 and 5.5, respectively. These mechanisms enhance the supply chain resilience mainly by raising order fulfillment and lowering recovery time by forecasting future-state and modeling risk propagation.

The penalty of constraint accounts for 6.8 percent, mostly due to the fact that it decreases the number of infeasible actions and constraint violations. The nodes on the right summarize the main performance indicators of the whole model, such as the fulfillment rate of 95.2, the recovery time of 7.1 days, the constraint violation rate of 0.5, and composite resilience of 0.892. In total, the network shows that the performance benefits of TMARL-ESCR are not due to any individual element, but rather due to the synergistic influence of predictive awareness, cross-actor coordination, and operational feasibility control.

3.4 Robustness, Generalization, and Computational Efficiency

In order to assess the stability of the model in a more complicated disruption scenario, the severity of the combined disruption is changed.

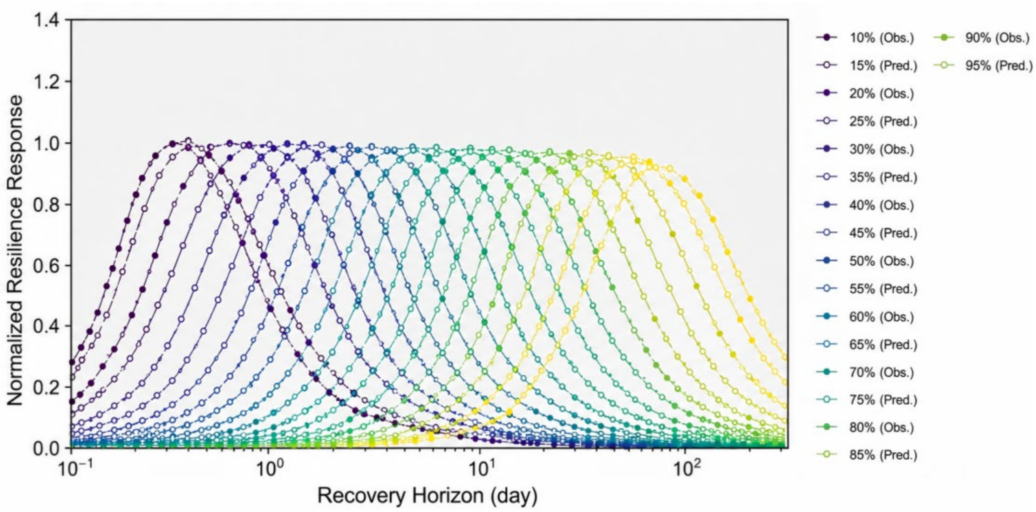

Figure 10. Dynamic recovery-response profiles of TMARL-ESCR under different disruption intensities

Figure 10 shows standardized dynamic recovery responses recorded when re-playing the trained TMARL-ESCR policy with various levels of disruption intensity. Every curve is related to a particular level of disruption intensity. The peak location shows the timespan during which the main recovery response takes place, whereas the curve width denotes the spread of recovery times among simulation cycles.

At increased intensities of disruption (10 -95%) the peaks move to longer recovery times and the response distributions widen at medium and high intensity. It is shown that critical shocks elongate the recovery period and add more uncertainty on the recovery results. Filled markers indicate average observations of repeated simulations and empty markers indicate smoothed fitted values based on these observations. The present figure describes the dynamic recovery behavior of the whole decision policy instead of considering resilience response as another direct prediction objective of the Transformer.

