A Bi-Level Optimization Method Integrating Evolutionary Game Theory and Deep Reinforcement Learning: A Novel Intelligent Dispatch Model for Ride-Hailing

Liping Yan *, Peiran Wu, Shaofeng Wang, Haojie Jia, Jingkai Huang

Posted Date: 25 March 2026

doi: 10.20944/preprints202603.1961.v1

Keywords: ride-hailing vehicle; evolutionary game theory; deep reinforcement learning; intelligent dispatching

 doi:10.20944/preprints202603.1961.v1

Article

A Bi-Level Optimization Method Integrating Evolutionary Game Theory and Deep Reinforcement Learning: A Novel Intelligent Dispatch Model for Ride-Hailing

Liping Yan [1,*], Peiran Wu [1], Shaofeng Wang [2], Haojie Jia [1] and Jingkai Huang [1]

[1] School of Information and Software Engineering, East China Jiaotong University, Nanchang, Jiangxi 330013, China

[2] MOE Engineering Research Center of Railway Environmental Vibration and Noise, East China Jiaotong University, Nanchang, Jiangxi 330013, China

* Correspondence: yanliping@ecjtu.edu.cn

Abstract

Ride-hailing dispatch systems face significant challenges under fluctuating demand and dynamic traffic conditions, where efficient coordination is essential for both platform performance and driver income among large-scale ride-hailing vehicles. This paper constructs a grid-based ride-hailing vehicle dispatch decision model (GRV-DDM), which provides a structured and quantifiable representation of vehicles and orders, effectively capturing spatio-temporal heterogeneity in dynamic traffic environments. Based on this model, a bi-level optimization multi-dimensional dispatch decision algorithm (BO-MDDA) is proposed. At the macro level, evolutionary game theory is employed to adaptively guide collective vehicle strategies toward supply–demand equilibrium, while at the micro level, deep reinforcement learning optimizes individual drivers' real-time dispatch decisions to maximize long-term profits. A bidirectional feedback mechanism is further designed to integrate macro-level collective intelligence with micro-level individual decision-making. Experimental results across diverse traffic scenarios demonstrate that the proposed approach outperforms classical dispatch algorithms in terms of efficiency and robustness.

Keywords: ride-hailing vehicle; evolutionary game theory; deep reinforcement learning; intelligent dispatching

1. Introduction

Ride-hailing services have become an integral component of modern urban transportation systems, driven by rapid urbanization and the widespread adoption of smart mobility solutions [1]. By offering flexible and on-demand travel services, ride-hailing platforms contribute to alleviating traffic congestion and improving overall travel efficiency [2]. However, the rapid growth in user demand poses a critical challenge: how to efficiently and accurately dispatch large-scale vehicle fleets under highly dynamic traffic conditions. Traditional dispatch methods, which often rely on static or simplified optimization assumptions, struggle to cope with complex traffic environments, fluctuating demand patterns, and multi-objective optimization requirements, thereby limiting their effectiveness in modern ride-hailing systems [3].

Early studies on ride-hailing dispatch primarily focused on improving system efficiency through rational resource allocation mechanisms [4]. To model strategic interactions among heterogeneous participants, game theory was subsequently introduced into dispatch decision-making processes [5]. Within this framework, ride-hailing dispatch is formulated as a strategic interaction among agents,

 doi:10.20944/preprints202603.1961.v1

such as vehicles and passengers, aiming to achieve globally optimal outcomes through the analysis of competition and cooperation [6]. While game-theoretic models provide valuable theoretical insights into balancing platform-level efficiency and individual benefits, their applicability is often constrained in highly dynamic environments characterized by rapidly changing traffic conditions and demand uncertainty [7].

To overcome these limitations, evolutionary game theory (EGT) has been adopted as an extension of classical game theory to better capture agents' adaptive behaviors and strategy evolution in dynamic systems. By incorporating mechanisms of individual adaptation and population-level evolution, EGT enables agents to adjust their strategies in response to environmental feedback, offering greater flexibility in addressing uncertainty and heterogeneity. As a result, EGT has demonstrated clear advantages in modeling coordination and competition in large-scale multi-agent systems, particularly in scenarios requiring long-term strategic adaptation.

In parallel, reinforcement learning (RL) has emerged as a prominent approach for ride-hailing dispatch due to its capability for self-learning and sequential decision optimization [8]. Through continuous interaction between agents and the environment, RL can adapt dispatch strategies based on real-time traffic states and passenger demand, making it well suited for dynamic decision-making problems [9]. With advances in deep learning, deep reinforcement learning (DRL) combines neural networks with RL to handle high-dimensional state spaces and complex feature representations, showing strong potential in large-scale traffic systems [10,11]. Nevertheless, in multi-agent settings that involve both cooperation and competition, DRL often suffers from challenges such as increased optimization complexity, training instability, and low sample efficiency [12]. These issues have motivated the development of multi-agent deep reinforcement learning (MADRL), which seeks to improve system performance through coordinated learning among multiple agents [13,14].

Despite its theoretical advantages, MADRL still faces significant challenges when applied to large-scale ride-hailing dispatch systems [15]. First, the complex interdependencies and information asymmetry among numerous agents substantially increase the difficulty of coordination and learning. Second, agents are required to make rapid and reliable decisions under highly dynamic traffic conditions. In large-scale scenarios, common training issues—such as strategy oscillation, slow convergence, and inefficient sample utilization—can lead to reduced dispatch efficiency and suboptimal decision accuracy. Consequently, improving the stability and efficiency of MADRL while effectively addressing multi-agent coordination remains an open research problem.

To address these challenges, this paper proposes a ride-hailing vehicle dispatch strategy algorithm, which integrates multi-agent deep reinforcement learning with evolutionary game theory. Specifically, DRL is employed to handle high-dimensional state spaces and optimize agents' local decision-making processes, while EGT is used to model the adaptive evolution of collective strategies and mitigate issues related to global coordination and local optima. Through a bi-level optimization mechanism, the proposed framework achieves coordinated optimization between system-level objectives and individual-level decision-making, thereby enhancing overall dispatch efficiency while improving long-term driver revenue.

The main contributions of this paper are summarized as follows: (i) A grid-based ride-hailing vehicle dispatch decision model (GRV-DDM) is constructed to provide a structured and quantifiable representation of the real-time states of large-scale ride-hailing vehicles and orders. This model effectively captures spatio-temporal heterogeneity in dynamic traffic environments and establishes a solid modeling foundation for efficient dispatch optimization. (ii) A bi-level optimization multi-dimensional dispatch decision algorithm (BO-MDDA) is proposed. At the macro level, it guides the adaptive evolution of collective vehicle strategies toward system-level supply–demand equilibrium. At the micro level, it optimizes drivers' real-time dispatch decisions to maximize long-term earnings, supporting sustainable platform operations. (iii) A bidirectional feedback mechanism is designed to tightly couple macro-level collective intelligence with micro-level individual decision-making. This mechanism improves learning convergence and significantly enhances the system's adaptability and collaborative performance in complex dynamic environments

 doi:10.20944/preprints202603.1961.v1

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the GRV-DDM model. Section 4 presents the BO-MDDA algorithm. Section 5 reports and analyzes experimental results. Section 6 concludes the paper and discusses future research directions.

2. Related Work

With the rapid expansion of ride-hailing services, optimizing dispatch systems under dynamic traffic conditions, fluctuating demand, and complex agent interactions has become a central research focus [16]. Early studies primarily relied on game theory and conventional optimization techniques to improve dispatch efficiency. With advances in learning-based methods, RL—particularly DRL and MADRL—has been increasingly applied to address dispatch problems characterized by dynamic demand and complex traffic environments [17]. Although these approaches substantially enhance system adaptability, achieving an effective balance between global system objectives and local agent interests in large-scale multi-agent settings remains a critical and unresolved challenge.

Building on early optimization-based frameworks, subsequent research explored advanced mathematical models, including nonlinear mixed-integer programming and dynamic programming, to tackle core dispatch problems [19]. For example, Wu et al. (2023) proposed a contract-based incentive mechanism to optimize resource allocation while simultaneously reducing carbon emissions and alleviating traffic congestion [18]. Despite their theoretical rigor, such model-driven approaches often suffer from high computational complexity and limited scalability, making them difficult to deploy in large-scale, real-world ride-hailing systems. As a result, these methods frequently fail to respond efficiently to real-time demand fluctuations and rapidly changing traffic conditions [20].

