

CoRide: Joint Order Dispatching and Fleet Management for Multi-Scale Ride-Hailing Platforms

Jiarui Jin¹, Ming Zhou¹, Weinan Zhang¹, Minne Li², Zilong Guo¹, Zhiwei Qin³, Yan Jiao³,
Xiaocheng Tang³, Chenxi Wang³, Jun Wang², Guobin Wu⁴, Jieping Ye³

¹Shanghai Jiao Tong University, ²University College London, ³DiDi AI Labs, ⁴DiDi Research
{jinjiarui97, mingak, wnzhang, gzlong}@sjtu.edu.cn, {minne.li, jun.wang}@cs.ucl.ac.uk, {qinziwei, yanjiao, tangxiaocheng, wangchenxi, wuguobin, yejieping}@didiglobal.com

ABSTRACT

How to optimally dispatch orders to vehicles and how to trade off between immediate and future returns are fundamental questions for a typical ride-hailing platform. We model ride-hailing as a large-scale parallel ranking problem and study the joint decision-making task of order dispatching and fleet management in online ride-hailing platforms. This task brings unique challenges in the following four aspects. First, to facilitate a huge number of vehicles to act and learn efficiently and robustly, we treat each region cell as an agent and build a multi-agent reinforcement learning framework. Second, to coordinate the agents from different regions to achieve long-term benefits, we leverage the geographical hierarchy of the region grids to perform hierarchical reinforcement learning. Third, to deal with the heterogeneous and variant action space for joint order dispatching and fleet management, we design the action as the ranking weight vector to rank and select the specific order or the fleet management destination in a unified formulation. Fourth, to achieve the multi-scale ride-hailing platform, we conduct the decision-making process in a hierarchical way where a multi-head attention mechanism is utilized to incorporate the impacts of neighbor agents and capture the key agent in each scale. The whole novel framework is named as *CoRide*. Extensive experiments based on multiple cities real-world data as well as analytic synthetic data demonstrate that *CoRide* provides superior performance in terms of platform revenue and user experience in the task of city-wide hybrid order dispatching and fleet management over strong baselines.

CCS CONCEPTS

• **Computing methodologies** → **Multi-agent reinforcement learning**; • **Theory of computation** → *Multi-agent reinforcement learning*; • **Applied computing** → *Transportation*.

KEYWORDS

Hierarchical Reinforcement Learning; Multi-agent Reinforcement Learning; Ride-Hailing; Order Dispatching; Fleet Management

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM '19, November 3–7, 2019, Beijing, China

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6976-3/19/11...\$15.00

<https://doi.org/10.1145/3357384.3357978>

ACM Reference Format:

Jiarui Jin, Ming Zhou, Weinan Zhang, Minne Li, Zilong Guo, Zhiwei Qin, Yan Jiao, Xiaocheng Tang, Chenxi Wang, Jun Wang, Guobin Wu and Jieping Ye. 2019. CoRide: Joint Order Dispatching and Fleet Management for Multi-Scale Ride-Hailing Platforms. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*, November 3–7, 2019, Beijing, China. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3357384.3357978>

1 INTRODUCTION

Online ride-hailing platforms such as Uber and Didi Chuxing have substantially transformed our lives by sharing and reallocating transportation resources to highly promote transportation efficiency. In a general view, there are two major decision-making tasks for such ride-hailing platforms, namely (i) order dispatching, i.e., to match the orders and vehicles in real time to directly deliver the service to the users [24, 43, 45], and (ii) fleet management, i.e., to reposition the vehicles to certain areas in advance to prepare for the later order dispatching [15, 21, 26].

Apparently, the decision-making of matching an order-vehicle pair or repositioning a vehicle to an area needs accounting for the future situation of the vehicle's location and the orders nearby. Thus, much of work has modeled order dispatching and fleet management as a sequential decision-making problem and solved it with reinforcement learning (RL) [15, 30, 36, 39]. Most of the previous work deals with either order dispatching or fleet management without regarding the high correlation of these two tasks, especially for large-scale ride-hailing platforms in large cities, which leads to sub-optimal performance. In order to achieve near-optimal performance, inspired by thermodynamics, we simulate the whole ride-hailing platform as dispatch (order dispatching) and reposition (fleet management). As illustrated in Figure 1, we resemble vehicle and order as different molecules and aim at building up the system stability via reducing their number by dispatch and reposition. To address this complex criterion, we provide two novel views: (i) interconnecting order dispatching and fleet management, and (ii) joint considering intra-district (grid-level) and inter-district (district-level) allocation. With such a practical motivation, we focus on modeling joint order dispatching and fleet management with multi-scale decision-making system. There are several significant technical challenges to learn highly efficient allocation policy for the real-time ride-hailing platform:

Large-scale Agents. A fundamental question in any ride-hailing system is how to deal with a large number of orders and vehicles. One alternative is to model each available vehicle as an agent [21,

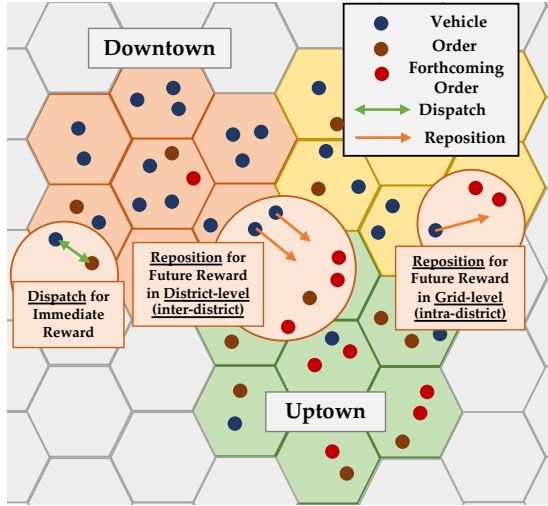


Figure 1: Ride-hailing task in thermodynamics view.

37, 39]. However, such setting needs to maintain thousands of agents interacting with the environment, which brings a huge computational cost. Instead, we utilize the region grid world (as will be further discussed in Figure 2), which regards each region as an agent, and naturally model ride-hailing system in a hierarchical learning setting. This formulation allows decentralized learning and control with distributed implementation.

Immediate & Future Rewards. A key challenge in seeking an optimal control policy is to find a trade-off between immediate and future rewards in terms of accumulated driver income (ADI). Greedily matching vehicles with long-distance orders can receive high immediate gain at a single order dispatching stage, but it would harm order response rate (ORR) and future revenue especially during rush hour because of its long drive time and unpopular destination. Recent attempts [21, 37, 39] deployed RL to combine instant order reward from online planning with future state-value as the final matching value. However, the coordination between different regions is still far from optimal. Inspired by hierarchical RL [34], we introduce the geographical hierarchical structure of region agents. We treat large district as *manager* agent and small grid as *worker* agent, respectively. The *manager* operates at a lower spatial and temporal dimension and sets abstract goals which are conveyed to its *workers*. The *worker* takes specific actions and interacts with environment coordinated with *manager*-level goal and *worker*-level message. This decoupled structure facilitates very long timescale credit assignment [34] and guarantees balance between immediate and future revenue.

Heterogeneous & Variant Action Space. Traditional RL models require a fixed action space [17]. If we model picking an order as an RL action, there is no guarantee of a fixed action space as the available orders keep changing. Zhang et al. [43] proposed to learn a state-action value function to evaluate each valid order-vehicle match, then use a combinatorial optimization method such as Kuhn-Munkres (KM) algorithm [18] to filter the matches. However, such a method faces another important challenge that order dispatching and fleet management are different tasks, which results in heterogeneous action spaces. To address this issue, we redefine action as

the weight vector for ranking orders and fleet management, where the fleet controls are regarded as fake orders, and all the orders are ranked and matched with vehicles in each agent. Thus, it bypasses the issue of heterogeneous and variant action space as well as high computational costs.

