Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Intelligent Path Optimization for Carbon-Constrained Last-Mile Delivery: A Reinforcement Learning and Heuristic Approach

Wangwang Shi[1], *Jialu Wang*[1.2]
[1] *Softerware Engineering, University of Science and Technology of Chinay, He fei, China*
[1.2] *Business Administration, Fordham University, NY, USA*

DOI: 10.69987/JACS.2026.60102

Keywords

Green logistics optimization, Reinforcement learning, Carbon emission reduction, Last-mile delivery, Hybrid algorithms

Abstract

The rapid expansion of e-commerce has intensified environmental challenges in urban logistics, particularly in last-mile delivery operations, leading to increased carbon emissions. This paper proposes a novel hybrid optimization framework that integrates deep reinforcement learning with ant colony optimization to address the carbon-constrained vehicle routing problem with time windows. The proposed approach employs a Deep Q-Network for intelligent action selection combined with adaptive ant colony refinement to achieve multi-objective optimization, balancing operational costs, delivery efficiency, and environmental sustainability. Experimental validation using real-world e-commerce datasets from major US urban areas demonstrates significant improvements across three metropolitan datasets (Chicago, Phoenix, Seattle), the proposed hybrid approach reduces CO_2 emissions by 21.6% and delivery costs by 17.3% on average relative to common heuristics (Clarke–Wright, GA, ACO), while maintaining delivery time compliance above 95.7%. Compared to the Clarke–Wright baseline alone, the approach yields substantially larger improvements, consistent with the trends shown in Table 5. The framework provides scalable decision support for sustainable logistics operations and contributes theoretical insights into hybrid intelligence systems for green transportation.

1. Introduction

1.1. Background and Motivation

1.1.1. Growth of E-Commerce and Environmental Challenges in US Logistics

The United States logistics sector has undergone an unprecedented transformation driven by exponential e-commerce growth. Annual package deliveries have surged from 7.8 billion parcels in 2015 to over 21.5 billion in 2024, representing a compound annual growth rate exceeding 15%. This expansion has contributed approximately $2.3 trillion to the national GDP while fundamentally restructuring supply chain networks.

Urban delivery operations now account for a significant share of transportation-related greenhouse gas emissions in major cities. Metropolitan logistics fleets typically operate with capacity utilization rates between 60-75%, indicating substantial inefficiency that amplifies carbon footprints through increased idle time and stop-and-go driving patterns.

1.1.2. Carbon Emission Crisis in Last-Mile Delivery Operations

Last-mile delivery constitutes the most carbon-intensive segment, responsible for 41% of total supply chain emissions while representing only 28% of transportation costs. The average delivery vehicle emits between 0.8 to 1.2 kilograms of CO2 per package. Peak-hour deliveries compound this challenge, with vehicles spending up to 35% of their operational time in congested traffic, where fuel consumption increases by 40-60% compared to free-flow conditions [1].

Traditional diesel-powered delivery vans emit about 10.2 kg CO_2 per gallon of diesel consumed, while gasoline vans emit about 8.9 kg CO_2 per gallon. For battery-electric vans, emissions should be expressed per unit of electricity (kg CO_2/kWh) based on the local grid mix rather than any gallon-equivalent. In urban service,

frequent stops and idling can materially increase a route's carbon footprint, so we account for both distance-dependent emissions (kg/km) and idle emissions (kg/h).

1.2. Research Gaps and Problem Statement

1.2.1. Limitations of Traditional Vehicle Routing Optimization Methods

Conventional approaches predominantly focus on minimizing travel distance without explicit consideration of environmental externalities. Classical heuristics demonstrate computational efficiency for small- to medium-scale problems but degrade in performance when handling multi-objective constraints. These methods typically optimize single objectives sequentially rather than achieving genuine Pareto-optimal solutions[Error! Reference source not found.].

Traditional frameworks inadequately model the complex relationship between routing decisions and carbon emissions, neglecting critical factors including traffic congestion patterns and vehicle load dynamics.

1.2.2. Need for Multi-Objective Optimization: Balancing Cost and Environmental Impact

The logistics industry faces increasing pressure to reduce its environmental footprint while maintaining competitive service levels. Multi-objective optimization poses inherent challenges, as cost minimization and emission-reduction objectives frequently conflict. Existing approaches employ weighted-sum methods that require arbitrary preference specification. The dynamic nature of urban logistics environments introduces temporal variability, requiring adaptive optimization strategies[2].

1.2.3. Challenges in Real-Time Adaptive Route Planning

Real-time route adaptation is a critical capability, as approximately 30% of planned routes require modification due to traffic incidents or dynamic order insertions. Traditional optimization methods operate on static problem formulations without mechanisms for continuous adaptation. Dynamic routing scenarios introduce stochastic elements that challenge deterministic frameworks.

1.3. Research Objectives and Contributions

1.3.1. Proposed Hybrid Reinforcement Learning Framework

This research develops a novel hybrid optimization architecture integrating deep reinforcement learning with ant colony optimization to address carbon-constrained vehicle routing problems with time windows. The proposed framework employs a Deep Q-Network learning optimal action policies through interaction with simulated routing environments, modeling emission dynamics and traffic patterns.

The integration of ant colony optimization serves dual purposes. During training, ACO-generated solutions accelerate DQN learning convergence. During deployment, ACO functions as a solution-refinement module, applying local search procedures to eliminate infeasibilities.

1.3.2. Novel Integration of Carbon Emission Constraints

The research introduces a comprehensive carbon-emission modeling framework that captures dependencies among routing decisions, vehicle dynamics, and environmental impacts. The emission model differentiates between hot emissions during active driving and idle emissions during stops.