To overcome the limitations of traditional optimization methods, RL and its advanced variants have been progressively introduced into ride-hailing dispatch research. Early RL-based studies focused on vehicle repositioning strategies to maximize long-term profitability. Subsequent work employed DRL to address issues such as reward sparsity—sometimes incorporating techniques like Generative Adversarial Networks (GANs)—thereby improving training stability in large-scale urban environments [21]. More recently, MADRL has been adopted to jointly optimize order matching and vehicle dispatch [22], as well as to integrate spatial information (e.g., road networks and grid-based representations) to enhance real-time responsiveness. While these approaches significantly improve system adaptability and cooperative efficiency, existing MADRL frameworks continue to face challenges such as training instability, slow convergence, and the absence of systematic mechanisms for coordinating global efficiency with individual agent benefits.

In recent years, integrating game-theoretic concepts—such as Nash equilibrium and EGT—into multi-agent reinforcement learning (MARL) has emerged as a promising paradigm for improving robustness and strategic coordination in complex multi-agent systems. This hybrid approach has demonstrated effectiveness across a range of domains, including dynamic traffic dispatch [23], equilibrium optimization [24], large-scale non-cooperative game learning [25], and information diffusion modeling in social networks [26]. Game theory offers a principled framework for analyzing strategic interactions and equilibrium behavior, while reinforcement learning enables agents to adaptively optimize decisions in dynamic environments. Their integration supports the simultaneous optimization of individual objectives and collective system performance in high-dimensional, dynamically evolving traffic scenarios.

Based on these challenges, this paper proposes the BO-MDDA to address the coordination–competition dilemma in ride-hailing dispatch systems while jointly optimizing platform-level efficiency and driver income. BO-MDDA integrates evolutionary game theory with multi-agent deep reinforcement learning to balance global and local objectives under dynamic traffic and demand conditions. Furthermore, the GRV-DDM is constructed as the operational framework, enhancing the adaptability and robustness of the proposed approach in practical, large-scale applications.

 doi:10.20944/preprints202603.1961.v1

3. Problem Model

Efficiently dispatching idle vehicles toward high-demand areas is essential for ride-hailing system performance, yet conventional methods struggle with dynamic urban traffic and insufficiently model interactions among vehicles. To address these challenges, this paper constructs the GRV-DDM, which integrates fine-grained spatial partitioning into a multi-agent learning framework. GRV-DDM supports adaptive dispatch decisions that enhance supply-demand matching, system efficiency, and platform profitability under diverse demand patterns. Section 3.1 presents the model's core components, while Section 3.2 formulates the dispatch process as a POMDP.

3.1. Core Components of GRV-DDM

In ride-hailing system, the GRV-DDM uses three core entities: ride-hailing vehicles, passengers, and spatial grids. Let the set of ride-hailing vehicles be denoted as V, and the set of passengers as P. Initially, vehicles and passengers are randomly distributed across the urban area. Each ride-hailing vehicle can be in one of three operational states: idle (awaiting an order), occupied (serving a passenger), or in-dispatch (being relocated by the dispatch algorithm).

To manage spatial complexity and mitigate supply–demand imbalance, the urban area is partitioned into a grid set $G=\{g_1, g_2, \cdots, g_W\}$ composed of multiple adjacent hexagonal cells. Each grid forms a local neighborhood with its six adjacent grids. Both vehicles and passengers are mapped onto this grid space.

At time step $t\in\{1, 2, \cdots, T\}$, the number of passenger requests and idle vehicles in grid g_k are denoted as $D_{g_k}^t$ and $U_{g_k}^t$, respectively. The platform first matches passenger requests with idle vehicles located within the same grid. For remaining idle vehicles, the dispatch algorithm determines relocation directions. At boundary grids, dispatch options are constrained due to limited feasible movement directions, as shown in Figure 1. If a vehicle $v\in V$ is dispatched from grid g_k to grid g_j, this action is recorded as $d_{g_j}^{g_k}$.

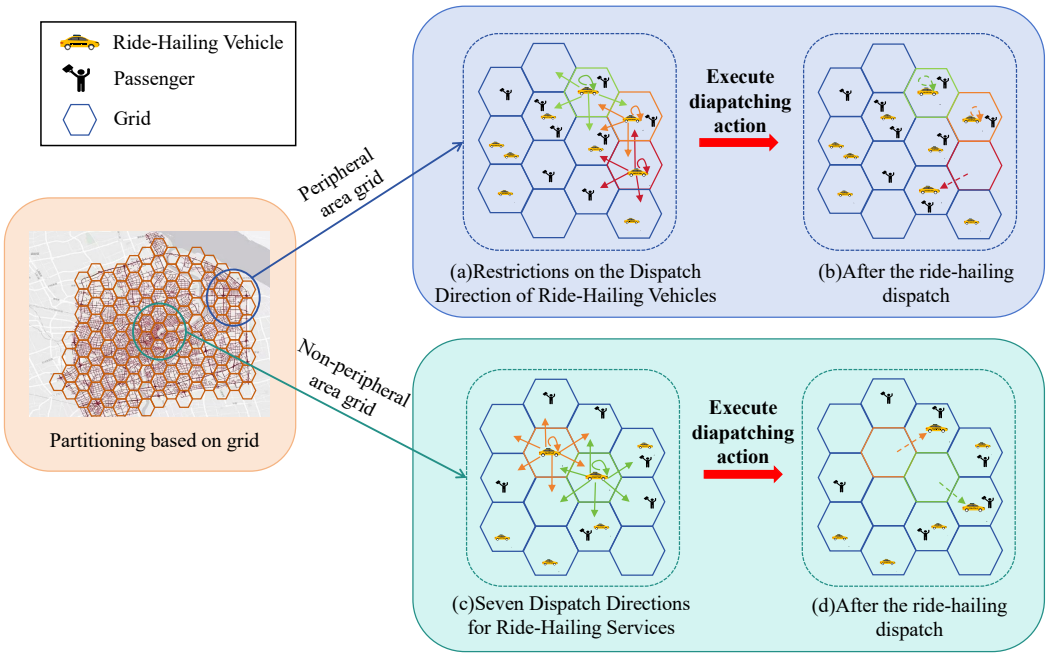

Figure 1. Illustration of grid-based dispatch and directional constraints in the GRV-DDM.

When vehicle v is occupied, its origin and destination grids are denoted as $ori(v)$, $dest(v)\in G$. To model heterogeneous ride-hailing services (e.g., express, economy, and carpooling), let the service type of vehicle v be denoted as $\mu = type(v)$, with corresponding pricing parameters including base fare b_μ, distance rate ρ_μ, time rate ω_μ, and platform commission rate κ_μ. For a trip completed at

 doi:10.20944/preprints202603.1961.v1

time step t, with travel distance $Le_v^t = dis(ori(v), dest(v))$ and duration $Ti_v^t = time(ori(v), dest(v))$, the gross fare revenue is:

$$Fare^t(v) = b_\mu + \rho_\mu Le_v^t + \omega_\mu Ti_v^t . \quad (1)$$

The driver's net income is defined as:

$$R_v^t = (1-\kappa_\mu) \cdot Fare^t(v) - Z_v^t , \quad (2)$$

where Z_v^t represents potential penalties or additional fees. The cost of dispatching vehicle v from grid g_k to grid g_j at time step t is defined as:

$$C_v^t(k, j) = \alpha_1 \cdot dis(k, j) + \alpha_2 \cdot time(k, j) + \alpha_3 \cdot type(v), \quad (3)$$

where α_1 and α_2 weight distance and travel time, respectively, and α_3 accounts for service-type-related cost differences. If a vehicle remains in its current grid, $dis(k, k) = 0$, and the dispatch cost reduces to a fixed waiting overhead. The overall platform objective is to maximize net system revenue $\sum_{t=1}^{T} \sum_{v \in V} (R_v^t - C_v^t)$ over the entire dispatching horizon by optimizing dispatch decisions. Occupied vehicles generate trip revenue and are excluded from idle dispatch decisions. This objective function captures the trade-off between platform revenue, driver income, and spatial supply–demand balance across heterogeneous service types.

3.2. POMDP-Based Dispatch Decision Formulation

The ride-hailing dispatch problem exhibits a high-dimensional state space and significant dynamic uncertainty, making it challenging to simultaneously balance regional supply-demand equilibrium and maximize system-wide revenue. To address these challenges, the GRV-DDM dispatch process is formulated as a multi-agent partially observable Markov decision process (POMDP). Within this framework, each vehicle agent makes decisions based on local observations while collectively contributing to global optimization, as illustrated in Figure 2. The overall decision-making process is governed by the proposed BO-MDDA dispatch strategy algorithm, which will be described in Section 4 in detail.

