Intelligent Electric Vehicle Charging Recommendation Based on Multi-Agent Reinforcement Learning

Weijia Zhang[1†], Hao Liu[2*], Fan Wang[3], Tong Xu[1], Haoran Xin[1], Dejing Dou[2], Hui Xiong[4*]
[1] University of Science and Technology of China, [2] Business Intelligence Lab, Baidu Research,
[3] Baidu Inc., [4] Rutgers University
{wjzhang3,xinhaoran}@mail.ustc.edu.com, {liuhao30,wangfan04,doudejing}@baidu.com,
tongxu@ustc.edu.com, hxiong@rutgers.edu

ABSTRACT

Electric Vehicle (EV) has become a preferable choice in the modern transportation system due to its environmental and energy sustainability. However, in many large cities, EV drivers often fail to find the proper spots for charging, because of the limited charging infrastructures and the spatiotemporally unbalanced charging demands. Indeed, the recent emergence of deep reinforcement learning provides great potential to improve the charging experience from various aspects over a long-term horizon. In this paper, we propose a framework, named *Multi-Agent Spatio-Temporal Reinforcement Learning* (Master), for intelligently recommending public accessible charging stations by jointly considering various long-term spatiotemporal factors. Specifically, by regarding each charging station as an individual agent, we formulate this problem as a multi-objective multi-agent reinforcement learning task. We first develop a multi-agent actor-critic framework with the centralized attentive critic to coordinate the recommendation between geo-distributed agents. Moreover, to quantify the influence of future potential charging competition, we introduce a delayed access strategy to exploit the knowledge of future charging competition during training. After that, to effectively optimize multiple learning objectives, we extend the centralized attentive critic to multi-critics and develop a dynamic gradient re-weighting strategy to adaptively guide the optimization direction. Finally, extensive experiments on two real-world datasets demonstrate that Master achieves the best comprehensive performance compared with nine baseline approaches.

KEYWORDS

Charging station recommendation, multi-agent reinforcement learning, multi-objective optimization

ACM Reference Format:
Weijia Zhang[1†], Hao Liu[2*], Fan Wang[3], Tong Xu[1], Haoran Xin[1], Dejing Dou[2], Hui Xiong[4*]. 2021. Intelligent Electric Vehicle Charging Recommendation Based on Multi-Agent Reinforcement Learning. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442381.3449934

[*] Corresponding author.
[†] The research was done when the first author was an intern in Baidu Research under the supervision of the second author.

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3449934

1 INTRODUCTION

Due to the low-carbon emission and energy efficiency, Electric vehicles (EVs) are emerging as a favorable choice in the modern transportation system to meet the increasing environmental concerns [10] and energy insufficiency [30]. In 2018, there are over 2.61 million EVs on the road in China, and this number will reach 80 million in 2030 [48]. Although the government is expanding the publicly accessible charging network to fulfill the explosively growing on-demand charging requirement, it is still difficult for EV drivers to charge their vehicles due to the fully-occupied stations and long waiting time [3, 16, 35, 39]. Undoubtedly, such an unsatisfactory charging experience raises undesirable charging cost and inefficiency, and may even increase the EV driver's "range anxiety", which prevents the further prevalence of EVs. Therefore, it is appealing to provide intelligent charging recommendations to improve the EV drivers' charging experience from various aspects, such as minimize the charging wait time (CWT), reduce the charging price (CP), as well as optimize the charging failure rate (CFR) to improve the efficiency of the global charging network.

The charging recommendation problem distinguishes itself from traditional recommendation tasks [7, 42] from two perspectives. First, the number of charging spots in a target geographical area may be limited, which may induce potential resource competition between EVs. Second, depending on the battery capacity and charging power, the battery recharging process may block the charging spot for several hours. As a result, the current recommendation may influence future EV charging recommendations and produce a long-term effect on the global charging network. Previously, some efforts [1, 8, 35, 38, 39, 43] have been made for charging recommendations via greedy-based strategies by suggesting the most proper station for the current EV driver in each step concerning a single objective (*e.g.*, minimizing the overall CWT). However, such an approach overlooks the long-term contradiction between the space-constrained charging capacity and the spatiotemporally unbalanced charging demands, which leads to sub-optimal recommendations from a global perspective (*e.g.*, longer overall CWT and higher CFR).

Recently, Reinforcement Learning (RL) has shown great advantages in optimizing sequential decision problems in the dynamic environment, such as order dispatching for ride-hailing and shared-bike rebalancing [15, 17]. By interacting with the environment, the

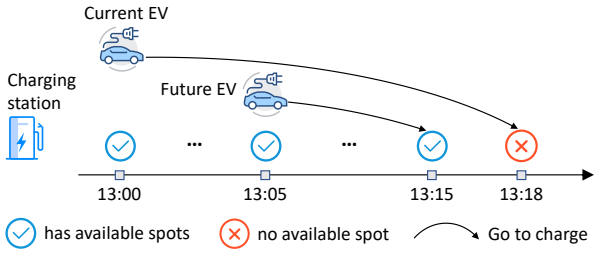


Figure 1: Illustrative example of future potential charging competition between EVs. The last available charging spot is preoccupied by a future EV, leading to extra CWT for current EV.

agent in RL learns the policy to achieve the global optimal long-term reward [4]. Therefore, it is intuitive for us to improve charging recommendations based on RL, with long-term goals such as minimizing the overall CWT, the average CP, and the CFR. However, there exist several technical challenges in achieving this goal.

The first challenge comes from the *large state and action space*. There are millions of EVs and thousands of publicly accessible charging stations in a metropolis such as Beijing. Directly learning a centralized agent system across the city requires handling large state and action space and high-dimensional environment, which will induce severe scalability and efficiency issues [2, 28]. Moreover, the "single point of failure" [25] in the centralized approach may fail the whole system [15]. As an alternative, an existing approach Wang et al. [36] tried to model a small set of vehicles as multiple agents and maximize the cumulative reward in terms of the number of served orders. However, in our charging recommendation task, most charging requests are ad-hoc and from non-repetitive drivers, which renders it impossible to learn a dedicated policy for each individual EV. To address the first challenge, we regard each charging station as an individual agent and formulate EV charging recommendation as a multi-agent reinforcement learning (MARL) task, and we propose a multi-agent actor-critic framework. In this way, each individual agent has a quantity-independent state action space, which can be scaled to more complicated environments and is more robust to other agents' potential failure.

The second challenge is the *coordination and cooperation in the large-scale agent system*. For a charging request, only one station will finally serve for charging. Different agents should be coordinated to achieve better recommendations. Moreover, the cooperation between agents is the key for long-term recommendation optimization. For example, consider a heavily occupied charging station with a large number of incoming charging requests. Other stations with sufficient available charging spots can help balance the charging demands via cross-agent cooperation. To tackle this challenge, we analogy the process of agents taking actions to a bidding game [47], and propose a tailor designed centralized attentive critic module to stimulate multiple agents to learn globally coordinated and cooperative policies.

