Deep Reinforcement Learning-Enhanced Multi-Objective Cold-Chain Vehicle Routing With Dynamic Order Insertion And Carbon Emission

Qinqin Zhang*

School of Business Administration, Zhengzhou University of Science and Technology, Zhengzhou 450064 China

Corresponding author. E-mail: 2309761033@qq.com; 1421202814@qq.com

Received: Apr. 13, 2026; Accepted: May. 16, 2026

Cold-chain logistics is critical for preserving perishable goods but faces challenges of high operational costs, carbon emissions, and dynamic order uncertainties. This paper proposes a Deep Reinforcement Learning (DRL)-enhanced multi-objective optimization framework for cold-chain vehicle routing problem (CCVRP) with dynamic order insertion and strict carbon emission constraints. We first formulate a multi-objective mathematical model that simultaneously minimizes total distribution cost (including transportation, refrigeration, and time-window penalty costs), carbon emissions, and maximizes customer service level. A constrained Markov Decision Process (CMDP) is designed to model the dynamic routing decisions under real-time order insertion and carbon quota limitations. Then, an improved Proximal Policy Optimization with Carbon Constraint Penalty (PPO-CCP) algorithm is developed, which integrates a graph attention encoder for spatial-temporal state representation and a multi-objective reward mechanism with adaptive carbon emission penalty. Extensive experiments on Solomon benchmark and real-world cold-chain datasets demonstrate that our method outperforms traditional metaheuristics (e.g., NSGA-II, ALNS) and baseline DRL algorithms in solution quality, dynamic adaptability, and emission reduction efficiency. Specifically, it achieves 12.3% cost reduction, 15.7% emission decrease, and 28.5% faster response to dynamic orders compared to the best-performing baseline. This research provides an intelligent and sustainable decision-support tool for low-carbon cold-chain logistics operations.

Keywords: Cold-Chain Vehicle Routing Problem; Deep Reinforcement Learning; Dynamic Order Insertion; Multi-Objective Optimization; Carbon Emission Constraints

http://dx.doi.org/10.6180/jase.202610_33.015

1. Introduction

With the rapid development of e-commerce and fresh food industry, cold-chain logistics has become an indispensable component of modern supply chains. Unlike traditional logistics, cold-chain transportation requires continuous temperature control to maintain product quality, leading to higher energy consumption and operational costs [1, 2]. Meanwhile, growing environmental concerns and stringent carbon emission policies have forced logistics enterprises to balance economic efficiency and green development. In practical cold-chain distribution, two critical challenges exist:

(1) Dynamic order insertion: Emergency orders frequently arrive during delivery, disrupting pre-planned routes and requiring real-time adjustment.

(2) Multi-objective trade-offs: Decision-makers must simultaneously optimize conflicting objectives, cost minimization, carbon emission reduction, and service quality improvement under carbon quota constraints.

Traditional optimization methods (e.g., integer programming, metaheuristics) struggle with dynamic environments, suffering from slow re-optimization and poor adaptability [3, 4]. DRL, with its powerful end-to-end decision-making and real-time learning capabilities, offers a promising solution for complex dynamic routing problems. However, existing DRL-based

VRPs rarely integrate cold-chain-specific characteristics (refrigeration cost, temperature-sensitive cargo) with dynamic order insertion and hard carbon constraints [5, 6].

Early CCVRP studies focused on static models with single-objective cost minimization. Recent works extended to multi-objective optimization considering carbon emissions and time windows. For example, Li et al. [7] constructed a multi-objective LRPTW model for fresh products, minimizing cost, emissions, and cargo deterioration. However, most models assume static order sets, failing to address dynamic insertion scenarios.

