E-Commerce Letters 电子商务评论, 2026, 15(5), 245-255
Published Online May 2026 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2026.155512

Hans汉斯

基于DQN的电商即时配送双层决策模型

俞梦楠，袁鹏程

上海理工大学管理学院，上海

收稿日期：2026年3月8日；录用日期：2026年3月20日；发布日期：2026年5月20日

摘　要

随着电子商务的快速发展，城市外卖等即时配送业务已成为支撑线上交易履约与提升城市物流效率的重要环节。尤其在电商订单高频到达与时效要求不断提高的背景下，配送调度面临多主体协同、多约束耦合及多目标优化等复杂问题。传统单层优化模型及启发式算法难以在动态环境下兼顾实时响应能力与系统整体收益优化。针对上述问题，本文构建了一个面向城市电商物流即时配送的双层决策优化模型。上层由平台统筹订单分配与配送路径规划，以最大化系统整体收益；下层刻画配送员接单收益响应机制与消费者履约满意度函数，实现多主体协同约束下的反馈优化结构。在算法层面，引入深度Q网络(DQN)算法，在高维动态状态空间中进行策略学习。仿真实验结果表明，所提方法在平台收益、配送员收入、消费者满意度与订单履约率等关键指标上均优于多种传统启发式算法。敏感性分析进一步验证了模型在不同成本与时效参数变化条件下的鲁棒性与适应能力。

关键词

电子商务，外卖即时配送，双层决策，深度强化学习，配送调度

Bi-Level Decision-Making Model for E-Commerce Instant Delivery Based on Deep Q-Network (DQN)

Mengnan Yu, Pengcheng Yuan

Business School, University of Shanghai for Science & Technology, Shanghai

Received: March 8, 2026; accepted: March 20, 2026; published: May 20, 2026

Abstract

With the rapid development of e-commerce, instant delivery services such as urban food delivery have become a critical component in supporting online order fulfillment and improving urban logistics efficiency. Particularly under the conditions of high-frequency order arrivals and increasingly stringent timeliness requirements, delivery scheduling faces complex challenges, including

文章引用: 俞梦楠, 袁鹏程. 基于 DQN 的电商即时配送双层决策模型[J]. 电子商务评论, 2026, 15(5): 245-255.
DOI: 10.12677/ecl.2026.155512

multi-agent coordination, multi-constraint coupling, and multi-objective optimization. Traditional single-layer optimization models and heuristic algorithms often struggle to balance real-time responsiveness with overall system profit maximization in dynamic environments. To address these challenges, this study develops a bi-level decision-making optimization model for urban e-commerce instant delivery systems. At the upper level, the platform coordinates order allocation and delivery route planning to maximize overall system revenue. At the lower level, the model captures couriers' order acceptance behavior and consumers' fulfillment satisfaction through utility-based response mechanisms, forming a feedback-driven optimization structure under multi-agent interactions. Algorithmically, a Deep Q-Network (DQN) approach is introduced to learn optimal scheduling strategies within a high-dimensional dynamic state space. Simulation results demonstrate that the proposed method outperforms several traditional heuristic algorithms in terms of platform revenue, courier income, consumer satisfaction, and order fulfillment rate. Sensitivity analysis further confirms the robustness and adaptability of the model under varying cost and time-related parameter settings.

Keywords

E-Commerce, Food Instant Delivery, Bi-Level Decision-Making, Deep Reinforcement Learning, Delivery Scheduling

1. 引言

近年来，随着移动互联网与电子商务的快速发展，线上交易规模持续扩大，即时配送逐渐成为电商履约体系的重要组成部分。特别是在外卖、本地生活服务及即时零售等业务场景中，消费者对配送时效性与服务质量的要求不断提高，促使电商平台不断优化其末端配送体系。外卖即时配送作为连接线上交易与线下履约的关键环节，其运行效率直接影响平台收益、配送员收入以及消费者体验，是电商生态系统中具有战略意义的核心能力[1]。