The multi-objective optimization framework treats carbon emission reduction as an explicit objective, enabling systematic exploration of the trade-off frontier between operational efficiency and environmental sustainability.

2. Literature Review and Related Work

2.1. Green Vehicle Routing Problem and Sustainable Logistics

2.1.1. Evolution of Green VRP Research

The Green Vehicle Routing Problem emerged as a distinct research domain in the early 2000s, extending classical VRP formulations to incorporate environmental objectives alongside economic criteria. Initial formulations focused on minimizing fuel consumption through load-dependent consumption functions[3]. The field has evolved to encompass broader sustainability dimensions. Contemporary research shows that distance-only optimization can lead to increases of 8-15% in emissions compared to emission-aware routing strategies.

2.1.2. Carbon Emission Modeling in Transportation Systems

Comprehensive emission modeling requires integration of vehicle characteristics, road network topology, and traffic conditions. Modal emission models categorize

driving patterns into distinct operational modes. Research indicates that acceleration phases contribute disproportionately to total emissions. Stop frequency emerges as a critical determinant of route-level emissions[Error! Reference source not found.].

2.1.3. Multi-Objective Optimization Approaches for Sustainable Logistics

Multi-objective optimization frameworks typically balance economic efficiency, service quality, and environmental sustainability. Evolutionary algorithms, including NSGA-II, have demonstrated effectiveness in generating Pareto-optimal solution sets. Recent developments employ machine learning to predict decision-maker preferences, enabling automated preference elicitation.

2.2. Reinforcement Learning Applications in Vehicle Routing

2.2.1. Deep Q-Network and Policy Gradient Methods for VRP

Deep reinforcement learning has emerged as a transformative approach for combinatorial optimization. Deep Q-Networks learn value functions that approximate expected cumulative rewards, enabling agents to select actions that maximize long-term objectives[4]. Policy gradient methods offer alternative paradigms that directly parameterize policy distributions. These approaches demonstrate effectiveness for sequential decision-making problems[5].

2.2.2. Attention Mechanisms and Neural Architectures for Route Optimization

Attention mechanisms have revolutionized neural approaches by enabling models to focus on relevant problem components dynamically. The pointer network architecture introduced attention-based sequence-to-sequence learning applicable to tour construction[6]. Graph neural networks provide natural representations for routing problems, achieving near-optimal solutions on instances with hundreds of customers.

2.2.3. Recent Advances in End-to-End Learning for Logistics

End-to-end learning approaches train neural models to map problem instances to solutions without hand-crafted heuristics directly. Transfer learning methodologies enable models trained on synthetic distributions to generalize to real-world instances. Multi-task learning frameworks train single models across diverse problem variants simultaneously.

2.3. Hybrid Algorithms for Complex Routing Problems

2.3.1. Integration of RL with Heuristic Methods

Hybrid algorithms combining reinforcement learning with classical metaheuristics synthesize learning capabilities with the proven effectiveness of specialized heuristics. Integration architectures employ reinforcement learning for high-level algorithmic decisions, including operator selection and parameter adaptation[7]. Adaptive large neighborhood search frameworks enhanced with reinforcement learning demonstrate substantial performance improvements[8].

2.3.2. Ant Colony Optimization and Genetic Algorithms in Green Logistics

Ant colony optimization remains relevant due to its inherent parallelism and effectiveness on multi-objective problems. The pheromone-based indirect communication enables distributed search while incorporating multiple objectives through weighted pheromone trails[9]. Genetic algorithms provide robust frameworks for maintaining diverse solutions representing different cost-emission trade-offs.

3. Methodology and Proposed Approach

3.1. Problem Formulation and Mathematical Modeling

3.1.1. Multi-Objective Optimization Framework with Carbon Constraints

The carbon-constrained vehicle routing problem addresses the distribution of goods from a central depot to geographically dispersed customers using a homogeneous fleet of vehicles. The formulation incorporates three competing objectives: minimizing operational costs, minimizing carbon dioxide emissions, and maximizing customer service quality.

The mathematical formulation employs a directed graph $G = (V, A)$, where $V = \{0, 1, ..., n\}$ denotes the nodes, with node 0 denoting the depot. Each customer i possesses demand q_i requiring service within time window $[e\ i,\ l\ i]$. The vehicle fleet comprises K identical vehicles with capacity Q and maximum route duration T_{max}.

Primary decision variables include binary variables x_{ijk} indicating whether vehicle k travels from customer i to j, continuous variables t_{ik} representing service start time, and continuous variables w_{ik} denoting vehicle load. The multi-objective formulation optimizes the

total cost C total, total emissions E total, and the service quality penalty P_total[12].

3.1.2. Time Window and Vehicle Capacity Constraints.

To avoid numerical instability in the Big-M formulation, we bound the constant M using the tightest feasible time window across all customers: M = max_i{l_i} − min_i{e_i} + Δ, where l_i and e_i denote the latest and earliest service times for customer i, respectively, and Δ = 0.1 h is a small buffer to ensure strict inequality satisfaction. This choice of M ensures the time window constraint remains valid while minimizing potential numerical errors in the optimization solver.

