Charge or Pick Up? Optimizing E-Taxi Management: A Dual-Stage Heuristic Coordinated Reinforcement Learning Approach

Donghe Li, *Member, IEEE*, Chunlin Hu, Qingyu Yang, *Senior Member, IEEE*, Pengtao Song, Feiye Zhang, and Dou An

***Abstract*— In recent years, the rapid adoption of electric vehicles (EVs) in the taxi industry has transformed traditional taxi-hailing systems into electric taxi (E-taxi) hailing systems. As a result, it is crucial to develop effective strategies for optimizing E-taxi management by considering both passenger-taxi matching and charging planning. In this paper, we first formalize the E-taxi management optimization problem as a Markov decision problem with dynamic state and heterogeneous action. We then propose a dual-stage heuristic coordinated reinforcement learning (RL) approach that incorporates advanced feature selection and heuristic allocation strategies. Our approach consists of two main stages. In the first stage, we introduce the feature-guided state dimensionality stabilization proximal policy optimization (PPO) method to address dynamic state dimensions by a feature selection method, and enabling E-taxis to decide whether to charge or pick up passengers. In the second stage, we propose a heuristic coordinated assignment method to further allocate charging stations and passengers for the E-taxis, and provide the RL network in the first stage with rewards based on the results. This approach effectively tackles the challenge of RL processing of heterogeneous action spaces (charge and pick up). We evaluate our proposed method in a real-world E-taxi environment and find that it significantly enhances the experience for both E-taxis and passengers. Specifically, due to our method's rational planning for passenger pick-up and charging, E-taxis can increase their revenue by 20% compared to traditional RL methods or random scheduling approaches. As for passengers, since the taxis have more efficiently planned their charging behavior, the probability of their orders being answered increases by 15%, while their waiting time is reduced by 55%. These achievements contribute to the advancement of E-taxi management strategies and promote the widespread adoption of electric vehicles, ultimately supporting the transition to a more sustainable transportation system.**

***Note to Practitioners*—The increasing adoption of electric vehicles in the taxi industry has led to the need for effective E-taxi management strategies that consider both passenger-taxi matching and charging planning. In this study, we introduce a dual-stage heuristic coordinated reinforcement learning approach that addresses these challenges by integrating a feature-guided state dimensionality stabilization proximal policy optimization method and a heuristic coordinated assignment method. Our approach offers several practical benefits for E-taxi service providers, drivers, and passengers. For E-taxi service providers, the proposed method improves E-taxi dispatch efficiency, resulting in a more effective use of available resources and potentially increasing overall revenue. For E-taxi drivers, our approach leads to better planning of charging and passenger pick-up decisions, increasing their earnings by 20% compared to traditional methods, and reducing the average occurrence of low battery status from more than 4 times every 10 hours to less than 1 time. Passengers, on the other hand, experience improved service quality due to the more efficient E-taxi management. The probability of their orders being answered increases by 15%, and their waiting time is reduced by 100%. These improvements contribute to an enhanced user experience and may encourage further adoption of E-taxis as a sustainable transportation solution. The proposed method can be integrated into existing E-taxi hailing platforms, such as DiDi and Uber, to enhance their dispatch and charging management capabilities. As the global trend towards sustainable transportation continues to grow, our approach provides valuable insights and a practical solution for the efficient management of E-taxi fleets in modern urban environments.**

***Index Terms*— E-taxi, resource allocation, reinforcement learning, proximal policy optimization, feature selection.**

Received 11 March 2024; revised 21 May 2024, 6 September 2024, and 6 October 2024; accepted 21 October 2024. This article was recommended for publication by Associate Editor N. Frigerio and Editor X. Xie upon evaluation of the reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under Grant 62203350, Grant 62173268, and Grant 62373297; in part by the Key Program of the National Natural Science Foundation of China under Grant 61833015; and in part by Industrial Field Project—Key Industrial Innovation Chain (Group) of Shaanxi Province under Grant 2022ZDLGY06-02. *(Corresponding author: Qingyu Yang.)*

Donghe Li and Dou An are with the School of Automation Science and Engineering and the MOE Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: lidonghe2020@xjtu.edu.cn; douan2017@xjtu.edu.cn).

Chunlin Hu, Pengtao Song, and Feiye Zhang are with the School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: hucl0918@stu.xjtu.edu.cn; songpengtao@stu.xjtu.edu.cn; zhangfy19970324@stu.xjtu.edu.cn).

Qingyu Yang is with the SKLMSE Laboratory and the School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: yangqingyu@mail.xjtu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TASE.2024.3486342.

Digital Object Identifier 10.1109/TASE.2024.3486342

I. Introduction

A. Motivations

THE swift advancements in intelligent terminals, real-time communication and positioning technologies have significantly enhanced people's daily life patterns [1], [2]. One of the most prominent manifestations of these advancements is in the online resource allocation system [3], which includes

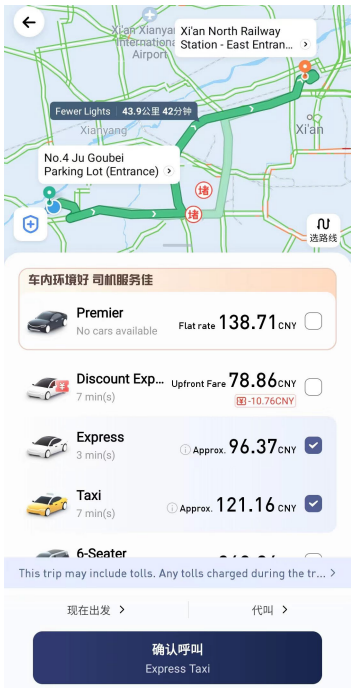


Fig. 1. Taxi reservation App (From DiDi Chuxing).

services such as medical appointments [4], taxi hailing [5], and charging reservations [6], [7]. These applications have revolutionized the traditional "first-come, first-served" service model into a reservation-based bidding service model, leading to a remarkable improvement in service efficiency and customer satisfaction.

In recent times, the online taxi-hailing system serves as an exemplary model of an online reservation system, integrating positioning and communication technologies [8]. This transportation mode has been widely adopted by numerous taxi-hailing companies, including Uber and DiDi Chuxing, extending its reach to over 60 countries worldwide [9]. Fig. 1 shows the operating interfaces of DiDi Chuxing. The system is responsible for assigning the most suitable taxi to passengers. It has become the primary mode of transportation in China [10]. The online taxi-hailing system facilitates the aggregation of data regarding taxis and passengers, with the primary goal of enabling efficient matching strategies. In recent years, researchers have focused on developing real-time matching algorithms that improve matching efficacy and profitability for both vehicle owners and passengers [11], [12].

Concurrently, the growing awareness of environmental protection has prompted a shift towards sustainable modes of transportation [13], [14]. EVs have emerged as a prominent solution in this regard and have gained significant traction in the taxi industry, where they now constitute a majority of the market share [15], [16], [17]. This has led to the transformation of the traditional taxi-hailing system into an E-taxi hailing system. However, the process of recharging EVs is significantly slower and less efficient compared to traditional refueling methods, posing a challenge for the efficient operation of the E-taxi hailing system [18]. If drivers are given the freedom to decide when to charge based on their individual experiences, there is a high probability that they will opt for charging only when the power level is extremely low or when the charging station is in close proximity. This tendency could lead to a surge in the number of E-taxis seeking charging within a brief timeframe, causing undue strain on the charging stations and a shortage of available E-taxis. Conversely, frequent recharging of an E-taxi not only occupies charging resources but also diminishes the driver's income, posing a long-term risk to the battery's safety. Therefore, in addition to passenger and taxi matching, the E-taxi hailing system must also incorporate charging behavior planning to ensure optimal operational efficiency. Currently, numerous studies focus on taxi matching and EV charging planning. However, it is unfortunate that only a few works have considered these aspects simultaneously and conducted preliminary research on their integration. For example, a request-vehicle assignment scheme for taxi matching is presented based on the learning value attained from vehicle routing in [19] but it does not consider the situation of charging. And regarding to E-taxis, an optimal assignment and scheduling approach is designed in [20] but it is only for E-taxis' charging planning problem not E-taxi hailing problem. In addition, many researchers also study the siting and size selection of charging stations from the perspectives of power grid or platform, but they fail to consider the charging planning from the user side and the service side [21], [22], [23]. Strategically planning the charging and picking up behavior of E-taxis not only aids in controlling the number of vacant E-taxis but also enhances drivers' income. Simultaneously, it alleviates pressure on charging stations, aligning with the demand for high-quality transportation. Therefore, in order to fill the gap between charging assignment and order assignment, it is essential to investigate E-taxi dispatch strategies that incorporate a combination of charging and passenger pick-up decisions.

The E-taxi hailing dispatch problem, as mentioned earlier, can indeed be formalized as a dynamic programming problem. However, identifying an effective solution method remains a crucial problem. Traditional optimization methods have achieved satisfactory results when applied to simple environments but perform poorly in large-scale dynamic settings. To address such complex and dynamic optimization decision-making problems, Deep Reinforcement Learning (DRL) has demonstrated impressive results [24]. DRL technology iteratively improves the agent's strategy through continuous interaction between the agent and the environment, enabling the agent to make decisions that maximize rewards in the face of uncertain future environments [25]. The most successful application of DRL is in the recommendation system [26], such as electric vehicle charging recommendation [27], [28] and taxi matching recommendation [29], [30], [31] related to this study. For instance, in [28], a DRL-based method, charging control deep deterministic policy gradient (DDPG), is proposed to learn the optimal charging control strategy. DiDi had also revealed some of the RL algorithms they employ in their taxi-hailing service [29].

It is evident that DRL serves as an effective method for addressing large-scale dynamic optimization problems.

However, when applying DRL to the E-taxi hailing scenario's optimization problem, certain challenges still need to be overcome. First of all, the large and heterogeneous action space poses great difficulties for conventional RL algorithms to make decisions. Specifically, the agent is generally designed to make decisions for the charging behavior of E-taxis and the matching behavior of passengers within the platform. However, there are conflicts in the action selection of taxis themselves, that is, charging cannot pick up passengers. This implies that two mutually exclusive behaviors of the agent, interacting with its environment, will also receive rewards in entirely different forms. Comparing these two rewards using the same standard type is challenging, making it difficult to evaluate both behaviors effectively. There are also certain conflicts in the actions between each taxi, that is, each charging station can accommodate only the specified number of E-taxis simultaneously, and two E-taxis cannot accept the same order from passengers at the same time. Conventional RL algorithms such as Deep-Q Network (DQN), DDPG [32], and PPO [33] cannot constrain these conflicts in the action space, which will bring great challenges to RL training. Currently, there are few works that aim to solve this type of action space, especially when it is constrained by punishment or other methods, the efficiency of the RL algorithm will become very poor. This arises because it influences not only a single behavior of the agent entity but also multiple behaviors of the agent entity over time, which greatly prolongs the learning time of the agent. Additionally, this method does not entirely eliminate the occurrence of similar situations. Secondly, the uncertain state space dimension also poses great challenges for the training of RL. Specifically, the information of passengers in the E-taxi hailing system is input as state information, but the number of passengers is uncertain. Considering that the input dimension of neural networks in RL is fixed, in order to deal with the problem of dynamic changes in the state space dimension, the state space dimension is generally fixed, and padding or pooling methods are adopted. However, this is not a fundamental solution to the problem, and it may lead to a model that is too large or ignore some important state information, which may result in a decrease in the performance of the RL algorithm. In summary, how to solve the conflict action space and dynamic state space is a key issue for RL in the E-taxi hailing system.

B. Contributions

In this paper, we propose a dual-state integrated RL approach for E-taxi management, which effectively addresses the challenges of uncertain state dimensions and heterogeneous action spaces. Our contributions can be summarized as follows:

- We present a novel E-taxi hailing model that seamlessly integrates passenger-taxi matching and charging planning, providing a comprehensive framework for optimizing E-taxi management. Specifically, we first formulate the E-taxi hailing management problem as a multi-objective optimization problem. Subsequently, a Markov Decision Process (MDP) with dynamic state space and heterogeneous action space is employed to model the decision-making process within the E-taxi hailing management system.
- We develop a dual-stage heuristic coordinated reinforcement learning approach for the E-taxi hailing management system addressing the problems of dynamic state and heterogeneous action. In the first stage, we propose a feature-guided state dimension stabilization method, which filters out redundant passengers by scoring their states, thereby fixing the input to the state dimension of the RL network. In the second stage, we introduce a heuristic coordinated assignment method to further refine the actions generated in the first stage and match E-taxis with passengers and charging stations. The reward generated by the final action serves as the reward for the RL network in the first stage and is used in the network update process.
- We evaluate our proposed dual-stage heuristic coordinated RL approach using a real-world E-taxi hailing environment. Results show a 20% increase in E-taxi profits per 10 hours. Passenger order acceptance ratio improves by 15%, and waiting time decreases by 55%. Our method outperforms existing techniques across multiple performance indicators, while ablation experiments and a time cost analysis demonstrate its effectiveness in addressing dynamic state dimensions and heterogeneous actions with consistent time consumption.

The organization of this paper is as follows: First, the related work is given in Section II. And the models of the E-taxi hailing system are described in Section III. In Section IV, our proposed dual-stage heuristic coordinated reinforcement learning approach is described. In Section V, the performance evaluations are given. Finally, in Section VI, we conclude and discuss this paper.