然而，城市电商即时配送系统具有显著的动态性与复杂性。一方面，订单呈现高频、随机到达特征，且存在严格的时间窗约束；另一方面，平台、配送员与消费者之间存在利益博弈与行为响应机制，形成多主体互动结构[2]。现有研究多基于单层优化模型或启发式算法对配送调度问题进行求解，例如，Luo 等[3]利用蚁群算法的正反馈机制，有效解决了冷链物流中温度敏感型路径的规划问题。Tang 等[4]针对众包物流环境下的动态任务分配与路径协同问题，设计了一种基于博弈论的激励机制，有效提升了社会运力的利用率。在类似的拼车问题中，Ma 等[5]的研究表明，基于起讫点的动态定价策略能够降低乘客出行成本，同时减少司机绕路行为，提高匹配效率。王建等[6]的研究通过动态匹配模型分析了网约车市场在不同时段的供需变化，并将乘客自主议价引入模型，结果表明乘客议价能够推动市场向供需平衡移动，并在高峰时段提升平台利润和社会福利。随着人工智能技术的发展，深度强化学习在复杂决策问题中的应用逐渐增多，其通过与环境交互不断学习最优策略的特性，为解决动态配送调度问题提供了新的技术路径[7]。

基于此，本文从电商即时配送视角出发，构建了一个平台－配送员－消费者三方互动的双层决策优化模型。上层由平台统筹订单分配与路径规划，以最大化系统整体履约收益；下层刻画配送员接单收益响应机制与消费者满意度函数，实现多主体反馈下的协同优化结构。在算法设计方面，引入深度 Q 网络(DQN)方法，在高维动态状态空间中进行策略学习，实现订单分配与路径决策的自适应优化。

2. 模型与算法求解

2.1. 问题描述

本文研究城市电商即时配送场景下的动态订单调度问题。系统由电商平台、配送员以及消费者构成。订单从商家发出，经配送员完成配送至消费者。研究问题场景的基本假定包括：

假设 1：每个配送订单只能由一名配送员完成。

假设 2：每名配送员具有最大载货容量限制。

假设 3：仅当配送员收益为正且消费者满意度不低于零时，订单履约才视为有效完成。

假设 4：平台按照匹配成功后的配送员得到的订单实际收益按比例抽取利润。

假设 5：所有订单均包含取货节点(商家)与送达节点(消费者)，并设有配送时间窗约束。

2.2. 符号说明

本文涉及到的符号如表 1 所示。

Table 1. Symbol description
表 1. 符号说明

符号	含义
集合	
K	配送员集合，$k \in K$，$K=\{1, 2, 3,\cdots,k,\cdots n_k\}$
P	订单集合，$p \in P$，$P=\{1, 2, 3,\cdots,p,\cdots n_p\}$
变量	
N	N 为所有节点集合(包括上下车点)
γ	平台佣金比例
C	平台的运营成本
f_i^k	配送员 k 接到订单 i 的收入
d_i	订单 i 的行驶距离
f_p^1	单位距离的支付价格
f_p^2	单位时间不准时的补偿价格
S_k	车辆消耗
s_k	车辆单位运行成本
$t_{i,pick}^k$	配送员 k 接到订单 i 的时间
ET_i	订单 i 的配送时间窗口的开始时间
LT_i	订单 i 的配送时间窗口的结束时间
D_k	配送员 k 的收益
Q_k	配送员 k 的最大载客量
q_i	订单的消费者数
U_i	消费者 i 的效用
L_i^*	单独 i 的最短出行距离
f_0^1	最大单位距离价格
F_i	配送员配送订单 i 的实际出行费用

续表

决策变量	
x_i^k	$x_i^k \in \{0,1\}$，表示订单 i 是否分配给配送员 k
y_{ij}^k	$y_{ij}^k \in \{0,1\}$，表示配送员 k 是否从节点 i 行驶到节点 j
$z_{i,drop}^k$	$z_{i,drop}^k \in \{0,1\}$，表示配送员 k 是否承担了订单 i 的取货任务
$z_{i,pick}^k$	$z_{i,pick}^k \in \{0,1\}$ 表示配送员 k 是否承担了订单 i 的送货任务
$z_i^k(t)$	$z_i^k(t) \in \{0,1\}$，t 时刻订单 i 由配送员 k 配送