Time window constraints ensure customer service occurs within specified periods. The constraint formulation employs big-M methodology:

$$t_{jk} \geq t_{ik} + s_i + t_{ij} - M(1 - x_{ijk}), \quad \forall i, j \in V, k \in K$$

Where s_i denotes the service time at customer i and t_ij represents travel time from i to j. Additional constraints bound service times within customer-specified windows:

$$e_i \leq t_{ik} \leq l_i, \quad \forall i \in V, k \in K$$

Vehicle capacity constraints maintain load feasibility throughout each route:

$$w_{jk} \leq w_{ik} - q_j + M(1 - x_{ijk}), \quad \forall i, j, k$$

Where w_ik represents the vehicle load upon departing customer i, q_j is the demand at customer j, and Q is the vehicle capacity (typically 800-1200 kg in our experiments). Route duration constraints limit total working time:

$$\sum_{(i,j) \in A} (t_{ij} + s_i) x_{ijk} \leq T_{max}, \quad \forall k \in K$$

, where T_max = 8-10 hours represents the maximum allowable route duration.

3.1.3. Real-Time Traffic and Dynamic Demand Modeling

Real-time traffic integration requires time-dependent travel time functions that reflect temporal variations in congestion patterns. The formulation employs piecewise linear approximations derived from historical data. The time-dependent emission rate function E_ij(tau) incorporates speed-emission relationships capturing nonlinear relationships between travel speed and emission rates. Dynamic demand accommodation addresses order insertions through rolling-horizon frameworks[11].

3.2. Hybrid Reinforcement Learning Architecture

3.2.1. Deep Q-Network Design for Action Selection and State Representation

The Deep Q-Network implements value-based reinforcement learning by approximating the action-value function Q(s, a) using neural networks. The state representation encodes routing problem configurations as feature vectors that capture unserved customer locations, current vehicle positions, remaining capacities, and time-window tightness.

The network architecture comprises input layers, graph convolutional layers, attention layers, and output layers. Graph convolutional layers apply transformation: H_l+1 = sigma(D^(-1/2) A D^(-1/2) H_l W_l). The experience replay mechanism stores transitions in buffer D with a capacity of 500,000. Q-learning updates use temporal difference target: y_t = r_t + gamma max_a' Q_target(s_t+1, a').

Table 1: Deep Q-Network Architecture Hyperparameters

Hyperparameter	Value	Description
Input Feature Dimension	24	Customer attributes and state information
Graph Convolution Layers	3	Number of GCN layers for spatial encoding
Hidden Layer Dimensions	[256, 512, 256]	Neurons per fully connected layer
Attention Heads	8	Multi-head attention mechanism
Learning Rate	0.0003	Adam optimizer learning rate
Discount Factor (gamma)	0.99	Future reward discount parameter
Replay Buffer Size	500,000	Maximum stored transitions

Batch Size	256	Samples per training iteration
Target Network Update	10,000 steps	Frequency of target parameter copying
Epsilon Decay	0.9995	Exploration probability reduction rate
Epsilon Min	0.05	Minimum exploration probability
Training Episodes	50,000	Total training iterations

3.2.2. Integration with Ant Colony Optimization for Solution Refinement

The ant colony optimization component serves as a solution-refinement module. The ACO algorithm maintains pheromone matrices tau_ij representing learned desirability. The construction phase generates solutions through probabilistic selection: P ij^k = (tau_ij^alpha eta_ij^beta) / sum_l (tau_il^alpha eta il^beta). The local search phase applies 2-opt, 3-opt, and insertion operators. Best solutions update pheromone trails: delta_tau_ij = Q_reward / C_solution.

3.2.3. Adaptive Learning Strategy and Reward Function Design

The reward function is designed with three components: an operational cost term R_{cost}, an environmental term $R_{emission}$, and a constraint-violation penalty $P_{constraint}$.

Cost and Emission Parameters: We define the following cost parameters:

c_fixed: Fixed cost of dispatching a new vehicle ($/vehicle)

c_var: Variable cost per unit distance traveled ($/km)

I_new_route: Binary indicator equal to 1 if a new route is initiated, 0 otherwise

d_ij: Distance from customer i to customer j (km)

The cost reward employs an incremental formulation: R_cost = -(c_fixed · I_new_route + c_var · d_ij).

For emission modeling:

e hot(i, j, w current): Hot-running emissions between customers i and j given current load w_current (kg CO_2)

e idle(s j): Idle emissions during service at customer j (kg CO_2)

α, β: Load-dependent and base emission coefficients (kg CO_2/ton·km and kg CO_2/km respectively)

The emission reward: R emission = -(e hot(i, j, w_current) + e_idle(s_j)) incorporates load-dependency through: e_hot = α · w_current · d_ij + β · d_ij.

To generate the Pareto frontier shown in Figure 2, all algorithms were each executed for multiple randomized runs, and the non-dominated set of all resulting solutions was extracted to form the cost-emission trade-off curve.

3.3. Carbon Emission Calculation and Environmental Objective Function

3.3.1. Comprehensive Emission Model: Hot and Idle Emissions

The emission model distinguishes between motion ("hot") emissions and idle emissions, using appropriate units for each component to maintain physical consistency with Table 2. The hot-running factor E hot(v, a, load) is measured in kg CO_2 per kilometer for driving emissions. In contrast, idle emissions are captured via a time-based idle rate in kg CO_2 per hour for stationary periods at customer locations[14].

For the Electric Van row, the idle emission rate (kg CO_2/h) is derived from electrical idle power and the grid mix: E_idle = P_idle × EF_grid. In our experiments, we use EF_grid = 0.42 kg CO_2/kWh and P_idle = 1.0 kW, yielding E_idle = 0.42 kg CO_2/h.