II. Related Work

Extensive research has been conducted on devising optimal scheduling strategies for EV charging and taxi matching [34], [35], [36], [37], [38]. For example, Shen et al. [34] proposed a two-stage integrated scheduling strategy to determine the optimal charging load profile for various EVs. Park et al. [35] introduced an EV scheduling algorithm employing fuzzy logic control in a smart charging network to enhance charging performance. Considering customer engagement and satisfaction, Ma [36] proposed a multi-objective optimal approach for scheduling large-scale electric vehicles based on customer behavior prediction. Ding et al. [38] proposed a hierarchical and cooperative macroscopic and microscopic dynamic dispatching approach for real-time urban network taxis in a connected taxi information environment. Abid et al. [39] modeled the taxi dispatch system as a multicriteria decision-making problem, incorporating user preferences in finalizing a taxi for a given passenger travel request. The performance results demonstrate that the method reduces passenger complaints. These methods exhibit good performance in relatively stable environments, but their effectiveness diminishes when the environment undergoes rapid state changes.

In recent years, DRL has garnered increasing interest for addressing EV charging and taxi matching scheduling problems [40], [41], [42], [43]. Wang et al. [40], for example, formulated the ride dispatching problem as a MDP and employed DQN with action search to optimize the dispatching policy for drivers on ride-sharing platforms. Zhao et al. [41] proposed an RL-based algorithm to optimize operation strategies for different types of EVs. Silva et al. [42] developed an intelligent E-taxi ride-hailing service controller that maximizes passenger satisfaction while ensuring reliable charging for each E-taxi. Haliem et al. [43] presented a dynamic, demand-aware, and pricing-based vehicle-passenger matching and route planning framework using DRL that generates optimal routes for each vehicle and predicts demand allocation based on online demand. For DRL, it is important to adopt strategies to avoid unnecessary trials. Du et al. [44] designed a DDPG with external knowledge algorithm to improve ride comfort, significantly increasing computational efficiency. Under a multi-agent DRL framework, Zhang et al. [45] designed a modified exploration strategy to direct agent training and avoid unnecessary trials. These research efforts have contributed to the development of effective and efficient scheduling strategies for taxis. However, the aforementioned works confirm the effectiveness of traditional optimization-based methods and DRL-based intelligent methods for single tasks but do not combine multi assignments, such as EV charging and taxi-matching tasks.

In fact, some researchers have recognized the demand for E-taxis in charging and matching, and have explored methods that can simultaneously study charging behavior and taxi matching. For instance, Lin et al. [46] proposed a management architecture combining the charging network and the car-hailing operating network of E-taxis while considering both passenger pickup and charging processes. The results indicate that the architecture effectively lowers E-taxi charging costs by reducing charging queue times at charging stations. Wang et al. [47] approached the scheduling problem from the perspective of E-taxi drivers and formulated their decision-making as a multi-agent reinforcement learning (MARL) problem, proposing a novel multi-agent mean field hierarchical RL framework to provide charging and relocation recommendations for E-taxi drivers. However, the aforementioned literature oversimplifies the process of picking up passengers by treating it as a mere subprocess of the charging action, which deviates significantly from real-life scenarios. Simultaneously, the RL networks presented in these studies fail to address the heterogeneous actions and dynamic states inherent in the E-taxi hailing problem, as proposed in the introduction of this paper.

In summary, efficient charging and matching algorithms within the E-taxi hailing system are of paramount importance. Although some previous studies have tackled individual challenges of charging or matching, few have addressed the coordination of these tasks, resulting in limitations in computation speed and matching accuracy. To bridge this gap, this paper proposes to investigate novel taxi recommendation approaches for the e-taxi hailing system, focusing on the integration of both charging and passenger matching tasks.

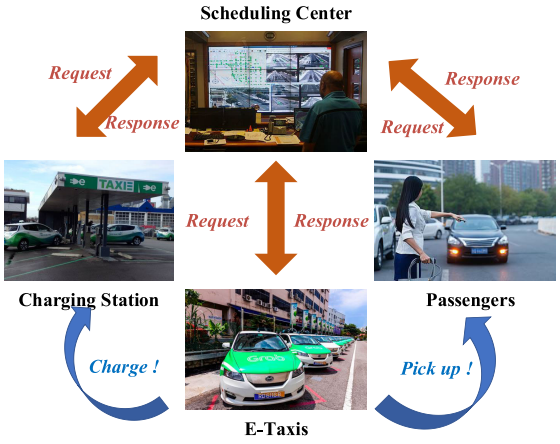


Fig. 2. System model.

This approach aims to enhance the overall performance of the system while maintaining high levels of customer satisfaction and operational efficiency.

III. Models in E-Taxi Hailing System

In this section, we first present the system models of the E-taxi hailing system, followed by an introduction to the key notations. Subsequently, we formalize the E-taxi scheduling problem and provide a detailed explanation of the optimization objective utilized in this paper.

A. System Model

Fig. 2 illustrates the system model proposed in our paper, which consists of four key components: Scheduling Center (SC), E-taxis, passengers/orders (One order may include many passengers), and charging stations. Within this E-taxi hailing framework, the SC, managed by service providers like DiDi and Uber, handles requests from both passengers and E-taxis and makes decisions utilizing the dual-stage heuristic coordinated reinforcement learning approach introduced in our paper. E-taxis act as service providers, responding to passenger requests and submitting information about their current status and charging requirements to the SC. Passengers, on the other hand, act as service consumers who request E-taxi rides and provide their preferences and pick-up locations. Charging stations are the infrastructure component of the system, providing charging services to E-taxis as needed. They communicate with the SC to share information about their availability and capacity, enabling the center to make informed decisions on charging allocations. Overall, the proposed system model effectively captures the interactions between the main components of an E-taxi hailing system, providing a foundation for the development and evaluation of the dual-stage heuristic coordinated reinforcement learning approach presented in this paper.

B. Preliminaries

Then we will introduce the notations in our paper, which can be concluded in Table I.

TABLE I
NOTATIONS

Symbols	Descriptions
T	Time slot set
V, P, CS	E-taxis, orders, charging stations sets
$a1^t_{i,j}$	Matching state between E-taxi i and order j
$a2^t_{i,k}$	Matching state between E-taxi i and charging station k
(xv^t_i, yv^t_i)	Location coordinates of E-taxi i
(xps^t_j, yps^t_j)	Starting location coordinates of order j
(xpd^t_j, ypd^t_j)	Destination location coordinates of order j
(xcs_k, ycs_k)	Location coordinates of charging station k
D_j	Passengers travelling distance
Q_j	Quoted price of order j
$DVP^t_{i,j}$	Distance between the E-taxi i and order j
$DVCS^t_{i,k}$	Distance between the E-taxi i and charging station k
$TVP^t_{i,j}$	Passenger waiting time
$TVCS^t_{i,k}$	Waiting time in charging station k
E^t_i	Remaining battery energy of the E-taxi i.
E_m	Max battery energy of an E-taxi
E_n	Punishment battery energy of an E-taxi
N^t_k	Number of waiting E-taxis in charging station k
C_{in}	Charging speed of charging stations

The time horizon of each day is divided into T time slots. Three integer sets, V, P, and CS, are used to represent E-taxis, passengers, and charging stations, respectively, while m, n, and q represent the number of E-taxis, orders from passengers, and charging stations, respectively. To distinguish between taxis, passengers, and charging stations, subscripts i, j, and k are used to represent their respective numbers. $a1^t_{i,j}$ and $a2^t_{i,k}$ indicate the matching status at time slot t between E-taxi i and order j, and E-taxi i and charging station k, respectively. (xv^t_i, yv^t_i) represents the location coordinates of E-taxi i at time slot t, and (xcs_k, ycs_k) represents the location coordinates of charging station k. (xps^t_j, yps^t_j) and (xpd^t_j, ypd^t_j) represent the location coordinates for the starting and ending points of order j at time slot t, the total travelling distance of order j is expressed as D_j, and Q_j denotes the quoted price of order j's order. Further, $DVP^t_{i,j}$ denotes the distance between E-taxi i and order j, while $DVCS^t_{i,k}$ denotes the distance between E-taxi i and charging station k. $TVP^t_{i,j}$ denotes the waiting time of order j after being accepted by E-taxi i, while $TVCS^t_{i,k}$ denotes the waiting time of E-taxi after arriving at charging station k. E^t_i denotes the remaining battery energy of E-taxi i, while E_m and E_n indicate the max battery energy and punishment battery energy of an E-taxi. N^t_k represents the number of waiting taxis in charging station k. C_{in} denotes the charging speed of charging stations.

C. Optimization Formalization

In previous research, the picking up behavior of E-taxis was commonly influenced by factors like passengers' waiting time, and travelling distance [19]. Charging behavior, often correlated with factors such as SoC, distance to charging stations, and the operational status of charging stations [34]. Influenced by the above factors, we introduce two indicators to evaluate the quality of charging and picking up, which are referred to Charging Trend (CT) and Picking up Trend (PT) in this paper. CT reflects a driver's willingness to go to a charging station, primarily influenced by SoC, distance to the charging station, and waiting time at the charging station. Similarly, PT reflects a willingness to accept an order at the current moment, primarily determined by the profit and distance to passengers. Therefore, the E-taxi hailing management problem can be formalized as solving the CT and PT maximization problem.

1) Charging Trend (CT): When the E-taxi i charges at charging station k, the CT is mainly influenced by three indicators: remaining battery energy E^t_i, charging distance $DVCS^t_{i,k}$, and waiting time $TVCS^t_{i,k}$.

Specifically, remaining battery energy E^t_i indicates the E-taxi i's state of charge (SoC). E-taxis are expected to be able to charge at a lower SoC, which would give them more pickup revenue.

Charging distance $DVCS^t_{i,k}$ represents the travelled distance between E-taxi i and charging station k. Considering the city roads are criss-crossed, the absolute distance more approximates the actual journey than Euclidean distance, which could be calculated as follows:

$$DVCS^t_{i,k} = \left|xv^t_i - xcs_k\right| + \left|yv^t_i - ycs_k\right| \tag{1}$$

$TVCS^t_{i,k}$ represents the waiting time when E-taxi i charges at charging station k, which can be calculated as:

$$TVCS^t_{i,k} = \frac{\lambda N^t_k}{C_{in}} \tag{2}$$

where λ represents the average charging volume.

To obtain better optimization performance, the above three indicators need to be normalized. After that, the Charging Trend $CT^t_{i,k}$ when E-taxi i charge in charging station k at time slot t can be expressed as:

$$CT^t_{i,k} = -(\mu_1 E^t_i + \mu_2 DVCS^t_{i,k} + \mu_3 T^t_{i,k}) \tag{3}$$

where μ_1, μ_2, μ_3 represent the weights.

As introduced before, obviously, E-taxis prefer lower remaining battery, lower charging distance, and lower waiting time. Therefore, a larger $CT^t_{i,k}$ indicates that the E-taxi i is more satisfactory when charging in charging station k at time slot t.

2) Picking Up Trend (PT): When the E-taxi i picks up passengers j, the PT is mainly influenced by two indicators: quoted price Q_j and picking distance $DVP^t_{i,j}$.

Quoted price Q_j represents the total revenue from order j's trip. Note that it relates only to passengers j, not to the taxi that picks up it. In reality, Q_j is a piecewise function about travelling distance D_j of passengers j, which can be expressed as:

$$Q_j = \begin{cases} p_b, & D_j \le D_{min} \\ p_b + p \cdot (D_j - D_{min}), & D_j > D_{min} \end{cases} \tag{4}$$

where D_j can be calculated by the starting and ending locations (xps^t_j, yps^t_j), (xpd^t_j, ypd^t_j). p_b represents the starting price, D_{min} denotes the starting distance, and p denotes the price per unit distance.

Picking distance $DVP^t_{i,j}$ refers to the travelled distance between E-taxi i and order j, which can be calculated as:

$$DVP^t_{i,j} = \left|xv^t_i - xps^t_j\right| + \left|yv^t_i - yps^t_j\right| \tag{5}$$

Also after normalization, the Picking up Trend $PT^t_{i,j}$ when E-taxi i picks up passengers j can be expressed as:

$$PT^t_{i,j} = \mu_4 Q_j - \mu_5 DVP^t_{i,j} \tag{6}$$

where μ_4, μ_5 represent the weights.

Similarly, E-taxis prefer lower picking distance and higher quoted price. Therefore, a larger $PT^t_{i,j}$ indicates that E-taxi i is more satisfactory when picking up passengers j.

3) Optimization Model: As introduced above, E-taxis prefer higher CT and PT, therefore the E-taxi management problem can be constructed as the following multi-objective optimization problem:

$$\max_{a1,a2} \sum_{t=1}^{T} [\sum_{i\in V}\sum_{k\in CS} a1^t_{i,k} CT^t_{i,k} + \sum_{i\in V}\sum_{j\in P} a2^t_{i,j} PT^t_{i,j}] \tag{7}$$

$$\text{subject to} \quad a1^t_{i,k},\ a2^t_{i,j} = \{0,1\},\ \forall i \in V,\ j \in P,\ k \in CS \tag{8}$$

$$\sum_{k\in CS} a1^t_{i,k} + \sum_{j\in P} a2^t_{i,j} \le 1, \forall i \in V, \quad \forall t \in T \tag{9}$$

$$\sum_{i\in V} a2^t_{i,j} \le 1, \forall j \in P, \quad \forall t \in T \tag{10}$$

where $a1^t_{i,k}$, $a2^t_{i,j}$ are the optimization parameters, and represent the matching status at time slot t between E-taxi i and charging station k, E-taxi i and order j respectively. i, j, k represent the ID of E-taxi, order, and charging station respectively.

The first and second terms of the objective function (Equation (7)) represent the Charging and Picking up Trends, respectively, of all E-taxis in the taxi-hailing system across the total T time slots. The first constraint (Equation (8)) denotes that the optimization parameter value is either 0 or 1, where 0 indicates no match and 1 indicates successful match. The second constraint (Equation (9)) stipulates that E-taxi i can choose at most one action (either picking up passengers or charging) at a given time t. The third constraint (Equation (10) specifies that order j can only be matched with one E-taxi at time slot t. Just to mention, a charging station could be matched with not only one E-taxi, but the extra E-taxis must wait for charging.