The third challenge is the *potential competition of future charging requests*. As illustrated in Figure 1, the competition comes from the temporally distributed charging requests for the limited charging resource. In the real world, the potential competition of future charging requests may happen in arbitrary charging stations, leading to problems such as extra CWT and charging failure. However, it is hard to quantify the influence of such future charging requests in advance. To this end, we integrate the future charging competition information into centralized attentive critic module through a delayed access strategy and transform our framework to a centralized training with decentralized execution architecture to enable online recommendation. In this way, the agents can fully harness future knowledge in training phase and take actions immediately without requiring future information during execution.

Finally, it is challenging to *jointly optimize multiple optimization objectives*. As aforementioned, it is desirable to simultaneously consider various long-term goals, such as the overall CWT, average CP, and the CFR. However, these objectives may in different scales and lie on different manifold structures. Seamlessly optimizing multiple objectives may lead to poor convergence and induce sub-optimal recommendations for certain objectives. Therefore, we extend the centralized attentive critic module to multi-critics for multiple objectives and develop a dynamic gradient re-weighting strategy to adaptively guide the optimization direction by forcing the agents to pay more attention to these poorly optimized objectives.

Along these lines, in this paper, we propose the *Multi-Agent Spatio-Temporal Reinforcement Learning* (MASTER) framework for intelligent charging recommendation. Our major contributions are summarized as follows: (1) We formulate the EV charging recommendation problem as a MARL task. To the best of our knowledge, this is the first attempt to apply MARL for multi-objective intelligent charging station recommendations. (2) We develop the multi-agent actor-critic framework with centralized training decentralized execution. In particular, the centralized attentive critic achieves coordination and cooperation between multiple agents globally and exploits future charging competition information through a delayed access strategy during model training. Simultaneously, the decentralized execution of multiple agents avoids the large state action space problem and enables immediate online recommendations without requiring future competition information. (3) We extend the centralized attentive critic to multi-critics to support multiple optimization objectives. By adaptively guiding the optimization direction via a dynamic gradient re-weighting strategy, we improve the learned policy from potential local optima induced by dominating objectives. (4) We conduct extensive experiments on two real-world datasets collected from BEIJING and SHANGHAI. The results demonstrate our model achieves the best comprehensive performance against nine baselines.

2 PROBLEM DEFINITION

In this section, we introduce some important definitions and formalize the EV charging recommendation problem.

Consider a set of N charging stations $C = \{c^1, c^2, \ldots, c^N\}$, by regarding each day as an episode, we first define the charging request as below.

DEFINITION 1. ***Charging request***. *A charging request $q_t = \langle l_t, T_t, T_t^c \rangle \in Q$ is defined as the t-th request (i.e., step t) of a day. Specifically, l_t is the location of q_t, T_t is the real-world time of the step t, and T_t^c is the real-world time q_t finishes the charging request.*

We say a charging request is finished if it successfully takes the charging action or finally gives up (*i.e.*, charging failure). We further denote $|Q|$ as the cardinality of Q. In the following, we interchangeably use q_t to denote the corresponding EV of q_t.

Definition 2. ***Charging wait time (CWT)***. *The charging wait time is defined as the summation of travel time from the location l_t of charging request q_t to the target charging station c^i and the queuing time at c^i until q_t is finished.*

Definition 3. ***Charging price (CP)***. *The charging price is defined as the unit price per kilowatt hour (kWh). In general, the charging price is a combination of electricity cost and service fee.*

Definition 4. ***Charging failure rate (CFR)***. *The charging failure rate is defined as the ratio of the number of charging requests who accept our recommendation but fail to charge over the total number of charging requests who accept our recommendation.*

Problem 1. ***EV Charging Recommendation***. *Consider a set of charging requests Q in a day, our problem is to recommend each $q_t \in Q$ to the most proper charging station $rc_t \in C$, with the long-term goals of simultaneously minimizing the overall CWT, average CP, and the CFR for the $q_t \in Q$ who accept our recommendation.*

3 METHODOLOGY

In this section, we present the MARL formulation for the EV charging recommendation task and detail our Master framework with centralized training decentralized execution (CTDE). Moreover, we elaborate on the generalized multi-critic architecture for multiple objectives optimization.

3.1 MARL Formulation

We first present our MARL formulation for the EV Charging Recommendation task.

- **Agent** c^i. In this paper, we regard each charging station $c^i \in C$ as an individual agent. Each agent will make timely recommendation decisions for a sequence of charging requests Q that keep coming throughout a day with multiple long-term optimization goals.
- **Observation** o_t^i. Given a charging request q_t, we define the observation o_t^i of agent c^i as a combination of the index of c^i, the real-world time T_t, the number of current available charging spots of c^i (supply), the number of charging requests around c^i in the near future (future demand), the charging power of c^i, the estimated time of arrival (ETA) from location l_t to c^i, and the CP of c^i at the next ETA. We further define $s_t = \{o_t^1, o_t^2, \ldots, o_t^N\}$ as the state of all agents at step t.
- **Action** a_t^i. Given an observation o_t^i, an intuitional design for the action of agent c^i is a binary decision, *i.e.*, recommending q_t to itself for charging or not. However, because one q_t can only choose one station for charging, multiple agents' actions may be tied together and are difficult to coordinate. Inspired by the bidding mechanism [47], we design each agent c^i offers a scalar value to "bid" for q_t as its action a_t^i. By defining $u_t = \{a_t^1, a_t^2, \ldots, a_t^N\}$ as the joint action, q_t will be recommended to the agent with the highest "bid" value, *i.e.*, $rc_t = c^i$, where $i = \arg\max(u_t)$.

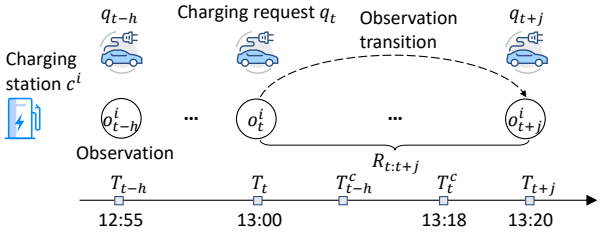


Figure 2: Illustrative example of transition in Master.