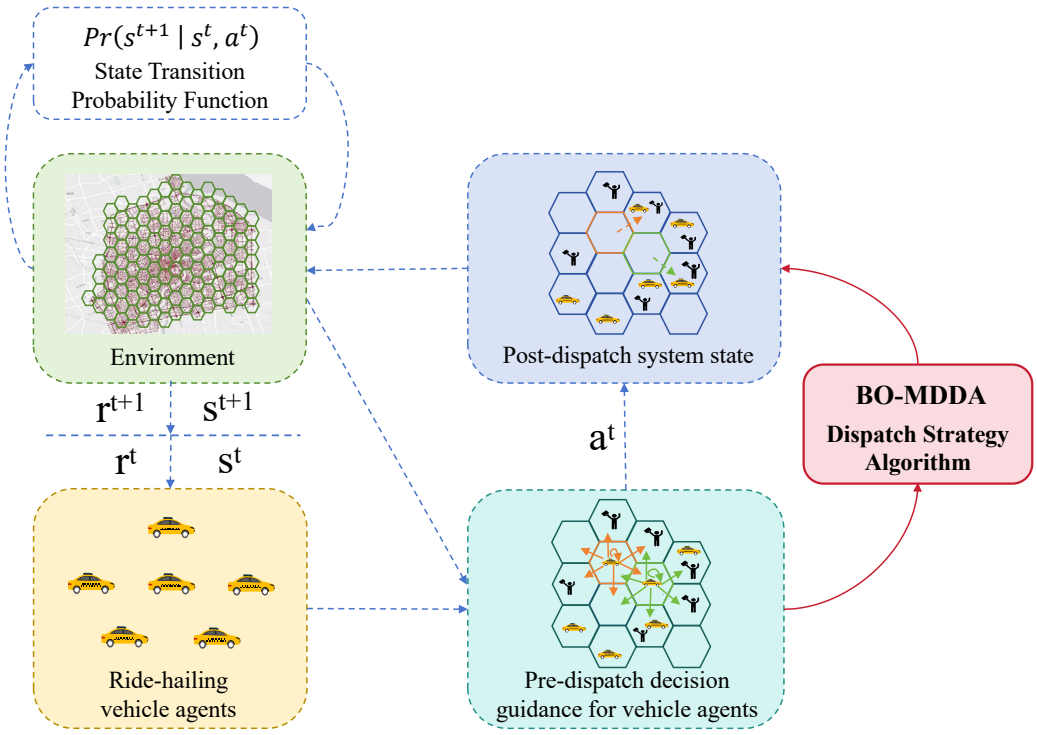

Figure 2. POMDP-based operational framework of the GRV-DDM.

 doi:10.20944/preprints202603.1961.v1

The multi-agent POMDP is defined by the tuple<S, A, R, P, G, N, γ>, representing the state space, action space, reward function, state transition probability, grid set, number of agents, and discount factor.

Agent. Each ride-hailing vehicle is modeled as an agent $i \in \{1, 2, \ldots, N\}$. At each time step $t \in \{1, 2, \ldots, T\}$, agent i observes its state $s_i^t \in S_i$, selects an action $a_i^t \in A_i$ according to policy π_i, receives reward r_i^t, and transitions to the next state s_i^{t+1} based on the state transmission probability.

State. The state consists of global and local components. At time step t, the global state is defined as $o^t = <t, day, UD^t>$, where day is the day of the week, and the supply-demand imbalance vector is given by $[UD^t] = [(U_{g_1}^t - D_{g_1}^t), \ldots, (U_{g_W}^t - D_{g_W}^t)]$, capturing the real-time imbalance across all grids. The local state of agent i is defined as $s_i^t = < g_i, t, day, f_i^t, UD_{\Omega_{g_i}}^t >$, where g_i is the current grid, $f_i^t \in \{0,1,2\}$ denotes the operational state (idle, occupied, in-dispatch), and $[UD_{\Omega_{g_i}}^t] = [(U_{g_1}^t - D_{g_1}^t), \ldots, (U_{|\Omega_{g_i}|}^t - D_{|\Omega_{g_i}|}^t)]$ captures the supply-demand imbalance within the neighborhood Ω_{g_i}.

Action. The action space is restricted to movements between adjacent grids. For agent i located in g_i, the action at time step t is $a_i^t = d_{g_i'}^{g_i}$, where $g_i' \in \Omega_{g_i}$, and $g_i = g_i'$ corresponds to remaining in the current grid.

Reward. The reward function balances supply-demand equilibrium and dispatch cost. The local equilibrium metric is defined as:

$$h_i^t = \sum_{j \in \Omega_{g_i}} \frac{|U_j^t - D_j^t|}{max\{U_j^t, D_j^t\}}, \tag{4}$$

and the dispatch distance and time are $Len_i^t = dis(g_i, g_i')$ and $Tim_i^t = time(g_i, g_i')$. The instantaneous reward is then given by:

$$r_i^t = -\lambda_1 h_i^t - \lambda_2 Len_i^t - \lambda_3 Tim_i^t, \tag{5}$$

where λ_1, λ_2, and λ_3 are weighting parameters.

Transmission probability. Given joint state s^t and joint action a^t, the state transition probability is $Pr(s^{t+1} | s^t, a^t)$.

Discount factor. A discount factor $\gamma \in [0,1]$ balance immediate and future rewards. The objective is to maximize the cumulative expected reward $\max_{\pi} E\left[\sum_{i=1}^{N} \sum_{t=1}^{T} \gamma^t r_i^t\right]$ for optimizing the joint policy $\pi = \{\pi_1, \pi_2, \ldots, \pi_N\}$.

This section introduces the GRV-DDM and models the dispatch process as a multi-agent POMDP to capture spatial coupling and partial observability. The resulting formulation provides a unified decision framework for vehicle relocation under dynamic supply-demand imbalance. Based on this model, the next section introduces a bi-level optimization algorithm to solve the GRV-DDM efficiently.

4. Bi-Level Optimization Multi-Dimensional Dispatch Decision Algorithm

This section presents the Bi-Level Optimization Multi-Dimensional Dispatch Decision Algorithm (BO-MDDA), designed to simultaneously maximize drivers' long-term earnings and mitigate regional supply-demand imbalances. Specifically, Section 4.1 introduces the multi-head neural network and experience buffer mechanisms adopted in BO-MDDA, and Section 4.2 details the algorithmic implementation of the global optimizer, local optimizer, and their bidirectional interaction process.

4.1. Multi-Head Neural Network and Experience Buffer Design

To enable efficient feature sharing while maintaining policy independence, BO-MDDA employs a multi-head neural network combined with agent-specific experience replay buffers as shown in Figure 3. The shared backbone extracts common spatio-temporal features of the urban traffic environment, including regional supply–demand distributions, temporal demand patterns, and

 doi:10.20944/preprints202603.1961.v1

traffic dynamics, providing a unified high-level representation. Each agent $i \in N$ then uses an independent head network to map its local state s_i^t to its dispatch policy $\pi_i(a_i^t \mid s_i^t)$, allowing individualized decision-making. This decoupling of feature extraction and policy learning balances parameter sharing with policy diversity, enhancing generalization across traffic conditions while mitigating interference and conflicts among agents in large-scale dispatch scenarios.

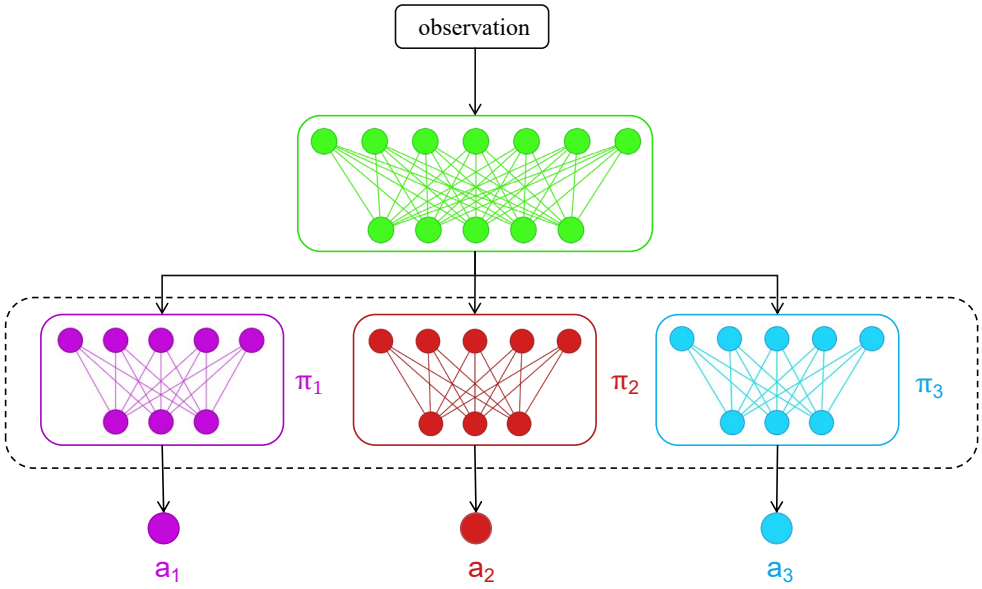

Figure 3. Multi-head neural network framework diagram.