Multi-Scale Ride-Hailing. Xu et al. [39] introduced a policy evaluation based RL method to learn the dynamics for each grid. As its result shows, orders and vehicles often centralize at different districts (e.g. uptown and downtown in Figure 1). How to combine large hotspots in the city (inter-district) with small ones in districts (intra-district) is another challenge without much attention. In order to take both inter-district and intra-district allocation into consideration, we adopt and extend attention mechanism in a hierarchical way (as will be further discussed in Figure 3). Compared with learning value function for each grid homogeneously [39], this attention-based structure can not only capture the impacts of neighbor agents, but also learn to distinguish the key grid and district in *worker* (grid) and *manager* (district) scales respectively.

Wrapping all modules together, we propose *CoRide*, a hierarchical multi-agent reinforcement learning framework to resolve the aforementioned challenges. The main contributions are listed as follows:

- We propose a novel framework that learns to collaborate in hierarchical multi-agent setting for ride-hailing platform.
- We conduct extensive experiments based on real-world data of multiple cities, as well as analytic synthetic data, which demonstrate that *CoRide* provides superior performance in terms of ADI and ORR in the task of city-wide hybrid order dispatching and fleet management over strong baselines.
- To the best of our knowledge, *CoRide* is the first work (i) to apply the hierarchical reinforcement learning on ride-hailing platform; (ii) to address the task of joint order dispatching and fleet management of online ride-hailing platforms; (iii) to introduce and study multi-scale ride-hailing task.

This structure conveys several benefits: (i) In addition to balancing long-term and short-term reward, it also facilitates adaptation in a dynamic real-world situation by assigning different goals to *worker*. (ii) Instead of considering all of the matches between available orders and vehicles globally, these tasks are distributed to each *worker* and *manager* agent and fulfilled in a parallel way.

2 RELATED WORK

Decision-making for Ride-hailing. Order dispatching and fleet management are two major decision-making tasks for online ride-hailing platforms, which have acquired much attention from academia and industry during the recent few years.

To tackle these challenging transportation problems, automatically ruled-based approaches addressed order dispatching problem with either centralized or decentralized settings. To improve global performance, Zhang et al. [43] proposed a novel model based on centralized combinatorial optimization by concurrently matching multiple vehicle-order pairs within a short time window. However, this approach needs to compute all available vehicle-order matches and requires feature engineering, which would be infeasible and prevent it to be adopted in the large-scale taxi-order dispatching

situation. With the decentralized setting, Seow et al. [24] addressed this problem with a collaborative multi-agent taxi dispatching system. However, this method requires rounds of direct communications between agents, so it is limited to a local area with a small number of vehicles.

Instead of rule-based approaches, which require additional hand-crafted heuristics, the current trending direction is to incorporate reinforcement learning algorithms in complicated traffic management problems. Xu et al. [39] proposed a learning and planning method based on reinforcement learning to optimize vehicle utilization and user experience in a global and more farsighted view. In [21], the authors leveraged the graph structure of the road network and expanded distributed DQN formulation to maximize entropy in the agents’ learning policy with soft Q-learning, to improve performance of fleet management. Wei et al. [37] introduced a reinforcement learning method, which takes the uncertainty of future requests into account and can make a look-ahead decision to help the operator improve the global level-of-service of a shared-vehicle system through fleet management. To capture the complicated stochastic demand-supply variations in high-dimensional space, Lin et al. [15] proposed a contextual multi-agent actor-critic framework to achieve explicit coordination among a large number of agents adaptive to different contexts in fleet management system.

Different from all aforementioned methods, our approach is the first, to the best of our knowledge, to consider the joint modeling of order dispatching and fleet management and also the only current work introducing and studying the multi-scale ride-hailing task.

Hierarchical Reinforcement Learning. Hierarchical reinforcement learning (HRL) is a promising approach to extend traditional reinforcement learning (RL) methods to solve tasks with long-term dependency or multi-level interaction patterns [5, 6]. Recent works have suggested that several interesting and standout results can be induced by training multi-level hierarchical policy in a multi-task setup [8, 25] or implementing hierarchical setting in sparse reward problems [23, 34].

The options framework [22, 28, 29] formulates the problem with a two-level hierarchy, where the low-level - option - is a sub-policy with a termination condition. Since traditional options framework suffers from prior knowledge on designing options, jointly learning high-level policy with low-level policy has been proposed by [2]. However, this actor-critic HRL approach needs to either learning sub-policies for each time step or one sub-policy for the whole episode. Therefore, the performance of the whole module often prone to learning useful sub-policies. To guarantee gaining effective sub-policies, one alternative direction is to provide auxiliary rewards for low-level policy: hand-designed rewards based on prior domain knowledge [11, 31] or mutual information [4, 7, 10]. Given having access to one well-designed and suitable reward is often a luxury, Vezhnevets et al. [34] proposed FeUdal Networks (FuN), where a generic reward is utilized for low-level policy learning, thus avoid designing hand-craft rewards. Several works extend and improve FuN with off-policy training [19], form of hindsight [12] and representation learning [20].

Our work is also developed from FuN [34], originally inspired by feudal RL [5]. FuN employs only one pair of *manager* and *worker*

and connects them with a parameterized goal and intrinsic reward. Instead, we model *CoRide* with multiple *managers*. Unlike our method, in FuN the *manager* and *worker* modules are set to one-to-one, share the same observation, and operate at the different temporal but same spatial resolution. In *CoRide*, there are multiple *workers* learning to collaborate under one *manager* while the *managers* are also coordinating. The *manager* takes a joint observation of all *workers*, and each *worker* produces action based on specific observation and sharing goal. In other words, FuN [34] is actually a special case of *CoRide*, where a single *manager* along with its only *worker* is employed. Stepping on this one-to-many setting, the *manager* can not only operate with long timescale credit but act at a lower spatial resolution. Recently, Ahilan and Dayan [1] introduced a novel architecture named FMH for cooperation in multi-agent RL. Different from this proposed method, *CoRide* not only extends the scale of the multi-agent environment but also facilitates communication through multi-head attention mechanism, which computes influences of interactions and differentiates the impacts to each agent. Yet, the majority of current HRL methods require careful task-specific design, making them difficult to apply in real-world scenarios [19]. To the best of our knowledge, *CoRide* is the first work to apply hierarchical reinforcement learning on the ride-hailing problem.

3 PROBLEM FORMULATION

We formulate the problem of controlling large-scale homogeneous vehicles in online ride-hailing platforms, which combines order dispatching system with fleet management system with the goal of maximizing the city-level ADI and ORR. In practice, vehicles are divided into two groups: order dispatching (OD) group and fleet management (FM) group. For OD group, we match these vehicles with available orders pair-wisely; whereas for FM group, we need to reposition them to the locations or dispatch orders to them (same as OD group). The illustration of the problem is shown in Figure 2. We use the hexagonal-grid world to represent the map and take a grid as an agent. Considering that only orders within the pick-up distance can be dispatched to vehicle, we set distance between grids based on the pick-up distance. Given that, in our setting, vehicles in the same spatial-temporal node are homogeneous, i.e., the vehicles located at the same grid share the same setting. As such, we can model order dispatching as a large-scale parallel ranking task, where we rank orders and match them with homogeneous vehicles in each grid. The fleet control for fleet management, i.e. repositioning the vehicle to neighbor grids or staying at the current grid, is treated as fake orders (as will be further discussed in Section 6) and conducted in the ranking procedure as same as order dispatching.