- **Transition**. For each agent c^i, its observation transition is defined as the transition from the current charging request q_t to the next charging request q_{t+j} after q_t is finished. Let's elaborate on this via an illustrative example as shown in Figure 2. Consider a charging request q_t arises at T_t (13:00). At this moment, each agent c^i takes action a_t^i based on its observation o_t^i and jointly decide the recommended station rc_t. After the request finish time T_t^c (13:18), the subsequent charging request q_{t+j} will arise at T_{t+j} (13:20). In this case, the observation transition of agent c^i is defined as $(o_t^i, a_t^i, o_{t+j}^i)$, where o_t^i is the current observation, o_{t+j}^i is the observation corresponding to q_{t+j}.
- **Reward**. In our MARL formulation, we propose a lazy reward settlement scheme (*i.e.*, return rewards when a charging request is finished), and integrate three goals into two natural reward functions. Specifically, if a charging request q_t succeeds in charging, then the environment will return the negative of CWT and negative of CP as the part of reward $r^{cwt}(s_t, u_t)$ and reward $r^{cp}(s_t, u_t)$ respectively. For the case that the CWT of q_t exceeds a threshold, the recommendation will be regarded as failure and environment will return much smaller rewards as penalty to stimulate agents reducing the CFR. Overall, we define two immediate rewards function for three goals as

$$r^{cwt}(s_t, u_t) = \begin{cases} -CWT, & charging\ success \\ \epsilon_{cwt}, & charging\ failure \end{cases}, \tag{1}$$

$$r^{cp}(s_t, u_t) = \begin{cases} -CP, & charging\ success \\ \epsilon_{cp}, & charging\ failure \end{cases}, \tag{2}$$

where ϵ_{cwt} and ϵ_{cp} are penalty rewards. All agents in our model share the unified rewards, means agents make the recommendation decisions cooperatively. Since the observation transition from o_t^i to o_{t+j}^i may cross multiple lazy rewards (*e.g.*, T_{t-h}^c and T_t^c as illustrated in Figure 2), we calculate the cumulative discounted reward by summing the rewards of all recommended charging requests $q_{t'}$ (*e.g.*, q_{t-h} and q_t) whose $T_{t'}^c$ is between T_t and T_{t+j}, denoted by

$$R_{t:t+j} = \sum_{T_t < T_{t'}^c \le T_{t+j}} \gamma^{(T_{t'}^c - T_t - 1)} r(s_{t'}, u_{t'}), \tag{3}$$

where γ is the discount factor, and $r(\cdot, \cdot)$ can be one of the two reward functions or the average of them depending on the learning objectives.

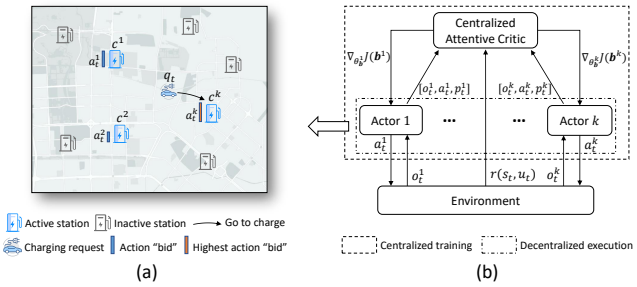


Figure 3: (a) Decentralized execution of active agents. The charging request will be recommended to the active charging station with the highest action "bid". (b) The centralized training decentralized execution process of MASTER.

3.2 Centralized Training Decentralized Execution

Centralized training decentralized execution is a class of methods in MARL to stimulate agents to learn coordinated policies and address the non-stationary environment problem [6, 11, 24]. In MASTER, the CTDE consists of three modules, the centralized attentive critic, the delayed access information strategy to integrate future charging competition, and the decentralized execution process. The advantages of CTDE for EV charging recommendation are two folds. On one hand, the centralized training process can motivate multiple agents to learn cooperation and other specific policies by perceiving a more comprehensive landscape and exploiting future information in hindsight. On the other hand, the execution process is fully decentralized without requiring complete information in the training phase, which guarantees efficiency and flexibility in online recommendation.

3.2.1 ***Centralized Attentive Critic.*** To motivate the agents to make recommendations cooperatively, we devise the multi-agent actor-critic framework with a centralized attentive critic for deterministic policies learning. A similar MARL algorithm with CTDE architecture is proposed in [24], which incorporates the full state s_t and joint action u_t of all agents into the critic to motivate agents to learn coordinated and cooperative policies. However, such an approach suffers from the large state and action space problem in our task.

In practice, the EVs tend to go to nearby stations for charging. Based on this fact, given a charging request q_t, we only activate the agents nearby q_t (*e.g.*, top-k nearest q_t) to take actions, denoted as C_t^a. We set other agents who are far away from q_t inactive and don't participate in the recommendation for q_t, as illustrated in Figure 3(a). In this way, only a small number of active agents are involved to learn cooperation for better recommendations. However, one intermediate problem is that the active agents of different q_t are usually different. To this end, we propose to use the attention mechanism which is permutation-invariant to integrate information of the active agents. Specifically, the attention mechanism automatically quantifies the influence of each active agents by

$$e_t^i = \mathbf{v}^\top \tanh\left(\mathbf{W}_a \left[o_t^i \oplus a_t^i \oplus p_t^i\right]\right), \tag{4}$$

$$\alpha_t^i = \frac{exp(e_t^i)}{\sum_{j \in C_t^a} exp(e_t^j)}, \tag{5}$$

where $\mathbf{v}$ and $\mathbf{W}_a$ are learnable parameters, $\oplus$ is the concatenation operation, and p_t^i are future information that will be detailed in Section 3.2.2. Once the influence weight α_t^i of each active agent $c^i \in C_t^a$ is obtained, we can derive the attentive representation of all active agents by

$$x_t = \text{ReLU}\left(\mathbf{W}_c \sum_{i \in C_t^a} \alpha_t^i \left[o_t^i \oplus a_t^i \oplus p_t^i\right]\right), \tag{6}$$

where $\mathbf{W}_c$ are learnable parameters.

Given the state s_t^a, joint action u_t^a, and future information p_t of active agents, the actor policy of each agent $c^i \in C_t^a$ can be updated by the gradient of the expected return following the chain rule, which can be written by

$$\nabla_{\theta_{\boldsymbol{b}}^i} J(\boldsymbol{b}^i) = \mathbb{E}_{s_t^a, p_t \sim D}\left[\nabla_{\theta_{\boldsymbol{b}}^i} \boldsymbol{b}^i(a_t^i|o_t^i) \nabla_{a_t^i} Q_{\boldsymbol{b}}(x_t)|_{a_t^i = \boldsymbol{b}^i(o_t^i)}\right], \tag{7}$$

where $\theta_{\boldsymbol{b}}^i$ are learnable parameters of actor policy $\boldsymbol{b}^i$ of agent c^i, D is the experience replay buffer containing the transition tuples $(s_t^a, u_t^a, p_t, s_{t+j}^a, p_{t+j}, R_{t:t+j})$. Then, each agent updates its policy by the gradients propagated back from the centralized attentive critic. The entire process is shown in Figure 3(b). As the centralized attentive critic perceived more complete information of all active agents, it can motivate the agents to learn policies in a coordinated and cooperative way. The centralized attentive critic $Q_{\boldsymbol{b}}$ is updated by minimizing the following loss:

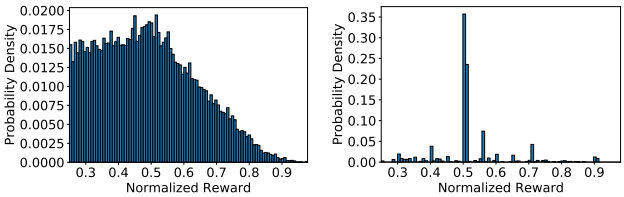
$$L(\theta_Q) = \mathbb{E}_{s_t^a, u_t^a, p_t, s_{t+j}^a, p_{t+j}, R_{t:t+j} \sim D}\left[\left(Q_{\boldsymbol{b}}(x_t) - y_t\right)^2\right], \tag{8}$$

$$y_t = R_{t:t+j} + \gamma^{(T_{t+j} - T_t)} Q'_{\boldsymbol{b}'}(x_{t+j})|_{a_{t+j}^i = \boldsymbol{b}'^i(o_{t+j}^i)}, \tag{9}$$

where θ_Q are the learnable parameters of critic $Q_{\boldsymbol{b}}$, $\boldsymbol{b}'^i$ and $Q'_{\boldsymbol{b}'}$ are the target actor policy of c^i and target critic function with delayed parameters $\theta_{\boldsymbol{b}}'^i$ and θ_Q'.

3.2.2 ***Integrate Future Charging Competition.*** The public accessible charging stations are usually first-come-first-served, which induces future potential charging competition between the continuously arriving EVs. Recommending charging requests without considering future potential competition will lead to extra CWT or even charging failure. However, it is a non-trivial task to incorporate the future competition since we can not know the precise amount of forthcoming EVs and available charging spots at a future step in advance.

In this work, we further extend the centralized attentive critic with a delayed access strategy to harness the future charging competition information, so as to enable the agents learning policies with future knowledge in hindsight. Specifically, consider a charging request q_t, without changing the execution process, we postpone the access of transition tuples until the future charging competition information in transition with respect to q_t is fully available. Concretely, we extract the accurate number of available charging spots of c^i at every next d minutes after T_t to reflect the future potential charging competition for q_t, denoted as I_t^i. Note that we erase the influence of q_t for I_t^i, and the number of available charging spots can be negative, means the number of EVs queuing at the station.



(a) Reward about CWT (b) Reward about CP

Figure 4: Distributions of two normalized rewards.

Overall, we obtain the future charging competition information of each c^i for q_t via a fully-connected layer

$$p_t^i = \text{ReLU}\left(\mathbf{W}_p I_t^i\right), \tag{10}$$

where $\mathbf{W}_p$ are the learnable parameters. The p_t^i is integrated into the centralized attentive critic (Eq. (4)~Eq. (6)) as the enhanced information to facilitate the agents' cooperative policy learning.

3.2.3 ***Decentralized Execution.*** The execution process is fully decentralized, by only invoking the learned actor policy with its own observation. Specifically, for a charging request q_t, the agent $c^i \in C_t^a$ takes action a_t^i based on its o_t^i by

$$a_t^i = \boldsymbol{b}^i(o_t^i). \tag{11}$$

And q_t will be recommended to the active agent with the highest a_t^i among all the actions of C_t^a. The execution of each agent is lightweight and does not require future charging competition information. More importantly, the large-scale agent system is fault-tolerant even part of the agents are failing.

3.3 Multiple Objectives Optimization

The goal of Electric Vehicle Charging Recommendation task is to simultaneously minimize the overall CWT, average CP and the CFR. We integrate these goals into two objectives corresponding to two reward functions, *i.e.*, r^{cwt} and r^{cp} as defined in Eq. (1) and Eq. (2). Figure 4 depicts the normalized reward distributions of r^{cwt} and r^{cp} by a random policy executing on a set of charging requests that charging success. As can be seen, the distribution of different objectives can be very different. More importantly, the optimal solution of different objectives may be divergent, *e.g.*, a cheaper charging station may be very popular and need a longer CWT. The above facts imply one policy that performs well on one objective may get stuck on another objective. A recommender that is biased to a particular objective is risky to induce unsatisfactory experience to most users.

A simple way to optimize multiple objectives is to average the reward of multiple objectives with a set of prior weights and maximize the combined reward as a single objective. However, such a static approach is inefficient for convergence and may induce a biased solution to a particular objective. To this end, we develop a dynamic gradient re-weighting strategy to self-adjust the optimization direction to adapt different training stages and enforce the policy to perform well on multiple objectives. Specifically, we extend the centralized attentive critic to multi-critics, where each critic is corresponding to a particular objective. Particularly, in our task, we learn two centralized attentive critics, $Q_{\boldsymbol{b}}^{cwt}$ and $Q_{\boldsymbol{b}}^{cp}$, which correspond to the expected returns of reward r^{cwt} and r^{cp}, respectively. Since the structure of two critics are identical, we only