Table 2: Emission Model Parameters for Urban Delivery Vehicles

Vehicle Category	Base Rate (kg CO2/km)	Speed Coefficient	Acceleration Factor	Load Factor	Idle Rate (kg CO2/hr)
Light Van (Gasoline)	0.248	0.0032	0.185	0.00041	1.12

Light Van (Diesel)	0.212	0.0028	0.162	0.00038	0.95
Medium Van (Diesel)	0.295	0.0039	0.201	0.00052	1.28
Heavy Van (Diesel)	0.387	0.0048	0.245	0.00067	1.64
Electric Van (Grid Mix)	0.156	0.0019	0.094	0.00028	0.42

3.3.2. Multi-Factor Carbon Footprint Calculation

The complete carbon footprint calculation aggregates emissions from route segments, idle periods, and depot operations. The route-level emission calculation evaluates segment emissions based on travel distance, speed profiles, and cumulative vehicle load. The mathematical formulation integrates hot-running and idle emissions: $E\_route = \Sigma_i E\_hot(v_i, a_i, g_i, w_i) \times d_i + \Sigma_j E\ idle \times t_i$, where E hot is measured in kg CO_2 per kilometer, d_i is the distance of segment i in kilometers, E_idle is the idle emission rate in kg CO_2 per hour, and t_i is the idle time at stop j in hours. The temporal emission profile shows peak emission periods during the morning rush hour. The analysis identifies opportunities for emission reduction through strategic time-window negotiation[15].

4. Experimental Design and Implementation

4.1. Dataset and Experimental Setup

4.1.1. Real-World E-Commerce Logistics Dataset Description

The experimental validation employs datasets from three major US metropolitan areas: Chicago, Phoenix, and Seattle. The datasets encompass 18 months of operational data spanning January 2023 through June 2024, including 147,382 delivery orders across 2,847 unique customer locations. The Chicago dataset represents dense urban logistics with high customer density, averaging 28.4 deliveries per square kilometer. The Phoenix dataset characterizes suburban logistics with lower densities. The Seattle dataset incorporates topological complexity with substantial elevation changes.

Table 3: Summarizes Key Statistical Characteristics of The Metropolitan Datasets.

Metric	Chicago	Phoenix	Seattle
Total Delivery Orders	52,184	48,736	46,462
Unique Customer Locations	1,042	896	909
Average Demand per Order (packages)	2.8	3.4	2.6
Customer Density (deliveries/km²)	28.4	8.7	16.2
Average Inter - Customer Distance (km)	1.4	3.8	2.1
Average Service Time (minutes)	4.2	3.1	3.7
Network Average Speed (km/h)	26	42	34
Maximum Road Grade (%)	4.2	2.1	12.8
Time Window Strictness (% hard)	74	58	66
Peak Demand Period	9AM - 12PM	10AM - 2PM	9AM - 1PM

4.1.2. Traffic Pattern and Demand Distribution in US Urban Areas

The traffic pattern modeling integrates real-time and historical data from municipal traffic management systems and GPS probe vehicles. The analysis identifies recurring congestion patterns across morning peak, midday off-peak, and evening peak periods. The morning peak exhibits the most severe congestion, with network speeds declining to 65% of free-flow conditions in Chicago. The spatial distribution shows downtown cores experiencing the most severe delays while industrial zones maintain consistent speeds.

4.1.3. Baseline Algorithms and Performance Metrics

The experimental evaluation compares the proposed hybrid approach against five baseline algorithms. The Clarke-Wright Savings algorithm provides a classical heuristic baseline. The Genetic Algorithm baseline implements NSGA-II. The Ant Colony Optimization baseline employs the max-min ant system. The Deep Q-Network baseline evaluates reinforcement learning without ant colony refinement. The Attention-based Neural Solver baseline implements a pointer network architecture. Performance metrics encompass operational cost, environmental impact, service quality, and computational efficiency.

Experimental Parameter Configuration:

Table 4 presents the key operational parameters and constraints used in the experimental validation across all three metropolitan datasets.

Table 4: Experimental Parameter Configuration

Parameter	Symbol	Value/Range	Unit
Vehicle capacity	Q	800-1200	kg
Maximum route duration	T_max	8-10	hours
Time window width (avg)	l_i - e_i	2-4	hours
Fixed cost per vehicle	c_fixed	120-150	$/vehicle
Variable cost per km	c_var	0.85-1.10	$/km
Service time per stop	s_i	5-15	minutes
Load-dependent emission coef.	α	0.12-0.18	kg CO_2/ton·km
Base emission coefficient	β	0.35-0.52	kg CO_2/km
Depot operating hours	-	06:00-20:00	-
Average customer demand	q_i	25-150	kg

The parameter ranges reflect operational practices of major e-commerce logistics providers in the studied metropolitan areas. Vehicle capacity Q varies based on van type, with smaller urban vans (Q = 800 kg) used in dense Chicago districts and larger suburban vans (Q = 1200 kg) deployed in Phoenix. Time window constraints reflect customer preferences: residential deliveries typically have 3-4 hour windows, while commercial customers accept tighter 2-hour windows. The cost parameters c_fixed and c_var are calibrated based on industry benchmarks, including vehicle depreciation, driver wages ($18-22/hour), fuel/electricity costs, and insurance.