2.3. 模型构建

上层模型由电商平台作为决策主体，负责订单分配与配送路径规划优化；下层模型刻画配送员接单行为与终端消费者履约满意度响应机制。具体建模过程如下。

2.3.1. 上层决策模型：平台决策

平台利润由订单履约收入扣除配送成本及平台运营成本构成，目标函数如下：

$$\max \sum_{k\in K}\sum_{i\in p}\gamma f_i^k \cdot x_{ik} - C \tag{1}$$

约束如下：

$$\sum_{k\in K} x_i^k \le 1,\ \forall i \in P \tag{2}$$

$$\sum_{j\in N} y_{ij}^k - \sum_{j\in N} y_{ji}^k = 0,\ \forall i \in N,\ k \in K \tag{3}$$

$$z_{i,drop}^k - z_{i,pick}^k = 0,\ \forall i \in P,\ k \in K \tag{4}$$

$$S_k = s_k \cdot d_i \tag{5}$$

$$\mathrm{D} = \sum_{i\in P_k}(1-\gamma) f_i^k \cdot d_i - S_k \tag{6}$$

$$t_{i,pick}^k \in [ET_i, LT_i],\ \forall i \in P,\ k \in K \tag{7}$$

$$\sum_{i\in P} q_i \cdot x_i^k \le Q_k,\ \forall t, k \in K \tag{8}$$

$$U_i \ge 0,\ \forall i \in P \tag{9}$$

其中，约束(2)为订单分配约束，要求每个订单最多由一名配送员承担配送任务。约束(3)为路径流量平衡约束，规定配送员到达某一配送节点后必须离开该节点，以保证配送路径的连续性与完整性。约束(4)为订单取送匹配约束，确保每个订单的取货节点与送达节点成对出现，且取货行为必须先于送达行为发生。约束(5)和(6)为配送成本与收益约束，分别表示配送员在执行路径过程中产生的运输成本以及由订单履约所获得的收益。约束(7)为时间窗约束，要求配送员到达订单送达节点的时间位于消费者期望的最早送达时间与最晚送达时间之间；若超过时间窗，则产生延误惩罚成本。约束(8)为配送容量约束，要求配送员在任一时刻所承载的订单数量不超过其最大配送容量。约束(9)为履约满意度约束，要求订单履约过程中消费者满意度不低于零，以保证服务质量水平。

2.3.2. 下层决策模型：配送员与消费者决策

在平台完成初步的订单分配与路径规划后，具体的履约执行依赖于配送员与消费者的响应行为。因

此，下层模型刻画配送员与消费者在接收到平台调度方案后的个体决策过程，以实现各自目标函数的最优化。

1) 配送员收益响应模型

配送员的目标是最大化其个人收益，即在接收到新订单后，通过判断插单后的路径变化与收益差值，决定是否接受该订单。配送员的总收益为所有订单收入减去总消耗：

$$\max \mathrm{D}_k = \sum_{i \in P_k} (1-\gamma) f_i^k \cdot d_i - S_k \tag{10}$$

2) 消费者决策模型

消费者的目标是保证自身出行效用。消费者的效用函数如下，其中，前一部分表示其愿意支付的价格(即期望效用)，后一部分为实际支付费用。

$$U_i = \rho f_0^1 L_i^* - F_i \tag{11}$$

$$F_i = f_p^1 \sum_{j \in O_p} \sum_{k \in O_p} d_{jk} - f_p^2 \Delta T \tag{12}$$

s.t.

$$t_{i,drop}^k \geq t_{i,pick}^k \tag{13}$$

$$\sum_{i \in P} q_i \cdot z_i^k(t) \leq Q_k, \ \forall t, k \in K \tag{14}$$

$$\sum_{j \in N} y_{ij}^k - \sum_{j \in N} y_{ji}^k = 0, \ \forall i \in N, \ k \in K \tag{15}$$

$$z_i^k(t) = \begin{cases} 1, \ t\text{时刻订单}i\text{由配送员}k\text{负责} \\ 0, \ \text{其他} \end{cases} \tag{16}$$

其中，订单 i 的配送费用 F_i 由基础配送费用 f_p^1 和延误惩罚成本 f_p^2 两部分组成。其中，基础配送费用 f_p^1 是绕行距离的递减函数。即配送员因为要配送其他订单而产生绕行时，消费者必定会减小单位距离的支付价格 f_p^1。因此，f_p^1 的计算如式(17)所示：

$$f_p^1 = -\theta_1 \Delta l_p + f_o^1, \ \forall p \in P \tag{17}$$

其中，θ_1 为单位绕行距离的价格弹性，表示每增加单位绕行距离，支付价格减少的程度，计算过程如式(18)所示：

$$\theta_1 = -\frac{\partial f_p^1}{\partial \Delta l_p} \tag{18}$$

其中，f_p^2 是随着不准时的时间损失 ΔT_p 的单调递增函数，具体来说，随着不准时服务时间的增加，消费者需要支付更多的补偿费用。因此，f_p^2 的计算如式(19)所示：