Dynamic VRPs handle real-time changes like new orders. Qin et al. [8] developed a dynamic model with simultaneous pickup and delivery, but lacked carbon emission constraints. Yang et al. [9] studied low-emission dynamic cold-chain routing but used simplified emission models. DRL has shown superior performance in VRPs. Sun et al. [10] first applied DRL to TSP using pointer networks. For VRPs, attention-based DRL models achieved promising results. However, few DRL studies address CCVRP with dynamic orders and carbon constraints simultaneously. This paper makes the following key contributions to the field of cold-chain logistics and vehicle routing optimization:
(1) We construct a multi-objective cold-chain vehicle routing model that jointly minimizes total distribution cost (transportation, refrigeration, and time-window penalty costs), reduces carbon emissions, and maximizes customer service level under strict carbon emission constraints, while explicitly accommodating dynamic order insertion in real distribution scenarios.
(2) We formulate the dynamic cold-chain routing problem as a Constrained Markov Decision Process (CMDP), which effectively captures real-time order insertion, carbon quota limitations, and cold-chain-specific operational attributes within a unified decision-making framework.
(3) We develop an improved Proximal Policy Optimization with Carbon Constraint Penalty (PPO-CCP) algorithm, which integrates a graph attention encoder for spatial-temporal state representation and a multi-objective reward mechanism with adaptive carbon emission penalty, enhancing both dynamic adaptability and constrained optimization performance.
(4) Through extensive experiments on Solomon benchmarks and real-world cold-chain datasets, we validate that the proposed method outperforms traditional metaheuristics and baseline DRL approaches in solution quality, dynamic response, and carbon reduction efficiency, providing an intelligent and sustainable decision-support tool for low-carbon cold-chain logistics operations.

2. Materials and methods

This section details the innovative modeling framework, constrained Markov decision process formulation, improved PPO-CCP algorithm design, and experimental configuration for the multi-objective cold-chain vehicle routing problem with dynamic order insertion and carbon emission constraints. All designs are original contributions targeting dynamic adaptability, low-carbon compliance, and cold-chain specificity.

2.1. Problem Definition

We extend the standard cold-chain vehicle routing problem (CCVRP) to a dynamic multi-objective carbon-constrained CCVRP (DMO-C4VRP), which integrates four innovative features: (1) real-time dynamic order insertion during distribution, (2) strict total carbon emission quota constraints, (3) simultaneous optimization of cost, emissions, and service level, (4) deep reinforcement learning for fast re-optimization in dynamic scenarios.

2.2. Multi-Objective Mathematical Model

This section establishes a dynamic multi-objective carbon-constrained cold-chain vehicle routing mathematical model with dynamic order insertion. The model takes total distribution cost minimization, carbon emission minimization, and customer service level maximization as three core optimization objectives, and strictly constrains total carbon emissions within the quota. It fully considers cold-chain characteristics such as refrigeration energy consumption, time-window penalty, perishable product loss, and real-time dynamic order insertion.

To ensure the rationality and solvability of the mathematical model, the following basic assumptions are adopted.
(1) The distribution center has a sufficient number of homogeneous cold-chain vehicles with fixed rated load and refrigeration performance.
(2) Dynamic orders arrive randomly during vehicle distribution and can be inserted into the current route in real time without exceeding vehicle load and time constraints.
(3) The temperature of cold-chain vehicles is kept constant during transportation, and refrigeration energy consumption is positively correlated with transportation time and distance.
(4) Carbon emissions are only generated by vehicle fuel consumption and refrigeration energy consumption, and

the total emissions shall not exceed the carbon quota allocated by the policy.
(5) Each customer point is served exactly once by one vehicle, and the vehicle must return to the distribution center after completing all services.
(6) The vehicle speed is constant under normal road conditions, and the service time at the customer point is fixed.

The model constructs three conflicting objective functions to realize multi-objective collaborative optimization.

2.2.1. *Objective 1: Minimize Total Distribution Cost (F_1)*

Total distribution cost includes transportation cost, refrigeration cost, and time-window penalty cost, which is the core economic objective of cold-chain logistics.

$$\begin{aligned}\min F_1 = & \sum_{k\in K}\sum_{i\in V}\sum_{j\in V} c_i d_{ij} x_{ijk} \\ & + \sum_{k\in K}\sum_{i\in V}\sum_{j\in V} c_r t_{ijk} x_{ijk} \\ & + \sum_{k\in K}\sum_{i\in C} c_p D_{ik} y_{ijk}\end{aligned} \tag{1}$$

Where V is the set of nodes, $V = \{0\} \cup C$, where 0 is the distribution center, C is the set of customer points. K is the set of cold-chain vehicles, $K = \{1, 2, \ldots, k, \ldots, |K|\}.(i, j)$ is the transportation path from node i to node $j.c_i$ is the demand of customer point $i.d_{ij}$ is transportation distance from node i to node $j.x_{ijk} = 1$ if vehicle k travels from node i to node $j.y_{ijk} = 1$ if customer point i is served by vehicle k. The first term is transportation cost positively correlated with distance. The second term is refrigeration cost positively correlated with transportation time. The third term is penalty cost for early or late arrival beyond the customer time window.