In addition, each agent i's trajectory, $traj(i)=\{s_i^1,a_i^1,r_i^1,s_i^2,a_i^2,r_i^2,\ldots,s_i^T,a_i^T,r_i^T\}$, is stored in its corresponding replay buffer E_i. During training, mini-batches are randomly sampled from E_i to update the agent's policy and value networks. Compared to a shared buffer, this design reduces temporal correlations, prevents negative interference from heterogeneous agent experiences, and significantly improves the stability and efficiency of policy optimization in multi-agent learning environments.

4.2. BO-MDDA Algorithm Implementation

To balance system-wide efficiency and individual driver objectives in dynamic ride-hailing environments, BO-MDDA adopts a dual-optimization architecture composed of a global optimizer and a local optimizer as shown in Figure 4. The global optimizer regulates the collective spatial distribution of vehicles to alleviate supply-demand imbalance (shown in Figure 4a), while the local optimizer optimizes real-time dispatch decisions under partial observability and maximize drivers' long-term earnings (shown in Figure 4b). A bidirectional feedback mechanism tightly couples the two layers, enabling coordinated population evolution and individual policy learning within a unified framework.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1961.v1

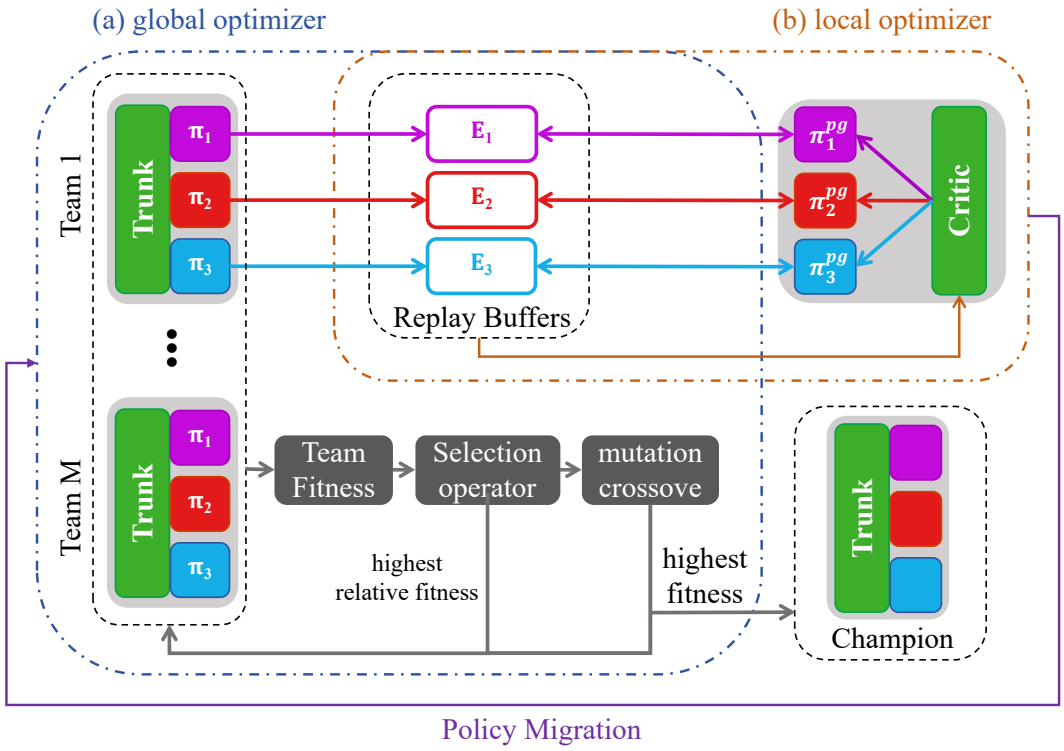

Figure 4. BO-MDDA framework diagram.

4.2.1. Global Optimizer Based on Evolutionary Algorithms

The global optimizer aims to guide the evolutionary dynamics of collective dispatch strategies among ride-hailing drivers, steering population-level behavior toward a stable equilibrium that can efficiently respond to dynamic supply-demand conditions. Within BO-MDDA, this objective is realized through a neuroevolutionary optimization mechanism that explicitly incorporates principles from evolutionary game theory. In particular, the global optimizer embeds a frequency-dependent selection mechanism into the fitness evaluation process, enabling strategy adaptation through natural selection, crossover, and mutation across competing strategy teams as shown in Figure 4a.

In each evolutionary generation, a population $pop = \{Team_1, Team_2, ..., Team_M\}$ comprising M strategy teams is maintained. Each team contains of x agents that share an identical neural network architecture but are initialized with distinct parameter vectors δ, representing alternative collective behavior patterns. All teams are initialized with zero fitness.

To reduce stochastic variance in fitness evaluation, φ independent population-based simulation rounds are conducted. In each round, M teams are randomly sampled with replacement from the population to form a dispatch environment consisting of N ride-hailing vehicles, where the condition $x \times M \approx N$ generally holds. Team $Team_m$ controls x vehicles, and its cumulative fitness is computed as:

$$fitness = \frac{1}{\varphi}\sum_{j=1}^{\varphi}\sum_{v=1}^{x}(R_v^j - C_v^j). \quad (6)$$

This formulation captures the relative competitiveness of a team's strategy under the current population distribution, reflecting the frequency-dependent nature of evolutionary game dynamics.

After fitness evaluation, population strategies evolve through the following evolutionary operations:

Selection: The top e teams with the highest fitness values are selected as the elite set L and are directly preserved for the next generation. From the remaining $M-e$ teams, tournament selection is applied to construct a parent candidate set F for crossover.

 doi:10.20944/preprints202603.1961.v1

Crossover: One team $Team_l$ is randomly selected from the elite set L, and another team $Team_f$ is selected from set F. A single-point crossover operation is performed on their parameter vectors to generate offspring teams. This process is repeated until the size of set F reaches $M-e$.

Mutation: For each offspring team in set F, a small random perturbation $\delta' = \delta + \epsilon$ is applied to its parameters with probability p_m, where $\epsilon \sim N(0, \sigma_1{}^2 I)$, I denotes the identity matrix and $\sigma_1{}^2$ controls the mutation intensity.

Population Update: The elite set L and the newly generated offspring teams are merged to form the updated population pop for the next generation.

Through iterative evolution, the population converges toward an evolutionarily stable strategy, providing global strategic guidance for dispatch coordination in dynamic environments. The complete global optimization procedure is summarized in Algorithm 1.

Algorithm 1: Global Optimizer based on Evolutionary Algorithm

Input: Population size M; Number of ride-hailing agents N; Evaluation trials φ; Number of elites e; Mutation probability p_m; Mutation strength σ; Experience buffer set $\{E_1, ..., E_N\}$

Output: Global optimal team strategy (Champion)

Initialize population *pop*, containing M candidate teams, each team $Team_m$ contains x agents sharing parameters δ_m;

for Generation number = 1 to Maximum iteration number **do**

  for Each team $Team_m \in pop$ **do**

    $fitness_m \leftarrow 0$;

    for Trial = 1 to φ **do**

      Randomly sample M teams from *pop* (with replacement);

      Instantiate simulation environment where:

        - $Team_m$ controls x vehicles

        - The remaining $N - x$ vehicles are controlled by sampled teams;

      Run simulation until episode end, record accumulated net profit G_sum of vehicles in $Team_m$;

    $fitness_m \leftarrow fitness_m + G\_sum$

  $fitness_m \leftarrow fitness_m/\varphi$;

  Sort by fitness, keep top e elites in set L;

  Initialize empty set F;

  Select $(M - e)$ teams from $pop \setminus L$ using tournament selection and add to set F;

  while $|F| < (M - e)$ **do**

    Randomly sample an elite team $Team_l \in L$ and a non-elite team $Team_f \in F$;

    perform single-point crossover, generate new offspring team and add to F;

  for Each offspring team Team$\in$F **do**

    if mutate with probability p_m **then**

      $\delta_m \leftarrow \delta_m + \epsilon$, where $\epsilon \sim N(0, \sigma_1{}^2 I)$;

  Update population $pop \leftarrow L \cup F$;

  Inject good strategies from local optimizers into *pop*;

  Write high-fitness team trajectories into $\{E\}$ for use by local optimizers;

4.2.2. Local Optimizers Based on Policy Gradients

The local optimizer employs policy gradient to learn individualized dispatch strategies for each vehicle agent by maximizing the discounted cumulative reward:

$$J(\theta_i) = E_{\pi_{\theta_i}}[\sum_{t=0}^{T} \gamma^t r_i^t], \tag{7}$$

where θ_i denotes the policy parameters of agent i.

 doi:10.20944/preprints202603.1961.v1

BO-MDDA adopts an Actor-Critic architecture implemented under centralized training with decentralized execution. During execution, each agent selects dispatch actions based solely on its local observation s_i^t through its Actor network. During training, the centralized Critic with access to the global state o^t and joint actions evaluates action quality, providing informative gradient signals for policy updates. The state-action value function estimated by the Critic is defined as:

$$Q(o, a) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r^{t+k} \mid o^t=o, a^t=a\right], \tag{8}$$

where $k = 0,1,2,\ldots$ denotes the future time step index.