Algorithm 1 MASTER algorithm

Randomly initialize critic networks $Q_{\boldsymbol{b}}^{cwt}$, $Q_{\boldsymbol{b}}^{cp}$ and each actor network $\boldsymbol{b}^i$ with weights θ_Q^{cwt}, θ_Q^{cp} and $\theta_{\boldsymbol{b}}^i$.
Initialize target networks $Q_{\boldsymbol{b}'}^{cwt'}$, $Q_{\boldsymbol{b}'}^{cp'}$ and $\boldsymbol{b}'^i$ with weights $\theta_Q^{cwt'} \leftarrow \theta_Q^{cwt}$, $\theta_Q^{cp'} \leftarrow \theta_Q^{cp}$, $\theta_{\boldsymbol{b}}'^i \leftarrow \theta_{\boldsymbol{b}}^i$.
Initialize objective-specific optimal networks $Q_{\boldsymbol{b}_{cwt}^*}^{cwt^*}$, $Q_{\boldsymbol{b}_{cp}^*}^{cp^*}$, $\boldsymbol{b}_{cwt}^{*i}$ and $\boldsymbol{b}_{cp}^{*i}$ with well-trained weights $\theta_Q^{cwt^*}$, $\theta_Q^{cp^*}$, $\theta_{\boldsymbol{b}_{cwt}}^{*i}$ and $\theta_{\boldsymbol{b}_{cp}}^{*i}$.
Initialize replay buffer D.
for $m = 1$ to max-iterations **do**
  Reset environment.
  for $t^- = 1$ to $|Q|$ **do**
    for agent $c^i \in C_{t^-}^a$ **do**
      Take action $a_{t^-}^i = \boldsymbol{b}^i(o_{t^-}^i)$ for charging request q_{t^-}.
    end for
    Store available delayed transition tuples $(s_{t'}^a, u_{t'}^a, p_{t'}, s_{t'+j'}^a, p_{t'+j'}, R_{t':t'+j'}^{cwt}, R_{t':t'+j'}^{cp})$ into D.
    Sample a random minibatch of M transitions $(s_t^a, u_t^a, p_t, s_{t+j}^a, p_{t+j}, R_{t:t+j}^{cwt}, R_{t:t+j}^{cp})$ from D.
    Set $y_t^{cwt} = R_{t:t+j}^{cwt} + \gamma^{(T_{t+j}-T_t)} Q_{\boldsymbol{b}'}^{cwt'}(x_{t+j})|_{a_{t+j}^i = \boldsymbol{b}'^i(o_{t+j}^i)}$.
    Set $y_t^{cp} = R_{t:t+j}^{cp} + \gamma^{(T_{t+j}-T_t)} Q_{\boldsymbol{b}'}^{cp'}(x_{t+j})|_{a_{t+j}^i = \boldsymbol{b}'^i(o_{t+j}^i)}$.
    Update Critic $Q_{\boldsymbol{b}}^{cwt}$ and $Q_{\boldsymbol{b}}^{cp}$ by minimizing the losses:
      $L(\theta_Q^{cwt}) = \frac{1}{M}\sum_t \left(Q_{\boldsymbol{b}}^{cwt}(x_t) - y_t^{cwt}\right)^2$.
      $L(\theta_Q^{cp}) = \frac{1}{M}\sum_t \left(Q_{\boldsymbol{b}}^{cp}(x_t) - y_t^{cp}\right)^2$.
    Compute β_t through Eq. (13) and Eq. (14).
    for agent $c^i \in C_t^a$ **do**
      Update actor by the sampled policy gradient:
        $\nabla_{\theta_{\boldsymbol{b}}^i} J(\boldsymbol{b}^i) \approx \frac{1}{M}\sum_t \Big(\beta_t \nabla_{\theta_{\boldsymbol{b}}^i} \boldsymbol{b}^i(a_t^i|o_t^i) \nabla_{a_t^i} Q_{\boldsymbol{b}}^{cwt}(x_t)$
        $+(1-\beta_t)\nabla_{\theta_{\boldsymbol{b}}^i} \boldsymbol{b}^i(a_t^i|o_t^i) \nabla_{a_t^i} Q_{\boldsymbol{b}}^{cp}(x_t)\Big)|_{a_t^i = \boldsymbol{b}^i(o_t^i)}$.
        $\theta_{\boldsymbol{b}}^i \leftarrow \theta_{\boldsymbol{b}}^i + \eta \nabla_{\theta_{\boldsymbol{b}}^i} J(\boldsymbol{b}^i)$.
      Update target actor networks:
        $\theta_{\boldsymbol{b}}'^i \leftarrow \tau\theta_{\boldsymbol{b}}^i + (1-\tau)\theta_{\boldsymbol{b}}'^i$.
    end for
    Update target critic networks:
      $\theta_Q^{cwt'} \leftarrow \tau\theta_Q^{cwt} + (1-\tau)\theta_Q^{cwt'}$.
      $\theta_Q^{cp'} \leftarrow \tau\theta_Q^{cp} + (1-\tau)\theta_Q^{cp'}$.
  end for
end for

present $Q_{\boldsymbol{b}}^{cwt}$ for explanation, which is formulated as

$$Q_{\boldsymbol{b}}^{cwt}(x_t) = \mathbb{E}_{s_{t+j}^a, p_{t+j}, R_{t:t+j}^{cwt} \sim E} \Big[R_{t:t+j}^{cwt} + \gamma^{(T_{t+j}-T_t)} Q_{\boldsymbol{b}'}^{cwt'}(x_{t+j})|_{a_{t+j}^i = \boldsymbol{b}'^i(o_{t+j}^i)} \Big], \tag{12}$$

where E denotes the environment, and $R_{t:t+j}^{cwt}$ is the cumulative discounted reward (defined in Eq. (3)) with respect to r^{cwt}.

To quantify the convergence degree of different objectives, we further define two centralized attentive critics associated with two objective-specific optimal policies with respect to reward r^{cwt} and r^{cp}, denoted as $Q_{\boldsymbol{b}_{cwt}^*}^{cwt^*}$ and $Q_{\boldsymbol{b}_{cp}^*}^{cp^*}$. The corresponding optimal policies of each c^i are denoted as $\boldsymbol{b}_{cwt}^{*i}$ and $\boldsymbol{b}_{cp}^{*i}$, respectively. Above objective-specific optimal policies and critics can be obtained by pre-training Master on a single reward. Then, we quantify the gap ratio between multi-objective policy and objective-specific optimal policy by

$$g_t^{cwt} = \frac{Q_{\boldsymbol{b}_{cwt}^*}^{cwt^*}(x_t)|_{a_t^i = \boldsymbol{b}_{cwt}^{*i}(o_t^i)} - Q_{\boldsymbol{b}}^{cwt}(x_t)|_{a_t^i = \boldsymbol{b}^i(o_t^i)}}{Q_{\boldsymbol{b}_{cwt}^*}^{cwt^*}(x_t)|_{a_t^i = \boldsymbol{b}_{cwt}^{*i}(o_t^i)}}. \tag{13}$$

The gap ratio g_t^{cp} can be derived in the same way. Intuitively, a larger gap ratio indicates a poorly-optimized objective and should be reinforced with a larger update weight, while a small gap ratio indicates a well-optimized objective can be fine-tuned with a smaller step size. Thus, we derive dynamic update weights to adaptively adjust the step size of the two objectives, which is learned by the Boltzmann softmax function,

$$\beta_t = \frac{exp(g_t^{cwt}/\sigma)}{exp(g_t^{cwt}/\sigma) + exp(g_t^{cp}/\sigma)}, \tag{14}$$

where σ is the temperature controls the adjustment sensitivity.

With the above two critics and adaptive update weights, the goal of each agent $c^i \in C_t^a$ is to learn an actor policy to maximize the following return,

$$J(\boldsymbol{b}^i) = \mathbb{E}_{s_t^a, p_t \sim D} \Big[\Big(\beta_t Q_{\boldsymbol{b}}^{cwt}(x_t) + (1-\beta_t) Q_{\boldsymbol{b}}^{cp}(x_t) \Big) |_{a_t^i = \boldsymbol{b}^i(o_t^i)} \Big]. \tag{15}$$

The complete learning procedure of Master is detailed in Algorithm 1. Note that for the consideration of scalability, we share the parameters of actor and critic networks by all agents.

4 EXPERIMENTS

4.1 Experimental setup

Table 1: Statistics of datasets.