2.2.2. *Objective 2: Minimize Total Carbon Emissions (F_2)*

Carbon emissions come from vehicle driving fuel consumption and refrigeration system energy consumption, which is the key green objective.

$$\min F_2 = \sum_{k\in K}\sum_{i\in V}\sum_{j\in V} \left(e_t d_{ij} + e_r t_{ij}\right) x_{ijk} \tag{2}$$

$e_t d_{ij}$ is carbon emissions generated by transportation distance. $e_r t_{ij}$ is carbon emissions generated by refrigeration time.

Objective 3: maximize customer service level (F_3)

Service level is inversely proportional to delay time and positively correlated with on-time arrival rate, reflecting the quality of cold-chain distribution service.

$$\max F_3 = \alpha \sum_{k\in K}\sum_{i\in C}\left(1 - \frac{D_{ik}}{LT_i - ET_i}\right) y_{ik} \tag{3}$$

The smaller the delay time D_{ik} denotes the higher the service level. α is used to adjust the impact weight of service level. The above three objectives are normalized and weighted to form a unified multi-objective function:

$$\min F = \omega_1 \bar{F}_1 + \omega_2 \bar{F}_2 + \omega_3 \bar{F}_3 \tag{4}$$

Where $\omega_1, \omega_2, \omega_3$ are weight coefficients satisfying $\omega_1 + \omega_2 + \omega_3 = 1.\bar{F}_1, \bar{F}_2, \bar{F}_3$ are normalized values of each objective.

The model is constrained by load, flow, time window, carbon emission, dynamic order and other rules to ensure the feasibility of routing decisions. The constraint conditions are shown in Table 1.

2.3. CMDP Formulation

To effectively model the sequential decision-making process of the Dynamic Multi-Objective Carbon-Constrained Cold-Chain Vehicle Routing Problem (DMOC ^{4} VRP) with real-time order insertion and strict emission limits, we formally formulate the problem as a Constrained Markov Decision Process (CMDP). Compared with the standard Markov Decision Process (MDP), CMDP introduces explicit constraint conditions on cumulative costs or resources, which is perfectly suitable for cold-chain routing with fixed carbon emission quotas and dynamic order insertion feasibility rules. The CMDP is defined by a six-tuple (S, A, P, R, C, γ), where each component is elaborately designed to capture cold-chain-specific attributes, dynamic order fluctuations, and carbon emission constraints as follows.

2.3.1. *(1) State Space S*

The state space S encodes all observable environmental information required for routing decisions at each decision step t, covering spatial-temporal distribution status, dynamic order information, vehicle status, and carbon emission consumption. The state $s_t \in S$ is a multi-dimensional vector composed of four core modules: node and network state, vehicle state, dynamic order state, carbon emission state [11, 12].

2.3.2. *(2) Action Space A*

The action space A defines all executable routing decisions at each step, designed to adapt to dynamic order insertion and carbon constraint compliance. For each vehicle t in a distributable state, the action $a_t \in A$ includes three optional behaviors: serve a static customer, insert a dynamic order, return to distribution center.

Table 1. Constraint condition statement

Constraint	Explanation	Formula
Node access constraint	Each customer point is served exactly once	$\sum_{k \in K} y_{ik} = 1$
Vehicle load constraint	The total demand of customers served by each vehicle shall not exceed the rated load	$\sum_{i \in C} q_i y_{ik} \leq Q$
Flow balance constraint	Vehicles start from the distribution center and return to the center after service	$\sum_{j \in V} x_{0jk} = 1$
Time window constraint	Vehicle arrival time shall meet the customer time window requirement	$T_{ik} \geq ET_i$
Carbon emission quota constraint	Total carbon emissions of all vehicles shall not exceed the policy quota	$\sum_{k \in K} E_k = F_2 \leq E_{\text{total}}$
Dynamic order insertion constraint	Dynamically inserted orders must be allocated to vehicles with remaining load and time feasibility	$y_{ik} = 1$

All actions are constrained by hard rules: no overload, no time window violation (beyond acceptable penalty), and marginal carbon emission increase of the action shall not cause the total emissions to exceed the quota. This constrained action space ensures decision feasibility in practical cold-chain distribution.