To improve learning stability and sample efficiency, each agent i maintains an independent experience replay buffer E_i. During training, agent i interacts with the environment by observing its local state s_i^t, executing action a_i^t, and receiving reward r_i^t. These transitions are stored in E_i to reduce temporal correlations and mitigate interference among heterogeneous experiences. Mini-batches sampled from E_i are then used to update both Actor and Critic networks as shown in Figure 4b.

To encourage sufficient exploration and avoid premature convergence, Gaussian noise is added to the Actor's output during action selection, such that the executed action is given by $a_i = \Pi_{A_i}(\pi_{\theta_i}(s_i) + \varepsilon)$, where $\varepsilon \sim N(0, \sigma_2{}^2 I)$ denotes Gaussian noise and Π_{A_i} is the projection operator that ensures the action remains within the feasible action space.

The local optimizer is trained by alternately updating the Actor and Critic networks, as illustrated in Figure 5. The Critic network aims to approximate the action-value function $Q_\phi(o, a)$ by minimizing the temporal-difference (TD) error. For a mini-batch of size T, the Critic loss is defined as:

$$L(\phi) = \frac{1}{T}\sum_{t=1}^{T}\left(y_i^t - Q_\phi(o_i^t, a_i^t)\right)^2, \tag{9}$$

where the TD target is computed as:

$$y_i^t = r_i^t + \gamma Q_{\phi'}(o_i^{t+1}, \pi_{\theta'}(s_i^{t+1})). \tag{10}$$

Here, ϕ' and θ' denote the parameters of the target Critic and Actor networks, respectively. By employing slowly updated target networks, the training process is stabilized and oscillations in value estimation are effectively mitigated.

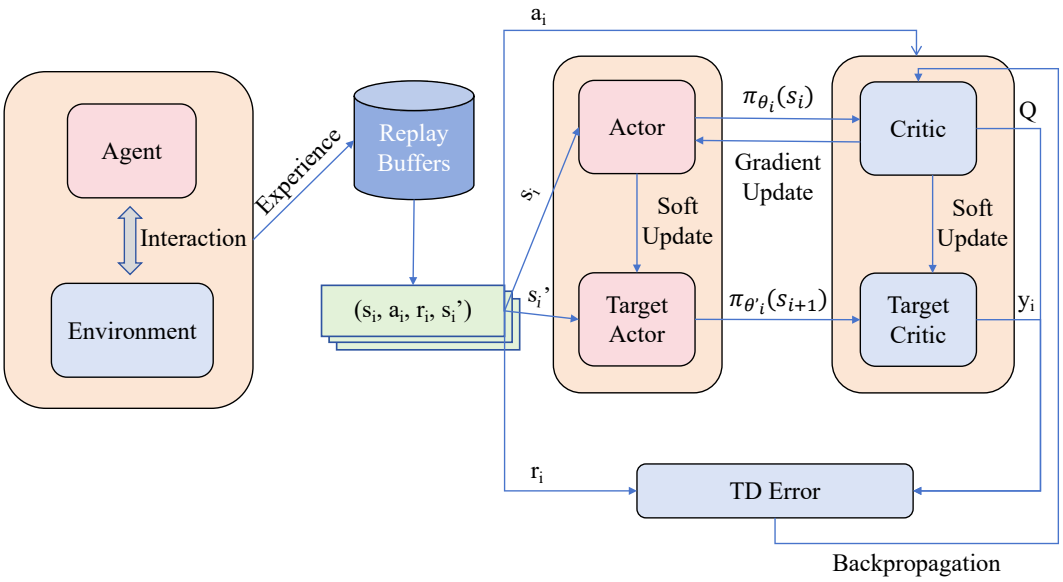

Figure 5. Alternating update network structure diagram for Critic and Actor.

 doi:10.20944/preprints202603.1961.v1

The Actor network is updated using gradient information provided by the Critic. According to the Deterministic Policy Gradient (DPG) theorem, the policy gradient for agent i is given by:

$$\nabla_{\theta_i} J(\theta_i) = \frac{1}{T}\sum_{t=1}^{T} \nabla_{\theta_i} \pi_{\theta_i}(s_i^t) \nabla_{a_i} Q_\phi(o_i^t, a_i)|_{a_i=\pi_{\theta_i}(s_i^t)}. \quad (11)$$

To maintain consistency between online and target networks while avoiding instability caused by abrupt synchronization, BO-MDDA adopts a soft update strategy:

$$\phi' \leftarrow \tau\phi + (1-\tau)\phi', \quad (12)$$

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta', \quad (13)$$

where $\tau \in (0,1)$ is the update rate. Smaller values of τ ensure smoother target network evolution, which improves convergence speed and final policy performance in complex multi-agent environments.

4.2.3. Bidirectional Feedback Mechanism

A key innovation of BO-MDDA lies in its bidirectional feedback mechanism, which enables structured information exchange between the global evolutionary optimizer and local reinforcement learning optimizers.

After each evolutionary generation, trajectories generated by the top-performing elite teams are injected into agents' replay buffers as high-quality experiences, guiding local policy learning toward globally advantageous behaviors and accelerating convergence. Conversely, the best-performing policy parameters θ_{best} learned by local optimizers are periodically encapsulated into new strategy teams and injected into the evolutionary population, replacing low-fitness teams and enriching the global search space.

Through this cyclic exchange, population-level evolution provides strategic guidance for individual learning, while locally optimized policies continuously enhance evolutionary exploration. This deep coupling allows BO-MDDA to effectively integrate evolutionary game dynamics with multi-agent reinforcement learning, achieving coordinated optimization of global efficiency and driver profitability. The complete BO-MDDA algorithm is summarized in Algorithm 2.

Algorithm 2: BO-MDDA Algorithm (Bi-level Optimization Framework)

Input: Number of agents N; Global optimizer parameters (M, φ, e, Crossover and Mutation parameters); Local optimizer parameters (γ, τ, σ_2); Experience buffer set $\{E_1, ..., E_N\}$
Output: Global optimal team strategy (Champion), Optimized individual policy parameters θ_π
Initialize global optimizer population *pop*, containing M candidate teams;
Initialize local optimizer (policy network and critic network) and target networks;
Initialize experience buffer $E_i \leftarrow \emptyset$ for each agent $i = 1, \ldots, N$;
for *Iteration steps* = 1 to *Maximum iteration number* **do**
  for Each team $\pi \in$ pop **do**
    $fitness_\pi \leftarrow 0$;
    for *Trial* = 1 to φ **do**
      Interact with the environment, accumulate team reward g;
      Store each agent's interaction (s_i, a_i, r_i, s_i') into corresponding experience buffer E_i;
      $fitness_\pi \leftarrow fitness_\pi + G\_sum$;
    $fitness_\pi \leftarrow fitness_\pi/\varphi$;
  Sort by fitness and keep top e elites in set L;
  The remaining individuals undergo selection, crossover, and mutation to form the next generation *pop*;
  for Agent i = 1 to N **do**
    Sample mini-batch (s, a, r, s') from experience buffer E_i, obtaining global

 doi:10.20944/preprints202603.1961.v1

states o, o';
Construct TD target $y = r + \gamma Q'(o', \pi(s'/\theta_\pi'))$;
Update critic: Minimize $L(\phi) = \frac{1}{T}\Sigma\left(y - Q(o, a/\phi)\right)^2$;
Update policy: Use deterministic policy gradient $\nabla_\theta J$;
Soft update target networks: $\phi' \leftarrow \tau\phi + (1-\tau)\phi'$, $\theta' \leftarrow \tau\theta + (1-\tau)\theta'$;
if Iteration steps $modT_{inject}=0$ **then**
Select optimal policy θ_π^{best} from local optimizer and inject into global optimizer population pop;
Write high-fitness team trajectories from the global optimizer into $\{E_i\}$ for local optimizer sampling use;

5. Experiment

This section conducts a comprehensive experimental evaluation of the proposed BO-MDDA framework to examine its effectiveness under dynamic and heterogeneous traffic conditions. A SimMobility-based simulation platform driven by real-world ride-hailing data is employed to reproduce realistic urban traffic dynamics. Through comparative experiments across multiple traffic scenarios, the performance of BO-MDDA is systematically assessed and compared with representative heuristic- and learning-based dispatch algorithms, highlighting its advantages in multi-agent coordination and global optimization.