Description	Beijing	Shanghai
# of charging stations	596	367
# of supplies records	38,620,800	23,781,600
# of charging requests	152,889	87,142

4.1.1 Data description. We evaluate Master on two real-world datasets, Beijing and Shanghai, which represent two metropolises in China. Both datasets are ranged from May 18, 2019, to July 01, 2019. All real-time availability (supplies) records, charging prices and charging powers of charging stations are crawled from a publicly accessible app [46], in which all charging spot occupancy information is collected by real-time sensors. The charging request data is collected through Baidu Maps API [22, 45]. We split each city as $1 \times 1\text{km}^2$ grids, and aggregate the number of future 15 minutes charging requests in the surrounding area (*i.e.*, the grid the station locates in and eight neighboring grids) as the future demands of the corresponding charging stations. We take the first 28 consecutive days data as the training set, the following three days data as validation set, and the rest 14 days data for testing. All real-world data are loaded into an EV Charging Recommendation simulator (refer to Appendix A and source code for details) for simulation. The spatial distribution of charging stations on Beijing is shown in Figure 5. The statistics of two datasets are summarized in Table 1. More statistical analysis results of the dataset can be found in Appendix B.

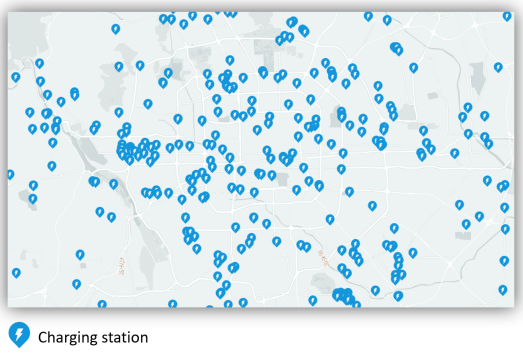


Figure 5: Spatial distribution of charging stations on Beijing.

4.1.2 Implementation details. All experiments are performed on a Linux server with 26-core Intel(R) Xeon(R) Gold 5117 CPU @ 2:00 GHz and NVIDIA Tesla P40 GPU. We set $d = 30$ minutes for charging competition modeling, set temperature $\sigma = 0.2$ for updated weights adjustment, and select discount factor $\gamma = 0.99$ for learning all RL algorithms. The actor networks and critic networks both consist of three linear network layers with dimension 64 and ReLU activation function for hidden layers. The replay buffer size is 1000, batch size is 32, and we set $\tau = 0.001$ for target networks soft update. We employ Adam optimizer for all learnable algorithms and set learning rate as 5e−4 to train our model. We carefully tuned major hyper-parameters of each baseline via a grid search strategy. All RL algorithms are trained to recommend the top-50 nearest charging stations for 60 iterations and chosen the best iteration by validation set for testing. Detailed settings of the simulator can be found in Appendix A.

4.1.3 Evaluation metrics. We define four metrics to evaluate the performance of our approach and baseline recommendation algorithms. Define Q^a as the set of charging requests who accept our recommendations. We further define $Q^s \subseteq Q^a$ as the set of charging requests who accept our recommendations and finally succeed in charging. $|Q^a|$ and $|Q^s|$ are the cardinalities of Q^a and Q^s, respectively.

To evaluate the overall charging wait time of our recommendations, we define Mean Charging Wait Time (MCWT) over all charging requests $q_t \in Q^a$:

$$\text{MCWT} = \frac{\sum_{q_t \in Q^a} CWT(q_t)}{|Q^a|},$$

where $CWT(q_t)$ is the charging wait time (in minute) of charging request q_t.

To evaluate the average charging price, we define Mean Charging Price (MCP) over all charging requests $q_t \in Q^s$:

$$\text{MCP} = \frac{\sum_{q_t \in Q^s} CP(q_t)}{|Q^s|},$$

where $CP(q_t)$ is the charging price (in CNY) of q_t.

We further define the Total Saving Fee (TSF) to evaluate the average total saving fees per day by comparing our recommendation algorithm with the ground truth charging actions:

$$\text{TSF} = \frac{\sum_{q_t \in Q^s} (RCP(q_t) - CP(q_t)) \times CQ(q_t)}{N_d},$$

where $RCP(q_t)$ is the charging price of ground truth charging action and $CQ(q_t)$ is the electric charging quantity of q_t, and N_d is the number of evaluation days. Note the TSF can be a negative number which indicates how many fees overspend comparing with the ground truth charging actions.

We finally define the Charging Failure Rate (CFR) to evaluate the ratio of charging failures in our recommendations:

$$\text{CFR} = \frac{|Q^a| - |Q^s|}{|Q^a|}.$$

4.1.4 Baselines. We compare our approach with the following nine baselines and one basic variant of Master:

- **Real** is the ground truth charging actions of charging requests.
- **Random** randomly recommends charging stations for charging.
- **Greedy-N** recommends the nearest charging station.
- **Greedy-P-5** recommends the least expensive charging station among the top-5 nearest stations.
- **Greedy-P-10** recommends the least expensive charging station among the top-10 nearest stations.
- **CDQN** [27] is a centralized deep q-network approach, all charging stations are controlled by a centralized agent. CDQN makes recommendation based on the state of all charging stations. The action-value function is a 3 layers MLP with dimension 256 and ReLU activation for hidden layers. The replay buffer size is 2000 and batch size is 64. The learning rate is set to 1e−3 and we use the decayed ϵ-greedy strategy for exploration.
- **CPPO** [31] is a centralized policy gradient approach, the recommendation mode is the same as CDQN. The policy network and value function network both consist of 3 layers MLP with dimension 256 and ReLU activation for hidden layers. The learning rate for policy and value networks are both set to 5e−4, the ϵ for clipping probability ratio is 0.2.
- **IDDPG** [19] is a straight-forward approach to achieve MARL using DDPG, where all agents are completely independent. The critic network approximates the expected return only bases on agent-specific observation and action. Other hyper-parameter settings are the same as Master.
- **MADDPG** [24] is a state-of-the-art algorithm for cooperative MARL. The actor takes action based on agent-specific observation, but the critic can access full state and joint action in training. The critic consists of 3 layers MLP with dimension 256 and ReLU activation for hidden layers. Other hyper-parameter settings are the same as Master. For scaling MADDPG to the large-scale agent system, we share actor and critic networks among all agents.
- **MASTER-ATT** is a basic variant of Master without charging competition information and multi-critic architecture.

Table 2: Overall performance evaluated by MCWT, MCP, TSF and CFR on Beijing.

Algorithm	Beijing			
	MCWT	MCP	TSF	CFR
Real	21.51	1.749	-	25.9%
Random	38.77	1.756	-447	52.9%
Greedy-N	20.27	1.791	-2527	31.3%
Greedy-P-5	23.40	1.541	9701	35.4%
Greedy-P-10	26.03	**1.424**	14059	39.9%
CDQN	19.24	1.598	9683	7.4%
CPPO	17.67	1.639	6707	5.3%
IDDPG	13.26	1.583	11349	2.2%
MADDPG	14.01	1.570	12033	4.7%
Master-ATT	11.30	1.562	12661	1.8%
Master	**10.46**	1.512	**16219**	**0.9%**

Table 3: Overall performance evaluated by MCWT, MCP, TSF and CFR on Shanghai.