Since each agent can only reposition vehicles located in the managing grid, we propose to formulate the problem using *Partially Observable Markov Decision Process (POMDP)* [27] in a hierarchical multi-agent reinforcement learning setting for both order dispatching and fleet management. Thus, we decompose the original complicated tasks into many local ones and transform a high-dimensional problem into multiple low-dimensional problems.

Formally, we model this task as a Markov game \mathcal{G} for N agents, which is defined by a tuple $\mathcal{G} = (N, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $N, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma$ are the number of agents, set of states, set of actions,

state transition probability, reward function, and a future reward discount factor, respectively. The definitions of major components are as follows.

Agent. We consider each region cell as an agent identified by $i \in \mathcal{I}$, where $\mathcal{I} = \{i \mid i = 1, \dots, N\}$. In detail, a single grid represents a *worker* agent, a district containing multiple grids represents a *manager* agent. An example is presented in Figure 2. Each individual grid serves as a *worker* agent with 6 neighbor grids, as shown in the same color, composes a *manager* agent. Note that although the number of vehicles and orders varies over time, the number of agents is fixed.

State $s_t \in \mathcal{S}$, **Observation** $o_t \in \mathcal{O}$. Although there are two different types of agents - *manager* and *worker*, their observations only differ in scale. Observation of each *manager* is actually a joint observation of its *workers*. At each timestep t , agent i draws private observations $o_t^i \in \mathcal{O}$ correlated with the environment state $s_t \in \mathcal{S}$. In our setting, the state input used in our method is expressed as $\mathcal{S} = \langle N_v, N_o, E, N_f, D_o \rangle$, where inner elements represent the number of vehicles, number of order, entropy, number of vehicles in FM group and distribution of order features (e.g. price, duration) in current grid respectively. Note that both dispatching and repositioning belong to resource allocation similar to the thermodynamic system (Figure 1), and once order dispatching or fleet management occurs, dispatched or fleted items slip out of the system. Namely, only idle vehicles and available orders can contribute to disorder and unevenness of the ride-hailing system. Therefore, we introduce and extend the concept of entropy here, defined as:

$$E = -k_B \cdot \sum_i \rho_i \log \rho_i := -k_B \cdot \rho_0 \log \rho_0 \quad (1)$$

where k_B is a Boltzmann constant, and ρ_i means probability for each state: ρ_1 for dispatched and fleted, ρ_0 elsewhere. As aforementioned analysis, once order and vehicle combined as order-vehicle pairs, both order and vehicle are out of the ride-hailing platform. Therefore, we choose to ignore items at state 1 (ρ_1) and compute ρ_0 as proportion of available order-vehicle pairs in all potential pairs:

$$\rho_0 = \frac{\min(N_v, N_o) \times \min(N_v, N_o)}{N_v \times N_o} = \frac{N_v \times N_v}{N_v \times N_o} = \frac{N_v}{N_o} \quad (2)$$

which is conditioned in $N_v < N_o$ situation and straightforward to transform to other situations.

Action $a_t \in \mathcal{A}$, **State Transition** \mathcal{P} . In our hierarchical RL setting, *manager's* action is to generate abstract and intrinsic goal to its *workers*, and each *worker* needs to provide a ranking list of relevant real orders (OD) and fleet control served as fake orders (FM). Thus, the action of *worker* is defined as the weight vector for the ranking features. Changing an action of the *worker* means to change the weight vector for the ranking features (as will be further discussed in Section 6). Each timestep the whole multi-agent system produces a joint action for each *manager* and *worker* $a_t \in \mathcal{A}$, where $\mathcal{A} := \mathcal{A}_1 \times \dots \times \mathcal{A}_N$, which induces a transition in the environment according to the state transition $\mathcal{P}(s_{t+1}|s_t, a_t)$.

Reward \mathcal{R} . Like previous hierarchical RL settings [34], only *manager* interacts with the environment and receives feedback from it. This extrinsic reward function determines the direction of optimization and is proportional to both immediate profit and potential

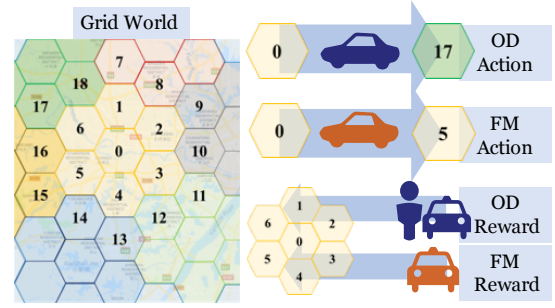


Figure 2: Illustration of the grid world and problem setting.

value; while the intrinsic reward is set to encourage the *worker* to follow the instruction from the *manager*. The details will be further discussed in Eq. (4) and Eq. (6).

More specifically, we give a simple example based on the above problem setting in Figure 2. At time $t = 0$, the *worker* agent 0 ranks available real orders and potential fake orders for fleet control, and selects the Selected-2 (as will be further discussed in Eq. (9)) options: a real order from grid 0 to grid 17, a fake order from grid 0 to grid 5. After the driver finished, the *manager* agent, whose sub-*workers* maintain the *worker* agent 0, received corresponding reward.

4 METHODOLOGIES

4.1 Overall Architecture

As shown in Figure 3, *CoRide* employs two layers of agents, namely the layer of *manager* agents and the layer of *worker* agents. Each agent is associated with a communication component for exchanging messages. As both agent and decision-making process conduct in a hierarchical way, multi-head attention mechanism served for communication is extended into multi-layer setting.

Compared with traditional one-to-one *manager-worker* control in hierarchical RL [34], we design one-to-many *manager-worker* pattern, extend the scale, and learn to collaborate on two layers of agents. The *manager* internally computes a latent state representation h_t^M as an input to the *manager-level* attention network, and outputs a goal vector g_t . The *worker* produces action and input for *worker-level* attention conditioned on its private observation o_t^W , peer message m_{t-1}^W , and the *manager's* goal g_t . The *manager-level* and *worker-level* attention networks share the same architecture

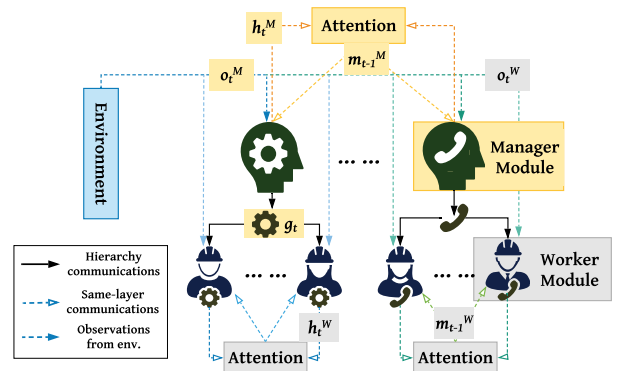


Figure 3: Overall architecture of *CoRide*.

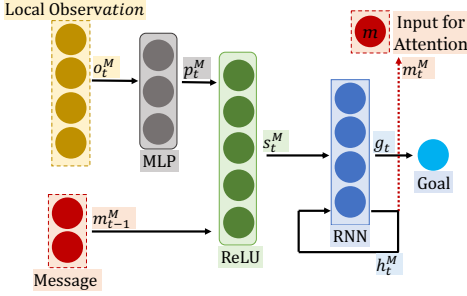


Figure 4: Manager Module.

introduced in Section 4.4. The details and training procedure for *manager* and *worker* are given in following parts.