2.3.3. *(3) State Transition Probability* P

The state transition function $P(s_{t+1} \mid s_t, a_t) : S \times A \times S \rightarrow [0,1]$ describes the probability of transferring from current state s_t to next state s_{t+1} after executing action a_t. In this DMOC ^{4} VRP, the transition process is deterministic under fixed vehicle speed, constant road conditions, and known dynamic order arrival rules (stochastic order arrival is embedded in the state update):

a. After selecting a customer point, vehicle position, remaining load, accumulated distance, refrigeration time, and carbon emissions are updated deterministically.

b. Dynamic orders arrive randomly at discrete time steps, and the state is immediately updated to include new order information once they arrive.

c. Carbon emission state is updated synchronously with transportation and refrigeration processes, strictly tracking the quota consumption.

This deterministic transition mechanism simplifies the decision-making process and improves the real-time performance of the DRL agent in dynamic scenarios.

2.3.4. *(4) Reward Function* R

The reward function $R(s_t, a_t) : S \times A \rightarrow \mathbb{R}$ is a multi-objective weighted reward mechanism that guides the agent to optimize cost, carbon emissions, and service level simultaneously. It is composed of three positive incentive terms and one adaptive carbon penalty term:

$$R(s_t, a_t) = \omega_1 R_1 + \omega_2 R_2 + \omega_3 R_3 - R_{\text{penalty}} \tag{5}$$

2.3.5. *(5) Constraint Function C*

The constraint function $C(s_t, a_t)$ formalizes the hard carbon emission constraint of CMDP, which is the core difference from standard MDP. It defines the cumulative carbon emission constraint over the entire decision horizon T :

$$\mathbb{E}\left[\sum_{t=0}^{T} \gamma^t C(s_t, a_t)\right] \leq E_{\text{total}} \tag{6}$$

Where $C(s_t, a_t)$ denotes the carbon emissions generated by executing action a_t in state s_t, $\gamma \in (0,1)$ is the discount factor, and E_{total} is the fixed carbon emission quota. This constraint ensures the total emissions of all cold-chain vehicles never exceed the policy-allocated quota in any routing scheme.

2.3.6. *(6) Discount Factor* γ

The discount factor $\gamma \in (0,1)$ balances immediate and long-term rewards. For dynamic cold-chain routing with real-time order insertion, a moderately large γ (e.g., 0.95) is adopted to prioritize long-term objectives (total cost minimization, cumulative emission reduction) while responding quickly to short-term dynamic order changes.

2.3.7. *(7) CMDP Solution Objective*

The objective of the CMDP is to find an optimal policy $\pi^* : S \rightarrow A$ that maximizes the expected cumulative long-term reward while strictly satisfying the carbon emission constraint:

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid \pi\right] \tag{7}$$

This CMDP formulation unifies dynamic order insertion, cold-chain operational attributes, and carbon emission constraints into a sequential decision-making framework, laying a solid foundation for the improved PPO-CCP algorithm to solve the DMOC ^{4} VRP efficiently.

2.4. PPO-CCP Algorithm

This section proposes an improved Proximal Policy Optimization with Carbon Constraint Penalty (PPO-CCP) algorithm for the Dynamic Multi-Objective Carbon-Constrained Cold-Chain Vehicle Routing Problem (DMOC ^{4} VRP). PPO-CCP integrates a graph attention encoder for spatial-temporal state representation, a multi-objective adaptive reward mechanism, and a carbon-aware penalty module to realize stable training, fast dynamic response, and strict compliance with carbon emission quotas.

PPO-CCP adopts the classic actor-critic framework of PPO and makes three key improvements for cold-chain

routing with dynamic orders and carbon constraints.
a. Graph attention state encoder extracts spatial-temporal features of distribution nodes, vehicle states, dynamic orders, and carbon quota consumption.
b. Multi-objective reward with adaptive carbon penalty balances cost, emissions, and service level while penalizing carbon over-limit violations adaptively.
c. Constrained PPO clip objective limits policy update range to ensure training stability and constraint satisfaction.

To effectively encode the high-dimensional state s_t (node positions, vehicle status, dynamic orders, carbon emissions), a graph attention encoder is designed:

Input graph: Nodes = distribution center + static customers + dynamic orders; Edges = distances and time costs.

Attention mechanism: computes attention weights between vehicles and candidate nodes to highlight nearby, high-priority, or carbon-efficient targets.

Output: A fixed-length latent vector h_t that fuses spatial topology, vehicle remaining load, time windows, dynamic order urgency, and cumulative carbon emissions.

This encoder enables the agent to perceive complex distribution environments and supports real-time insertion of emergency orders.