$$f_p^2 = \theta_2 \Delta T_p, \ \forall p \in P \tag{19}$$

其中，θ_2 为单位不准时价格弹性，表示每增加单位时间损失 ΔT_p，消费者需要支付的补偿价格增加的程度，计算过程如式(20)所示：

$$\theta_2 = \frac{\partial f_p^2}{\partial \Delta T_p} \tag{20}$$

约束(13)为时效性约束。约束(14)是配送员的容量限制。约束(15)保证了在所有节点上的流量一致性，即流量平衡约束。约束(16)是订单与配送员车辆匹配约束。

3. 算法求解

3.1. 状态空间定义

平台状态空间 s 表示调度系统在任一时刻的环境信息，用于引导智能体决策。平台环境包含随机生成的订单和配送员，每个决策时刻具有如下状态变量：o_i^t：节点 i 当前待分配订单数量；d_i^t：节点 i 当前空闲配送员数量；s_i^t：节点 i 的历史订单完成率或平均等待时间。因此，系统状态 s_t 在第 t 步被编码为如下结构向量：

$$s_t = \left[o_1^t, \cdots, o_{24}^t, d_1^t, \cdots, d_{24}^t, r_{t-1}, \varepsilon_t\right] \tag{21}$$

其中，r_t 表示上一轮调度获得的即时奖励(作为策略历史反馈)；ε_t：当前探索率，反映系统决策偏随机性。

3.2. 动作空间定义

动作空间为平台可选择的调度决策，定义为配送员的选择向量，长度等于配送员数量，每一维代表平台选择某一配送员作为调度或服务对象的概率或决策输出，即 $A=\{0, 1, \cdots, action\_size-1\}$。平台每一步从当前状态中选择一个最优配送员进行订单分配或路径更新。

3.3. 奖励函数设计

为了模拟平台的决策效果，采用基于平台收益的奖励函数。该奖励值可视为调度后整体平台收益的近似反馈信号，间接包含配送员收益和消费者满意度。若订单成功分配，奖励为平台收益；若分配失败，奖励为−5。长期收益通过折扣因子 $\gamma(\gamma=0.95)$ 累加，用于优化策略的稳定性和长期效益权衡。

3.4. Q 值计算与网络训练

DQN 使用神经网络近似 Q 值函数，目标是学习一个最优动作值函数 $Q^*(s,a)$，满足贝尔曼最优方程：

$$Q(s,a;\theta) = E\left[r + \gamma \max_{a'} Q(s',a';\theta') \middle| s,a\right] \tag{22}$$

其中，s 为当前状态，a 为当前动作，s′ 为下一个状态，γ 为折扣因子，θ 和 θ' 分别为当前网络和目标网络的参数。训练过程通过最小化以下损失函数进行参数更新：

$$L(\theta) = E_{(s,a,r,s')\sim D}\left[\left(r + \gamma \max_{a'} Q(s',a';\theta') - Q(s,a;\theta)\right)^2\right] \tag{23}$$

其中，D 为经验回放池中的样本集合。

3.5. 智能体定义

智能体为平台调度系统，其目标是通过学习一套基于环境状态的调度策略，使得每次调度决策最大化长期累计收益。智能体在每一轮交互中执行以下步骤：

接收当前系统状态 s_t：包括订单分布与配送员分布；

根据策略 $\pi(a|s)$ 选择调度节点(动作 a_t)；

获取环境反馈：包括奖励 r_t 和新状态 s_{t+1}；

存储经验 (s_t, a_t, r_t, s_{t+1}) 到经验池；

利用 Q 学习更新策略。

4. 数值试验

本文以 2022 年 1 月 1 日至 12 月 31 日美国芝加哥市的网约车订单数据为基础(数据源：

https://www.chicago.gov.cityofchicago/)，尽管原始数据来源于网约车订单，其本质反映的是城市内部高频OD (Origin-Destination)需求的时空分布特征，可用于刻画城市订单在空间与时间维度上的分布规律。因此，本文将该数据作为城市配送需求场景的构建基础。