4.2 Manager Module

The architecture of the *manager* module is presented in Figure 4. The *manager* network is a two layer Preceptron (MLP) and a dilated RNN [34]. Note that the structure of *CoRide* and formula of the RNN enable *manager* operate both at lower spatial resolution via taking joint observation of its *workers* and lower temporal resolution via dilated convolutional network [41].

At timestep t , the agent receives an observation o_t^M from the environment and feeds into the dilated RNN with peer messages m_{t-1}^M . Goal g_t and input for *manager*-level attention h_t^M are generated as output of the RNN, governed by the following equations:

$$h_t^M, \hat{g}_t = \text{RNN}(s_t^M, h_{t-1}^M; \theta^{rnn}); \quad g_t = \hat{g}_t / \|\hat{g}_t\| \quad (3)$$

where θ^{rnn} is the parameters of the RNN network. The environment responds with a new observation o_{t+1}^M and a scalar reward r_t . The goal of the agent is to maximize the discounted return $\mathcal{R}_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}^M$ with $\gamma \in [0, 1]$. Specifically, in the ride-hailing setting, we design our global reward taking both ADI and ORR into account, which can be formulated as:

$$r_t^M = r_{ADI} + r_{ORR} \quad (4)$$

where r_{ADI} denotes accumulated driver income, computed according to the price of each served order; while r_{ORR} encourages ORR, and is calculated with several correlative factors as:

$$r_{ORR} = \sum_{grid} (E - \bar{E})^2 + \sum_{area} D_{KL}(\mathcal{P}_t^o \| \mathcal{P}_t^v) \quad (5)$$

where E, \bar{E} are the *manager*'s entropy and global average entropy respectively. Area, different from grid, often means a certain region which needs to be taken more care of. In our experiment, we select several grids whose entropy largely differs from the average as the area. $D_{KL}(\mathcal{P}_t^o \| \mathcal{P}_t^v)$ denotes Kullback-Leibler (KL) divergence which shows the margin between vehicle and order distributions of certain area at timestep t . \mathcal{P}_t^o and \mathcal{P}_t^v are realized with Poisson distribution, a common distribution for vehicle routing [9] and arriving [16]. In practice, this distribution parameters can be estimated from real trip data by the mean and std of orders and vehicles in each grid at each timestep. Such a combined ORR reward design helps optimization both globally and locally.

4.3 Worker Module

We adopt the goal embedding from Feudal Networks (FuN) [34] in our *worker* framework (see Figure 5), where w_t is generated as goal-embedding vector via linear projection. At each timestep t , the agent receives an observation o_t^W from the environment and feeds into a regular RNN with peer message m_{t-1}^W . As Figure 5 shows, the output of RNN u_t^W together with w_t generates the primitive action - ranking weight vector ω_t .

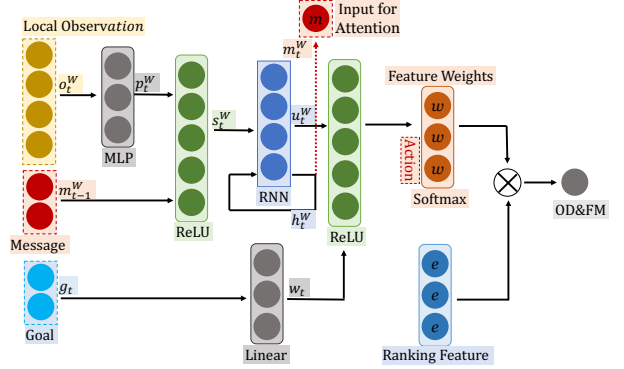


Figure 5: Worker Module.

Given that *worker* needs to be encouraged to follow the goal generated by its *manager*, we adopt the intrinsic reward proposed by [34], defined as:

$$r_t^W = \frac{1}{c} \sum_{i=1}^c d_{cos}(o_t^W - o_{t-i}^W, g_{t-i}), \quad (6)$$

where $d_{cos}(\alpha, \beta) = \alpha^T \beta / (|\alpha| \cdot |\beta|)$ is the cosine similarity between two vectors. Notice that g_t now represents an advantageous direction in the latent state space at a horizon c [34]. Such intrinsic reward design would provide directional shift for *workers* to follow.

Different from traditional FuN [34], procedure of our *worker* module produces action consists of two steps: (i) parameter generating, and (ii) action generating, inspired by [44]. We utilize state-specific scoring function f_{θ^W} in parameter generating setup to map the current state o_t^W to a list of weight vectors ω_t as:

$$f_{\theta^W} : o_t^W \rightarrow \omega_t, h_t^W \quad (7)$$

which is calculated based on neural network shown in Figure 5. In action generating setup, note that it is straightforward to extend linear relations with non-linear ones, we formulate that the scoring function parameter ω_t and the ranking feature e_i for order i as:

$$score_i = \omega_t^T e_i \quad (8)$$

The detailed formulation of e_i will be discussed in Section 5. Then, we build and add real orders and potential fleet control - repositioning to neighbor grids and staying at the current grid - as fake orders into item space \mathcal{I} . After computing scores for all available options in \mathcal{I} , instead of directly ranking and selecting Top- k items for order dispatching and fleet management, we adopt Boltzmann softmax selector to generate Selected- k items:

$$\text{Selected-}k = \frac{\exp(score_i/\tau)}{\sum_{i=1}^M \exp(score_i/\tau)} \quad (9)$$

where $k = \min(N_o, N_o)$, τ denotes temperature hyper-parameter to control the exploration rate, and M is the number of scored order candidates. In practice, we set the initial temperature as 1.0, then gradually reduce the temperature until 0.01 to limit exploration. This approach not only equips the action selection procedure with controllable exploration but also diversify the policy’s decision to avoid choosing groups of vehicles flected to the same destination.

Algorithm 1 *CoRide* for joint multi-scale OD & FM

Require: current observations o_t^M, o_t^W ; mutual communication messages m_{t-1}^M, m_{t-1}^W .

- 1: **for** each *manager* in grid world **do**
- 2: Generate g_t, h_t^M according to Eq. (3).
- 3: **for** each *worker* of the *manager* **do**
- 4: Generate ω_t, h_t^W according to Eq. (7).
- 5: Add orders and fleet control items to item space \mathcal{I} .
- 6: Rank items in \mathcal{I} according to Eq. (8).
- 7: Generate Selected- k items according to Eq. (9).
- 8: **end for**
- 9: *worker*-level attention mechanism generates m_t^W according to Eq. (12).
- 10: *manager* receives extrinsic reward r_t^M , and its *workers* receive intrinsic reward r_t^W according to Eq. (4) and Eq. (6) respectively.
- 11: **end for**
- 12: *manager*-level attention mechanism generates m_t^M according to Eq. (12).
- 13: Update parameters according to Algorithm 2.