IV. Dual-Stage Heuristic Coordinated Reinforcement Learning Approach for E-Taxi Management

In this section, we will introduce the proposed dual-stage heuristic coordinated reinforcement learning approach.

A. Markov Decision Process With Dynamic State and Heterogeneous Action Space

Regarding to the above optimization problem, RL is more suitable due to the dynamic and uncertain nature of the taxi-hailing environment comparing with solving it directly. In real-world scenarios, the state of the system continuously changes as passengers request rides, E-taxis become available or occupied, and charging stations experience varying levels of utilization. Moreover, the preferences of passengers and the operational decisions of E-taxis are also subject to change over time. Traditional optimization techniques often struggle to cope with such highly dynamic and uncertain systems, as they rely on fixed parameters and assumptions. In contrast, RL can adapt to these changes in real-time by learning from interactions with the environment and continuously updating its decision-making strategy. This adaptive behavior enables RL to better capture the complex, evolving relationships between the various components in the E-taxi hailing system, ultimately leading to more effective and efficient management solutions.

In the E-taxi hailing system, the MDP is employed to model the decision-making process and capture the intricate interactions between various components of the system, including passengers, E-taxis, and charging stations. Specifically, the SC acts as an agent responsible for making decisions on behalf of the E-taxis. The agent first observes the environment as the MDP state, which includes passenger information (numbers, location, destination, etc.), charging station information (location, occupancy, charging speed, etc.), and E-taxi information (location, status, etc.). The number of E-taxis and charging stations is fixed, leading to a determined number of states for them. While the number of order from passengers in the environment at each moment cannot be predetermined, resulting in an uncertain state space dimension for passengers and a dynamic state space overall.

The agent then provides actions for the E-taxis, determining whether they should charge or pick up passengers. Subsequently, a reward is calculated to evaluate the quality of the action. The two actions, charging and picking up passengers, are distinct and conflicting types of actions with entirely different reward calculation methods and state transition methods, constituting a heterogeneous action space. Finally, as the environment changes, the agent transitions to the next state. The agent repeats this process, ultimately obtaining the optimal policy. Thus, the E-taxi hailing management system can be formulated as an MDP with dynamic state and heterogeneous action space.

B. Design Rational and Overview

As previously discussed, we have formulated the E-taxi hailing management system as an MDP to address the limitations of traditional optimization algorithms when dealing with dynamic environments. However, the dynamic state space and heterogeneous action space pose challenges for the application of existing RL techniques. Conventional RL algorithms, such as Q-learning, SARSA, and Actor-Critic, are well-suited for small-scale systems with well-defined state and action spaces but struggle to handle uncertain state dimensions and heterogeneous action spaces. PPO is an on-policy RL method, which performs well in solving dynamic and complex problems. Compared to value-based networks like DQN, the clipping mechanism utilized by the PPO method allows for the reuse of sample data multiple times, resulting in a higher utilization of

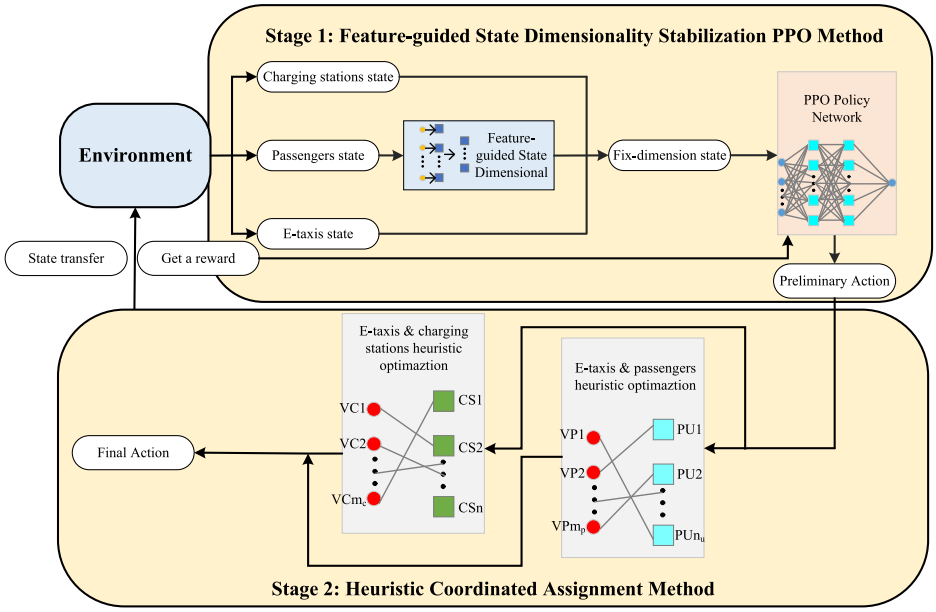


Fig. 3. Overall algorithm framework.

playback samples than the empirical pools used by DQN. PPO exhibits broader applicability compared to other Actor-Critic networks such as DDPG, as it can handle not only continuous problems but also discrete types of problems. Moreover, PPO employs the clipping mechanism to directly optimize the policy while constraining the magnitude of policy updates, thereby eliminating the need for designing and training the value function, which simplifies the operation and improves stability of system. The E-taxi scheduling problem addressed in this paper is a discrete dynamic decision-making problem, challenging to accurately design its value function due to the complexity and variability of the environment. However, the discrete PPO method can directly optimize the charging quantity strategy of E-taxis and subsequently optimize their charging behavior and pick-up behavior. This approach is theoretically feasible and efficient. Similar to other RL methods, PPO fails to train on dimensionally varying data. For instance, an E-taxi at a charging station cannot simultaneously pick up passengers, resulting in action selection conflicts for the taxis themselves. Moreover, two taxis cannot charge at the same charging station in the meantime or pick up the same passengers simultaneously, leading to conflicts between the actions of different taxis. Traditional PPO algorithms cannot efficiently resolve these conflicts, resulting in suboptimal resource utilization and extended passenger wait times.

To address these challenges, we propose a dual-stage heuristic coordinated reinforcement learning approach. The overall algorithm framework is depicted in Fig. 3. In general, we construct a DRL algorithm based on PPO, which is well-suited for complex and dynamic environments. We then enhance the standard PPO mechanism with a dual-stage decision process to address the challenges posed by dynamic state space and heterogeneous action space. Specifically, in the first stage, we introduce a feature-guided state dimensionality stabilization method that filters passenger feature scores to fix the input state dimension for the RL network. This approach addresses the uncertainty stemming from the variable number of passengers. Upon receiving the state, the agent's preliminary action only decides the number of E-taxis will charge. Based on that, the agent could obtain whether each E-taxi should charge or pick up passengers and send the results to the second stage. In the second stage, we propose a heuristic coordinated assignment method to help the agent refine the actions generated by the PPO network in the first stage. We formulate the matching problem between E-taxis and passengers, as well as between E-taxis and charging stations, as two static optimization problems, and the final actions are obtained through a heuristic algorithm. Ultimately, the PPO reward is based on the actions derived from the heuristic algorithm. As the action obtained by PPO can acquire the optimal action through the heuristic algorithm, we believe that the reward generated by the final action can judge the action directly output by PPO. Consequently, the second stage addresses the challenge of RL algorithms not directly generating heterogeneous actions.

In comparison to traditional RL methods, our approach not only broadens the application scope of RL by stabilizing the dimensionality of the environment state but also integrates heuristic algorithms to enhance the optimization of individual E-taxi strategies, thereby accelerating training speed and reducing training costs. Compared to the two-stage RL-RL method, this approach significantly reduces training costs and avoids issues such as oscillation and non-convergence. Contrasted with direct heuristic method, our approach simplifies the model construction process significantly by incorporating PPO method, leading to a considerable reduction in the final solution space and hastening computation. In the first stage, the feature-guided state dimensionality stabilization method

efficiently identifies high-value orders for drivers, contributing to increased drivers' profit and reduced passengers' waiting time. Subsequently, based on the RL-derived suggestion for the number of charging E-taxis, the system classifies E-taxis, enabling the selection of appropriate behaviors for each E-taxi to coordinate traffic. In the second stage, the application of the heuristic optimization method facilitates the pairing of E-taxis with suitable charging stations or orders, thereby enhancing the utilization of charging stations and further improving driver income. Therefore, the approach presented in this paper showcases remarkable efficiency and effectiveness from a theoretical perspective.

C. Stage 1: Feature-Guided State Dimensionality Stabilization PPO Method

1) State: The overall environment state at time slot t includes three parts: E-taxi state, passenger state and charging station state, which can be expressed as:

$$s_t = \left\{SV_1^t, \ldots, SV_m^t, SP_1^t, \ldots, SP_n^t, SCS_1^t, \ldots, SCS_q^t\right\} \quad (11)$$

where SV_i^t represents the state of E-taxi i, including the location coordinates (xv_i^t, yv_i^t), remaining battery energy E_i^t, and statuses KV_i^t and $KTAR_i^t$, thus its expression is formulated as $\left\{xv_i^t, yv_i^t, E_i^t, KV_i^t, KTAR_i^t\right\}$. SP_j^t represents the state of order j, including the starting and ending location coordinates (xps_j^t, yps_j^t), (xpd_j^t, ypd_j^t), and status KP_j^t, thus its formalization is designed as $\left\{xps_j^t, yps_j^t, xpd_j^t, ypd_j^t, KP_j^t\right\}$. SCS_k^t represents the state of charging station k, which is designed as $\left\{KCS_{k1}^t, KCS_{k2}^t, \ldots, KCS_{kg_k}^t\right\}$. g_k denotes the number of E-taxis that can be serviced simultaneously by charging station k.

Next, we introduce the concept of statuses. As time progresses to the next time slot, some E-taxis may haven't completed their previous assignments, which can include not arriving at the destinations of their orders and having unsatisfactory remaining battery energy. To address this issue, four statuses are introduced. The value of KV_i^t is contained within the set $\{-1, 0, 1, 2\}$. -1 denotes that the E-taxi is charging at a charging station. 0 indicates that the E-taxi is idle. If the E-taxi accepts an order, the process of picking up passengers can be divided into two parts. The first part represents the period from accepting an order to arriving at the order's starting location, and the second part represents the period from the starting location to the order's destination. 1 and 2 denote that the E-taxi is in the first and second processes after accepting an order, respectively. To record the matching relationship, $KTAR_i^t$ is introduced to store order's index associated with KV_i^t. The value of $KTAR_i^t$ denotes the target of E-taxi i and is contained within the set $\{0, 1, 2, \cdots, q)\}$. For example, if $KV_i^t = 1$ and $KTAR_i^t = 1$, it represents that E-taxi i is in the first process of picking up the first order. The value of KP_j^t is contained in the set $\{0, 1\}$, where 0 denotes that the order is unaccepted and 1 denotes that the order has been accepted. This prevents two E-taxis from accepting the same order. The value of KCS_k^t is contained in the set $\{0, 1, \cdots, m\}$, indicating the index of the E-taxi charging at charging station k. This prevents too many E-taxis for charging at charging station k in the meantime, with 0 denoting that no E-taxi is charging.

The status changes of E-taxis, orders, and charging stations during interaction with the environment proceed as follows: at any given moment, for E-taxis, only the E-taxis with $KV_i^t = 0$ can take an action. On the one hand, when an E-taxi is matched with a charging station and successfully enters the charging status, its status KV_i^t transitions from 0 to -1. Upon completion of the charging process, KV_i^t changes back from -1 to 0, during which the E-taxi cannot take any action. On the other hand, when an E-taxi i is matched with an order j, its status KV_i^t changes from 0 to 1, and $KTAR_i^t$ transitions from 0 to j at once. As the E-taxi arrives at the passengers' location and prepares to travel to the destination, KV_i^t changes from 1 to 2. Once the E-taxi successfully delivers the passenger to the destination, its status KV_i^t changes from 2 back to 0, and $KTAR_i^t$ transitions from j to 0, allowing it to undertake a new task as it interacts with the environment. For orders and passengers, only orders with a status of $KP_j^t = 0$ can be accepted by an E-taxi. Newly generated orders and those not previously accepted by an E-taxi have a status of $KP_j^t = 0$. Once an order is accepted, the order's status KP_j^t changes from 0 to 1, preventing it from being matched with other E-taxis. The order remains in the environment state until the E-taxi successfully delivers its passengers to the destination, at which point the order is deleted immediately. For charging stations, only charging stations with elements in SCS_k^t that are not all 0 can provide charging services for E-taxis. For instance, if KCS_{k1}^t and KCS_{k2}^t are both 0 at time step t, and there are two E-taxis i_1 and i_2 (assuming E-taxi i_1 arrives first) matched with charging station k, then KCS_{k1}^t changes to i_1 immediately after E-taxi i_1 arrives, and similarly, KCS_{k2}^t changes to i_2 after E-taxi i_2 reaches. Until E-taxi i_1 finishes charging and leaves, KCS_{k1}^t changes from i_1 to 0. The same occurs when E-taxi i_2 leaves, and KCS_{k2}^t changes from i_2 to 0. When elements in SCS_k^t are all 0, although charging station k could be matched with E-taxis, it cannot provide immediate charging service to the E-taxis because its charging ports are all occupied.