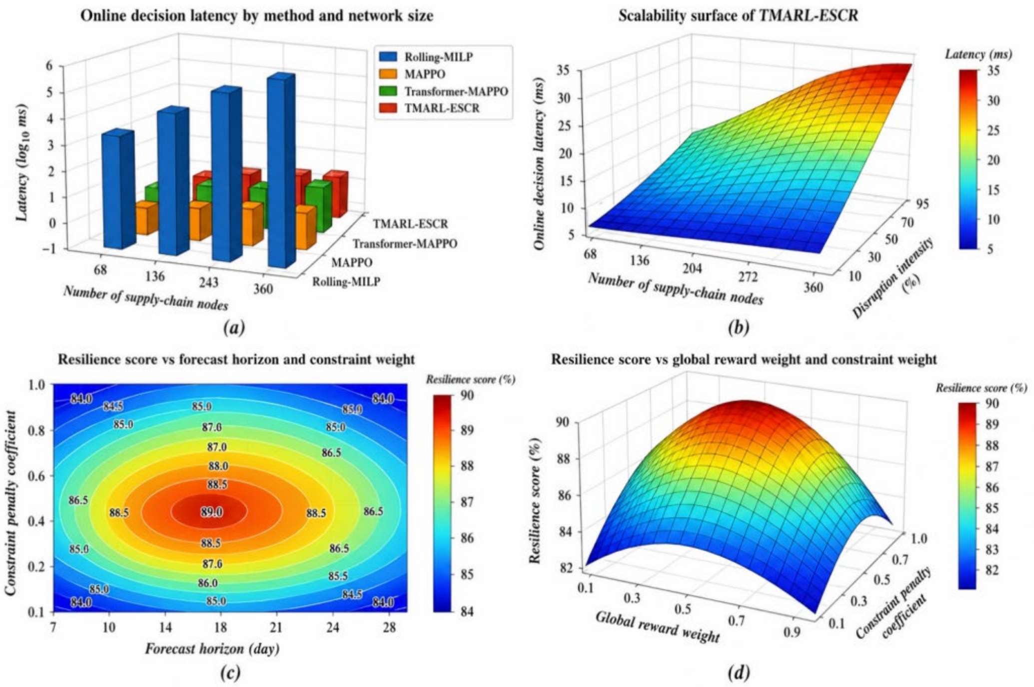

Figure 11. Scalability and parameter sensitivity analysis of TMARL-ESCR

As shown in Figure 11(a), the online solution time of Rolling-MILP becomes significantly larger as the network size increases, but TMARL-ESCR continues to be able to have millisecond-level decision latency even in a network of 360 nodes. Figure 11(b) also illustrates that simultaneous growths in network size and intensity of disruption result in higher online inference latency but still fall within the limits necessary to respond in real-time. This scalability experiment extends the simulated network to 360 nodes, but it does not substitute for validation on real enterprise networks; transfer across enterprise data distributions and operating rules remains to be established.

The results of parameter sensitivity are depicted in Figure 11(c) and 11(d). Resilience scores are higher with a forecast horizon of about 14-18 days and a constraint penalty coefficient of about 0.4-0.5. An underweight global reward weight reduces cooperation between agents, whereas a high-weight global reward weight limits the flexibility of local policies. Intermediate weight values thus give the best overall performance.

Because the forecasts are part of each agent's observation, prediction errors can propagate into anticipatory procurement, capacity, transport, and inventory actions. The dependence analysis in Figure 12 indicates that prediction-related features materially affect the resilience score; systematic bias in demand, lead-time, capacity, or disruption-risk estimates can therefore weaken proactive allocation. Current-state inputs and operational constraints still provide feasibility safeguards, but explicit forecast-noise and bias stress tests remain an important direction for future robustness evaluation.