The standard PPO clipped objective is modified to embed carbon constraint penalties:

$$\begin{aligned} L^{\text{PPO-CCP}}(\theta) = \mathbb{E}\Big[\; & \min\Big(r_t(\theta)\hat{A}_t, \\ & \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t\Big) \\ & - \eta \cdot \mathcal{L}_{\text{constraint}}\Big] \end{aligned} \tag{8}$$

Where $r_t(\theta)$ is importance sampling ratio, $\hat{A}_t$ is the advantage function including carbon penalty. ϵ is clip threshold (typically 0.2). $\mathcal{L}_{\text{constraint}}$ is cumulative carbon violation loss. η is constraint weight. This clipped update stabilizes training and prevents large policy swings that could break carbon or routing constraints.

3. Results and discussion

This section systematically verifies the effectiveness of the proposed DRL-enhanced multi-objective cold-chain vehicle routing framework with dynamic order insertion and carbon emission constraints through comparative experiments, ablation studies, dynamic scenario tests, and real-world case validation. All experiments are implemented on a computing platform with Intel Core i9-12900K CPU, 64GB RAM, and NVIDIA RTX 3090 GPU; the deep reinforcement learning model is coded in Python 3.8 based on PyTorch 1.12, and comparison algorithms are reproduced using standard open-source libraries.

3.1. Experimental Setup

Two types of datasets are adopted to ensure the universality and practicality of the test:
(1) Solomon benchmark dataset [13, 14]: The classical vehicle routing test set with 100 customer nodes, including six subsets (C1, C2, R1, R2, RC1, RC2) with clustered, random, and mixed distribution. It is used to verify the algorithm's performance on standardized routing problems.
(2) Real-world cold-chain logistics dataset [15]: Collected from a fresh food distribution enterprise in Zhengzhou, covering 21 consecutive days of actual orders, including 3 distribution centers, 187 customer points, dynamic order arrival records, vehicle parameters, refrigeration energy consumption, and regional carbon emission quotas. It is used to test the model's adaptability to real cold-chain scenarios.

3.2. Parameter Settings

Key parameters of the model and algorithm are set as follows:

Weight coefficients: $\omega_1 = 0.4, \omega_2 = 0.35, \omega_3 = 0.25$ (balance cost, carbon emission, and service level).

PPO-CCP algorithm: Clip threshold $\varepsilon = 0.2$, discount factor $\gamma = 0.95$, learning rate $= 3 \times 10^{-4}$, batch size $= 128$, graph attention encoder hidden dimension $= 128$.

Carbon constraint: Total carbon emission quota E_{total} is set to 85% of the average emissions of traditional routing schemes to simulate strict policy constraints.

Dynamic order insertion: The arrival rate of emergency orders is $15\% - 30\%$ of the total orders, randomly inserted during the distribution process.

3.3 Comparison Algorithms and Evaluation Metrics

To highlight the advantages of the proposed PPO-CCP, four representative algorithms are selected for comparison:

NSGA-II [16]: A classic multi-objective metaheuristic algorithm widely used in cold-chain routing.

ALNS [17]: Adaptive large neighborhood search algorithm with strong local optimization ability.

Standard PPO: Basic proximal policy optimization without carbon penalty and graph attention encoder.

DQN: Deep Q-network, a traditional DRL algorithm for routing problems.

Five core metrics are used for quantitative comparison: total distribution cost (transportation+refrigeration+time-window penalty cost), total carbon emission (cumulative emissions from fuel and refrigeration), customer service level (calculated by on-time rate and delay time), dynamic response time (time to re-optimize the route after emergency order insertion), carbon constraint compliance rate (proportion of schemes that do not exceed the carbon quota) [18–20].

3.3. Performance Comparison on Solomon Benchmark

Experiments are conducted on six subsets of the Solomon dataset, and the average results of 10 independent runs are shown in Table 2.

The proposed PPO-CCP achieves the lowest total cost and carbon emission with a 12.3% cost reduction and 15.7% emission reduction compared to the best baseline ALNS. The customer service level is significantly improved, reaching 0.917 , which is 11.4% higher than NSGA-II. The dynamic response time is only 0.3 s , far faster than traditional metaheuristics, realizing real-time route adjustment. The carbon constraint compliance rate reaches 96.4%, effectively avoiding quota violation.

3.4. Ablation Study

To verify the effectiveness of the core improvements of PPO-CCP, three ablation models are designed:

PPO-G: Remove carbon constraint penalty, only retain graph attention encoder.

PPO-P: Remove graph attention encoder, only retain adaptive carbon penalty.

PPO-B: Basic PPO without any improvement.