2) Feature-Guided State Dimensionality Stabilization: In our model, the number of new orders generated from the last time step to current time step is uncertain, which results in a variable state dimension and heterogeneous action spaces. To maintain a fixed state input dimension for the RL network, we have designed a feature-guided state dimensionality stabilization method. In our method, the order data must undergo a pretreatment before utilizing by the PPO algorithm in the first stage. When the number of all orders is more than m, the SC will score all new orders and add a certain number of order data with high scores to environment state. And when the number of all orders is less than m, the SC will also add some meaningless order data to environment state to maintain state dimension, which does not match with E-taxis in the second stage. The detailed operation of feature-guided state dimensionality stabilization method consists of the following two steps:

Step 1. AT of Orders. In this paper, we employ a filter method to handle the original data. The principle of the filter

method is to evaluate certain features and retain those with high scores. AT is considered as an evaluation factor in this paper. First, a sufficiently large set is defined to store the data information of all new orders, which can be represented as $PN = \{PN_1, PN_2, \cdots, PN_{n_p^t}\}$, where n_p^t denotes the number of new orders. Considering the inherent randomness and uncertainty associated with the number of new orders in real-world scenarios, this paper stipulates that the generation of new orders adheres to a random distribution. Specifically, n_p^t follows

$$n_p^t \sim U(NL_p^t, NH_p^t) \tag{12}$$

where NL_p^t, NH_p^t indicate the lower and upper bounds of the uniform distribution, respectively. Orders are often accepted by near available E-taxis, which helps reduce waiting time of passengers and improve efficiency of E-taxis. Therefore, we consider the total distance from the nearest m_0 available E-taxis except the quoted price. Considering that there may not be m_0 spare E-taxis, whose number is m_s^t at some times, the number m_0^t of spare E-taxis chosen is as follows:

$$m_0^t = \min\{m_0, m_s^t\} \tag{13}$$

A large quoted price and a short total distance both contribute to the attractiveness of an original order. The set of the nearest m_0^t available E-taxis from the order PN_j can be represented as $VN_j = \left\{V_j^1, V_j^2, \cdots, V_j^{m_0^t}\right\}$. The AT of the order PN_j is defined as:

$$AT_j = \mu_4 m_0^t \cdot Q_j - \mu_5 \cdot \sum_{i=1}^{m_0^t} DVPN_{i,j}^t \tag{14}$$

where $DVPN_{i,j}^t$ represents the distance between E-taxi V_j^i and order PN_j, which can be calculated by Equation (5).

Step 2. Choosing AT. It is assumed that there are n_0^t uncompleted orders from the last time step at step t, then the number Δn of new orders needed by environment can be formulated as:

$$\Delta n = m - n_0^t \tag{15}$$

Considering the number of new orders is random which may be less than Δn, the number n_t^t of new orders can be transferred from set PN to P, is formulated as:

$$n_t^t = \min\{\Delta n, n_p^t\} \tag{16}$$

If $n_t^t = \Delta n$, the SC will choose the largest Δn values from all AT_j and extracts the data of corresponding orders into environment state. If $n_t^t = n_p^t$, the SC will extract data of all new orders and add meaningless information into environment state to keep dimensions. Finally, the orders which is taken as passenger state can be divided into three parts: old orders from the last time step, new generated orders, and meaningless orders. The optimization objective of feature-guided state dimensionality stabilization can be formulated as follows:

$$\max_{a_j}(\sum_{j=1}^{n_p^t} a_j AT_j) \tag{17}$$

Algorithm 1 Feature-Guided State Dimensionality Stabilization

Input: old order set PO, new order set PN, E-taxi set V
Output: order set P

```
a, b=len(PO),len(PN);
Initialize the set ATV;
if b ≤ m − a then
    P = PO + PN;
    Add meaningless data to P
else
    for j = 0 to b do
        Obtain a set VN_j;
        Calculate AT_j using Equation (14);
        Record AT_j in ATV;
    end
    for Index = 0 to (m − a) do
        Obtain the maximum AT_j in ATV;
        Obtain the order PN_j corresponding to AT_j;
        Add PN_j to P;
        Delete AT_j in ATV;
    end

end
return The order set P
```

subject to:

$$a_j = \{0, 1\} \tag{18}$$

$$\sum_{j=1}^{n} a_j = n_t^t \tag{19}$$

where a_j is the optimization parameter with $j \in PN$ and indicates whether the order PN_j passes pretreatment. The objective function (Equation (17)) represents the total AT of orders passing pretreatment. The first constraint (Equation (18)) denotes the value of a_j is either 0 or 1, where 0 indicates not pass and 1 indicates pass. The second constraint (Equation (19)) denotes the number of orders transported from PN must be n_t^t. The algorithm process of feature-guided state dimensionality stabilization is shown in Algorithm 1.

3) Preliminary Action: After pretreatment, the SC takes the best action according to the environment state at each time slot.

$a_t \in A$ denotes the action that the agent takes at time step t. In the first stage, we treat the SC as the agent and model the decision-making process of selecting the number of E-taxis to be charged as a MDP. Therefore, the action a_t represents the number of E-taxis that will charge in the time slot advised by SC. Since the number of E-taxis is m, the action space A in this paper is designed as:

$$A = \{0, 1, 2, \cdots, m\} \tag{20}$$

4) E-Taxis to Charge: After obtaining an action a_t, SC needs to decide which E-taxis should charge or pick up passengers. The filter method is utilized again to choose a_t E-taxis to charge. Considering that E-taxi drivers prefer to charge their vehicles in a near charging station, the total CT

between an spare E-taxi and its nearest n_c charging stations is considered as the evaluation factor to choose E-taxis. The total CT_i of spare E-taxi i can be formulated as

$$CT_i = \sum_{k=1}^{n_c} CT^t_{i,i_k} \tag{21}$$

where i_k denotes the index of a charging station, which is one of charging stations nearest to the E-taxi i. Then SC chooses largest a_t values from all CT_i and arranges them from the largest value. The corresponding E-taxis will be advised to charge in the second stage. To express clearly, all spare E-taxis are classified into two categories: the E-taxis to match with charging stations which are set as VC, and the E-taxis to match with passengers which are set as VP. m^t_c and m^t_p represent the number of E-taxis in VC and VP in time step t, respectively.

Algorithm 2 Matching Strategy Between E-Taxis and Charging Stations

Input: E-taxi set V_c, charging station set CS

```
c = len(CS);
for V_{c_i} ∈ V_c do
    Initialize the set CT_v;
    for k = 0 to c do
        Calculate CT^t_{c_i,k} using Equation (3);
        Add CT^t_{c_i,k} to CT_v;
    end
    Obtain the maximum CT^t_{c_i,k} in CT_v;
    Match the c_ith with the kth CS;
    Update assignment state of the CS k;
end
return match result
```

D. Stage 2: Heuristic Coordinated Assignment Method

In this paper, the best matching strategy is hard to obtain directly using solvers or existing assignment algorithms due to the dynamic and comprehensive environment. Fortunately, the heuristic optimization method can generate an optimized matching strategy through adaptive spatial search and comparison, while also permitting some deviation from the optimal solution. Compared to solvers and existing assignment algorithms, the heuristic optimization method does not ensure the discovery of the globally optimal solution. However, it typically yields a relatively effective solution within a short timeframe, which can achieve the benefits of simple design and computational efficiency. The features designed are CT and PT mentioned in Section III. Correspondingly, there are two main optimization objectives in this section: one is maximizing the total CT, and the other is maximizing the total PT. The specific optimization matching strategy is as follows:

1) Optimization Assignment Between E-Taxi and Charging Stations: The first optimization strategy is between the E-taxis VC and charging stations. Considering E-taxis can wait in a busy charging station, we match the charging station from the largest CT_i for each E-taxi in VC in turn. First, the E-taxi with the largest CT_i is matched with each charging station and all CT can be obtained. If $CT^t_{i,k}$ is the largest CT, the SC will match the E-taxi i with charging station k. At the same time, N^t_k increases by 1 with the matching times of the charging station k. Then the SC will match the E-taxi with the second largest CT_i like the process above, which continues until all E-taxis in VC are matched. The total matching time is m^t_c. The first optimization objective can be formulated as:

$$\max_{a_{i,k}} \Big(\sum_{i \in VC} \sum_{k=1}^{q} a_{i,k} CT^t_{i,k} \Big) \tag{22}$$

$$\text{subject to:} \quad a_{i,k} = \{0, 1\} \tag{23}$$

$$\sum_{k=1}^{q} a_{i,k} = 1 \tag{24}$$

where $a_{i,k}$ is the optimization parameters with $i \in VC, k \in CS$ and represents whether an E-taxi is matched with a charging station. The objective function (Equation (22)) represents the total CT of E-taxis in VC. The first constraint (Equation (23)) denotes $a_{i,k}$ is either 0 or 1, where 0 denotes no match and 1 denotes successful match. The second constraint (Equation (24)) denotes an E-taxi could be only matched with one charging station. The algorithm process of matching strategy between E-taxis and charging stations is shown in Algorithm 2.

2) Optimization Assignment Between E-Taxis and Passengers: There is an obvious difference between two optimizations: a charging station can be matched with multi E-taxis while an order can't. So not each E-taxi or order can be matched if the other is less. First, all meaningful and unaccepted orders in P are extracted and set as PU, where n^t_u is introduced to represent the number of orders in PU. To maximize the total profit of drivers and minimize the total waiting time of passengers as much as possible, the matching begins from the order with the largest AT_j. Next, the order with the largest AT_j is matched with each E-taxi in VP and all PT can be obtained. If $PT^t_{i,j}$ is the largest PT, the SC will match E-taxi i with order j. Then the SC will match the order with the second largest AT_j like the process above, which continues until all E-taxis in VC or all orders in P_u are matched. All matching times M_{vp} can be formulated as:

$$M_{vp} = \min\{m^t_p, n^t_u\} \tag{25}$$

The second optimization objective can be expressed as

$$\max_{a_{i,j}} \Big(\sum_{j=1}^{M_{vp}} \sum_{i \in VP} a_{i,j} PT^t_{i,j} \Big) \tag{26}$$

$$\text{subject to:} \quad a_{i,j} = \{0, 1\} \tag{27}$$

$$\sum_{i=1}^{m^t_p} a_{i,j} = 1 \tag{28}$$

$$\sum_{j=1}^{n^t_u} a_{i,j} \le 1 \tag{29}$$

where $a_{i,j}$ is the optimization parameter with $i \in VP$, $j \in PU$ and indicates whether an E-taxi is matched with an order. The objective function (Equation (26)) represents the total PT of E-taxis in VP. The first constraint (Equation (27)) denotes $a_{i,j}$

Algorithm 3 Matching Strategy Between E-Taxis and Orders

Input: E-taxi set V_p, order set P_u
a, b = len(V_p),len(P_u);
$l = \min(a, b)$;
for $j = 1$ *to* l **do**
  Initialize the set PT_v;
  for $V_{p_i} \in V_p$ **do**
    Calculate $PT^t_{p_i,j}$ using Equation (6);
    Add $PT^t_{p_i,j}$ to PT_v;
  end
  Match the p_ith with the jth order;
  Update assignment state of the p_ith E-taxi;
  Update assignment state of the jth order;
end
return match result

is either 0 or 1, where 0 indicates no match and 1 indicates successful match. The second constraint and third constraints (Equation (28) and (29)) denote an order in PU can be only matched with one E-taxi and an E-taxi in VP can be only matched with one order at most. The algorithm process of matching strategy between E-taxis and orders is shown in Algorithm 3.

3) Final Action: The behaviors of all E-taxis are clear after two optimization assignments by utilizing the heuristic method. a'_t is defined to represent the actions of all E-taxis, which is a set with m dimensions. The transfer process from the action a_t of the SC to the actions a'_t of E-taxis can be formulated as:

$$a_t \rightarrow a'_t \tag{30}$$

where the formulation of a'_t is designed as follows:

$$a'_t = \left\{a^1_t, a^2_t, \cdots, a^m_t\right\} \tag{31}$$

where a^i_t denotes the action of E-taxi i at the time step t, determining the charging station or order with which taxi i will be matched. The constriction of a^i_t is as follows:

$$a^i_t \in \{-q, \cdots, -2, -1, 0, 1, 2, \cdots, n\} \tag{32}$$

It's assumed that $a^i_t = l$ ($l \in Z$). If $l < 0$, it represents that the E-taxi i is matched with the charging station $-l$ to charge at the time step t. If $l > 0$, it represents that the E-taxi i is matched with the order l at the time step t. There are three situations when an E-taxi chooses the action 0: the E-taxi is charging in a charging station, the E-taxi is on the way to compete the last order, and the E-taxi does nothing at the time step t.

4) Reward: r^i_t denotes the reward that the E-taxi i gets after taking an action at the time step t. In this paper, we establish that when the E-taxi i is paired with the charging station k, it receives a reward $CT^t_{i,k}$. When the E-taxi i is matched with the order j, it receives a reward $PT^t_{i,j}$, and when it neither picks up passengers nor opts to charge, it receives a reward of 0. Therefore, the reward of each E-taxi in this paper can be formulated as follows:

$$r^i_t = \begin{cases} \mu_1(E_m - E^i_v) - \mu_2 DVCS^t_{i,-l} - \mu_3 TVCS^t_{i,-l}, & l < 0 \\ 0, & l = 0 \\ \mu_4 Q_j - \mu_5 DVP^t_{i,l}, & l > 0 \end{cases} \tag{33}$$

We hope that the E-taxis can keep working in the environment. But the remaining battery energy of some E-taxis is almost zero in certain situations. To avoid this situation, a punishment is designed when an E-taxi' energy falls below E_n, and is formulated as:

$$r^i_t = PUN \tag{34}$$

where PUN denotes the value of punishment when E-taxi i's energy is lower than E_n.

r_t denotes the reward that the agent gets after an action at the time step t. r_t is the sum of the all E-taxis' reward, which indicates the quality of the action taken by agent under the environment state. r_t is formulated as:

$$r_t = \sum_{i=0}^{m} r^i_t \tag{35}$$

Then the cumulative reward of one day can be formulated as follows:

$$r = \sum_{t=0}^{T} r_t = \sum_{t=0}^{T} \sum_{i=0}^{m} r^i_t \tag{36}$$

The expression of r is consistent with the optimization objective as shown in Equation (7). Therefore, the optimization objective in this paper can be expressed as the maximum cumulative reward in one day.