5.1. Simulation Platform, Dataset, and Evaluation Metrics

Experiments are conducted on a SimMobility-based simulator that incorporates dynamic traffic variations reconstructed from real-world taxi trajectory data. By adjusting background traffic intensity, the simulator reproduces realistic congestion patterns and order fulfillment times, enabling robust evaluation under varying traffic conditions.

The dataset is provided by Didi Chuxing and contains ride-hailing orders and vehicle trajectories from Chengdu over four consecutive weeks in November 2016. Each order record includes origin, destination, trip duration, and fare, while trajectory data record vehicle ID, timestamp, and geographic coordinates. The urban area is discretized into a hexagonal grid with an approximate radius of 1.3 km. After removing isolated regions based on road network topology, 142 grids covering the city center are retained. Vehicle and passenger locations are mapped to grids via unique grid identifiers.

In order to evaluate whether the proposed BO-MDDA framework effectively reconciles global system coordination and individual driver optimization, algorithm performance is assessed using three complementary metrics.

Order Response Rate (ORR): Measures dispatch efficiency by the proportion of completed orders, defined as $\frac{N_{done}}{N_{total}} \times 100\%$, where N_{done} and N_{total} denote the number of completed and total orders, respectively.

Accumulated Driver Income (ADI): Captures the total income accrued by drivers over a period period, computed as $\sum_{t=1}^{T}\sum_{i\in N}\left(R_i^t - C_i^t\right)$.

Average Idle Time (AIT): Represents the mean duration vehicles remain idle, expressed as $1-\frac{\sum_{t=1}^{T}\sum_{i=1}^{N} W_i^t}{T}$, where $W_i^t = \begin{cases} 0 & f=0 \text{ or } 2 \\ travel\ time & f=1 \end{cases}$. Travel time refers to the driving time during the entire process of accepting an order for a ride-hailing vehicle, and W_i^t denotes the working time of agent i.

In order to highlight the effectiveness of integrating evolutionary game theory with deep reinforcement learning for enhancing multi-agent coordination, BO-MDDA is compared with representative heuristic-based and reinforcement-learning-based dispatch algorithms that lack explicit population-level strategic evolution.

 doi:10.20944/preprints202603.1961.v1

Greedy [27]: Assigns idle vehicles to regions with the highest immediate demand, serving as a heuristic baseline. It evaluates whether BO-MDDA can outperform purely myopic dispatch decisions that ignore long-term system dynamics.

MADDPG: A multi-agent reinforcement learning approach emphasizing cooperative policy learning. It is adopted as a representative MARL baseline to examine the performance gains of BO-MDDA over cooperative learning frameworks without evolutionary global optimization.

COX [28]: Combines temporal graph convolution for demand prediction with DQN-based dispatching. It is used for comparison to assess the advantage of BO-MDDA over prediction-driven dispatch strategies that rely on explicit demand forecasting.

RBDQN: Integrates graph convolutional networks with DQN to model road grids for vehicle dispatching. This baseline is selected to evaluate the robustness of BO-MDDA in capturing spatial correlations and handling complex urban traffic dynamics.

5.2. Performance Evaluation

Ride-hailing vehicles are initially distributed randomly across the road network, while passenger orders follow real-world demand patterns. To evaluate adaptability under different traffic conditions, three scenarios are considered: (i)Free-flowing traffic, which represents an idealized environment with smooth traffic and balanced supply-demand, providing a baseline for evaluating dispatch efficiency. (ii)Real-world traffic, which mirrors actual urban traffic fluctuations, capturing unpredictable variations in demand and congestion that challenge dispatch algorithms. (iii)Congested traffic, which simulates peak-hour conditions with severe traffic delays and pronounced supply-demand imbalances, testing algorithm robustness under extreme conditions. For each scenario, the number of ride-hailing vehicles is set to 3,000, 6,000, and 9,000, respectively.

5.2.1. Performance Analysis

Table 1 summarizes the ORR, ADI, and AIT of BO-MDDA in comparison with Greedy, MADDPG, COX, and RBDQN across three traffic congestion scenarios. Overall, BO-MDDA consistently achieves superior performance across all three metrics, indicating its robustness under varying traffic conditions.

Table 1. Experimental performance comparison of BO-MDDA and four other contrast algorithms.

Methods		Greedy	MADDPG	COX	RBDQN	BO-MDDA
Free-Flowing	ORR(%)	75.23	78.31	82.58	83.64	86.11
	ADI	1,015,483	1,537,922	1,589,308	1,717,380	2,487,594
	AIT(%)	49.55	16.58	16.09	15.33	15.51
Real	ORR(%)	62.64	65.88	68.37	71.31	76.29
	ADI	1,577,296	2,428,668	2,409,751	2,526,019	2,583,447
	AIT(%)	59.72	32.99	33.21	28.84	24.03
Conges-tion	ORR(%)	49.20	56.58	57.94	57.25	59.97
	ADI	1.628.982	2,370,153	2,489,228	2,548,062	2,697,730
	AIT(%)	68.04	52.59	50.13	47.54	46.29

Under free-flowing traffic, BO-MDDA attains the highest order response rate and accumulated driver income while maintaining a low average idle time. This suggests that even when supply–demand imbalance is relatively mild, the proposed bi-level optimization framework can still exploit fine-grained coordination among agents to improve both system efficiency and driver profitability.

In real-world traffic conditions, where demand fluctuations and congestion introduce higher uncertainty, the advantages of BO-MDDA become more pronounced. Compared with heuristic and learning-based baselines, BO-MDDA exhibits a clear improvement in order response rate and driver income, accompanied by a substantial reduction in vehicle idle time. This demonstrates its stronger adaptability to dynamic and partially observable environments.

 doi:10.20944/preprints202603.1961.v1

Under congested traffic scenarios, all methods experience performance degradation due to severe traffic delays and intensified supply–demand mismatch. Nevertheless, BO-MDDA maintains the best overall performance, achieving the highest driver income and the lowest idle time among all compared methods. These results indicate that the integration of evolutionary game-based global coordination with deep reinforcement learning at the individual level enables BO-MDDA to better mitigate congestion-induced inefficiencies.

In summary, the results in Table 1 confirm that BO-MDDA consistently outperforms both heuristic and state-of-the-art learning-based dispatch algorithms across diverse traffic scenarios, effectively balancing global dispatch efficiency and individual driver benefits.

Figures 6–8 further illustrates the comparative trends of ORR, ADI, and AIT across different dispatch methods. Overall, a clear monotonic pattern can be observed: as dispatch strategies evolve from heuristic-based to learning-based and coordinated approaches, order response rate and accumulated driver income increase steadily, while average idle time decreases. This inverse relationship between service efficiency and vehicle idleness indicates improved utilization of fleet resources rather than simple performance trade-offs.

Notably, learning-based baselines show incremental gains over the greedy strategy, whereas more coordinated methods achieve simultaneous improvements across all three metrics. The smoother ORR and ADI growth, together with the consistent decline in AIT, suggests that enhanced inter-agent coordination enables dispatch policies to align individual decisions with system-level objectives more effectively. These trends visually corroborate the quantitative results in Table 1 and highlight the role of coordinated decision-making in stabilizing performance under increasing operational complexity.

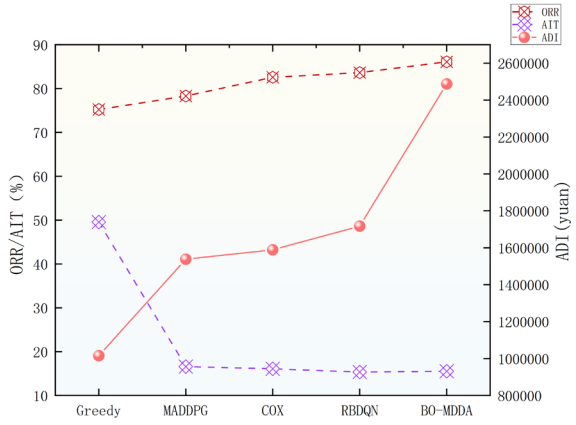

Figure 6. Performance comparison chart of algorithms in free-flowing traffic environments.