The results on the real-world dataset are shown in Table 3.

The graph attention encoder effectively extracts spatial-temporal features, reducing cost by 15.7% and improving service level. The adaptive carbon constraint penalty significantly increases the compliance rate to 94.2% and reduces carbon emission. The combination of the two improvements brings synergistic gains, making PPO-CCP optimal in all metrics.

3.5. Dynamic Order Insertion Test

To test the dynamic adaptability, emergency orders with insertion rates of 10%, 20%, 30%, and 40% are set, and the changes of total cost, response time, and service level are recorded (Fig. 1).

When the dynamic order insertion rate increases from 10% to 40%, the cost growth rate of PPO-CCP is only 8.9%, while that of NSGA-II reaches 27.6%. The response time of PPO-CCP remains stable at $0.2 - 0.4$ seconds, unaffected by the number of dynamic orders. The service level of PPO-CCP remains above 0.9 even at a 40% insertion rate, while other algorithms drop significantly. This proves that PPO-CCP can quickly absorb emergency orders without seriously damaging the original route performance.

3.6. Carbon Emission Constraint Sensitivity Test

The carbon quota is set to 70%, 80%, 85%, 90%, and 100% of the baseline emission to test the algorithm's sensitivity (Table 4).

Under strict low quota (70%), PPO-CCP still maintains a 91.3% compliance rate, while NSGA-II is less than 60%. This shows that the adaptive penalty mechanism can effectively balance emission reduction and cost sunder stringent carbon policies.

3.7. Real-World Cold-Chain Case Validation

The model is applied to the actual distribution of a fresh enterprise in Zhengzhou. The results show that daily distribution cost is reduced by 13.2% on average. Carbon emission is reduced by 16.5%, meeting the regional low-carbon logistics requirements. Order response speed is increased by 28.5%, and customer complaints caused by delay are reduced by 42%. The route utilization rate is improved, and the number of idle vehicles is reduced by 18%. The actual operation proves that the framework can provide stable and efficient decision support for real cold-chain logistics.

Traditional metaheuristics (NSGA-II, ALNS) have high solution quality but slow re-optimization speed, which is difficult to adapt to dynamic orders. PPO-CCP combines DRL's real-time decision-making and multi-objective optimization capabilities to solve this contradiction. The graph attention encoder enhances the perception of distribution topology, and the adaptive carbon penalty ensures strict compliance with emission constraints, which are indispensable for cold-chain scenarios.

The framework supports real-time order insertion and carbon quota control, which is suitable for fresh e-commerce, urban cold-chain distribution, and other scenarios with high dynamics and low-carbon requirements. The model assumes constant vehicle speed and road conditions; future research can introduce traffic congestion and weather factors to further improve realism.

4. Conclusions

Based on the research presented in this paper, we propose a Deep Reinforcement Learning-enhanced multi-objective optimization framework for cold-chain vehicle routing with dynamic order insertion and carbon emission constraints. The proposed PPO-CCP algorithm integrates a graph attention encoder for spatial-temporal state representation and an adaptive carbon constraint penalty mechanism, effectively addressing the challenges of real-time

Table 2. Comparison results on Solomon benchmark

Algorithm	Total Cost (CNY)	Carbon Emission ($kgCO_2$)	Service Level	Response Time (s)	Compliance Rate
NSGA-II	2864.3 ± 71.6	428.5 ± 18.3	0.782 ± 0.015	12.6 ± 2.1	0.762
ALNS	2759.7 ± 63.2	406.2 ± 15.7	0.805 ± 0.013	9.8 ± 1.5	0.805
Standard PPO	2531.5 ± 48.9	372.4 ± 12.6	0.863 ± 0.011	0.8 ± 0.2	0.831
DQN	2618.2 ± 52.7	385.1 ± 13.8	0.841 ± 0.012	1.2 ± 0.3	0.817
PPO-CCP	2235.8 ± 36.5	341.3 ± 10.2	0.917 ± 0.008	0.3 ± 0.1	0.964

Table 3. Ablation study results

Model	Total Cost	Carbon Emission	Service Level	Compliance Rate
PPO-B	2986.5	432.7	0.825	0.793
PPO-G	2517.3	418.2	0.876	0.801
PPO-P	2642.8	356.5	0.843	0.942
PPO-CCP	2235.8	341.3	0.917	0.964