5) State Transfer: According to state s_t and E-taxis' action a'_t, each E-taxi interacts with environment, and environment state will change with time. For E-taxis, the location (xv^t_i, yv^t_i), remaining battery energy E^t_i, statuses KV^t_i and $KTAR^t_i$ will be updated respectively to (xv^{t+1}_i, yv^{t+1}_i), E^{t+1}_i, KV^{t+1}_i and $KTAR^{t+1}_i$. For charging stations, the status KCS^t_k will be updated to KCS^{t+1}_k due to the departure and arrival of E-taxis. For passengers, some old orders are completed while others not. The later will be ranked again and reserved for the next step $t+1$. The uncompleted orders' starting location (xps^t_j, yps^t_j), ending location (xpd^t_j, ypd^t_j), status KP^t_j will be updated respectively to $(xps^{t+1}_{j^*}, yps^{t+1}_{j^*})$, $(xpd^{t+1}_{j^*}, ypd^{t+1}_{j^*})$ and $KP^{t+1}_{j^*}$, where j^* may not equal to j due ranking again. Then, the whole environment state s_t will be updated to s_{t+1} after pretreatment.

6) Training Process of Proximal Policy Optimization: Fig. 4 shows the overall structure of PPO algorithm in this paper. The PPO algorithm is an on-policy gradient algorithm, which includes an actor network and a critic network. At each time slot, according to current policy and environment state s_t, the actor network outputs an action distributed probability p_t and an action a_t, which determines an action array a'_t by using a heuristic coordinated assignment method. Then, the SC reaches next state s_{t+1} by interacting with environment, which must experience a pretreatment for utilizing, and obtains

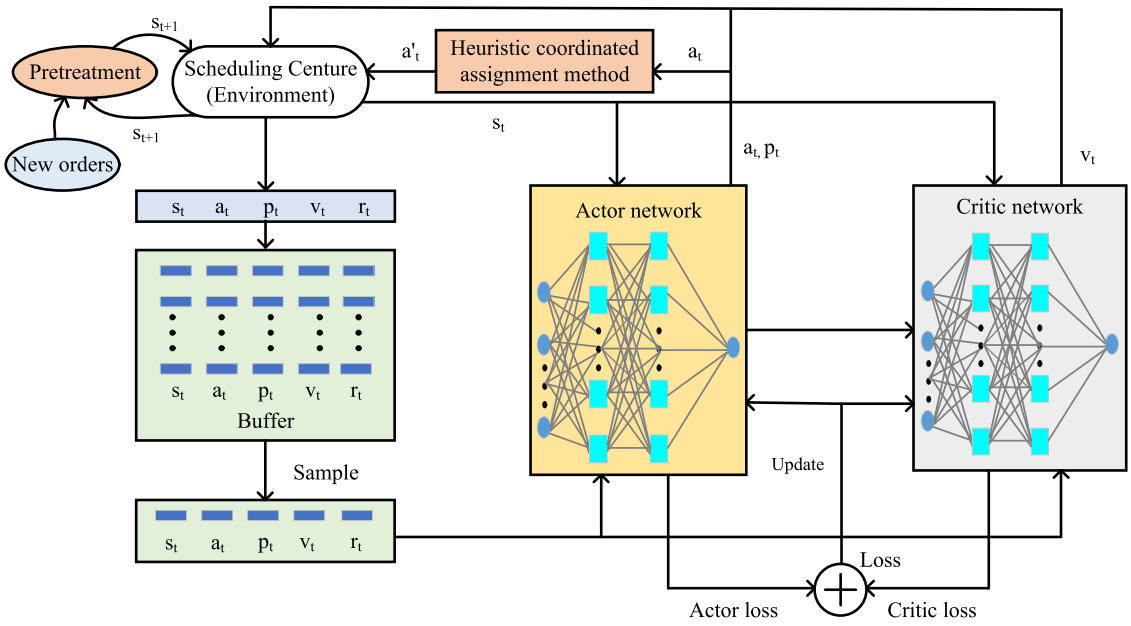


Fig. 4. PPO structure.

a reward r_t. Correspondingly, the critic network makes an evaluation v_t for s_t and a_t. Next, an array $\{s_t, a_t, p_t, v_t, r_t\}$ is stored in a buffer. When updating, a certain number of arrays are sampled from the buffer and transported to the actor network and the critic network to calculate actor loss and critic loss. Then the parameters of the actor network and the critic network will be updated by minimizing total loss, which is determined by actor loss and critic loss.

There are two policies: new policy $\pi_\theta(a_t|s_t)$ and old policy $\pi_{\theta_{old}}(a_t|s_t)$, where θ denotes the policy parameter vector. Then the probability ratio $r_t(\theta)$ can be expressed as:

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \tag{37}$$

One of the main characteristics of PPO is that the advantage function $\hat{A}_t$ is introduced, which can be expressed as:

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1} \tag{38}$$

$$\delta_t = Q(s_t) - V(s_t) = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{39}$$

where $Q(s_t)$ denotes the Q-value of state s_t, γ denotes discount factor, γ is a constant, $V(s_t)$ denotes the state-value function of state s_t. Different from trust region policy optimization (TRPO) algorithm using KL divergence to calculate the policy gradient, PPO introduces a clip operation based on TRPO to calculate the policy gradient, which can be expressed as:

$$L_\pi^{clip}(\theta) = \hat{E}_t[\min(r_t(\theta)\hat{A}_t, clipr_t(\theta), 1-\varepsilon, 1+\varepsilon] \tag{40}$$

where ε is a hyper parameter which is used to clip the value of $r_t(\theta)$. This approach could remove the incentive for the $r_t(\theta)$, which is not in interval $[1-\varepsilon, 1+\varepsilon]$. The whole MDP of E-taxis scheduling based on dual-stage heuristic coordinated reinforcement learning approach is shown in Algorithm 4.

V. Performance Evaluation

In this section, we assess the performance of our proposed dual-stage heuristic coordinated reinforcement learning approach using a real-world E-taxi hailing dataset. To validate the effectiveness and efficiency of our method, we compare it against existing common methods in the field. Additionally, we conduct a series of ablation experiments to demonstrate the impact of each component within our approach. Furthermore, a time cost analysis is performed to evaluate the scalability of our method as the size of the E-taxi hailing system increases. The evaluation metrics employed in this section include rewards, total profits of E-taxis, low-energy occurrences times, pickup count for E-taxis, passenger acceptance ratio, and the number of active charging stations.

A. Settings

In this paper, our experiments were conducted on a desktop computer equipped with an Intel Core i7-12700 processor. We used Python 3.9 as the programming language and Pytorch 1.12.1 as the deep learning framework. All experiments were conducted under the OptiPlex 7000 operating system. We apply a city-level real world taxi-hailing environment in Xi'an, China. We selected an area of 10km $\times$ 10km including 5 charging stations in Xi 'an, which is shown in Fig. 5. To mitigate an excessively large environmental state dimension, we opted for a representation of 20 E-taxis in the environment, a number lower than the actual count in the region. Accordingly, to maintain a balanced ratio of E-taxis to charging stations, as well as to accurately depict the authentic charging pressure on E-taxis, each charging station is equipped to accommodate only one E-taxi for charging at a time. To express location information clearly, the area is divide into a multi small areas of 100 $\times$ 100, where a unit length represents 100m. In the experiment, E-taxis work 10 hours

Algorithm 4 MDP of Dual-Stages Management Strategy Based on Heuristic Coordinated Reinforcement Learning Approach

Input: Episode number e, time slot number T for each episode, E-taxi number m

Load the parameters of trained actor network;
for $episode = 0$ *to* e **do**
  Initialize environment state s_0;
  Initialize total reward r of each episode;
  for $t = 0$ *to* T **do**
    Pretreat order state as shown in Algorithm 1;
    Obtain environment state s_t;
    Obtain an action a_t, a probability pro, a value val from agent;
    Match optimally between E-taxis and charging stations;
    Match optimally between E-taxis and orders;
    Obtain actions set a_t' of all E-taxis;
    for $i = 0$ *to* m **do**
      Interact with environment and get a reward r_t';
    end
    Calculate total reward r_t of all E-taxis;
    Update environment state and obtain s_{t+1};
    Store transition $\{s_t, a_t, prob, val, r_t\}$;
    $s_t = s_{t+1}$;
    $r = r + r_t$;
  end
  Update the actor network;
  Update the critic network;
end
return Scheduling result

Fig. 5. A real environment in Xi'an (From Google map).

every day, that represents $T = 60$, and the environment state updates every 10 minutes. The speed of all E-taxis follows a normal distribution $N(60, 5^2)$ km/h. At beginning of each day, the location of each E-taxi is set as a random coordinate (xv, yv) where xv and yv are both integers from 0 to 100. The initial remaining battery energy of each E-taxi is set as a random integer from 50 to 100. E_m and E_n are set to 100 and 5, respectively. In the majority of cases, (NL_p^t, NH_p^t) are assumed to (0, 30), and during peak hours, i.e., 8:00 ~ 9:00, 11:00 ~ 13:00, 18:00 ~ 20:00, (NL_p^t, NH_p^t) are set to (10, 60). When environment state updates, the starting location (xps, yps) and ending location (xpd, ypd) of each order are both random coordinates, where xps, yps, xpd and ypd are all integers from 0 to 100. If the total distance D_j of any order is lower than 500m, the order is meaningless and will be deleted. The relationship between quoted price Q_j and travelling distance D_j of passengers can be formulated as

$$Q_j = \begin{cases} 7.67, & 5 \le D_j \le 30 \\ 7.67 + 0.160 \cdot (D_j - 30), & 30 < D_j \le 80 \\ 15.67 + 0.255 \cdot (D_j - 80), & 80 < D_j \le 150 \\ 33.52 + 0.265 \cdot (D_j - 150), & D_j > 150 \end{cases} \tag{41}$$

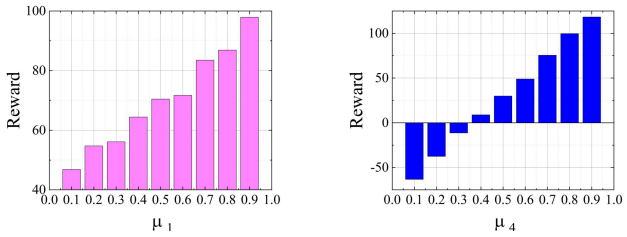
The profit of an E-taxi can obtain is Q_j Yuan when picking up passengers and -2 Yuan every 10 minutes when charging. Similar to real distribution in Xi'an, the detail locations of 5 charging piles are respectively as [45, 80], [35, 55], [15, 35], [65, 25] and [95, 60]. The charging speed C_{in} of all charging piles is set to 60 every hour. λ follows a normal distribution $N(50, 5^2)$. m_0 and n_c are both set to 3. We consider an E-taxi is in a bad situation when its SoC falls below 5, and the punishment PUN is set to -100.

PPO algorithm is utilized to train data which includes an actor network and a critic network. Two hidden layers are designed in each network. There are respectively 205 neurons, 128 neurons and 1 neuron in the input layer, hidden layer and output layer of each network. The number of training episodes e is set to 1000. There are 60 time steps in one day. The learning rate and discount factor are respectively set to 0.01 and 0.98. The memory size and batch size are respectively set to 10000 and 64. When training, the parameters of actor and critic are updated every 10 episodes.

A series of experiments is designed to explore the influence of parameters in Equation (3) and Equation (6). Considering E_i^t counts more than other two factors, we set that $\mu_1 + \mu_2 + \mu_3 = 1$ and $\mu_2 = \mu_3$. Similarly, $\mu_4 + \mu_5 = 1$. The overall convergence reward with μ_1 and μ_4 are respectively shown in Fig. 6-(a) and Fig. 6-(b). It is important to emphasize that our goal is to obtain a suitable set of parameters to fix the weights of the rewards, not necessarily the optimal ones. This is because the final performance of the system is determined by the mark in front of the parameters, with the size of the parameters having a relatively minor effect. To guarantee simultaneously that the final reward and all parameters are not too small, μ_1, μ_2, μ_3, μ_4 and μ_5 are respectively set to 0.6, 0.2, 0.2, 0.7 and 0.3 in this paper.

B. Reinforcement Learning Training Performance

In this section, we evaluate the training performance of our proposed dual-stage heuristic coordinated reinforcement



(a) Sensitivity of μ_1 (b) Sensitivity of μ_4

Fig. 6. Parameter sensitivity.

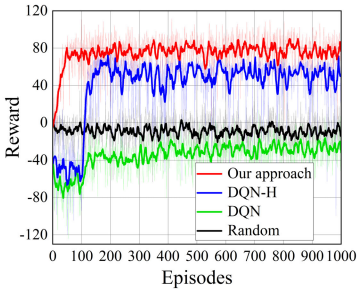
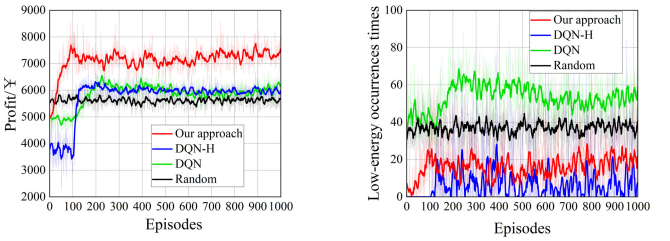
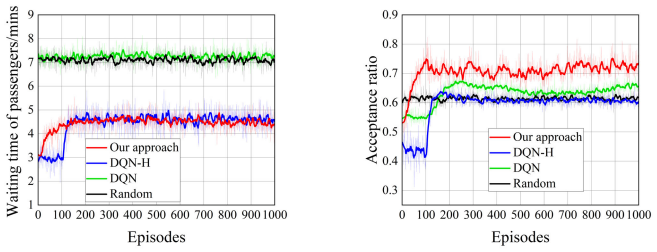


Fig. 7. Reward.



(a) Total profits of E-taxis (b) Low-energy occurrences times

Fig. 8. E-taxis performance.