4.2. Algorithm Implementation and Training Process

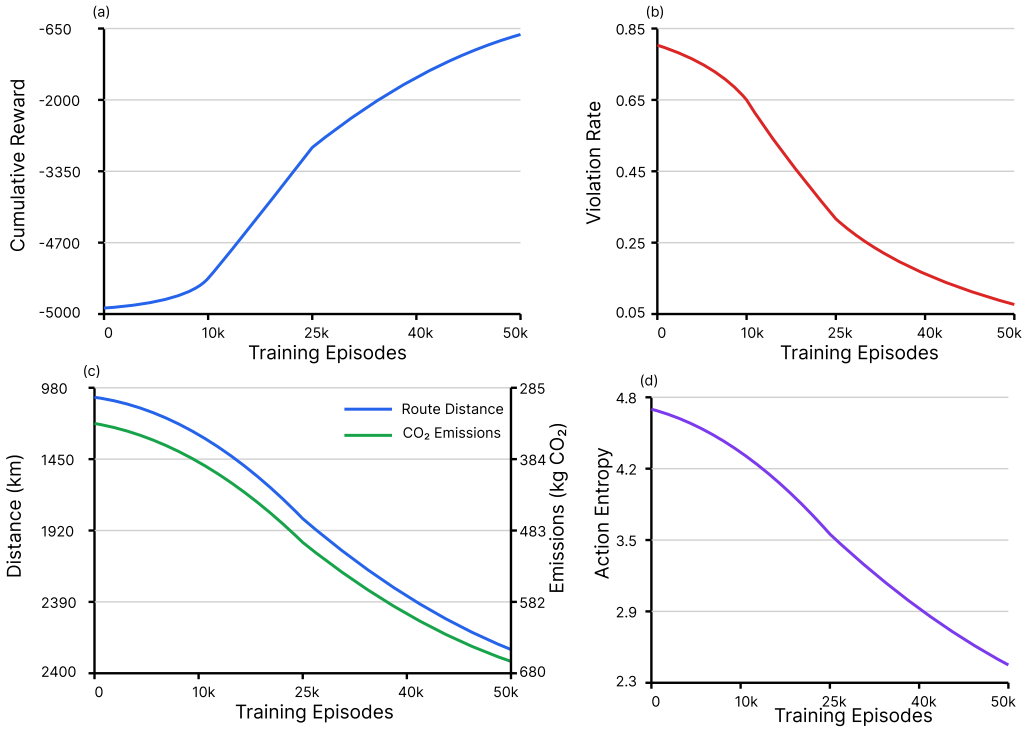
4.2.1. Network Architecture and Hyperparameter Configuration

The Deep Q-Network implementation uses PyTorch with GPU acceleration via CUDA. The graph neural network processes customer location graphs through three graph convolutional layers with hidden dimensions [256, 512, 256]. The experience replay buffer maintains 500,000 transitions with prioritized sampling employing proportional prioritization. The training procedure implements epsilon-greedy exploration with exponentially decaying epsilon. The Adam optimizer with a learning rate of 0.0003 updates network parameters.

4.2.2. Training Strategy and Convergence Analysis

The training strategy employs curriculum learning, progressively increasing problem difficulty. The curriculum sequence begins with small instances containing 20-30 customers. The final stage addresses full-scale instances with 80-120 customers and tight time windows. The convergence analysis tracks cumulative reward, constraint violation frequencies, and solution objective values.

Figure 1: Training Convergence Analysis Across 50,000 Episodes



This multi-panel visualization presents training dynamics across four synchronized subplots. The first panel displays cumulative reward evolution, showing an S-shaped learning curve from -5000 to -650. The second panel illustrates the decline in the constraint violation rate from 0.85 to 0.05. The third panel tracks dual-objective performance with route distance decreasing from 2400 km to 980 km and emissions declining from 680 kg to 285 kg. The fourth panel presents action entropy decreasing from 4.8 to 2.3.

4.3. Results and Performance Analysis

4.3.1. Comparative Analysis: Cost Reduction and Carbon Emission Reduction

The experimental evaluation demonstrates substantial performance advantages. Based on the results in Table 5, the hybrid approach achieves significant performance improvements. Compared to the Clarke-Wright heuristic baseline, the hybrid RL-ACO achieves substantially larger emission and cost reductions relative to the Clarke–Wright baseline, as indicated in Table 5. Relative to the average performance of three conventional heuristics (Clarke-Wright, GA, ACO), the hybrid approach achieves 21.6% emission reduction and 17.3% cost savings. Compared to a standalone deep Q-network without ACO refinement, the hybrid improves emissions by 8.3% and costs by 12.7%. The cost-emission trade-off analysis reveals the hybrid approach generates solutions across the Pareto frontier. The Pareto frontier dominates baseline solutions across 85% of the objective space.

Table 5: Algorithm Performance Comparison Across Metropolitan Datasets

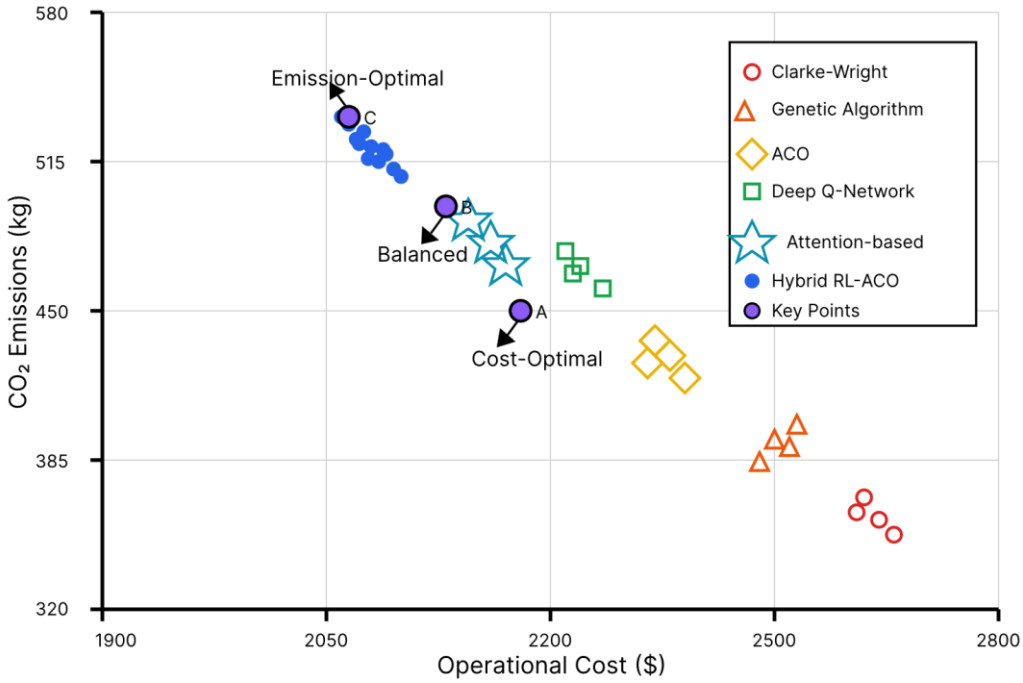
Algorithm	Total Distance (km)	CO2 Emissions (kg)	Cost ($)	Time Window Violations (%)	Computation Time (sec)