在模型构建过程中，本文对订单收益结构、配送成本函数及履约惩罚机制进行了电商化重构，数据仅用于生成城市级订单时空分布特征，并未沿用拼车定价或拼车博弈机制。基于此构建的仿真环境用于验证所提出双层决策模型及深度强化学习算法在城市电商物流最后一公里配送场景下的适用性与有效性。根据前序研究[8]，参数标定结果如表 2 所示。

Table 2. Specific numerical values of simulation parameters
表 2. 仿真参数具体数值

仿真参数	数值
单位距离最高价格 f_o^1	4.95 元/英里
车辆单位距离的费用 c	3.38 元/英里
消费者拼车需求强度参数 ρ	2.07
单位绕行距离的价格弹性 θ_1	−1.30
单位不准时价格弹性 θ_2	−0.14

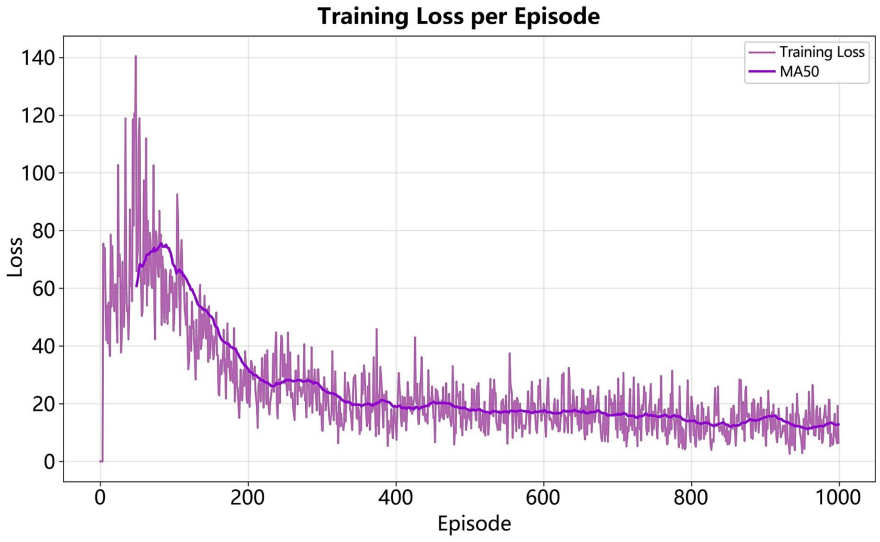
5. 结果分析

5.1. DQN 算法结果分析

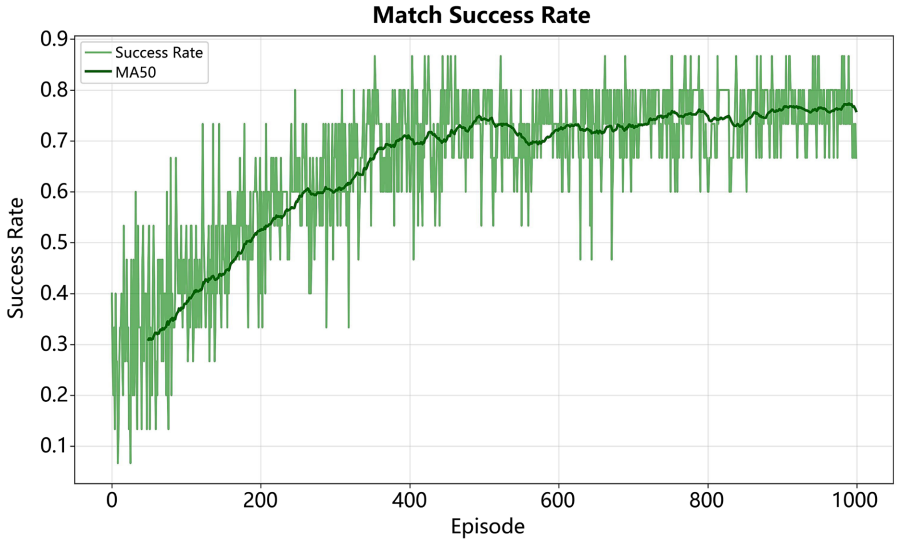
根据之前的研究，本文以硬时间窗中的算例[10, 5, 5, 15]为例进行说明，考察 DQN 算法训练误差(Q值差)与平台累计奖励和匹配率的变化趋势，以评估模型的收敛性与优化效果。