(a) Waiting time of passengers (b) Passenger acceptance ratio

Fig. 9. Passengers performance.

learning approach in comparison to typical RL method (DQN-H and DQN) and typical optimization method (Random). 'DQN-H' indicates that the platform employs the DQN method during the first stage and maintains the heuristic approach in the second stage. 'DQN' signifies that the platform directly applies the joint DQN method consisting of 20 value networks and 20 target networks, wherein the goal of each value network is to maximum overall cumulative rewards. At each time step, the value networks produce 20 actions concurrently to determine the behavior of all E-taxis. 'Random' denotes that the platform generates a random action between -5 and 20 for each E-taxi at each time slot. To avoid the randomness, the average of training results from 19 runs is used as the final data presented in this paper. The results corresponding to each of the six indicators are illustrated in Fig. 7, Fig. 8-(a) to (b), Fig. 9-(a) to (b) and Fig. 10.

Overall performance: (i) Reward (Fig. 7): Our proposed approach consistently achieves the highest rewards during training, indicating that the method is capable of learning an efficient policy for E-taxi scheduling. The upward trend in the graph demonstrates the effectiveness of the dual-stage heuristic coordinated reinforcement learning approach in optimizing the overall system performance. This phenomenon arises because each value network aims to maximize the overall cumulative reward by focusing solely on the behavior of an individual E-taxi, disregarding the impact of other value networks and the collective behavior of the E-taxis. Consequently, the agent lacks a holistic understanding of the environment, hindering the acquisition of an effective and comprehensive optimization strategy. The weak interconnections among the value networks result in the inability of the joint DQN method to adapt to complex and dynamic environments, leading to performance even lower than the 'Random' strategy.

E-taxis' perspective: (ii) Total profits of E-taxis (Fig. 8-(a)): The total profits generated by E-taxis using our method (about 7500) are considerably higher than those achieved by the other three methods (about 6000), illustrating the substantial economic benefits of our approach. (iii) Low-energy occurrences times (Fig. 8-(b)): Low-energy occurrences times refers to the times that the SoC falls below the set minimum limit during the operation of all the E-taxis. Regarding to this figure, the low-energy occurrences times obtained by our method after training stabilization were about 17 times, slightly higher than DQN-H (10 times), but much lower than DQN (70 times) and 'Random' method (37 times). The lowest convergence rewards for the DQN strategy can be attributed to the increased frequency of low-energy occurrences. From the above, it can be seen that the algorithm proposed in this paper can effectively improve the performance of E-taxis. Compared with other methods, it can maximize E-taxis' profits while substantially reducing the low-energy occurrences times, so the algorithm proposed in this paper can improve the satisfaction of E-taxis.

Passengers' perspective: (iv) Waiting times of passengers (Fig. 9-(a)): it indicates the pickup time after the passenger order is received by the taxi, and this indicator is related to the matching distance. As shown in the figure, compared with other methods, the results of our method show a minimum value in waiting time, with our method being able to achieve approximately 1/2 of DQN and 'Random'. (v) Passenger acceptance ratio (Fig. 9-(b)): Passenger acceptance ratio represents the probability that a passenger's order will be accepted when they first enter the market. The results obtained in this paper are significantly higher (0.72) than other three methods, reflecting an increase of 16.7% compared to DQN-H and 'Random' strategy, as illustrated in the figure. The above waiting times index and passenger acceptance ratio index

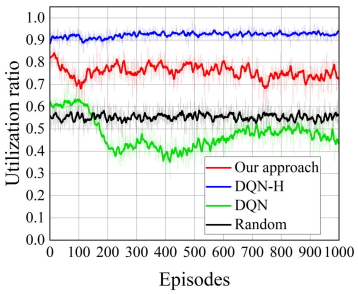
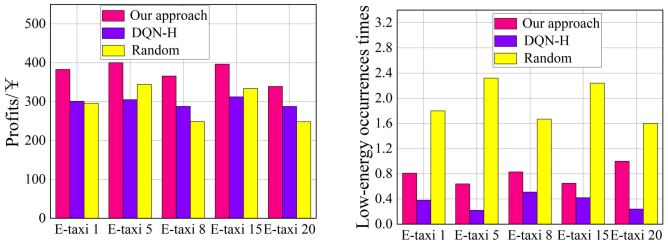


Fig. 10. Charging stations utilization ratio.



(a) Profits (b) Low-energy occurrences times

Fig. 11. Profits and low-energy occurrences times of five specific E-taxis.

judge the performance of the algorithm from the passenger's point of view, and it can be seen that our method also takes into account the passenger's satisfaction.

Charging stations' perspective: (vi) Charging stations utilization ratio (Fig. 10): Our method optimizes the use of charging stations, ensuring an appreciable utilization ratio, slightly lower than DQN-H, higher than DQN and 'Random' strategy, as illustrated in the figure. This result confirms the ability of our approach to efficiently manage the charging infrastructure and support the growing E-taxi ecosystem. Presumably, the DQN-H method's low profitability and low passenger acceptance ratio for E-taxis is attributed to the exceptionally high utilization of charging stations, which, in turn, results in a low occurrences times of low-energy.

In conclusion, the performance evaluation of our proposed dual-stage heuristic coordinated reinforcement learning approach, as shown in Figs. 7, Fig. 8-(a) to (b), Fig. 9-(a) to (b) and Fig. 10, achieves commendable results across all indicators compared to the DQN-H method, DQN method and 'Random' matching. It is worth noting that the performance of some indicators for DQN is even lower than 'Random' matching. The reason for this is that the scene is highly complex, with dynamic states and heterogeneous actions, making conventional RL algorithms unable to function effectively. Despite DQN-H exhibiting the lowest number of low-energy occurrences and the highest charging stations utilization, its profitability and passenger acceptance fall short of expectations. Overall, our approach effectively addresses the challenges of E-taxi scheduling, and optimizes rewards, profits, energy management, and customer satisfaction while efficiently utilizing charging stations.

C. E-Taxi Hailing System Performance

In the previous section, we analyzed the performance of each indicator during the training of the RL algorithm. In this section, we will focus on the performance of the proposed dual-stage heuristic coordinated reinforcement learning approach after convergence. We will examine specific aspects such as E-taxi charging and pickup patterns within a particular E-taxi hailing system, passenger satisfaction, and charging station occupancy. The data presented in this subsection has been averaged for each indicator after convergence. Given the poor performance of the 'DQN' method, we have chosen not to present its results in this context.

Table. II displays the states of various participants across some specific time periods in the taxi-hailing system when using our proposed dual-stage heuristic coordinated reinforcement learning approach, where LET, EwO, EiC, AR, WT, OR indicates 'Low-energy times', 'E-taxis with order', 'E-taxis in charging', 'Acceptance ratio', 'Waiting time' and 'Occupancy ratio', respectively. Overall, our proposed mechanism delivers superior performance compared to the DQN-H method and 'Random' matching method.

For E-taxis, our approach ensures that no more than 3 vehicles per hour have a battery level below the minimum threshold, while maintaining a higher SoC. This substantially reduces the operating time of E-taxis and enhances efficiency. During the 8-9 am peak period, our mechanism results in an average of 16 E-taxis picking up passengers, 4 E-taxis charging, and less than 1 E-taxi remaining idle. In contrast, the other two methods have a significant number of idle vehicles. This demonstrates that our mechanism prioritizes keeping E-taxis in active states, either picking up passengers or charging, to increase operational efficiency.

For passengers, our approach offers an order acceptance rate close to 70% during peak demand periods and as high as 75% during off-peak hours. In comparison, the other mechanisms exhibit acceptance rates of around 60% for both time periods. Our mechanism achieves these results by scheduling E-taxis for charging when passenger demand is low and making more E-taxis available when demand is high. This differs from other mechanisms that DQN-H prefers E-taxis to choose charging, whereas 'Random' tend to let E-taxis charge when they run out of power, resulting in insufficient service during peak periods. Furthermore, our mechanism keeps passengers' waiting times within 5 minutes, indicating effective matching between passengers and E-taxis based on their locations.

From the charging station's perspective, our mechanism maintains a stable utilization rate by preventing E-taxis from becoming idle. In conclusion, our proposed mechanism optimizes the use of charging stations at all times, significantly reduces low-energy occurrences times in E-taxis, and enhances both the operational profits of E-taxis and the acceptance rate of passengers' orders.

In the following analysis, we validate the advantages of our proposed algorithm in terms of E-taxi profits and the low-energy occurrences times from the perspectives of several specific E-taxis. Fig. 11-(a) and (b) respectively display the total daily profits (over 10 hours) and the total low-energy occurrences times for E-taxis 1, 5, 8, 15, and 20 when using our method, DQN, and the random method.

TABLE II
E-TAIX HAILING SYSTEM PERFORMANCE IN SPECIFIC TIME PERIOD

Indicators		Time period											
		8:00 to 9:00			11:00 to 12:00			16:00 to 17:00			17:00 to 18:00		
		Our	DQN-H	Random	Our	DQN-H	Random	Our	DQN-H	Random	Our	DQN-H	Random
E-taxis	SoC	70.50%	78.50%	76.50%	71.50%	84.62%	74.10%	47.07%	60.27%	48.16%	47.22%	55.60%	49.41%
	LET	1.55	0.53	3.85	1.75	0.65	3.83	2.00	0.73	3.95	2.28	0.53	3.88
	EwO	16.24	12.66	10.78	16.55	11.98	10.77	13.03	10.99	10.81	13.21	10.98	10.79
	EiC	3.51	4.25	3.55	3.44	4.79	3.68	6.35	7.58	3.25	6.51	7.56	3.72
Passengers	AR	70.39%	60.61%	61.24%	71.78%	60.87%	61.43%	75.29%	60.61%	61.39%	74.36%	60.62%	61.24%
	WT	4.56	4.63	7.11	4.51	4.60	7.08	4.46	4.60	7.09	4.47	4.63	7.09
Charging stations	OR	70.80%	92.79%	55.72%	68.80%	92.84%	55.32%	74.18%	92.79%	55.16%	73.08%	93.05%	55.34%

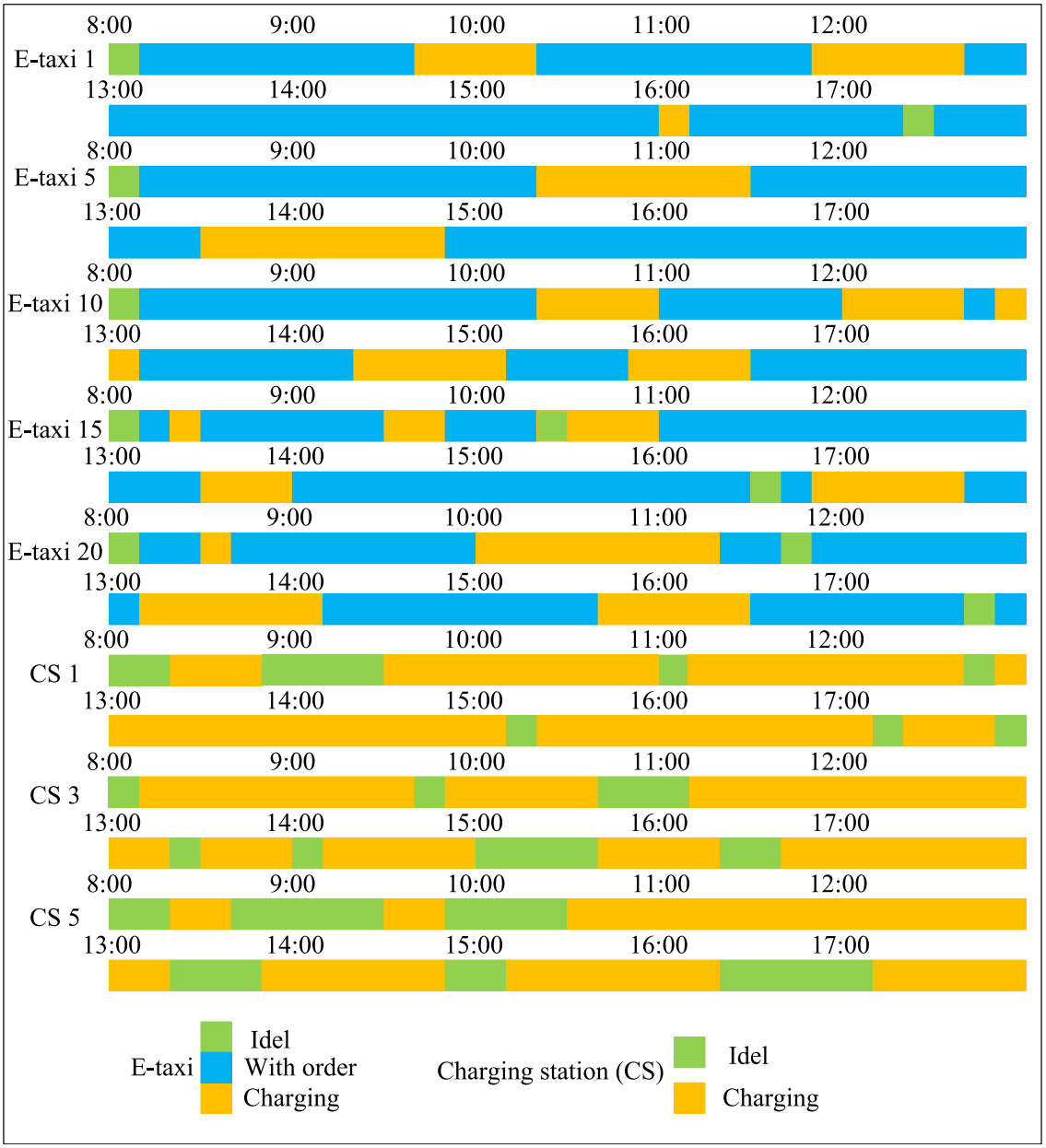


Fig. 12. E-taxis and charging stations' actions.