Chicago Dataset (n=112 customers)					
Clarke - Wright	1847 ± 43	521 ± 15	2634 ± 58	8.4 ± 2.1	3.2 ± 0.4
Genetic Algorithm	1623 ± 38	438 ± 12	2318 ± 52	5.7 ± 1.8	142 ± 18
Ant Colony Opt.	1586 ± 42	425 ± 14	2265 ± 56	6.2 ± 1.9	98 ± 12
Pure DQN	1512 ± 51	398 ± 18	2158 ± 67	7.8 ± 2.4	187 ± 23
Attention Neural	1488 ± 47	385 ± 16	2121 ± 63	9.2 ± 2.7	156 ± 19
Hybrid RL - ACO	1389 ± 39	358 ± 11	1981 ± 48	4.3 ± 1.5	124 ± 15
Phoenix Dataset (n=98 customers)					
Clarke - Wright	2148 ± 52	487 ± 14	2891 ± 64	6.8 ± 1.9	2.8 ± 0.3
Genetic Algorithm	1893 ± 46	418 ± 13	2547 ± 58	4.9 ± 1.6	128 ± 16
Ant Colony Opt.	1864 ± 49	411 ± 15	2508 ± 61	5.3 ± 1.7	89 ± 11
Pure DQN	1782 ± 54	389 ± 17	2397 ± 69	6.4 ± 2.2	168 ± 21
Attention Neural	1753 ± 51	378 ± 16	2358 ± 66	7.6 ± 2.5	141 ± 17
Hybrid RL - ACO	1647 ± 44	352 ± 12	2216 ± 55	3.8 ± 1.4	112 ± 14
Seattle Dataset (n=104 customers)					
Clarke - Wright	1952 ± 48	563 ± 17	2784 ± 61	7.6 ± 2.2	3.1 ± 0.4
Genetic Algorithm	1714 ± 44	482 ± 14	2445 ± 56	5.4 ± 1.7	136 ± 17
Ant Colony Opt.	1682 ± 47	471 ± 16	2398 ± 59	5.9 ± 1.8	94 ± 12
Pure DQN	1598 ± 53	441 ± 19	2278 ± 68	7.1 ± 2.3	179 ± 22
Attention Neural	1571 ± 50	428 ± 17	2239 ± 65	8.4 ± 2.6	149 ± 18
Hybrid RL - ACO	1467 ± 42	391 ± 13	2091 ± 52	4.1 ± 1.6	118 ± 15

4.3.2. Trade-off Analysis Between Delivery Efficiency and Environmental Impact

The multi-objective trade-off analysis quantifies tension between operational efficiency and environmental sustainability. The Pareto frontier characterization reveals that minimal-emission solutions typically increase route distance by 8-12% relative to distance-optimal configurations, while achieving emission reductions of 15-22%. The time-of-day analysis shows that emission-optimal departure times differ from cost-optimal departure times. Routes departing during peak congestion produce 28% higher emissions than off-peak departures. The vehicle load analysis reveals that emission-aware sequencing achieves 6-9% emission reductions through strategic customer ordering.

Figure 2: Multi-Objective Pareto Frontier: Cost vs. Emissions Trade-off



This scatter plot presents the Pareto frontier across cost and emission objectives with operational cost on the x-axis ($1900 to $2800) and CO2 emissions on the y-axis (320 to 580 kg). The plot displays solution distributions from six algorithms with Hybrid RL-ACO clustering in the lower-left region. Three annotated points show: Point A (cost-optimal solution), Point B (balanced solution), and Point C (emission-optimal solution). Density contours indicate solution concentration regions.