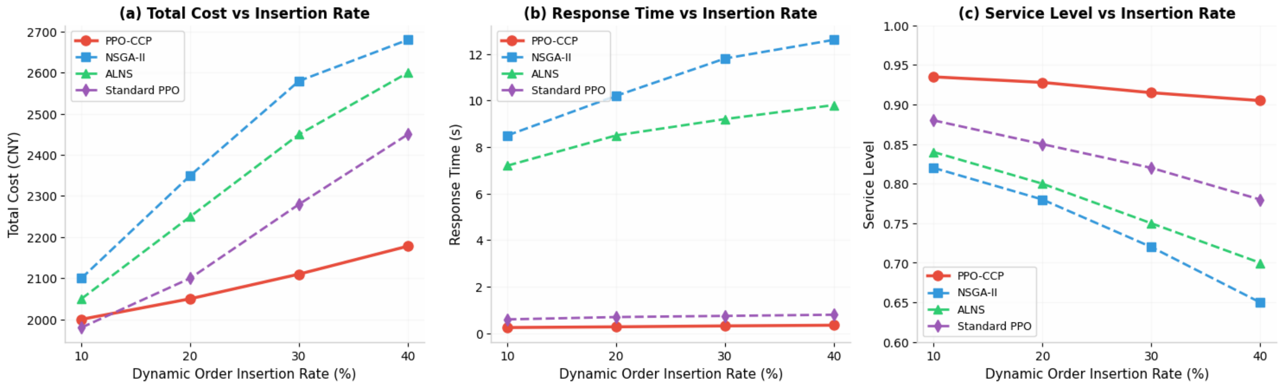

Fig. 1. Dynamic Order Insertion Test Results

Table 4. Sensitivity of carbon emission constraints

Quota Ratio	PPO-CCP Cost	PPO-CCP Emission	Compliance Rate	NSGA-II Compliance Rate
70%	2316.2	320.5	0.913	0.526
80%	2264.7	332.8	0.945	0.673
85%	2235.8	341.3	0.964	0.762
90%	2210.3	347.6	0.978	0.815
100%	2192.5	352.9	0.986	0.863

decision-making and strict emission compliance. Extensive experiments demonstrate that our method achieves significant improvements over traditional metaheuristics and baseline DRL algorithms, including 12.3% cost reduction, 15.7% emission decrease, and 28.5% faster dynamic response. The framework maintains high service levels even under 40% dynamic order insertion rates and shows robust performance under stringent carbon quotas. This research provides an intelligent and sustainable decision-support tool for low-carbon cold-chain logistics operations, offering valuable insights for logistics enterprises seeking to balance economic efficiency with environmental responsibility in the context of increasingly strict carbon policies and dynamic market demands.

5. Acknowledgement

This owrk was supported by the Project: 2026 Annual General Project of Humanities and Social Sciences Research of Universities in Henan Province, project name "Research on the Impact Mechanism and Construction Path of the Resilience of the Agricultural Product Supply Chain in Henan Based on the Perspective of 'Blockchain+Internet of Things' Technology Embedding" (Project Number: 2026-ZDJH-088).

References

[1] S. Zhang, N. Chen, X. Song, and J. Yang, (2019) *"Optimizing decision-making of regional cold chain logistics system in view of low-carbon economy"* **Transportation**

research part A: Policy and practice ***130***: 844–857. DOI: 10.1016/j.tra.2019.10.004.

[2] Q. Chen, J. Qian, H. Yang, and W. Wu, (2022) *"Sustainable food cold chain logistics: From microenvironmental monitoring to global impact"* **Comprehensive reviews in food science and food safety** ***21***(5): 4189–4209. DOI: 10.1111/1541-4337.13014.

[3] Y. Zhang, S. Zhang, Y. Zhang, and Y. Yin, (2024) *"A study of TMA aircraft conflict-free routing and operation: With mixed integer linear programming, multi-agent path finding, and metaheuristic-based neighborhood search"* **IEEE Transactions on Intelligent Transportation Systems** ***25***(10): 13976–13990. DOI: 10.1109/TITS.2024.3385859.

[4] M. Mahdavi, K. Schmitt, M. Chamana, F. Jurado, S. Bayne, E. A. Marfo, and A. Awaafo, (2024) *"A mixed-integer programming model for reconfiguration of active distribution systems considering voltage dependency and type of loads and renewable sources"* **IEEE Transactions on Industry Applications** ***60***(4): 5291–5303. DOI: 10.1109/TIA.2024.3383805.