Firstly, it is evident from Fig. 11-(a) that each E-taxi can achieve higher profits using our mechanism, averaging 360 per day. Moreover, the profits for each E-taxi are evenly distributed, avoiding situations where some taxis earn significantly more or less than others. In contrast, both DQN-H and the 'Random' matching method yield lower profits, averaging

TABLE III
PASSENGERS' ACTIONS

Passenger	Our approach		DQN-H		Random	
	Response	Waiting time	Response	Waiting time	Response	Waiting time
p1	1	4.5	1	4.8	1	7.5
p2	1	4.4	2	4.5	3	5.5
p3	1	3.5	1	3.8	3	3.5
p4	1	5.5	1	2.2	3	9.9
p5	2	3.7	3	4.8	1	6.8
p6	2	4.5	1	5.8	1	2.7
p7	1	5.5	1	3.2	2	8.4
p8	1	3.1	1	8.5	2	7.5
p9	1	2.9	2	5.1	3	5.8
p10	1	4.5	2	4.9	1	3.5

around 280. Similarly, as seen in Fig. 11-(b), the low-energy occurrences are extremely rare for E-taxis, with only E-taxi 0.7 times on average per day, slightly more than DQN-H (0.3) but much less than 'Random' strategy. This demonstrates that our mechanism can effectively schedule charging and passenger pick-up behaviors for E-taxis, avoiding low battery energy situations, increasing passenger pick-up efficiency, and subsequently enhancing E-taxi profits. Additionally, our mechanism considers all participants in the E-taxi hailing system as a whole, resulting in relatively small differences in profits and low-energy occurrences times for each E-taxi. This approach can effectively prevent malicious competition between E-taxis, leading to better service for passengers.

Fig. 12 further employs a Gantt chart to illustrate the actions (picking up passengers, idle, charging) of 5 E-taxis and 3 charging stations throughout an entire day (10 hours). Consistent with the previously obtained results, our mechanism consistently avoids keeping E-taxis in an idle state. In other words, when E-taxis are not engaged in picking up passengers, they opt for charging to prevent being in a low-energy state and unable to provide services when the number of passengers increases in the future.

Next, we also explore the specific waiting time of 10 orders and how many time slots they experience before being accepted. As shown in Table III, with our proposed method, most of the orders are responded in the first time slot, and only two orders are answered in the second scheduling (after 10 minutes). Meanwhile, the waiting time of passengers all falls below 6 minutes. While in the other two methods, approximately half of orders are accepted in the second time slot, even in the third times, and the maximum waiting time of passengers reach nearly 10 minutes. The cause is that both E-taxis' profits and passengers' satisfaction are considered in our approach to increase participation.

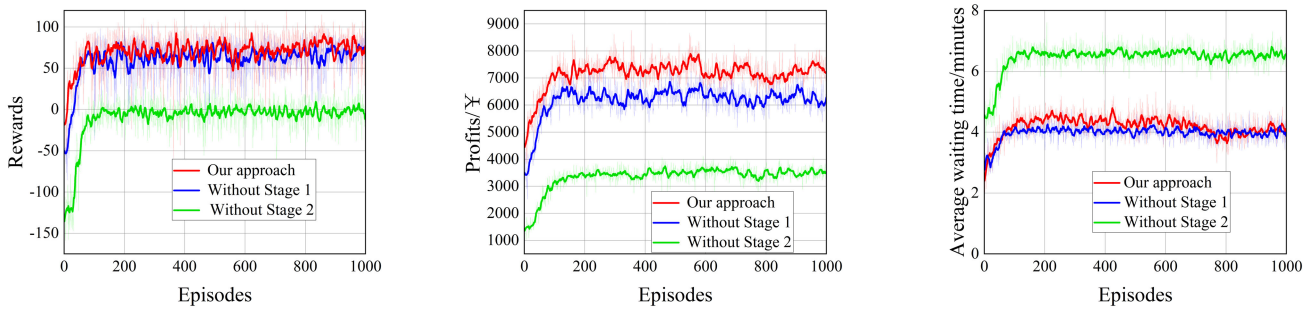
D. Ablation Experiment

To investigate the performance of our proposed dual-stage heuristic coordinated reinforcement learning approach in handling dynamic states and heterogeneous actions, we conduct ablation experiments for the first stage's feature-guided state dimensionality stabilization method and the second stage's heuristic coordinated assignment method. In Fig. 13-(a) to (c), "Without Stage 1" means that the feature preprocessing no longer uses our proposed feature-guided state dimensionality stabilization method, but adopts a random feature selection method, while "Without Stage 2" means that the matching between E-taxis, passengers, and charging stations no longer takes the proposed heuristic coordinated assignment method, but does random matching. Similar to Section V-B, the data represents the average of the results from 19 training sessions.

First, Fig. 13-(a) to (c) shows the convergence of rewards in our proposed method, as well as in cases without stage 1 and without stage 2. From the figure, it is evident that our method converges the fastest and achieves the highest final reward. For the case without stage 1, its convergence speed is similar to that of our proposed method, but the final reward is not as high, and the fluctuations after convergence are relatively larger. The reason for this is that without stage 1's feature selection, when there are too many passengers, some passengers may be randomly discarded to satisfy the input dimension of the RL state, and these discarded passengers could potentially bring higher profits. Regarding the case without stage 2, its performance in terms of convergence speed and final reward is the worst. The reason is that, as mentioned earlier, the actions in the E-taxi system have mutual interference. If the output actions of RL are not processed, this interference will severely affect the training process of the RL network, leading to slow convergence or even non-convergence. The smaller rewards observed here are also due to the fact that basic RL algorithms cannot learn feasible policies under such a high-load environment. Additionally, both stages contribute to the overall profitability of E-taxis. However, the second stage notably reduces passenger waiting time, while the impact of the first stage is minimal.

E. Comparisons

This section compares the results of the two-stage reinforcement learning heuristic algorithm proposed in this paper with existing algorithms, including (i) rewards: which intuitively reflect the overall advantages and disadvantages of the



(a) Rewards (b) Total profits of E-taxis (c) Average waiting time of passengers

Fig. 13. Ablation experiment.

TABLE IV
OVERALL STATISTICAL RESULT

Metrics	Our approach	Optimization	DQN	DQN-H	Without feature preporcessing	Without heuristic
Rewards	78.8	-11.3	-20.8	46.1	74.2	-2.1
Profits (¥)	7274	5575	5597	6020	6350	3605
Waiting time of passengers (minutes)	4.4	7.1	7.1	4.7	3.9	6.6
Low-energy occurrences times	18.6	38.4	44.3	11.0	10.6	23.5
Passenger acceptance ratio	72.1%	61.6%	60.8%	60.1%	72.9%	48.0%
Charging stations utilization ratio	70.7%	55.6%	53.6%	92.8%	70.1%	40.2%

methods, (ii) taxi revenue: reflecting the satisfaction level of the taxis in picking up passengers, (iii) passenger waiting time: reflecting the satisfaction level of the passengers, (iv) low battery occurrences: reflecting the satisfaction level of the taxis in terms of charging, (v) passenger acceptance rate: reflecting the satisfaction level of the passengers, and charging station utilization rate: reflecting the utilization of the charging stations. There are several comparison algorithms as follows, all of which are modified on the basis of the original algorithm to adapt to the complex scenarios studied in this paper: (i) Traditional optimization scheduling method: Using online optimization scheduling methods to directly solve the optimization equations. (ii) End-to-end RL (DQN): Directly using the DQN algorithm to match taxis with passengers and charging stations. (iii) Two-stage reinforcement learning (DQN+Heuristic): The first stage uses DQN to decide the number of taxis and passengers, and the second stage uses heuristic algorithms for solving. (iv) Our method without feature preprocessing: The first stage does not perform feature processing. (v) Our method without heuristic: After obtaining the number of taxis and passengers from the first stage RL, random matching is performed in the second stage. The results are shown in Table IV, where all results are averaged over 10 simulations.

From the results, it can be observed that the method proposed in this paper has certain advantages in all aspects, while the End-to-end RL method shows the worst results. This is because the E-taxi hailing system is a highly dynamic scenario, bringing about a dynamic state space and action space. Directly using reinforcement learning to match taxis, passengers, and charging stations is difficult to train effectively, with specific reasons discussed in Section IV-B. The traditional optimization scheduling method, due to its inability to obtain future states, also yields results inferior to the method proposed in this paper. The DQN+Heuristic and our method without Feature preprocessing have the same architecture as the method in this paper, but the performance gap in the algorithms used in the first stage leads to a certain gap in the final results compared to the results of this paper. Finally, our method without heuristic, which performs random matching in the second stage, also yields better results than the End-to-end RL method, demonstrating the advantage of the two-stage algorithm in the E-taxi hailing system.

In conclusion, our approach optimizes all aspects of the system in a systematic manner through its rational and well-designed dual-stage structure. In the first stage, the PPO method effectively increases the profitability of E-taxis. In the second stage, the heuristic method further improves the profitability of E-taxis and reduces passenger waiting time. Compared to other methods, our approach not only enhances the profitability of E-taxis, passenger acceptance ratio, and charging station utilization but also significantly reduces the number of low-energy occurrences and passenger waiting time, thereby achieving a high level of satisfaction for E-taxi drivers, passengers, and charging stations. Overall, following the initial exploration process using the RL method, the agent gradually learns the optimal number of E-taxis needed to charge in certain state. Once the E-taxis are assigned to charge or pick up passengers, the heuristic method is then employed to effectively match the E-taxis with orders and charging stations. RL and heuristic method work together to optimize the scheduling of E-taxis, guiding them to appropriate charging

stations when their battery levels are low and enabling them to serve high-value orders when their battery levels are sufficient.

VI. Conclusion and Discussion

A. Conclusion

In conclusion, this paper presents a novel approach to address the challenges of E-taxi management in the context of the rapidly growing adoption of EVs. By formalizing the E-taxi hailing management problem as a MDP with dynamic state and heterogeneous action, we provide a solid foundation for the development of advanced optimization algorithms. Our proposed dual-stage heuristic coordinated reinforcement learning approach consists of the feature-guided state dimensionality stabilization PPO method and a heuristic coordinated assignment method, effectively tackling the challenges of uncertain state dimensions and heterogeneous action spaces. The evaluation of our approach on a real-world E-taxi hailing environment demonstrates its effectiveness in significantly improving satisfaction of E-taxis, passengers and charging stations. Compared to traditional RL approach and random strategy, our approach has augmented daily E-taxi profits by nearly 20%, boosted order acceptance rates by approximately 15%, and slashed passenger waiting time by almost 55%, all while substantially enhancing charging station utilization. Consequently, our approach efficiently optimizes E-taxi scheduling, making it suitable for real-world E-taxi hailing applications.

B. Discussion and Future Works

Constraints execution: In operational research optimization systems, constraints are often imposed to mitigate the adverse effects of certain actions on the system. Traditional optimization solutions typically employ constraints to limit these actions. In reinforcement learning systems, scholars have designed various methods to avoid such adverse effects, such as the penalty-based soft constraint approach and the action masking strong constraint approach. This paper adopts a penalty-based method, which does not forcibly prohibit the agent from performing certain actions but instead makes judgments in conjunction with future rewards. For instance, during rush hours, to reduce passenger wait times and increase taxi revenue, taxis are encouraged to operate in low-energy consumption areas. The use of penalties can minimize the occurrence of low-energy operation for electric taxis while providing them with greater decision-making flexibility. The agent coordinates the remaining battery life, profitability, and passenger wait times of electric taxis through continuous interaction with the environment to make optimal decisions. On the other hand, action masking is also a common method for handling constraints. Unlike penalties, this method forcibly prohibits certain actions by assigning a probability of zero, which is also feasible in the scenarios considered in this paper. By forcibly prohibiting taxis from operating at low battery levels, the action space is simplified, reducing the instability of the training process and aiding in faster convergence. Both mainstream methods have their advantages; the penalty method is more flexible but cannot ensure that constraints are met, while the action masking method fundamentally satisfies constraints but cannot find optimal results. Future work in similar areas could combine the two approaches to leverage their respective strengths. For example, setting stricter action masking rules (forcing charging when battery levels are below 5%, currently at 10%) and imposing penalties between 5% and 10% could reduce the probability of action errors while maintaining a certain level of strategy flexibility. For instance, maintaining a penalty mechanism to restrict actions during the training phase ensures the agent's strategic flexibility, while increasing action masking during the usage phase prohibits the agent from making erroneous actions, ensuring system constraints are met.

End-to-end reinforcement learning is an algorithm capable of taking input states and directly outputting the final policy. In other words, the input consists of status information such as taxis and passengers, and the output is the matching information between taxis and passengers. However, in e-taxi hailing system, the number of taxis and passengers is constantly changing. This scenario, where the number of agents leads to a dynamic state and action space, presents a significant challenge in the field of reinforcement learning. Addressing such dynamic environments with varying agent populations is an area that requires further research and development in the domain of end-to-end RL algorithms. The primary methodologies we have contemplated include: Single-Agent Reinforcement Learning System: Here, the RL agent assumes the role of a dispatch center, facilitating real-time allocation and decision-making processes for taxis and passengers. To accommodate the variability in the quantity of taxis and passengers, it is imperative to configure the action and state spaces to be expansive, utilizing padding techniques to address any deficiencies. A significant challenge with this approach is the unpredictability of the number of taxis and passengers in future system iterations. Consequently, despite an extensive action space, it remains infeasible to make decisions for scenarios beyond the training environment. Multi-Agent Reinforcement Learning System: In this framework, RL agents embody taxis, making autonomous real-time decisions. Given the dynamic nature of the environment, the number of agents within the system is also subject to change. To surmount this, pre-training of the agents' strategy networks is essential, followed by retraining through the addition or subtraction of agents as the environment evolves. This method can markedly affect the stability of the training process. While both end-to-end RL methodologies possess some capacity to manage dynamic environments, they are not without significant challenges. To counteract this, our paper introduces a two-stage RL approach tailored for such settings: Stage 1: The RL system exclusively receives data concerning the quantity of taxis and passengers and their respective states, without differentiating between individuals, thereby mitigating the impact of quantity fluctuations on the matching process. Stage 2: A heuristic algorithm takes into account the number of charging taxis and taxis ready for passengers as determined by the RL, along with the current states of taxis and passengers, to formulate the optimal matching strategy. This dual-stage approach capitalizes on the advantages of RL in achieving long-term rewards while

circumventing its limitations in adapting to environmental shifts. Prospectively, end-to-end RL is poised to become a pivotal method for optimizing the scheduling within dynamic and intricate systems. For example, the application of multi-agent RL techniques to dynamically adjust the agent population could lead to a more generalized response to environmental variations. Alternatively, the deployment of large-scale RL models could diminish the model's environmental sensitivity and enhance its generalizability by augmenting its scale.