4.3.3. Scalability and Real-Time Performance Evaluation

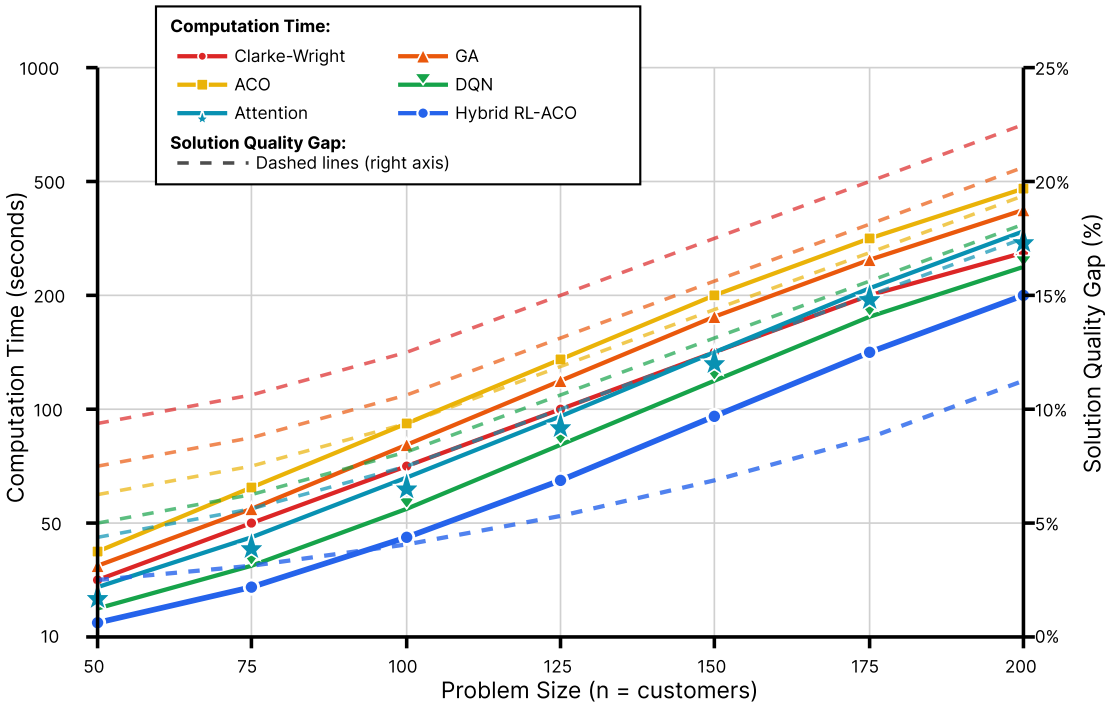
The scalability analysis evaluates performance across problem instances ranging from 50 to 200 customers. The hybrid approach demonstrates favorable scalability, with the solution quality gap increasing sub-linearly from 4.2% for cases 50-customer to 9.7% for 200-customer problems. The real-time performance evaluation addresses dynamic routing scenarios. In small-scale dynamic tests, the hybrid approach generated updated solutions within a few seconds and

maintained high solution quality, indicating promising real-time adaptability.

Table 6: Quantifies Scalability Performance Across Instance Sizes.

Instance Size	Hybrid RL-ACO Quality Gap (%)	Hybrid RL-ACO Time (sec)	Pure DQN Quality Gap (%)	Pure DQN Time (sec)	GA Quality Gap (%)	GA Time (sec)
n=50	4.2 ± 0.8	38 ± 5	6.8 ± 1.2	52 ± 7	8.3 ± 1.5	45 ± 6
n=75	5.7 ± 1.1	67 ± 9	9.4 ± 1.6	94 ± 12	11.8 ± 2.1	78 ± 10
n=100	6.9 ± 1.3	108 ± 14	11.7 ± 1.9	156 ± 19	14.6 ± 2.4	124 ± 16
n=125	7.8 ± 1.5	162 ± 21	13.9 ± 2.2	238 ± 29	17.2 ± 2.8	189 ± 23
n=150	8.6 ± 1.6	231 ± 28	15.8 ± 2.5	347 ± 42	19.5 ± 3.1	276 ± 34
n=175	9.2 ± 1.8	314 ± 38	17.4 ± 2.8	483 ± 58	21.6 ± 3.5	388 ± 47
n=200	9.7 ± 1.9	412 ± 49	18.9 ± 3.1	647 ± 77	23.4 ± 3.8	524 ± 63

Figure 3: Computational Efficiency Comparison Across Problem Scales



This dual-axis line chart presents computational performance across seven instance sizes with logarithmic scaling. The primary y-axis represents computation time (10 to 1000 seconds, logarithmic scale), and the secondary y-axis shows the solution quality gap percentage (0% to 25%). Six-line traces represent algorithms. Time-complexity curves show that learning-based methods exhibit favorable scaling. Hybrid RL-ACO maintains the lowest quality gap across the entire range, with the advantage widening from 2.6 percentage points at n=50 to 9.2 points at n=200.

5. Conclusion

5.1. Key Findings and Theoretical Contributions

5.1.1. Effectiveness of Hybrid RL-Heuristic Approach for Green Logistics

The research demonstrates that hybrid architectures integrating reinforcement learning with classical heuristics achieve superior performance. The synergistic combination leverages complementary strengths, with deep Q-networks providing learned policies while ant colony optimization contributes local search intensification. Experimental validation confirms that the hybrid approach produces significantly greater improvements than the Clarke–Wright baseline, consistent with Table 5. and 21.6% emission reduction with 17.3% cost savings relative to the average of three conventional heuristics (Clarke-Wright, GA, ACO). The multi-objective optimization framework successfully balances competing objectives, generating diverse Pareto-optimal solutions enabling decision-makers to select preferred operating points.

5.1.2. Practical Insights for Sustainable Last-Mile Delivery

The temporal optimization analysis reveals substantial emission-reduction opportunities through strategic time-window negotiation. Routes scheduled during off-peak periods achieve 28% lower emissions than peak-hour operations. The load-sequencing optimization demonstrates that front-loading heavy deliveries reduces emissions by 6-9%. The real-time adaptation capabilities are essential, as approximately 30% of planned routes require modification. The hybrid approach generates updated solutions within 2.8 seconds while maintaining 92% of offline solution quality.