4.4 Multi-head Attention for Coordination

Note that *manager* and *worker* share the same setting of multi-head attention mechanism, agent in this subsection can represent either of them. We utilize h_{t-1}^i to denote the cooperating information for i -th agent generated from RNN at $t-1$, and extend self-attention mechanism to learn to evaluate each available interaction as:

$$h_{t-1}^{ij} = (h_{t-1}^i W_T) \cdot (h_{t-1}^j W_S)^T \quad (10)$$

where $h_{t-1}^i W_T, h_{t-1}^j W_S$ are embedding of messages from target agent and source agent respectively. We can model h_{t-1}^{ij} as the value of communication between i -th agent and j -th agent. To retrieve a general attention value between source and target agents, we further normalize this value in neighborhood scope as:

$$\alpha_{t-1}^{ij} = \text{softmax}(h_{t-1}^{ij}) = \frac{\exp(h_{t-1}^{ij}/\iota)}{\sum_{j \in \mathcal{N}_i} \exp(h_{t-1}^{ij}/\iota)} \quad (11)$$

where \mathcal{N}_i is the neighborhood scope: the set of communication available for target agent, and ι denotes temperature factor. To jointly attend to the neighborhood from different representation subspaces at different grids, we leverage multi-head attention as in previous work [32, 33, 38, 42] to extend the observation as:

$$m_t^i = \sigma \left(W_q \cdot \left(\frac{1}{H} \sum_{n=1}^{n=H} \sum_{j \in \mathcal{N}_i} \alpha_{t-1}^{ijn} (h_{t-1}^j W_C^n) \right) + b_q \right) \quad (12)$$

where H is the number of attention heads, and W_T, W_S, W_C are multiple sets of trainable parameters. Thus, peer message m_t is generated and will be feed into the corresponding module to produce the cooperative information h_t . We present the overall *CoRide* for joint order dispatching and fleet management in Algorithm 1.

4.5 Training

As described in Algorithm 1, *managers* generate specific goals based on their observations and peer messages (line 2). The *workers* under the *manager* generate the weight vector according to private observation and sharing goal (line 4). We then build a general item space \mathcal{I} for order dispatching and fleet management (line 5), and rank items in \mathcal{I} (line 6). Considering that our action is conditional to the minimum of the number of vehicles and orders, we generate Selected- k items as the final action (line 7).

We extend learning approach from FuN [34] and HIRO [19] to train *manager* and *worker* module in the similar way. In *CoRide*, we utilize DDPG algorithm [14] to train the parameters for both *manager* and *worker* module for following reasons. Classically, the critic is designed to leverage an approximator, to learn an action-value function $Q(o_t, a_t)$, and to direct the actor updating its parameters. The optimal action-value function $Q^*(o_t, a_t)$ should follow the Bellman equation [3] as:

$$Q^*(o_t, a_t) = \mathbb{E}_{o_{t+1}} [r_t + \gamma \max_{a_{t+1}} Q^*(o_{t+1}, a_{t+1}) | o_t, a_t] \quad (13)$$

Algorithm 2 Parameters Training with DDPG

- 1: Randomly initialize Critic network $Q(m_{t+1}, o_t, a_t | \theta^Q)$ and actor $\mu(m_{t-1}, o_t | \theta^\mu)$ with weights θ^Q and θ^μ .
- 2: Initialize target network Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^Q, \theta^{\mu'} \leftarrow \theta^\mu$.
- 3: Initialize replay buffer R .
- 4: **for** each training episode **do**
- 5: **for** agent $i = 1$ to M **do**
- 6: $m_0 =$ initial message, $t = 1$.
- 7: **while** $t < T$ and $o_t \neq$ terminal **do**
- 8: Select the action $a_t = \mu(m_{t-1}, o_t | \theta^\mu)$ for active agent;
- 9: Receive reward r_t and new observation o_{t+1} ;
- 10: Generate message $m_t = \text{Attention}(h_{t-1}^0, h_{t-1}^1, \dots, h_{t-1}^K)$, where h_{t-1}^k is latent vector in RNN and K denotes the number of neighboring agents;
- 11: **end while**
- 12: Store episode $\{m_0, o_1, a_1, r_1, m_1, o_2, a_2, r_2, \dots\}$ in R .
- 13: **end for**
- 14: Sample a random minibatch of transitions $\mathcal{T} : \langle m_{t-1}, o_t, a_t, r_t, o_{t+1} \rangle$ from replay buffer R .
- 15: **for** each transition \mathcal{T} **do**
- 16: Set $y_t = r_t + \gamma Q'(m_t, o_{t+1}, \mu(m_t, o_{t+1} | \theta^{\mu'})) | \theta^{Q'}$;
- 17: Update Critic by minimizing the loss:
 $L(\theta^{Q'}) = (y_t - Q(m_{t-1}, o_t, a_t | \theta^Q))^2$;
- 18: Update Actor policy by maximizing the Critic:
 $J(\theta^\mu) = Q(m_{t-1}, o_t, a_t | \theta^Q) |_{a=\mu(m_{t-1}, o_t | \theta^\mu)}$;
- 19: Update communication component.
- 20: **end for**
- 21: **end for**

which requires $|\mathcal{A}|$ evaluations to select the optimal action. This prevents Eq. (13) to be adopted in real-world scenario, e.g. ride-hailing setting, with enormous state and action spaces. However, the actor architecture proposed in Section 4.3 generates a deterministic action for critic. Furthermore, Lillicrap et al. [14] proposed a flexible and practical method to use an approximator function to estimate the action-value function, i.e. $Q(o, a) \approx Q(o, a; \theta^\mu)$. In practice, we refer to leverage DQN: a neural network function approximator can be trained by minimizing a sequence of loss functions $L(\theta^\mu)$ as:

$$L(\theta^\mu) = \mathbb{E}_{s_t, a_t, r_t, o_{t+1}} [(y_t - Q(o_t, a_t; \theta^\mu))^2] \quad (14)$$

where $y_t = \mathbb{E}_{o_{t+1}} [r_t + \gamma Q'(o_{t+1}, a_{t+1}; \theta^\mu) | o_t, a_t]$ is the target for the current iteration. According to the aforementioned analysis, the general training algorithm for the *manager* and *worker* module is presented in Algorithm 2.

5 SIMULATOR

The trial-and-error nature of reinforcement learning requires a dynamic simulation environment for training and evaluation. Thus, we adopt and extend the grid-based simulator designed by Lin et al. [15] to joint order dispatching and fleet management.

5.1 Data Description

The real world data provided by Didi Chuxing[†] includes order information and trajectories of vehicles in the central area of three large cities with millions of orders in four consecutive weeks. Data of each day contains million-level orders and tens of thousands vehicles in each city. The order information includes order price, origin, destination, and duration. The trajectories contain the positions (latitude and longitude) and status (on-line, off-line, on-service) of all vehicles every few seconds. As the radius of grid is approximate 1.3 kilometers, the central area of the city is covered by a hexagonal grids world consisting of 182, 126, 112 grids in three cities respectively. In order to adapt to the grid-based simulator, we utilize unique gridID to represent position information.

5.2 Simulator Design

In the grid-based simulator, the city is covered by a hexagonal grid-world as illustrated in Figure 2. At each timestep t , the simulator provides an observation o_t with a set of idle vehicles and a set of available orders including real orders and aforementioned fake orders for fleet control. All such fake orders share the same attributes as real orders, except that some of attributes are set stationary like price. All these real orders are generated by bootstrapping from real-world dataset introduced above. More specifically, suppose the current timestep of simulator is t , we randomly sample real orders occurring in the same period, i.e. happening between $t_\Delta \times t$ to $t_\Delta \times (t + 1)$, where t_Δ denotes timestep interval. In practice, we set *sampling rate* 100%. Like orders, vehicles are set online and offline alternatively according to a distribution learned from real-world dataset via maximum likelihood estimation. Each order feature, i.e. ranking feature e_i in Eq. (8), includes the origin gridID, the destination gridID, price, duration and the type of order indicating real or fake order; while each vehicle takes its gridID as a feature, and

[†]Similar dataset supported by Didi Chuxing can be found via GAIA open dataset (<https://outreach.didichuxing.com/research/opendata/en/>).

vehicles located at the same grid are regarded as homogeneous. Moreover, as the travel distance between neighboring grids is approximately 1.3 kilometers and timestep interval t_Δ is 10 minutes, we assume that vehicles will not automatically move to other grids before taking another order. The ride-hailing platform then provides an optimal list of vehicle-order pairs according to current policy. After receiving the list, the simulator will return a new observation o_{t+1} and a list of order fees. Stepping on this feedback, rewards r_t for each agent will be calculated and the corresponding record (o_t, a_t, r_t, o_{t+1}) will be stored into a replay buffer. The whole network parameters will be updated using a batch of samples from replay buffer.