如图 1(a)所示，训练误差在初期波动较大，随着训练轮数增加，误差整体呈下降趋势，并在第 600 轮之后趋于平稳，说明模型已逐步学习到较优的订单分配策略。

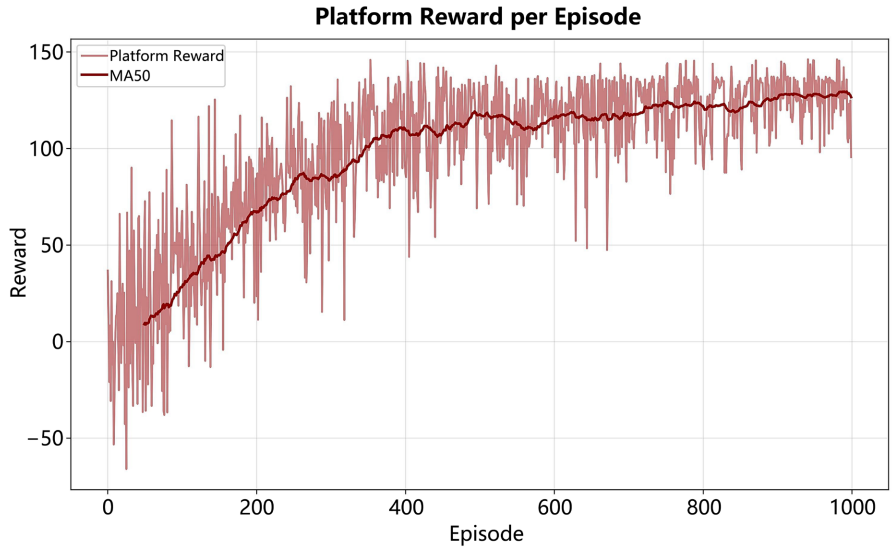
图 1(b)和图 1(c)呈现了平台匹配率和累计奖励随训练轮次的变化情况。从第 1 轮至第 400 轮，累计收益快速增长，在第 600 轮后趋于稳定，说明平台收益达到了局部最优点。



(a) 训练误差



(b) 匹配率



(c) 平台累计奖励

Figure 1. DQN training performance
图 1. DQN 训练表现

5.2. 对比算法结果分析

表 3 通过 10 个不同的算例运行平均结果展示了 DQN 与传统算法(Greedy, SA, Genetic, PSO)在关键指标上的对比结果。

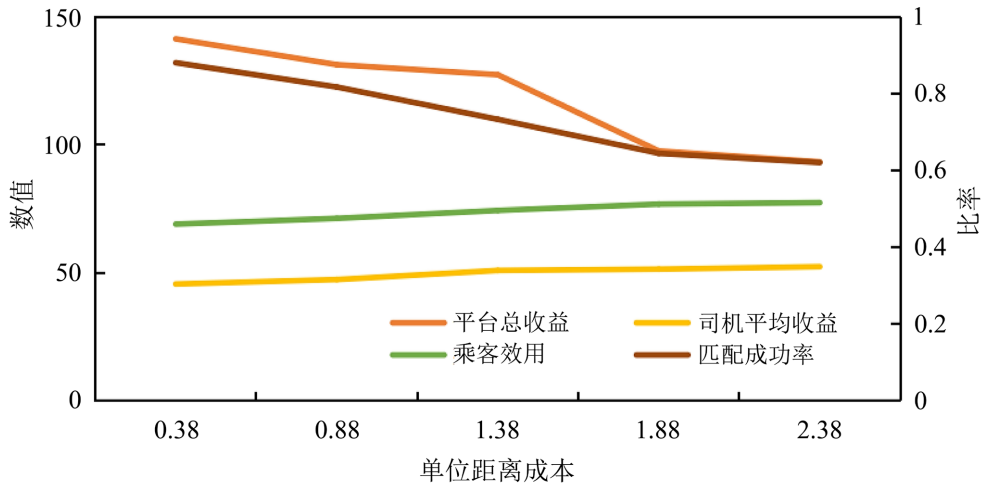
总体来看，DQN 在软时间窗和硬时间窗场景下均表现出明显优势。匹配率方面，DQN 分别达到 94.13%和 81.40%，显著高于其他算法；配送员收益与消费者效益指标上，DQN 同样保持最优或次优水平，体现出较强的综合优化能力。相较之下，SA 整体表现次优，Greedy、Genetic 与 PSO 在匹配率与稳定性方面存在一定不足。结果表明，DQN 能够在动态环境中有效协调订单分配与路径规划，在提升匹配效率的同时保障配送员收入与消费者体验，适用于电商即时配送等对实时性与优化质量要求较高的场景。