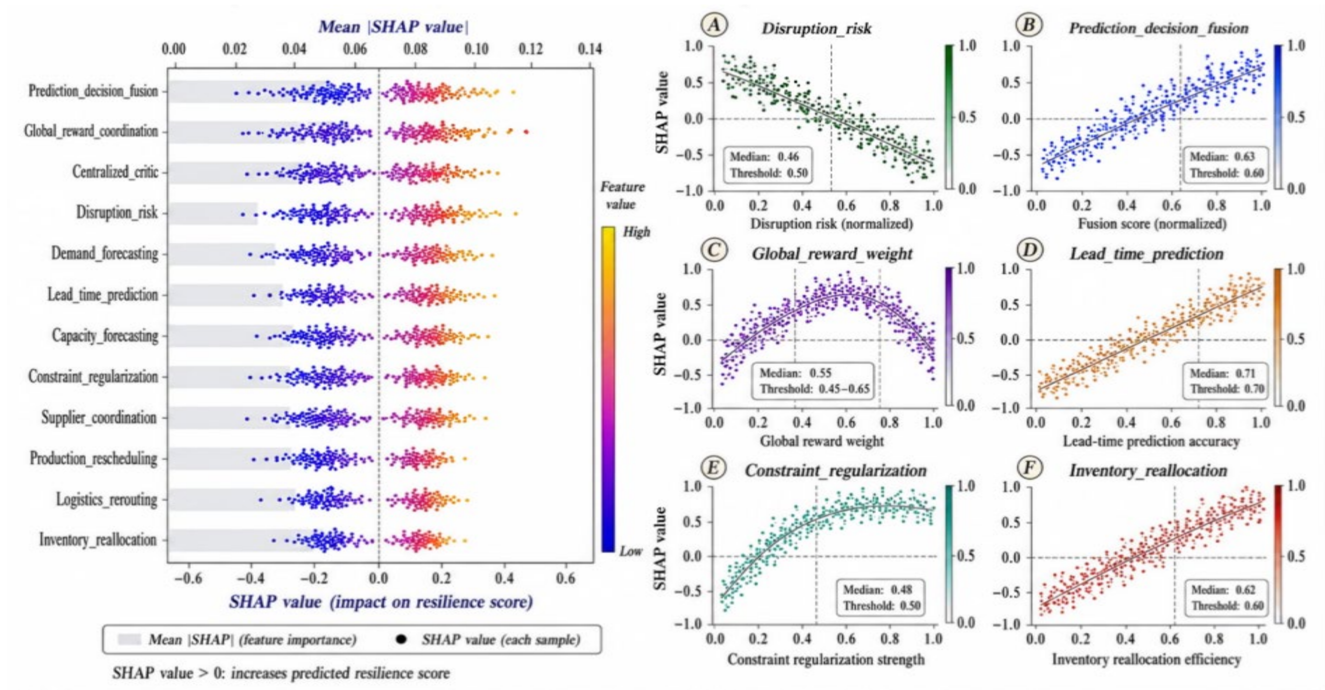

Figure 12. SHAP-based importance and dependence analysis of the comprehensive resilience score in TMARL-ESCR

Figure 12 goes on to consider how various state variables and decision mechanisms affect supply chain recovery time reduction. According to the SHAP beeswarm plot on the left, the highest mean absolute SHAP values are achieved by global reward coordination, the centralized critic and the Transformer prediction module. It is clear that inter-actor objective consistency, global value assessment, and predicting the future are the main aspects of speeding up the recovery of the supply chain.

According to the analysis of dependence, the achievable reduction in recovery time decreases significantly with the increase in the intensity of the disruption which means that the compound disruptions in the severe forms limit the recovery space of the policy. The forecast horizon has a nonlinear relationship with the most significant positive contribution at around 14 days. A shorter horizon does not provide enough lead-time, but a longer horizon is prone to greater errors accumulated over time.

The flexibility of backup suppliers, efficiency in rescheduling production, and ability to respond to the unexpected have a positive correlation with decreases in recovery time and hence supply substitution, capacity reconfiguration, and policy generalization may hasten post-disruption recovery. Improving recovery performance by increasing the constraint penalty weight in a relatively low range has a significant effect followed by a gradual stabilization of the effect. It implies that mild feasibility restrictions are useful in preventing non-productive actions, but too strict limitations cannot be used effectively to bring long-term extra profits.

4. Conclusion

The research paper talks about the problem of demand volatility, cross-node risk propagation, and poor coordination in interactor recovery decisions in educational equipment manufacturing supply chains through the development of TMARL-ESCR, a model that combines a Transformer with multi-agent reinforcement learning. The general findings are as follows.

A multitask prediction mechanism based on Transformers has been devised to predict the future state of the supply chain and the risk of disruptions. The model can be used to learn the long-range temporal dependencies as well as the cross-node interactions by introducing the temporal-position, node-type and product-category encodings and simultaneously learning the demand, logistics lead times, available capacity and disruption risk. The suggested predictor has demand WMAPE of 14.38 percent, logistics lead-time MAE of 2.31 hours, disruption-risk F1-score of 0.882, and AUC of 0.941 which is higher than LSTM, GRU, TCN, Informer, TFT, and the regular Transformer. These findings indicate that joint modeling of heterogeneous states and risks offer increased anticipatory state representations to the subsequent AI-driven decision-making than single-series demand forecasting.