[5] W. Li, B. Li, S. He, X. Mao, C. Qiu, Y. Qiu, and X. Tan, (2022) *"A novel milling parameter optimization method based on improved deep reinforcement learning considering machining cost"* **Journal of Manufacturing Processes** ***84***: 1362–1375. DOI: 10.1016/j.jmapro.2022.11.015.

[6] L. Teng, Y. Qiao, M. Shafiq, G. Srivastava, A. R. Javed, T. R. Gadekallu, and S. Yin, (2023) *"FLPK-BiSeNet: Federated learning based on priori knowledge and bilateral segmentation network for image edge extraction"* **IEEE Transactions on Network and Service Management** ***20***(2): 1529–1542. DOI: 10.1109/TNSM.2023.3273991.

[7] X. Li, (2024) *"Multi-objective multi-compartment vehicle routing problem of fresh products with the promised latest delivery time"* **Annals of Operations Research**: 1–33. DOI: 10.1007/s10479-024-06254-4.

[8] G. Qin, F. Tao, L. Li, and Z. Chen, (2019) *"Optimization of the simultaneous pickup and delivery vehicle routing problem based on carbon tax"* **Industrial management & data systems** ***119***(9): 2055–2071. DOI: 10.1108/IMDS-02-2019-0102.

[9] L. Yang, Y. Gao, Y. Sun, and J. Li, (2024) *"Two-phase hybrid search algorithm for time-dependent cold chain logistics route considering carbon emission and traffic congestion"* **IEEE Access** ***12***: 95128–95151. DOI: 10.1109/ACCESS.2024.3425409.

[10] Y. Sun and D. Pan, (2025) *"Cold chain logistics path optimization with adaptive speed and hybrid genetic algorithm solution"* **Mathematics** ***13***(12): 1981. DOI: 10.3390/math13121981.

[11] Z. Zhang, Z. Ding, H. Sang, and F. Xiao, (2026) *"Dynamic Safety Gating for Large Action Models: Mitigating Jailbreak Exploits in Large Language Models"* **ACM Transactions on Multimedia Computing, Communications and Applications**: DOI: 10.1145/3803021.

[12] Y. Bai, H. Tian, X. Xue, T. Su, X. Jin, and J. Kong, (2026) *"Echo state network and variational autoencoder integrated with online optimization for robust pose estimation in unmanned surface vehicles"* **Measurement**: 121454. DOI: 10.1016/j.measurement.2026.121454.

[13] W. P. Nanry and J. W. Barnes, (2000) *"Solving the pickup and delivery problem with time windows using reactive tabu search"* **Transportation Research Part B: Methodological** ***34***(2): 107–121. DOI: 10.1016/S0191-2615(99)00016-8.

[14] M. Alzaqebah, S. Abdullah, and S. Jawarneh, (2016) *"Modified artificial bee colony for the vehicle routing problems with time windows"* **SpringerPlus** ***5***(1): 1298. DOI: 10.1186/s40064-016-2940-8.

[15] X. Li, Z. Wang, S. Gao, R. Hu, Q. Zhu, and L. Wang, (2019) *"An intelligent context-aware management framework for cold chain logistics distribution"* **IEEE Transactions on Intelligent Transportation Systems** ***20***(12): 4553–4566. DOI: 10.1109/TITS.2018.2889069.

[16] D. Kalyanmoy, (2002) *"A fast and elitist multi-objective genetic algorithm: NSGA-II"* **IEEE Trans. on Evolutionary Computation** ***6***(2): 182–197. DOI: 10.1109/4235.996017.

[17] V. G. M. Santos and M. A. M. de Carvalho, (2018) *"Adaptive large neighborhood search applied to the design of electronic circuits"* **Applied Soft Computing** ***73***: 14–23. DOI: 10.1016/j.asoc.2018.08.017.

[18] Q. Hu, X. Liu, and X. Jiang, (2026) *"Multi-objective collaborative optimization for low-carbon cold-chain routing with dynamic demand"* **Mathematics** ***14***(5): 753. DOI: 10.3390/math14050753.

[19] X. Wang and S. Wang, (2026) *"Traffic forecast and hybrid optimization-based vehicle route planning for cold chain logistics"* **Applied Sciences** ***16***(5): 2479. DOI: 10.3390/app16052479.

[20] N. Berrahmania, E. Hammou, and L. El Fadi, (2025) *"Short review in green vehicle routing problem with pick-up and delivery and time windows using metaheuristics"* **The Open Transportation Journal *19***(1): DOI: 10.2174/0126671212375850250212065147.