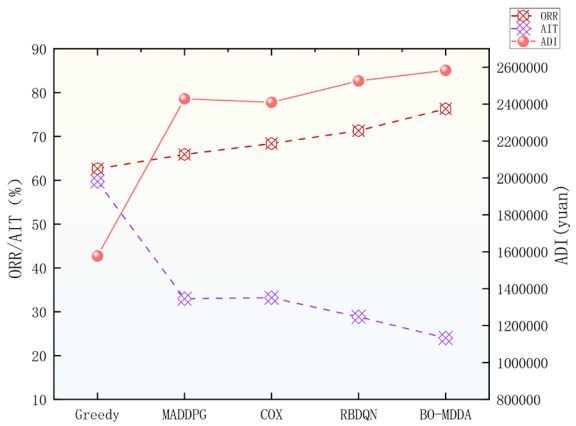

Figure 7. Performance comparison chart of algorithms in real-world traffic environments.

 doi:10.20944/preprints202603.1961.v1

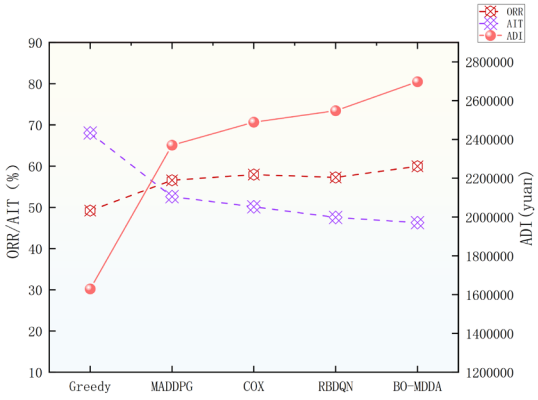

Figure 8. Performance comparison chart of algorithms in congestion traffic environments.

5.2.2. Ablation Experiment

Table 2 presents the ablation results comparing BO-MDDA with G-BOMAMR under three traffic conditions. The key difference between the two methods lies in agent modeling: BO-MDDA treats individual vehicles as decision-making agents, whereas G-BOMAMR aggregates decision-making at the grid level by modeling each grid cell as an agent characterized by local supply-demand gaps.

Table 2. Comparison of experimental performance results for BO-MDDA and G-BOMAMR algorithms.

Methods		BO-MDDA	G-BOMDDA
Free-Flowing	ORR(%)	86.11	73.97
	ADI	2,487,594	1,293,807
	AIT(%)	15.51	21.46
Real	ORR(%)	76.29	64.83
	ADI	2,583,447	2,412,971
	AIT(%)	24.03	39.23
Congestion	ORR(%)	59.97	59.55
	ADI	2,697,730	2,657,086
	AIT(%)	46.29	47.29

Under free-flowing traffic conditions, BO-MDDA significantly outperforms G-BOMAMR across all metrics. The order response rate increases from 73.97% to 86.11%, while accumulated driver income nearly doubles. Meanwhile, average idle time is reduced by approximately 6 percentage points. These results indicate that vehicle-level agent modeling enables finer-grained dispatch decisions, allowing individual vehicles to respond more flexibly to localized demand variations even when traffic conditions are relatively unconstrained.

In real-world traffic scenarios, where demand uncertainty and traffic dynamics are more pronounced, the advantages of BO-MDDA become even more evident. Compared with G-BOMAMR, BO-MDDA achieves a substantially higher order response rate and a marked reduction in vehicle idle time. This suggests that modeling individual vehicles as agents allows the dispatch policy to better exploit heterogeneous states and partial observations, which cannot be fully captured by grid-level aggregation.

Under congested traffic conditions, the performance gap between the two methods narrows, particularly in terms of order response rate and driver income. Nevertheless, BO-MDDA consistently maintains a slight advantage and achieves lower average idle time. This indicates that although severe congestion constrains overall system performance, vehicle-level agent modeling still contributes to more efficient utilization of limited mobility resources.

Overall, the ablation results demonstrate that replacing vehicle-level agents with grid-level agents leads to a consistent degradation in dispatch performance. This confirms that fine-grained

 doi:10.20944/preprints202603.1961.v1

agent modeling is a critical component of BO-MDDA, enabling more precise coordination between individual decision-making and system-level optimization across diverse traffic environments.

Figures 9–11 illustrates the performance differences between BO-MDDA and its grid-agent variant G-BOMAMR under varying traffic conditions. Across all scenarios, BO-MDDA consistently achieves higher order response rates and accumulated driver income, while maintaining lower average idle time. The performance gap is particularly pronounced under free-flowing and real-world traffic, indicating that vehicle-level agent modeling enables more fine-grained and effective dispatch decisions. Under congested conditions, the gap narrows, yet BO-MDDA still maintains a slight advantage, suggesting that explicit vehicle-level decision-making remains beneficial even when system dynamics are heavily constrained. Overall, the observed trends confirm the importance of individual vehicle agents in enhancing coordination efficiency and system robustness.

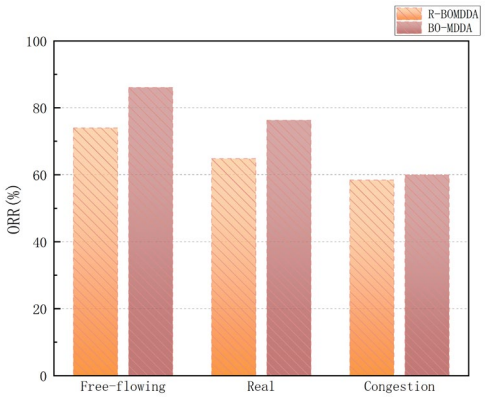

Figure 9. Comparison of ORR between G-BOMAMR and BO-MDDA under Various Traffic Environments.

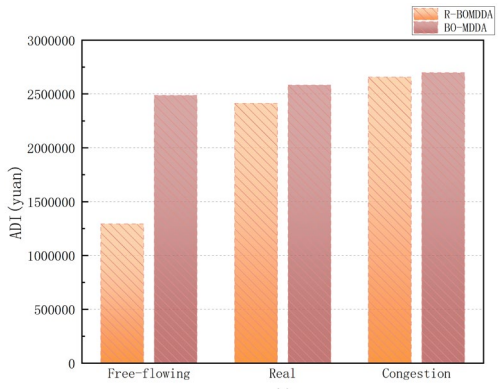

Figure 10. Comparison of ADI between G-BOMAMR and BO-MDDA under Various Traffic Environments.

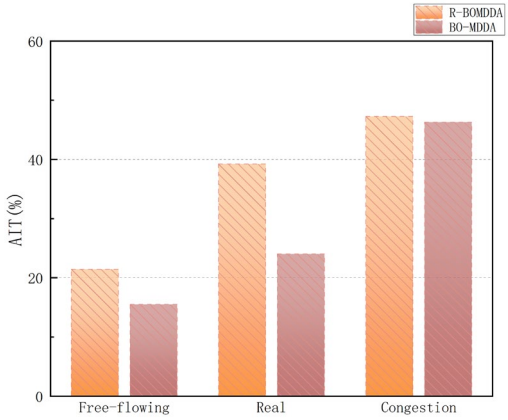

Figure 11. Comparison of AIT between G-BOMAMR and BO-MDDA under Various Traffic Environments.

 doi:10.20944/preprints202603.1961.v1

Overall, the experimental results consistently demonstrate that BO-MDDA outperforms both heuristic and learning-based baselines across diverse traffic conditions, confirming its effectiveness in jointly improving dispatch efficiency, driver income, and system robustness through bi-level optimization and coordinated multi-agent decision-making.

6. Conclusions

To jointly optimize global system efficiency and individual driver benefits under dynamic traffic conditions and fluctuating passengers demand, this study constructed a grid-based ride-hailing vehicle dispatch decision model (GRV-DDM), providing a structured modeling framework for dynamic supply-demand environments. Besides, a Bi-Level Optimization Multi-Dimensional Dispatch Decision Algorithm (BO-MDDA) was proposed, which leveraged evolutionary games to guide supply-demand equilibrium at the group level while employing deep reinforcement learning to optimize real-time driver profits at the individual level. A novel bidirectional feedback coordination mechanism was designed, enabling mutual guidance and co-evolution between these two optimization paradigms, thereby effectively balancing global system objectives with local agent interests. Extensive experimental results across multiple traffic scenarios demonstrated that BO-MDDA consistently outperforms representative dispatch baselines, including greedy heuristics and state-of-the-art reinforcement learning methods. Compared with these approaches, BO-MDDA achieved higher order response rates, greater accumulated driver income, and lower vehicle idle time.

In spite of the progress in this paper, challenges remain in enhancing convergence speed for large-scale systems, improving training stability, and adapting to increasingly complex traffic dynamics. Future research will focus on addressing these issues and exploring the extension of BO-MDDA to other ride-sharing platforms.