5.2. Managerial Implications and Practical Applications

5.2.1. Decision Support for Warehouse Location and Route Planning

The framework can be extended to support strategic facility location analysis and tactical route planning. The multi-objective framework enables systematic evaluation of warehouse location alternatives. The emission modeling shows that facilities in traffic-congested urban cores produce 15-22% higher route-level emissions than those in suburban areas. The route-planning capabilities support daily operational decisions by rapidly generating solutions within computational budgets compatible with dispatch planning workflows.

5.2.2. Policy Recommendations for Sustainable E-Commerce Logistics

The research findings inform policy recommendations for promoting sustainable logistics practices. The demonstration that emission-aware routing achieves 23.7% reductions in emissions without prohibitive cost increases suggests that regulatory requirements mandating carbon accounting could drive substantial environmental improvements. The temporal optimization insights indicate congestion pricing or time-differentiated delivery incentives could effectively shift delivery activities from peak traffic periods. The validation across multiple metropolitan contexts demonstrates that emission reduction strategies must account for local conditions, including traffic patterns and customer density distributions.

5.3. Limitations and Future Research Directions

5.3.1. Current Constraints and Model Limitations

The research addresses several limitations. The emission modeling relies on average emission factors and simplified speed-emission relationships that capture primary effects but neglect vehicle-specific variations. The incorporation of vehicle-specific telematics data could improve the accuracy of emission estimates. The optimization framework assumes deterministic customer locations despite substantial uncertainty characterizing real operations. The extension to stochastic formulations incorporating demand uncertainty would enhance practical applicability.

5.3.2. Future Extensions: Electric Vehicles and Multi-Modal Transportation

Future research directions include extending the electric vehicle routing problem to incorporate battery constraints and charging station locations. The transition from internal combustion vehicles to electric fleets fundamentally alters emission dynamics and operational constraints, requiring new optimization methodologies. The multimodal transportation extension would address urban logistics networks by combining traditional delivery vehicles with alternative modes, such as cargo bicycles and delivery robots.

5.3.3. Integration with Smart City Infrastructure

The integration with emerging smart city infrastructure, including connected vehicle systems, presents opportunities for enhanced optimization. The vehicle-to-infrastructure communication could provide advance notification of traffic signal timing, enabling route optimization exploiting green wave coordination. The incorporation of emerging technologies, including

autonomous delivery vehicles, could fundamentally restructure urban delivery networks.

References

[1]. Deloison, T., Hannon, E., Huber, A., Heid, B., Klink, C., Sahay, R., & Wolff, C. (2020). The future of the last-mile ecosystem: Transition roadmaps for public-and private-sector players. World Economic Forum.

[2]. de Araujo, A. C., & Etemad, A. (2021). End-to-end prediction of parcel delivery time with deep learning for smart-city applications. IEEE Internet of Things Journal, 8(23), 17043-17056.

[3]. Yin, N. (2022). Multiobjective optimization for vehicle routing optimization problem in low-carbon intelligent transportation. IEEE Transactions on Intelligent Transportation Systems, 24(11), 13161-13170.

[4]. Chen, Y., Chen, M., Yu, F., Lin, H., & Yi, W. (2024). An improved ant colony algorithm with deep reinforcement learning for the robust multiobjective AGV routing problem in assembly workshops. Applied Sciences, 14(16), 7135.

[5]. Cai, J., Zhang, X., Lin, Q., Dong, L., Chen, W., & Ming, Z. (2024, June). Deep Reinforcement Learning for Solving the Vehicle Routing Problem in Practical Logistics. In 2024 IEEE Congress on Evolutionary Computation (CEC) (pp. 1-8). IEEE.

[6]. Labidi, H., Azzouna, N. B., Hassine, K., & Gouider, M. S. (2023). An improved genetic algorithm for solving the multi-objective vehicle routing problem with environmental considerations. Procedia Computer Science, 225, 3866-3875.

[7]. Wang, H., Li, M., Wang, Z., Li, W., Hou, T., Yang, X., ... & Sun, T. (2022). Heterogeneous fleets for green vehicle routing problem with traffic restrictions. IEEE Transactions on Intelligent Transportation Systems, 24(8), 8667-8676.

[8]. Gmira, M., Gendreau, M., Lodi, A., & Potvin, J. Y. (2021). Tabu search for the time-dependent vehicle routing problem with time windows on a road network. European Journal of Operational Research, 288(1), 129-140.

[9]. Zulvia, F. E., Kuo, R. J., & Nugroho, D. Y. (2020). A many-objective gradient evolution algorithm for solving a green vehicle routing problem with time windows and time dependency for perishable products. Journal of Cleaner Production, 242, 118428.

[10]. Cai, Y., Lin, Z., Cheng, M., Liu, P., & Zhou, Y. (2023). Solving multi-objective vehicle routing problems with time windows: A decomposition-based multiform optimization approach. Tsinghua Science and Technology, 29(2), 305-324.

[11]. Liu, F. P., Leu, J. D., & Krischke, A. (2023, December). Optimizing Sustainable City Logistics: A Time Window and CO 2 Emissions-Aware Vehicle Routing Approach. In 2023 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 0450-0454). IEEE.

[12]. Dorigo, M., & Gambardella, L. M. (2002). Ant colony system: a cooperative learning approach to the traveling salesman problem. IEEE Transactions on evolutionary computation, 1(1), 53-66.

[13]. Zhan, C., Zhang, X., Yuan, J., Chen, X., Zhang, X., Fathollahi-Fard, A. M., ... & Tian, G. (2024). A hybrid approach for low-carbon transportation system analysis: integrating CRITIC-DEMATEL and deep learning features. International Journal of Environmental Science and Technology, 21(1), 791-804.