Algorithm	Shanghai			
	MCWT	MCP	TSF	CFR
Real	19.31	1.787	-	16.7%
Random	38.62	1.826	-698	46.1%
Greedy-N	14.44	1.838	-1252	16.9%
Greedy-P-5	16.65	1.502	10842	13.8%
Greedy-P-10	19.60	**1.357**	**15649**	16.5%
CDQN	20.74	1.686	4650	6.5%
CPPO	18.84	1.750	1918	4.6%
IDDPG	14.02	1.562	9720	2.6%
MADDPG	13.63	1.553	10209	2.8%
Master-ATT	12.24	1.548	10630	2.2%
Master	**11.80**	1.497	12497	**1.5%**

4.2 Overall Performance

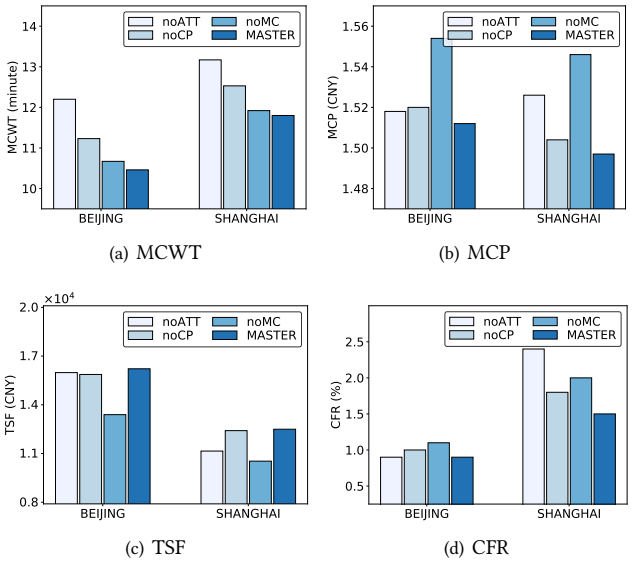
Table 2 and Table 3 report the overall results of our methods and all the compared baselines on two datasets with respect to our four metrics. As can be seen, overall, Master achieves the most well-rounded performance among all the baselines. Specifically, Master

reduces (51.4%, 13.6%, 96.5%) and (38.9%, 16.2%, 91.0%) for (MCWT, MCP, CFR) compared with the ground truth charging actions on Beijing and Shanghai, respectively. And through our recommendation algorithm, we totally help users saving 16, 219 and 12, 497 charging fees per day on Beijing and Shanghai. Besides, Master achieves (21.1%, 4.5%, 42.9%, 59.1%) and (25.3%, 3.7%, 34.8%, 80.9%) improvements for (MCWT, MCP, TSF, CFR) compared with two multi-agent baselines IDDPG and MADDPG on Beijing, and the improvements are (15.8%, 4.2%, 28.6%, 42.3%) and (13.4%, 3.6%, 22.4%, 46.4%) on Shanghai. All the above results demonstrate the effectiveness of Master. Look into Master-ATT and MADDPG, two MARL algorithms with centralized training decentralized execution. We can observe the Master-ATT with centralized attentive critic have all-sided improvements comparing with MADDPG, especially in the MCWT and CFR. This is because the critic in MADDPG integrates full state and joint action suffering from the large state and action space problem, while our centralized attentive critic can effectively avoid this problem in our task.

Looking further into the results, we observe centralized RL algorithms (*i.e.*, CDQN and CPPO) consistently perform worse than multi-agent approaches (*i.e.*, IDDPG, MADDPG, Master-ATT and Master), which validates our intuition that the centralized algorithms suffer from the high dimensional state and action space problem and hard to learn satisfying policies, while multi-agent approaches perform better because of their quantity-independent state action space in large-scale agents environment. We also observe that Greedy-P-10 has the best performance in MCP among all compared algorithms. This is no surprise, because it always recommends the least expensive charging stations without considering other goals, which causes that the Greedy-P-10 performs badly in MCWT and CFR. It is worth mentioning that Master unexpectedly exceeds Greedy-P-10 in TSF on Beijing for the reason that Master has a much lower CFR comparing with Greedy-P-10, more users are benefited from our recommendation.

4.3 Ablation Study

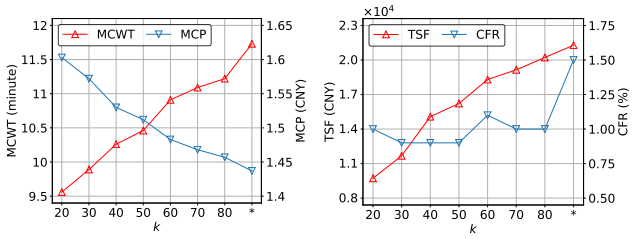
In this section, we conduct ablation studies on Master to further verify the effectiveness of each component. We evaluate the performance of Master and it's three variants for the four metrics on Beijing and Shanghai. Specifically, the three variants are (1) *noATT* removes centralized attentive critic, learning agents independently; (2) *noCP* removes the future potential charging competition information from centralized attentive critic; (3) *noMC* removes multi-critic architecture, learning policies by a single centralized attentive critic with the average combined reward. The ablation results are reported in Figure 6. As can be seen, removing centralized attentive critic (*noATT*) or charging competition information (*noCP*) have a greater impact on MCWT. This is because MCWT is highly related to the cooperation of agents and charging competition of EVs, whereas the removed modules exactly work on these two aspects. Removing multi-critic architecture (*noMC*) has a large performance degradation on MCP and TSF. This is because *noMC* suffers from the convergence problem and finally leads to performance variance on multiple objectives. All the above experimental results demonstrate the effectiveness of each component in Master.



(a) MCWT (b) MCP

(c) TSF (d) CFR

Figure 6: Ablation studies for MCWT, MCP, TSF and CFR on Beijing and Shanghai

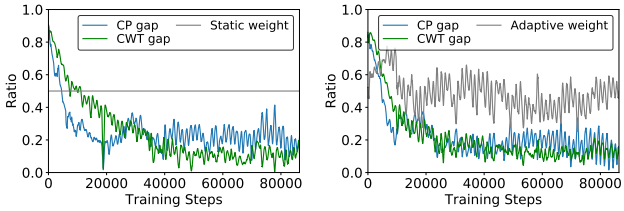


(a) MCWT and MCP (b) TSF and CFR

Figure 7: Effect of top-k nearest charging stations for recommendation performance.