The effectiveness of the grid-based simulator is evaluated based on the calibration against the real data in term of the most important performance measurement: accumulated driver income (ADI) [15]. The coefficient of determination r^2 between simulated ADI and real ADI is 0.9331 and the Pearson correlation is 0.9853 with p -value $p < 0.00001$.

6 EXPERIMENT

In this section, we conduct extensive experiments to evaluate the effectiveness of our proposed method in joint order dispatching and fleet management environment. Given that there are no published methods fitting our task. Thus, we first compare our proposed method with other models either widely used in the industry or published as academic papers based on a single order dispatching environment. Then, we further evaluate our proposed method in joint setting and compare with its performance in single setting.

6.1 Compared Methods

As discussed in [15], learning-based methods, currently regarded as state-of-the-art methods, usually outperform rule-based methods. Thus, we employ 6 learning-based methods and random method as the benchmark for comparison in our experiments.

- **RAN**: A random dispatching algorithm considering no additional information. It only assigns idle vehicles with available orders randomly at each timestep.
- **DQN**: Li et al. [13] conducted action-value function approximation based on Q -network. The Q -network is parameterized by a MLP with four hidden layers and we adopt the ReLU activation between hidden layers and to transform the final linear output of Q -network.
- **MDP**: Xu et al. [39] implemented dispatching through a learning and planning approach: each vehicle-order pair is valued in consideration of both immediate rewards and future gains in the learning step, and dispatch is solved using a combinatorial optimizing algorithm in planning step.
- **DDQN**: Wang et al. [36] introduced a double-DQN with spatial-temporal action search. The network architecture is similar to the one described in DQN except that a selected action space is utilized and network parameters are updated via double-DQN.
- **MFOD**: Li et al. [13] modeled the order dispatching problem with MFRL [40] and simplified the local interactions by taking an average action among neighborhoods.
- **CoRide**: Our proposed model as detailed in Section 4.

Table 1: Performance comparison of competing methods in terms of ADI and ORR with respect to the performance of RAN. For a fair comparison, the random seeds that control the dynamics of the environment are set to be the same across all methods.

City	City A		City B		City C	
Metrics	Normalized ADI	Normalized ORR	Normalized ADI	Normalized ORR	Normalized ADI	Normalized ORR
DQN	+5.71% ± 0.02%	+2.67% ± 0.01%	+6.30% ± 0.01%	+3.01% ± 0.02%	+6.11% ± 0.02%	+3.04% ± 0.01%
MDP	+7.11% ± 0.05%	+2.71% ± 0.03%	+7.89% ± 0.05%	+3.13% ± 0.04%	+7.53% ± 0.03%	+3.19% ± 0.03%
DDQN	+6.68% ± 0.04%	+3.19% ± 0.04%	+7.75% ± 0.06%	+4.06% ± 0.05%	+7.62% ± 0.04%	+4.58% ± 0.05%
MFOD	+6.62% ± 0.03%	+3.71% ± 0.02%	+7.91% ± 0.04%	+4.01% ± 0.02%	+7.32% ± 0.02%	+4.60% ± 0.01%
CoRide-	+9.27% ± 0.04%	+4.23% ± 0.03%	+8.73% ± 0.03%	+4.35% ± 0.02%	+9.06% ± 0.03%	+4.23% ± 0.04%
CoRide	+9.80% ± 0.04%	+4.81% ± 0.05%	+8.94% ± 0.06%	+4.89% ± 0.04%	+9.23% ± 0.05%	+5.19% ± 0.04%

- **CoRide-**: In order to further evaluate performance for hierarchical setting and agent communication, we set *CoRide* without multi-head attention mechanism as one of the baselines.

6.2 Result Analysis

For all learning methods, following [13], we run 20 episodes for training, store the trained model periodically, and conduct the evaluation on the stored model with 5 random seeds. We compare the performance of different models regarding two criteria, including ADI, computed as the total income in a day, and ORR, calculated by the number of orders taken divided by the number of orders generated.

Experimental Results and Analysis. As shown in Table 1, the performance surpasses the state-of-the-art models like DDQN and industry deployed model like MDP. DDQN along with DQN mainly limits in lack of interaction and cooperation in the multi-agent environment. MDP mainly focuses on order price but ignores other features of order like duration, which makes against finding a balance between getting higher income per order and taking more orders. Instead, our proposed method achieves higher growths in term of ADI not only by considering every feature of each order concurrently but through learning to collaborate hierarchically. MFOD tries to capture dynamic demand-supply variations by propagating many local interactions between vehicles and the environment among mean field. Note that the number and information of available grid are relatively stationary while the number and feature of active vehicles are more dynamic. Thus, *CoRide*, which takes grid as agent, is more likely and easier to learn to cooperate from interaction between agents.

Apart from cooperation, multi-head attention network also enables *CoRide* to capture demand-supply dynamics from both district (*manager*) and grid (*worker*) scale (as will be further discussed in Figure 6). Such a novel combined scale setting facilitates *CoRide* both effectively and efficiently.

Visualization Analysis. Except for quantitative results, we also analyze whether the learned multi-head attention network can capture the demand-supply relation (see Figure 6(b)) through visualization. As shown in Figure 3, our communication mechanism conducts in a hierarchical way: attention among the *managers* communicates and learns to collaborate abstractly and globally while peers in *worker*-layer operate and determine key grid locally.

The values of several example *managers* and a group of *workers* belonging to the same *manager* are visualized in Figure 6(a). By

taking a closer look at Figure 6, we can observe that the area with high demand-supply indeed centralized at certain places, which has been well captured in *manager*-scale. Such district-level attention value allows precious vehicles to be dispatched efficiently in a global view. Apart from *manager*-scale one, multi-head attention network also provides *worker*-scale attention value, which focuses on local allocation. Stepping on this multi-scale dispatching system design, *CoRide* could operate as a microscope, where coarse and fine focuses work together to obtain precise action.

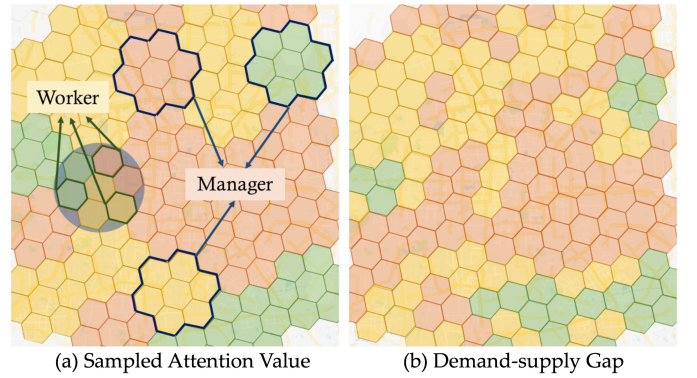


Figure 6: Sampled attention value and demand-supply gap in the city center during peak hours. Grids with more orders or higher attention value are shown in red (in green if opposite) and the gap is proportional to the shade of colors.