Table 3. Comparison of experimental results of different algorithms
表 3. 不同算法实验结果对比

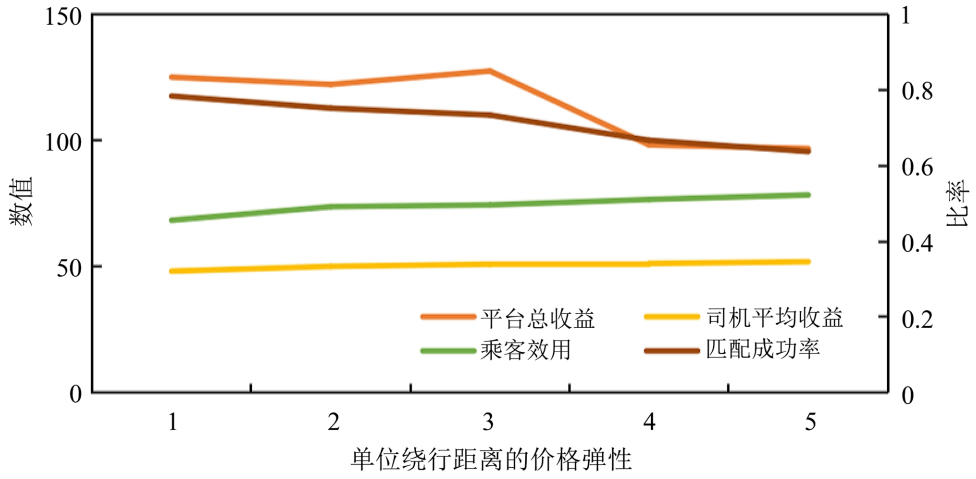
算例类型	模型	平均匹配率/%	平均配送员收益/美元	平均消费者效益
软时间窗	DQN	94.13	49.52	68.37
	Greedy	70.83	44.11	58.91
	Genetic	69.17	46.37	60.03
	SA	78.33	45.73	62.51
	PSO	66.67	48.37	60.26
硬时间窗	DQN	81.40	47.84	70.99
	Greedy	71.67	43.41	60.32
	Genetic	69.17	45.56	61.64
	SA	77.50	45.21	60.94
	PSO	64.17	47.97	65.19

5.3. 参数敏感性分析

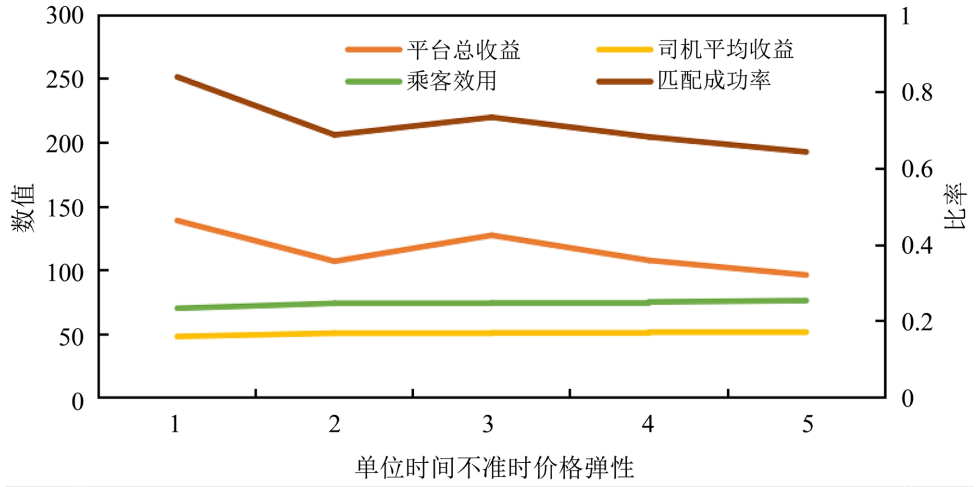
为检验模型对关键参数扰动的鲁棒性与响应特性，设计三组敏感性实验，分别为单位距离成本、单位绕行距离的价格弹性与单位时间不准时价格弹性，输出平台总收益、配送员平均收益、消费者效用与匹配成功率四个核心指标，结果如图 2(a)~(c)所示。结果表明，不同参数对系统绩效的影响具有明显差异。首先，单位距离成本上升会显著压缩平台总收益与匹配成功率，但配送员收益和消费者效用有所提升，说明高成本环境下平台更倾向于筛选高质量订单，整体服务规模随之收缩。其次，当消费者对绕行更加敏感时，平台收益与匹配率下降，而消费者效用明显提高，表明平台会通过优化路径减少绕行，但同时放弃部分低效率订单。再次，延误惩罚系数提高会降低平台收益和匹配规模，但有助于提升消费者满意度，并在一定程度上激励配送员提高履约效率。总体来看，成本因素主要影响平台收益与订单规模，而绕行与延误敏感度更多作用于用户体验和匹配效率，体现出系统在规模与服务质量之间的权衡关系。