Author Contributions: Conceptualization, L.Y. and P.W.; methodology, P.W.; software, P. W.; validation, L.Y., P.W. and S.W.; formal analysis, H.J.; investigation, J.H.; resources, S.W.; data curation, P.W.; writing—original draft preparation, P.W.; writing—review and editing, L.Y.; visualization, P.W.; supervision, L.Y.; project administration, S.W.; funding acquisition, L.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the National Natural Science Foundation of China (Grants No. 62362031, 52268066 and 62262022) and the Jiangxi Provincial Natural Science Foundation (Grants No. 20252BAC240353).

Data Availability Statement: The datasets analyzed during the current study are available from the DiDi Technology collaborative platform, https://outreach.didichuxing.com.

Acknowledgments: The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

EGT	Evolutionary Game Theory
DRL	Deep Reinforcement Learning
MADRL	Multi-Agent Deep Reinforcement Learning
GANs	Generative Adversarial Networks

 doi:10.20944/preprints202603.1961.v1

References

1. Rui Qiao, Yiyan Zhou, and Guohua Cheng. 2025. Research on the Impact Effect of Online Ride hailing Management Policies on Urban Public Transportation. Proceedings of the 4th Asia-Pacific Artificial Intelligence and Big Data Forum. Association for Computing Machinery, New York, NY, USA, 564–568.
2. Alminas Čivilis, Andrius Barauskas, Agnė Brilingaitė, Linas Bukauskas, and Simonas Šaltenis. 2024. Managing Advanced and Flexible Ride-Hailing Requests. In Proceedings of the 16th ACM SIGSPATIAL International Workshop on Computational Transportation Science (IWCTS '23). Association for Computing Machinery, New York, NY, USA, 62–69.
3. Shixiang Wan, Shikai Luo, and Hongtu Zhu. 2024. Causal Probabilistic Spatio-Temporal Fusion Transformers in Two-Sided Ride-Hailing Markets. ACM Trans. Spatial Algorithms Syst. 10, 3, Article 28 (September 2024), 18 pages.
4. C. V. Beojone, P. Zhu, I. I. Sirmatel and N. Geroliminis, "A Two-Layer Approach for Rebalancing Ride-Hailing Vehicles," 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 2023, pp. 2460-2465.
5. Taoyuan Yu. 2024. Optimizing Computational Efficiency in Autonomous Vehicles: Integrative Edge and Cloud Computing Strategies in Vehicular Networks. In Proceedings of the 2024 11th International Conference on Wireless Communication and Sensor Networks (icWCSN '24). Association for Computing Machinery, New York, NY, USA, 5–12.
6. Lefeng Zhang, Tianqing Zhu, Ping Xiong, Wanlei Zhou, and Philip S. Yu. 2021. More than Privacy: Adopting Differential Privacy in Game-theoretic Mechanism Design. ACM Comput. Surv. 54, 7, 136.
7. Panayiotis Danassis and Boi Faltings. 2019. Courtesy as a Means to Coordinate. In Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 665–673.
8. Florian Merkle, Gregor Blossey, Stefan Haeussler, and Manuel Schneckenreither. 2024. Reinforcement learning in autonomous multi-vehicle systems: A structured review. In Proceedings of the 2024 13th International Conference on Software and Computer Applications (ICSCA '24). Association for Computing Machinery, New York, NY, USA, 310–317.
9. Zhang, X.; Sun, J.; Gong, C.; Wang, K.; Cao, Y.; Chen, H.; Liu, Y. Mutual Information as Intrinsic Reward of Reinforcement Learning Agents for On-demand Ride Pooling. In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS '24). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2024, 2597–2599.
10. R. Taallah and I. C. Msadaa, "Optimizing Cluster-Based Vehicle Dispatching in Ride-Hailing Services: A Reinforcement Learning Approach to Demand-Supply Dynamics," 2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), Gwalior, India, 2025, pp. 1-6.
11. Bolong Zheng, Lingfeng Ming, Qi Hu, Zhipeng Lü, Guanfeng Liu, and Xiaofang Zhou. 2022. Supply-Demand-aware Deep Reinforcement Learning for Dynamic Fleet Management. ACM Trans. Intell. Syst. Technol. 13, 3, Article 37 (June 2022), 19 pages.
12. Kaixiang Lin, Renyu Zhao, Zhe Xu, and Jiayu Zhou. 2018. Efficient Large-Scale Fleet Management via Multi-Agent Deep Reinforcement Learning. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). Association for Computing Machinery, New York, NY, USA, 1774–1783.
13. Yao Jing, Bin Guo, Yan Liu, Daqing Zhang, Djamal Zeghlache, and Zhiwen Yu. 2024. Efficient Bike-sharing Repositioning with Cooperative Multi-Agent Deep Reinforcement Learning. ACM Trans. Sen. Netw. Just Accepted (January 2024).
14. Jiaxuan Jiang, Ling Pan, Lin Zhou, Longbo Huang, and Zhixuan Fang. 2025. Tackling Sparsity in Designated Driver Dispatch with Multi-Agent Reinforcement Learning. In Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '25). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1060–1069.

 doi:10.20944/preprints202603.1961.v1

15. J. Sun, H. Jin, Z. Yang and L. Su, "Optimizing Long-Term Efficiency and Fairness in Ride-Hailing Under Budget Constraint via Joint Order Dispatching and Driver Repositioning," in IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 7, pp. 3348-3362, July 2024.
16. Yan, Liping, Chan Peng, Yue Tang, Wen-Bin Zhang, Jing Wang and Yu Cai. "NCG-TSM: A Noncooperative Game for the Taxi Sharing Model in Urban Road Networks." *J. Adv. Transp.* (2023): n. pag.
17. Alberto Castagna, Maxime Guériau, Giuseppe Vizzari, Ivana Dusparic, Marin Lujak, Ivana Dusparic, Franziska Klügl, and Giuseppe Vizzari. 2021. Demand-responsive rebalancing zone generation for reinforcement learning-based on-demand mobility. AI Commun. 34, 1 (2021), 73–88.
18. M. Wu, G. Cheng, P. Prakash, R. Yu, X. Xiong and M. Pan, "Resting Your Car for Cash: Incentive Mechanism Based TNC Car Scheduling for Carbon Neutrality," 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 2023, pp. 1429-1434.
19. K. Lai, X. Liu and W. K. V. Chan, "The Benefits of Willingness-to-Pay-Based Incentive-Driven Rider Repositioning in Ride-Hailing Systems," 2023 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, Singapore, 2023, pp. 0632-0639.
20. Y. Huang, N. Zheng, E. Liang, S. -C. Hsu and R. Zhong, "An Approximate Dynamic Programming Approach to Vehicle Dispatching and Relocation Using Time-Dependent Travel Times," 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 2023, pp. 2652-2657.
21. Jiachong Tu, Jia Yu, Xiaohui Huang, and Junhang Zong. 2025. GAN-SAC: Multi-agent Reinforcement Learning with Generative Reward Estimation for Urban Ride-hailing Optimization. In Proceedings of the 2025 5th International Conference on Internet of Things and Machine Learning (IoTML '25). Association for Computing Machinery, New York, NY, USA, 247–252.
22. Xu, Mingyue, Peng Yue, Fan Yu, Can Yang, Mingda Zhang, Shangcheng Li, and Hao Li. 2022. "Multi-Agent Reinforcement Learning to Unify Order-Matching and Vehicle-Repositioning in Ride-Hailing Services." International Journal of Geographical Information Science 37 (2): 380–402.
23. Y. Jwa, M. Gwak, J. Kwak, C. W. Ahn and P. Park, "Scalable Robust Multi-Agent Reinforcement Learning for Model Uncertainty," 2023 62nd IEEE Conference on Decision and Control (CDC), Singapore, Singapore, 2023, pp. 3402-3407.
24. J. Wang, "Deep reinforcement learning based multi-agent non-cooperative game strategy approach," 2023 3rd International Symposium on Computer Technology and Information Science (ISCTIS), Chengdu, China, 2023, pp. 770-774.
25. T. Yilmaz and Ö. Ulusoy, "Misinformation Propagation in Online Social Networks: Game Theoretic and Reinforcement Learning Approaches," in IEEE Transactions on Computational Social Systems, vol. 10, no. 6, pp. 3321-3332, Dec. 2023.
26. Y. Zhang and E. Zhao, "Design of MADDPG Capture Algorithm for Multiple UAV Cooperation," 2023 IEEE International Conference on Mechatronics and Automation (ICMA), Harbin, Heilongjiang, China, 2023, pp. 2021-2026.
27. A. dos Santos Mignon and R. L. d. A. da Rocha, "An adaptive implementation of ε-greedy in reinforcement learning," Procedia Computer Science, vol. 109, pp. 1146–1151, 2017.
28. Z. Liu, J. Li, and K. Wu, "Context-aware taxi dispatching at city-scale using deep reinforcement learning," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 3, pp. 1996–2009, 2020.