Ablation Study. In this subsection, we evaluate the effectiveness of components of *CoRide*. Notice that *manager* and *worker* modules serve as key components and are integrated through the hierarchical multi-agent architecture, as Figure 3 shows. Thus, we choose to investigate the performance of multi-head attention network here and set *CoRide-* as a variation of proposed method. As shown in the last two rows in Table 1, *CoRide-* achieves significant advantages over the aforementioned baselines, especially in City A. Similar results occur with *CoRide*. This phenomenon can be explained from the fact that City A is the largest one according to Section 5.1, which requires frequent and large numbers of transportations among regions. Multi-scale guidance via multi-head attention network and hierarchical multi-agent architecture is therefore potentially helpful, especially at a large-scale case.

Case Study. The above experimental results show that the success rate of our model is significantly better than others in single

Table 2: Performance comparison of competing methods in terms of AST and TNF with three different *discounted rates* (DR). The numbers in Trajectory denote gridID in Figure 7 and its color denotes the district it located in. O and W mean the vehicle is On-service and Waiting at the current grid. Also, we use underlined number to present fleet management.

DR	20%			30%			40%		
Metrics	Trajectory	AST	TNF	Trajectory	AST	TNF	Trajectory	AST	TNF
RES	13,9,W,14,W,13,8,2,7,11	8	8	13,W,14,W,W,13,8,W,7,11	6	6	13,W,14,W,W,W,W,19,O,9	5	4
REV	O,O,15,W,20,O,O,O,O,11	9	3	O,O,15,W,W,W,O,O,O,11	7	2	O,O,15,W,W,W,20,W,O,14	6	3
CoRide	13,9,W,O,0,4,O,2,O,5	9	6	13,W,O,11,W,W,O,O,0,5	7	4	13,W,W,O,17,W,W,O,0,3	6	3
CoRide+	13,9,W,O,0,4,O,2,O,5	9	6	<u>8,3,0,2,O,4,O,2,O,5</u>	8	5	<u>8,3,0,2,O,4,O,2,O,5</u>	8	5

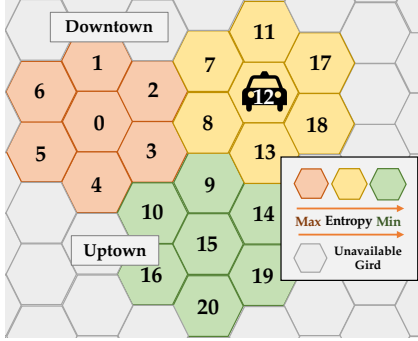


Figure 7: Illustration of the grid world in case study, where color of the grids denote their entropy.

dispatching task. In order to evaluate the performance of *CoRide* in joint dispatching and repositioning. Also, to further differ the formulations of the models, we constructed a synthetic dataset containing 3 districts with 21 grids, as showed in Figure 7. All these synthetic datasets are obtained via sampling real-world dataset supported by Didi Chuxing. More concretely, order distributions of all grids are sampled from the average distribution in real world dataset. Namely, order distributions in each grid are homogeneous. In order to differ downtown areas from uptown areas, we introduce *sampling rate* here. The *sampling rate* for each grid denotes popular rate in the real world. We set downtown (red grids in Figure 7) with stationary *sampling rate* 100%. The other regions are sampled with *sampling rate* := $100\% - \text{discounted rate}$ for yellow district and $100\% - 2 \times \text{discounted rate}$ for green district. Specifically we set *discounted rate* as 20%, 30% and 40% respectively, and further verify our proposed model by comparing against following 3 methods:

- **RES:** This response-based method aims to achieve higher ORR, which corresponds to Total Number of Finished orders (TNF) in this section. Orders with short duration will gain high priority to get dispatched first. Once there are multiple orders with the same estimated trip time, then orders with higher prices will be served first.
- **REV:** The revenue-based algorithm focuses on a higher ADI, which corresponds to Accumulated on-Service Time (AST) in this section, by assigning vehicles to high price orders. Following the similar principle as described above, the price and duration of orders will be considered as primary and secondary factors respectively.
- **CoRide+:** To distinguish *CoRide* running in different environment: single order dispatching, and joint order dispatching and fleet management, we sign the former one *CoRide* and latter one *CoRide+*.

In order to analyze these performances in a more straightforward way, we mainly employ rule-based methods here. Also, we introduce AST calculated as the accumulated on-service time and TNF computed as the total number of finished orders as the new metrics corresponding to ADI and ORR respectively. In order to further analyze these performances in a long-term way, we select one vehicle starting at grid 12, trace 10 timesteps and record its trajectory (as Figure 7 shows), then conclude these results in Table 2. Although we only record the first 10 timesteps, we can observe that our proposed methods, both *CoRide+* and *CoRide*, are guiding the vehicle to regions with larger entropy. This is benefit from architecture where the state of both *manager* and *worker*, and ranking feature e_i take the grid information into consideration. In contrast, other methods greedily optimize either AST (ADI) or TNF (ORR) and ignore these information. After taking a close look at Table 2, we can find that *CoRide* and *CoRide+* share the same trajectory on *discounted rate* 20% and differ greatly when *discounted rate* moves to 30% and 40%. This can be explained by regarding *CoRide+* as a combined design between our proposed model *CoRide* and joint order dispatching and fleet management setting. Namely, *CoRide* is actually a special case of *CoRide+*, where fleet management is unable. Equipped with fleet management, *CoRide+* allows the vehicle move to and serve order in the hotspots more directly than *CoRide*. Also, when *discounted rate* varies from 20% to 40%, fleet management enables *CoRide+* with better adaptation and achieve stable performance, even can ignore the dynamics of order distributions in some cases.

According to aforementioned analysis, we can conclude that (i) *CoRide+* achieves not only the state-of-the-art but a more stable result benefiting from joint order dispatching and fleet management setting; (ii) both *CoRide* and *CoRide+* can direct the vehicle to grids with larger entropy via taking grid information into consideration.

7 CONCLUSION AND FUTURE WORK

In this paper, we proposed *CoRide*, a hierarchical multi-agent reinforcement learning solution to combine order dispatching and fleet management for multi-scale ride-hailing platforms. The results on multi-city real-world data as well as analytic synthetic data show that our proposed algorithm achieves (i) a higher ADI and ORR than aforementioned methods, (ii) a multi-scale decision-making process, (iii) a hierarchical multi-agent architecture in the ride-hailing task and (iv) a more stable method at different cases. Note that *CoRide* could achieve fully decentralized execution and incorporate closely with other geographical information based model like estimating time of arrival (ETA) [35] theoretically. Thus, it's interesting to

conduct further evaluation and investigation. Also, we notice that applying hierarchical reinforcement learning in real-world scenarios is very challenge and our work is just a start. There is much work for future research to improve both stability and performance of hierarchical reinforcement learning methods on real-world tasks.

Acknowledgments. The corresponding author Weinan Zhang thanks the support of National Natural Science Foundation of China (Grant No. 61702327, 61772333, 61632017). We would also like to thank our colleague in DiDi for constant support and encouragement.