REFERENCES

[1] D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac, "Internet of Things: Vision, applications and research challenges," *Ad Hoc Netw.*, vol. 10, no. 7, pp. 1497–1516, Sep. 2012.

[2] M. Conti, A. Dehghantanha, K. Franke, and S. Watson, "Internet of Things security and forensics: Challenges and opportunities," *Future Generat. Comput. Syst.*, vol. 78, pp. 544–546, Jan. 2018.

[3] J. Zhang, N. Xie, X. Zhang, and W. Li, "An Online auction mechanism for cloud computing resource allocation and pricing based on user evaluation and cost," *Future Gener. Comput. Syst.*, vol. 89, pp. 286–299, Dec. 2018.

[4] A. H. C. Tsiamparlis-Wildeboer, E. I. F.-D. Jong, and F. Scheele, "Factors influencing patient education in shared medical appointments: Integrative literature review," *Patient Educ. Counseling*, vol. 103, no. 9, pp. 1667–1676, Sep. 2020.

[5] T. Lyu, P. Wang, Y. Gao, and Y. Wang, "Research on the big data of traditional taxi and online car-hailing: A systematic review," *J. Traffic Transp. Eng. (English Ed.)*, vol. 8, no. 1, pp. 1–34, Feb. 2021.

[6] X. Zhang, Y. Cao, L. Peng, J. Li, N. Ahmad, and S. Yu, "Mobile charging as a service: A reservation-based approach," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 4, pp. 1976–1988, Oct. 2020.

[7] C. Latinopoulos, A. Sivakumar, and J. W. Polak, "Response of electric vehicle drivers to dynamic pricing of parking and charging services: Risky choice in early reservations," *Transp. Res. C, Emerg. Technol.*, vol. 80, pp. 175–189, Jul. 2017.

[8] L. A. de Souza Silva, M. O. de Andrade, and M. L. A. Maia, "How does the ride-hailing systems demand affect individual transport regulation?" *Res. Transp. Econ.*, vol. 69, pp. 600–606, Sep. 2018.

[9] S. D. Contreras and A. Paz, "The effects of ride-hailing companies on the taxicab industry in Las Vegas, Nevada," *Transp. Res. A Policy Practice*, vol. 115, pp. 63–70, Sep. 2018.

[10] G. Zhu, H. Li, and L. Zhou, "Enhancing the development of sharing economy to mitigate the carbon emission: A case study of online ride-hailing development in China," *Natural Hazards*, vol. 91, no. 2, pp. 611–633, Mar. 2018.

[11] R. M. Mepparambath, Y. S. Soh, V. Jayaraman, H. E. Tan, and M. A. Ramli, "A novel modelling approach of integrated taxi and transit mode and route choice using city-scale emerging mobility data," *Transp. Res. A, Policy Pract.*, vol. 170, Apr. 2023, Art. no. 103615.

[12] G. Gao, M. Xiao, and Z. Zhao, "Optimal multi-taxi dispatch for mobile taxi-hailing systems," in *Proc. 45th Int. Conf. Parallel Process. (ICPP)*, 2016, pp. 294–303.

[13] S. F. Roselli, P.-L. Götvall, M. Fabian, and K. Åkesson, "A compositional algorithm for the conflict-free electric vehicle routing problem," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 3, pp. 1405–1421, Jul. 2022.

[14] M. P. Fanti, A. M. Mangini, M. Roccotelli, and B. Silvestri, "Innovative approaches for electric vehicles relocation in sharing systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 1, pp. 21–36, Jan. 2022.

[15] Y. Xiang, J. Yang, X. Li, C. Gu, and S. Zhang, "Routing optimization of electric vehicles for charging with event-driven pricing strategy," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 1, pp. 7–20, Jan. 2022.

[16] H. Cai, F. Wu, Z. Cheng, B. Li, and J. Wang, "A large-scale empirical study on impacting factors of taxi charging station utilization," *Transp. Res. D, Transp. Environ.*, vol. 118, May 2023, Art. no. 103687.

[17] L. Zhang, K. Leng, S. Li, and J. Wang, "Comparative analysis of comprehensive benefits of Beijing's taxi electrification paths," *Transp. Res. D, Transp. Environ.*, vol. 115, Feb. 2023, Art. no. 103612.

[18] Y. Yuan, D. Zhang, F. Miao, J. Chen, T. He, and S. Lin, "P^2 charging: Proactive partial charging for electric taxi systems," in *Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jul. 2019, pp. 688–699.

[19] G. Guo and Y. Xu, "A deep reinforcement learning approach to ride-sharing vehicle dispatching in autonomous mobility-on-demand systems," *IEEE Intell. Transp. Syst. Mag.*, vol. 14, no. 1, pp. 128–140, Jan. 2022.

[20] L. Yang, P. Yan, Y. Pan, and C. Chu, "Optimal assignment and scheduling approach for electric Taxis' charging problem," in *Proc. Int. Conf. Ind. Eng. Syst. Manage. (IESM)*, Sep. 2019, pp. 1–6.

[21] T.-Y. Zhang et al., "Deployment optimization of battery swapping stations accounting for taxis' dynamic energy demand," *Transp. Res. D, Transp. Environ.*, vol. 116, Mar. 2023, Art. no. 103617.

[22] L. Cilio and O. Babacan, "Allocation optimisation of rapid charging stations in large urban areas to support fully electric taxi fleets," *Appl. Energy*, vol. 295, Aug. 2021, Art. no. 117072.

[23] J.-M. Clairand, M. González-Rodríguez, R. Kumar, S. Vyas, and G. Escrivá-Escrivá, "Optimal siting and sizing of electric taxi charging stations considering transportation and power system requirements," *Energy*, vol. 256, Oct. 2022, Art. no. 124572.

[24] K. Li, T. Zhang, and R. Wang, "Deep reinforcement learning for multiobjective optimization," *IEEE Trans. Cybern.*, vol. 51, no. 6, pp. 3103–3114, Jun. 2020.

[25] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.

[26] X. Chen, L. Yao, J. McAuley, G. Zhou, and X. Wang, "Deep reinforcement learning in recommender systems: A survey and new perspectives," *Knowl.-Based Syst.*, vol. 264, Mar. 2023, Art. no. 110335.

[27] K. Park and I. Moon, "Multi-agent deep reinforcement learning approach for EV charging scheduling in a smart grid," *Appl. Energy*, vol. 328, Dec. 2022, Art. no. 120111.

[28] F. Zhang, Q. Yang, and D. An, "CDDPG: A deep-reinforcement-learning-based approach for electric vehicle charging control," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3075–3087, Mar. 2021.

[29] Z. Qin et al., "Ride-hailing order dispatching at DiDi via reinforcement learning," *INFORMS J. Appl. Anal.*, vol. 50, no. 5, pp. 272–286, Sep. 2020.

[30] C. Mao, Y. Liu, and Z.-J. Shen, "Dispatch of autonomous vehicles for taxi services: A deep reinforcement learning approach," *Transp. Res. C, Emerg. Technol.*, vol. 115, Jun. 2020, Art. no. 102626.

[31] Y. Liu, F. Wu, C. Lyu, S. Li, J. Ye, and X. Qu, "Deep dispatching: A deep reinforcement learning approach for vehicle dispatching on online ride-hailing platform," *Transp. Res. E, Logistics Transp. Rev.*, vol. 161, May 2022, Art. no. 102694.

[32] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.

[33] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[34] J. Shen, L. Wang, and J. Zhang, "Integrated scheduling strategy for private electric vehicles and electric taxis," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 1637–1647, Mar. 2021.

[35] J. Park, Y. Sim, G. Lee, and D.-H. Cho, "A fuzzy logic based electric vehicle scheduling in smart charging network," in *Proc. 16th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2019, pp. 1–6.

[36] S. Ma, "Multi-objective optimization of electric vehicle scheduling based on behavior prediction," in *Proc. IEEE 4th Conf. Energy Internet Energy Syst. Integr. (EI2)*, Oct. 2020, pp. 2832–2836.

[37] X. Liang and G. H. d. A. Correia, "An optimal charging location model of an automated electric taxi system considering two types of charging," in *Proc. Forum Integr. Sustain. Transp. Syst. (FISTS)*, Nov. 2020, pp. 264–271.

[38] H. Ding, J. Li, N. Zheng, X. Zheng, W. Huang, and H. Bai, "Dynamic dispatch of connected taxis for large-scale urban road networks with stochastic demands: An MFD-enabled hierarchical and cooperative approach," *Transp. Res. C, Emerg. Technol.*, vol. 142, Sep. 2022, Art. no. 103792.

[39] A. Abid, N. A. Nawaz, M. S. Farooq, U. Farooq, I. Abid, and I. Obaid, "Taxi dispatch optimization in smart cities using TOPSIS," *Secur. Commun. Netw.*, vol. 2022, pp. 1–10, Jan. 2022.

[40] Z. Wang, Z. Qin, X. Tang, J. Ye, and H. Zhu, "Deep reinforcement learning with knowledge transfer for online rides order dispatching," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 617–626.

[41] Y. Zhao, Y. Xu, Y. Guo, and Q. Guo, "Reinforcement learning based optimal operation strategy for electric taxis," in *Proc. IEEE 4th Conf. Energy Internet Energy Syst. Integr. (EI2)*, Oct. 2020, pp. 2942–2947.

[42] P. Silva, Y. J. Han, Y.-C. Kim, and D.-K. Kang, "Ride-hailing service aware electric taxi fleet management using reinforcement learning," in *Proc. 13th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Jul. 2022, pp. 427–432.

[43] M. Haliem, G. Mani, V. Aggarwal, and B. Bhargava, "A distributed model-free ride-sharing approach for joint matching, pricing, and dispatching using deep reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 12, pp. 7931–7942, Dec. 2021.

[44] Y. Du, J. Chen, C. Zhao, F. Liao, and M. Zhu, "A hierarchical framework for improving ride comfort of autonomous vehicles via deep reinforcement learning with external knowledge," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 38, no. 8, pp. 1059–1078, May 2023.

[45] X. Zhang, C. Zhao, F. Liao, X. Li, and Y. Du, "Online parking assignment in an environment of partially connected vehicles: A multi-agent deep reinforcement learning approach," *Transp. Res. C, Emerg. Technol.*, vol. 138, May 2022, Art. no. 103624.

[46] W. Lin, C. Zhu, W. Zhu, and S. Shen, "Charging scheduling strategies of cooperated car-hailing operating business for electric taxis," in *Proc. Int. Conf. Wireless Commun. Smart Grid (ICWCSG)*, Aug. 2021, pp. 407–413.

[47] E. Wang et al., "Joint charging and relocation recommendation for E-taxi drivers via multi-agent mean field hierarchical reinforcement learning," *IEEE Trans. Mobile Comput.*, vol. 21, no. 4, pp. 1274–1290, Apr. 2022.

Donghe Li (Member, IEEE) received the B.S. degree in automation and the Ph.D. degree in control science and engineering from Xi'an Jiaotong University, Xi'an, China, in 2015 and 2020, respectively. He is currently an Associate Professor with the School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University. His current research interests include cyber-physical systems, auction mechanism, energy trading market, and privacy preservation.

Chunlin Hu received the B.S. degree in automation from Jilin University, Changchun, China, in 2022. He is currently pursuing the M.S. degree with the Faculty of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China. His current research interests include smart grid energy trading, attack, and privacy security.

Qingyu Yang (Senior Member, IEEE) received the B.S. and M.S. degrees in mechatronics engineering and the Ph.D. degree in control science and engineering from Xi'an Jiaotong University, Xi'an, China, in 1996, 1999, and 2003, respectively. He is currently a Professor with the Faculty of Electronics and Information Engineering, Xi'an Jiaotong University, where he is also with the State Key Laboratory for Manufacturing System Engineering. His current research interests include cyber-physical systems, power grid security and privacy, control and diagnosis of mechatronic systems, and intelligent control of industrial process.

Pengtao Song received the B.S. degree in automation from Northeastern University, Qinhuangdao, China, in 2020. He is currently pursuing the Ph.D. degree with the Faculty of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China. His current research interests include cyber-physical systems, networked control, and machine learning.

Feiye Zhang received the B.S. degree in electronic science and technology from Xi'an Jiaotong University, Xi'an, China, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Automation Science and Technology, School of Electronics and Information Engineering. His current research interests include multi-agent systems, reinforcement learning, and auction mechanisms design for the smart grids.

Dou An received the Ph.D. degree in control science and engineering from Xi'an Jiaotong University, Xi'an, China, in 2017. He is currently an Associate Professor with the Department of Automation Science and Engineering, Faculty of Electronics and Information Engineering, Xi'an Jiaotong University. His current research interests include cyber-physical systems, the IoT security and privacy, and incentive mechanism design for smart grid and IoT.