A collaborative recovery system of multi-agent predictions is suggested to jointly optimize supply, manufacturing, logistics, and distribution decisions. TMARL-ESCR incorporates Transformer-based estimates of demand in the future, capacity, lead times, and risk probabilities into the local observations of each agent. Its goal is coordinated by centralised training and decentralised execution with a common local-global reward system. The full-blown model has a fulfillment rate of 95.2, a recovery time of 7.1 days, and a compound resilience measure of 0.892. In comparison to Transformer-MAPPO, it provides an improvement of the fulfillment rate by 1.8 percentage points, the reduction of recovery time by 19.3 percent, and the increase of the resilience score by 5.9 percent. The outcomes of ablation also establish the centralized critic and global reward as critical elements of the coordination between several actors. These results imply that AI-based prediction can be successfully translated into a useful supply chain recovery ability only in conjunction with cross-actor policy learning.

Resilience constraints and joint prediction-decision training enhance the feasibility of policy and the overall optimization performance. Policy constraints include supply capacity, production capacity, safety inventory and transportation feasibility whereas prediction loss, policy objectives, value estimation and constraint penalties are combined into a single optimization process. The overall model has a violation rate of constraints of 0.5% which becomes 6.5% when eliminating the constraint penalty. It is shown by this result that reinforcement learning policies optimized only to maximize cumulative reward can produce a significant amount of infeasible operational actions.

In summary, TMARL-ESCR creates a closed-loop system based on the supply chain state awareness and the prediction of the risk of disruption and coordinated multi-agent recovery by using AI. It offers a practical strategy to the transformation of educational equipment manufacturing supply chains that are reactive in nature into prediction-based and autonomously optimized ones. The next wave of work would be to include multisource operational data of actual businesses and quantify uncertainties as well as interact with online digital twin models to enhance cross-scenario transferability and instantaneous implementation in complex supply networks. Accordingly, the present framework should be viewed as a decision-support tool rather than an autonomous ordering system until enterprise-data retraining, sim-to-real validation, and stricter constraint verification are completed.

Acknowledgements.

This work was supported by a project grant from Supported by the Open Project Program of Guangdong Provincial Key Laboratory of Development and Education for Special Needs Children (Grant No.TJ202301), Guangdong Province Philosophy and Social Sciences Plan 2025 East and Northwest Research Project: Research on the Intelligent Supply of Educational Assistance Resources in Western Guangdong Islands under the Background of the "Hundred Million Project"(No. GD20XJY06).

References

[1] Grossman G M, Helpman E, Lhuillier H. Supply chain resilience: Should policy promote international diversification or reshoring?[J]. Journal of Political Economy, 2023, 131(12): 3462–3496.

[2] Safari A, Balicevac Al Ismail V, Parast M, et al. Supply chain risk and resilience in startups, SMEs, and large enterprises: A systematic review and directions for research[J]. The International Journal of Logistics Management, 2024, 35(2): 680–709.

[3] Tian S, Wu L, Ciano M P, et al. Enhancing innovativeness and performance of the manufacturing supply chain through datafication: The role of resilience[J]. Computers & Industrial Engineering, 2024, 188: 109841.

[4] Mukherjee S, Baral M M, Nagariya R, et al. Artificial intelligence-based supply chain resilience for improving firm performance in emerging markets[J]. Journal of Global Operations and Strategic Sourcing, 2024, 17(3): 516–540.

[5] Riad M, Naimi M, Okar C. Enhancing supply chain resilience through artificial intelligence: Developing a comprehensive conceptual framework for AI implementation and supply chain optimization[J]. Logistics, 2024, 8(4): 111.