(a) 单位距离成本敏感性分析



(b) 单位绕行距离的价格弹性敏感性分析结果



(c) 单位不准时价格弹性敏感性分析

Figure 2. Sensitivity analysis results
图 2. 敏感性分析结果

6. 结论

针对电商平台外卖即时配送中多主体、多约束的协同优化问题，本文构建了平台–配送员–消费者双层决策模型，并引入深度强化学习(DQN)方法，实现订单匹配、路径规划与个体行为响应的协同优化。主要结论如下：

1) 构建了平台–配送员–消费者三方博弈模型，上层处理订单匹配与路径规划，下层模拟个体行为，并通过 DQN 算法在动态环境中学习最优策略，具有良好的收敛性和泛化能力。

2) 与弹性、模拟退火、遗传和粒子群优化等方法相比，DQN 在平台收益、消费者匹配率和效用上表现出色，实现了收益、体验与效率的平衡。

3) 敏感性分析显示，单位配送成本对平台收益影响显著，而价格弹性与延误惩罚主要作用于用户侧效用与匹配效率。

尽管模型在仿真中有效，但未来可通过真实数据验证、探索多智能体强化学习及一体化优化、考虑不确定性和行为异质性，进一步拓展应用边界，为电商即时履约系统优化提供了可行的建模思路与算法支持。

参考文献

[1] 谢舒婷, 李金碧, 邓万琼, 等. 基于碳减排规制的城市生鲜农产品冷链物流配送路径优化研究[J]. 中国市场, 2026(5): 168-172.

[2] Leelertkij, T., Buddhakulsomsiri, J. and Huynh, V. (2025) A Multi-Thread Simulated Annealing for Multi-Objective Vehicle Routing Problem with Time Windows and Demand Priority. *Computers & Industrial Engineering*, **207**, Article ID: 111253. https://doi.org/10.1016/j.cie.2025.111253

[3] Luo, H., Liang, Z., Zhu, M., Hu, X. and Wang, G. (2018) Integrated Optimization of Unmanned Aerial Vehicle Task Allocation and Path Planning under Steady Wind. *PLOS ONE*, **13**, e0194690. https://doi.org/10.1371/journal.pone.0194690

[4] Tang, Y., Zhou, J., Hao, H., Hao, F. and Xu, H. (2022) Path Planning and Trajectory Tracking for Automatic Guided Vehicles. *Computational Intelligence and Neuroscience*, **2022**, Article ID: 8981778. https://doi.org/10.1155/2022/8981778

[5] Ma, J., Xu, M., Meng, Q. and Cheng, L. (2020) Ridesharing User Equilibrium Problem under OD-Based Surge Pricing Strategy. *Transportation Research Part B*: *Methodological*, **134**, 1-24. https://doi.org/10.1016/j.trb.2020.02.001

[6] 王健, 王慧, 胡晓伟, 等. 高峰期考虑乘客议价的网约车定价与平台收益及社会福利优化[J]. 交通运输系统工程与信息, 2022, 22(2): 54-63.

[7] Hou, L., Xu, Y., Ren, R., Yang, J. and Su, L. (2025) Optimization of Three-Dimensional Urban Underground Logistics System Alignment: A Deep Reinforcement Learning Approach. *Computers & Industrial Engineering*, **205**, Article ID: 111185. https://doi.org/10.1016/j.cie.2025.111185

[8] 俞梦楠, 袁鹏程, 徐凯. 考虑司机与乘客协同效益的拼车服务动态优化策略[J]. 交通运输工程与信息学报, 2025, 23(4): 181-195.