REFERENCES

- [1] Sanjeevan Ahilan and Peter Dayan. 2019. Feudal multi-agent hierarchies for cooperative reinforcement learning. *arXiv preprint arXiv:1901.08492* (2019).
- [2] Pierre-Luc Bacon, Jean Harb, and Doina Precup. 2017. The Option-Critic Architecture. In *AAAI* 1726–1734.
- [3] Richard Bellman. 2013. *Dynamic programming*. Courier Corporation.
- [4] Christian Daniel, Gerhard Neumann, and Jan Peters. 2012. Hierarchical relative entropy policy search. In *Artificial Intelligence and Statistics*. 273–281.
- [5] Peter Dayan and Geoffrey E Hinton. 1993. Feudal reinforcement learning. In *Advances in neural information processing systems*. 271–278.
- [6] Thomas G Dietterich. 2000. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research* 13 (2000), 227–303.
- [7] Carlos Florensa, Yan Duan, and Pieter Abbeel. 2017. Stochastic neural networks for hierarchical reinforcement learning. *arXiv preprint arXiv:1704.03012* (2017).
- [8] Kevin Frans, Jonathan Ho, Xi Chen, Pieter Abbeel, and John Schulman. 2017. Meta learning shared hierarchies. *arXiv preprint arXiv:1710.09767* (2017).
- [9] Gianpaolo Ghiani, Francesca Guerriero, Gilbert Laporte, and Roberto Musmanno. 2003. Real-time vehicle routing: Solution concepts, algorithms and parallel computing strategies. *European Journal of Operational Research* 151, 1 (2003), 1–11.
- [10] Xiangyu Kong, Bo Xin, Fangchen Liu, and Yizhou Wang. 2017. Effective master-slave communication on a multiagent deep reinforcement learning system. In *Hierarchical Reinforcement Learning Workshop at the 31st Conference on NIPS, Long Beach, CA, USA*.
- [11] Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. 2016. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *Advances in neural information processing systems*. 3675–3683.
- [12] Andrew Levy, George Konidaris, Robert Platt, and Kate Saenko. 2018. Learning multi-level hierarchies with hindsight. (2018).
- [13] Minne Li, Yan Jiao, Yaodong Yang, Zhichen Gong, Jun Wang, Chenxi Wang, Guobin Wu, Jieping Ye, et al. 2019. Efficient Ridesharing Order Dispatching with Mean Field Multi-Agent Reinforcement Learning. *arXiv* (2019).
- [14] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).
- [15] Kaixiang Lin, Renyu Zhao, Zhe Xu, and Jiayu Zhou. 2018. Efficient Large-Scale Fleet Management via Multi-Agent Deep Reinforcement Learning. *arXiv preprint arXiv:1802.06444* (2018).
- [16] Dominique Lord, Simon P Washington, and John N Ivan. 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention* 37, 1 (2005), 35–46.
- [17] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).
- [18] James Munkres. 1957. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics* 5, 1 (1957), 32–38.
- [19] Ofir Nachum, Shane Gu, Honglak Lee, and Sergey Levine. 2018. Data-Efficient Hierarchical Reinforcement Learning. *arXiv preprint arXiv:1805.08296* (2018).
- [20] Ofir Nachum, Shixiang Gu, Honglak Lee, and Sergey Levine. 2018. Near-Optimal Representation Learning for Hierarchical Reinforcement Learning. *arXiv preprint arXiv:1810.01257* (2018).
- [21] Takuma Oda and Yulia Tachibana. 2018. Distributed Fleet Control with Maximum Entropy Deep Reinforcement Learning. (2018).
- [22] Doina Precup. 2000. *Temporal abstraction in reinforcement learning*. University of Massachusetts Amherst.
- [23] Martin Riedmiller, Roland Hafner, Thomas Lampe, Michael Neunert, Jonas De-Grave, Tom Van de Wiele, Volodymyr Mnih, Nicolas Heess, and Jost Tobias Springenberg. 2018. Learning by Playing-Solving Sparse Reward Tasks from Scratch. *arXiv preprint arXiv:1802.10567* (2018).
- [24] Kiam Tian Seow, Nam Hai Dang, and Der-Horng Lee. 2010. A collaborative multiagent taxi-dispatch system. *IEEE Transactions on Automation Science and Engineering* 7, 3 (2010), 607–616.
- [25] Olivier Sigaud and Freek Stulp. 2018. Policy Search in Continuous Action Domains: an Overview. *arXiv preprint arXiv:1803.04706* (2018).
- [26] Hugo P Simao, Jeff Day, Abraham P George, Ted Gifford, John Nienow, and Warren B Powell. 2009. An approximate dynamic programming algorithm for large-scale fleet management: A case application. *Transportation Science* 43, 2 (2009), 178–197.
- [27] Matthijs TJ Spaan. 2012. Partially observable Markov decision processes. In *Reinforcement Learning*. Springer, 387–414.
- [28] Martin Stolle and Doina Precup. 2002. Learning options in reinforcement learning. In *International Symposium on abstraction, reformulation, and approximation*. Springer, 212–223.
- [29] Richard S Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112, 1-2 (1999), 181–211.
- [30] Xiaocheng Tang and Zhiwei Qin. 2018. A Deep Value-network Based Approach for Multi-Driver Order Dispatching. *Technical Report* (2018).
- [31] Chen Tessler, Shahar Givony, Tom Zahavy, Daniel J Mankowitz, and Shie Mannor. 2017. A Deep Hierarchical Approach to Lifelong Learning in Minecraft. In *AAAI*, Vol. 3. 6.
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.
- [33] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv* (2017).
- [34] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. 2017. Feudal networks for hierarchical reinforcement learning. *arXiv preprint arXiv:1703.01161* (2017).
- [35] Zheng Wang, Kun Fu, and Jieping Ye. 2018. Learning to estimate the travel time. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 858–866.
- [36] Zhaodong Wang, Zhiwei Qin, Xiaocheng Tang, Jieping Ye, and Hongtu Zhu. 2018. Deep Reinforcement Learning with Knowledge Transfer for Online Rides Order Dispatching. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 617–626.
- [37] Chong Wei, Yinhu Wang, Xuedong Yan, and Chunfu Shao. 2018. Look-Ahead Insertion Policy for a Shared-Taxi System Based on Reinforcement Learning. *IEEE Access* 6 (2018), 5716–5726.
- [38] Hua Wei, Nan Xu, Huichu Zhang, Guanjie Zheng, Xinshi Zang, Chacha Chen, Weinan Zhang, Yanmin Zhu, Kai Xu, and Zhenhui Li. 2019. CoLight: Learning Network-level Cooperation for Traffic Signal Control. *arXiv preprint arXiv:1905.05717* (2019).
- [39] Zhe Xu, Zhixin Li, Qingwen Guan, Dingshui Zhang, Qiang Li, Junxiao Nan, Chunyang Liu, Wei Bian, and Jieping Ye. 2018. Large-Scale Order Dispatch in On-Demand Ride-Hailing Platforms: A Learning and Planning Approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 905–913.
- [40] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. 2018. Mean Field Multi-Agent Reinforcement Learning (ICML).
- [41] Fisher Yu and Vladlen Koltun. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* (2015).
- [42] Huichu Zhang, Siyuan Feng, Chang Liu, Yaoyao Ding, Yichen Zhu, Zihan Zhou, Weinan Zhang, Yong Yu, Haiming Jin, and Zhenhui Li. 2019. CityFlow: A Multi-Agent Reinforcement Learning Environment for Large Scale City Traffic Scenario. *arXiv preprint arXiv:1905.05217* (2019).
- [43] Lingyu Zhang, Tao Hu, Yue Min, Guobin Wu, Junying Zhang, Pengcheng Feng, Pinghua Gong, and Jieping Ye. 2017. A taxi order dispatch model based on combinatorial optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2151–2159.
- [44] Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Dawei Yin, Yihong Zhao, and Jiliang Tang. 2017. Deep Reinforcement Learning for List-wise Recommendations. *arXiv preprint arXiv:1801.00209* (2017).
- [45] Qingnan Zou, Guangtao Xue, Yuan Luo, Jiadi Yu, and Hongzi Zhu. 2013. A novel taxi dispatch system for smart city. In *International Conference on Distributed, Ambient, and Pervasive Interactions*. Springer, 326–335.