[6] Samadhiya A, Yadav S, Kumar A, et al. The influence of artificial intelligence techniques on disruption management:

Does supply chain dynamism matter?[J]. Technology in Society, 2023, 75: 102394.
[7] Tirkolaee E B, Torkayesh A E, Tavana M, et al. An integrated decision support framework for resilient vaccine supply chain network design[J]. Engineering Applications of Artificial Intelligence, 2023, 126: 106945.
[8] Ashraf M, Eltawil A, Ali I. Disruption detection for a cognitive digital supply chain twin using hybrid deep learning[J]. Operational Research, 2024, 24: 23.
[9] Sadeghi R K, Ojha D, Kaur P, et al. Explainable artificial intelligence and agile decision-making in supply chain cyber resilience[J]. Decision Support Systems, 2024, 180: 114194.
[10] Jackson I, Ivanov D, Dolgui A, et al. Generative artificial intelligence in supply chain and operations management: A capability-based framework for analysis and implementation[J]. International Journal of Production Research, 2024, 62(17): 6120–6145.
[11] Singh D, Sharma A, Singh R K, et al. Augmenting supply chain resilience through AI and big data[J]. Business Process Management Journal, 2025, 31(2): 631–657.
[12] Jauhar S K, Harinath S, Krishnaswamy V, et al. Explainable artificial intelligence to improve the resilience of perishable product supply chains by leveraging customer characteristics[J]. Annals of Operations Research, 2025, 354: 103–142.
[13] Ceptureanu E G, Ferraro G, Ceptureanu S I, et al. Impact of artificial intelligence and blockchain on supply chain resilience under influence of change management[J]. Soft Computing, 2025, 29(7): 3617–3625.
[14] Singh D, Sharma A, Singh R K, et al. Artificial intelligence enabled supply chain resilience: Insights from FMCG industry[J]. Journal of Global Operations and Strategic Sourcing, 2025, 18(2): 414–441.
[15] Tang H, Wu K, Zhou J. Smarter supply chains, stronger resilience? The impact of AI on preparation, response, and recovery[J]. Economics Letters, 2025, 254: 112488.
[16] Guo X, Chen Y, Xie J, et al. Research on supply chain resilience mechanism of AI-enabled manufacturing enterprises—based on organizational change perspective[J]. Scientific Reports, 2025, 15: 31177.
[17] Yue X, Kang M, Zhang Y. The impact of artificial intelligence usage on supply chain resilience in manufacturing firms: A moderated mediation model[J]. Journal of Manufacturing Technology Management, 2025, 36: 759–776.
[18] Pan H, Zou N, Wang R, et al. Artificial intelligence usage and supply chain resilience: An organizational information processing theory perspective[J]. Systems, 2025, 13(9): 724.
[19] Nawaz W, Li Z. Complexity to resilience: Machine learning models for enhancing supply chains and resilience in the Middle Eastern trade corridor nations[J]. Systems, 2025, 13(3): 209.
[20] Wu D, Li T, Cai H, et al. Machine learning-based prediction of resilience in green agricultural supply chains: Influencing factors analysis and model construction[J]. Systems, 2025, 13(7): 615.
[21] Douaioui K, Oucheikh R, Benmoussa O, et al. Machine learning and deep learning models for demand forecasting in supply chain management: A critical review[J]. Applied System Innovation, 2024, 7(5): 93.
[22] Aldahmani E, Alzubi A, Iyiola K. Demand forecasting in supply chain using uni-regression deep approximate forecasting model[J]. Applied Sciences, 2024, 14(18): 8110.
[23] Oliveira J M, Ramos P. Evaluating the effectiveness of time series Transformers for demand forecasting in retail[J]. Mathematics, 2024, 12(17): 2728.
[24] Caetano R, Oliveira J M, Ramos P. Transformer-based models for probabilistic time series forecasting with explanatory variables[J]. Mathematics, 2025, 13(5): 814.
[25] Alrashdi I, Abd El-Aziz R M, Taloba A I, et al. Hybrid TCN-Transformer model for predicting sustainable food supply and ensuring resilience[J]. Alexandria Engineering Journal, 2025, 127: 380–391.
[26] Geevers K, van Hezewijk L, Mes M R K. Multi-echelon inventory optimization using deep reinforcement learning[J]. Central European Journal of Operations Research, 2024, 32: 653–683.
[27] Mousa M, van de Berg D, Kotecha N, et al. An analysis of multi-agent reinforcement learning for decentralized inventory control systems[J]. Computers & Chemical Engineering, 2024, 188: 108783.
[28] Zhang B, Tan W J, Cai W, et al. Leveraging multi-agent reinforcement learning for digital transformation in supply chain inventory optimization[J]. Sustainability, 2024, 16(22): 9996.
[29] Liu X, Hu M, Peng Y, et al. Multi-agent deep reinforcement learning for multi-echelon inventory management[J]. Production and Operations Management, 2025, 34(7): 1836–1856.
[30] Hu J, Xia L, Huang T, et al. A multi-agent deep reinforcement learning approach for multi-echelon inventory optimization and its application to the beer game[J]. Transportation Research Part E: Logistics and Transportation Review, 2025, 203: 104367.
[31] Zarghami S A, Dumrak J. Evaluating vulnerability of supply chain networks to capacity reduction[J]. Global Journal of Flexible Systems Management, 2024, 25: 629–646
[32] Yang S, Ogawa Y, Ikeuchi K, et al. Post-hazard supply chain disruption: Predicting firm-level sales using graph neural network[J]. International Journal of Disaster Risk Reduction, 2024, 110: 104664.
[33] Tang X, Wang Y, Liu X, et al. Federated graph neural network for privacy-preserved supply chain data sharing[J]. Applied Soft Computing, 2025, 168: 112475.
[34] Wang Z, Ran H, Ren J, et al. PWDformer: Deformable transformer for long-term series forecasting[J]. Pattern Recognition, 2024, 147: 110118.
[35] Zhao J, Chu F, Xie L, et al. A survey of transformer networks for time series forecasting[J]. Computer Science Review, 2026, 60: 100883.
[36] Dehaybe H, Catanzaro D, Chevalier P. Deep Reinforcement Learning for inventory optimization with non-stationary uncertain demand[J]. European Journal of Operational Research, 2024, 314(2): 433-445.
[37] Kaynov I, van Knippenberg M, Menkovski V, et al. Deep Reinforcement Learning for One-Warehouse Multi-Retailer inventory management[J]. International Journal of Production Economics, 2024, 267: 109088.
[38] Kim B, Kim J G, Lee S. A multi-agent reinforcement learning model for inventory transshipments under supply chain disruption[J]. IISE Transactions, 2024, 56(7): 715-728.
[39] Kotecha N, del Rio Chanona A. Leveraging graph neural networks and multi-agent reinforcement learning for inventory control in supply chains[J]. Computers & Chemical Engineering, 2025, 199: 109111.
[40] Burtea R, Tsay C. Constrained continuous-action reinforcement learning for supply chain inventory

management[J]. Computers & Chemical Engineering, 2024, 181: 108518.

[41] Stranieri F, Fadda E, Stella F. Combining deep reinforcement learning and multi-stage stochastic programming to address the supply chain inventory management problem[J]. International Journal of Production Economics, 2024, 268: 109099.

[42] Wang J, Swartz C L E, Huang K. Risk-averse supply chain management via robust reinforcement learning[J]. Computers & Chemical Engineering, 2025, 192: 108912.

[43] Temizöz T, Imdahl C, Dijkman R, et al. Deep Controlled Learning for Inventory Control[J]. European Journal of Operational Research, 2025, 324(1): 104-117.

[44] Yu C, Velu A, Vinitsky E, et al. The surprising effectiveness of PPO in cooperative multi-agent games[C]. Advances in Neural Information Processing Systems. 2022, 35.

[45] Wen M, Kuba J, Lin R, et al. Multi-agent reinforcement learning is a sequence modeling problem[C]. Advances in Neural Information Processing Systems. 2022, 35.

[46] Achiam J, Held D, Tamar A, et al. Constrained Policy Optimization[C]. Proceedings of the 34th International Conference on Machine Learning. PMLR, 2017, 70: 22–31.