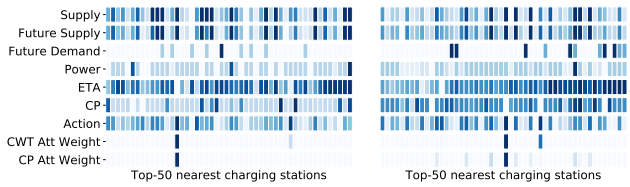
4.4 Effect of Recommendation Candidates

In this section, we study the effect of the number of top-k active agents for recommendations on Beijing. We vary k from 20 to the number of all the agents. The results are reported in Figure 7, where "$*$" denotes all the agents are activated for recommendation decision. As can be seen, with relaxing the recommendation constraint k, the MCWT increases while the MCP and TSF decrease, further validating the divergence of the optimal solutions for different objectives. This is possible because a more relaxed candidate number will expose more alternative economic but distant charging stations for recommendations. Even so, we should notice that the performance under the most strict constraint (*i.e.*, $k = 20$) or without constraint (*i.e.*, $k = *$) are varying in a small range, *i.e.*, the (MCWT, MCP) are (9.56, 1.603) and (11.73, 1.437), respectively, are not extreme and still acceptable for the online recommendation. The above results demonstrate that our model are well-rounded with different candidate numbers. It also inspires us that we can



(a) Static updated weight (b) Adaptive updated weight

Figure 8: Convergence curves by moving average for every 500 successive training steps.



(a) Case 1 (b) Case 2

Figure 9: Cases of the centralized attentive critics. For each factor (*i.e.*, each row), the darker color means a greater value. And each col denotes a top-50 nearest charging station for the charging request.

explore to make diversified recommendations that are biased to different objectives to satisfy personalized preferences in the future.

4.5 Convergence Analysis

In this section, we study the effect of our multi-objective optimization method by comparing the convergence of MASTER between the static and our adaptive strategy. Note the MASTER with static average updated weight is equivalent to MASTER-noMC (*i.e.*, average two rewards as a single reward for training). More details can be found in proposition C.1 in the Appendix C. The gap ratios (Eq. (13)) and dynamic weights (Eq. (14)) are reported in Figure 8. As can be seen in Figure 8(a), at the beginning of training, the gap ratio of the CP critic is decreasing faster than the CWT one. This is because the distribution of CP reward is more concentrated (as illustrated in Figure 4), it's easier to accumulate rewards comparing with CWT when agents with poor policies. In such a case, the static method keeps invariant update weights and results in slow convergence for CWT. In contrast, as shown in Figure 8(b), the dynamic re-weighting method in MASTER assigns a higher weight to the CWT objective when it falls behind the optimization of CP, and adjust them to a synchronous convergence pace. As the training step increase, the convergence of CWT overtakes the CP. This is because the distribution of CWT reward is more even, having larger room for improvement. And the optimization of CWT drags the CP objective to a sub-optimal point since the divergences between two objectives. However, the static method is helpless for such asynchronous convergence, but MASTER with the dynamic gradient re-weighting strategy weakens such asynchronism, and finally adjust them to be synchronous (*i.e.*, weight approaches 0.5) to learn well-rounded policies.

4.6 Case Study

In this section, we visualize the correlation among the attention weights and some input features of two centralized attentive critics (*i.e.*, Q_b^{cwt} and Q_b^{cp}), to qualitatively analysis the effectiveness of MASTER. Two cases are depicted in Figure 9. We can observe these two critics pay more attention to the charging stations with high action values. This makes sense, since these charging stations are highly competitive bidding participants. The charging request will be recommended to the charging station with the highest action, then environment returns rewards depending on this recommended station. Furthermore, we observe the action is highly correlated with factors such as supply, future supply, future demand, ETA and CP. A charging station with low ETA, CP and sufficient available charging spots (supplies) always has a high action value. Conversely, the charging stations with few available charging spots but high future demands usually derive a low action value for avoiding future charging competitions. The above observations validate the effectiveness of our model to mediate the contradiction between the space-constrained charging capacity and spatiotemporally unbalanced charging demands.

5 RELATED WORK

Charging Station Recommendation. In recent years, EVs have become an emerging choice in the modern transportation system due to their low-carbon emission and energy efficiency. Some efforts [1, 8, 14, 35–39, 44] have been made for charging station recommendation for EVs. In particular, most studies [1, 8, 35, 38, 39] focus on recommending charging station locations for EV drivers with the goal of time. For example, Guo et al. [8] propose to make recommendations of charging stations to minimize the time of traveling and queuing with a game-theoretical approach. Wang et al. [39] devise a fairness-aware recommender system to reduce the idle time based on the fairness constraints. Cao et al. [1] introduce the charging reservation information into the vehicle-to-vehicle system to facilitate the location recommendation. Different from charging location recommendation problem, another line of works [14, 36, 37, 44] investigate to handle more complicated scenarios, especially considering commercial benefits. Yuan et al. [44] propose a charging strategy that allows an electric taxi to get partially charged to meet the dynamic passenger demand. In particular, with the help of deep reinforcement learning which has been widely adopted for solving sequential decision-making problems, Wang et al. [36] design a multi-agent mean field hierarchical reinforcement learning framework by regarding each electric taxi as an individual agent to provide charging and relocation recommendations for electric taxi drivers, so as to maximize the cumulative rewards of the number of served orders. However, formulating each EV driver as an agent is not suitable for our problem, since most charging requests of a day in our task are ad-hoc and from non-repetitive drivers. Besides, the above works mainly focus on a single recommendation objective, which can not handle multiple divergent recommendation objectives simultaneously.

Multi-Agent Reinforcement Learning. Multi-agent Reinforcement Learning (MARL) is an emerging sub-field of reinforcement learning. Compared with traditional reinforcement learning, MARL expects the agents to learn to cooperate and compete with others. The simplest approach to realize a multi-agent system is learning agents independently [9, 33, 34]. However, the independent agents are not able to coordinate their actions, failing to achieve complicated cooperation[24, 26, 32]. A kind of natural approaches to achieve agents' cooperation is to learn communication among multiple agents [5, 12, 29, 32]. However, such approaches always lead to high communication overhead because of the large amount of information transfer. Alternatively, there are some works that employ the centralized training decentralized execution architecture to achieve agents' coordination and cooperation [6, 11, 24]. The advantage of such methods is that the agents can make decentralized execution without involving any other agents' information, which is lightweight and fault-tolerant in a large-scale agent system.

Multi-Agent Transportation Systems. In the past years, a few studies have successfully applied MARL for several intelligent transportation tasks. For example, [40, 41] use MARL algorithm for cooperative traffic signal control. [13, 15, 20, 49] apply MARL into the large-scale ride-hailing system to maximize the long-term benefits. Besides, MARL also has been adopted for shared-bike repositioning [17], express delivery-service scheduling [18]. However, we argue that our problem is inherently different from the above applications as a recommendation task, and the above approaches cannot be directly adopted for our problem.

6 CONCLUSION

In this paper, we investigated the intelligent EV charging recommendation task with the long-term goals of simultaneously minimizing the overall CWT, average CP and the CFR. We formulated this problem as a multi-objective MARL task and proposed a spatiotemporal MARL framework, Master. Specifically, by regarding each charging station as an individual agent, we developed the multi-agent actor-critic framework with centralized attentive critic to stimulate the agents to learn coordinated and cooperative policies. Besides, to enhance the recommendation effectiveness, we proposed a delayed access strategy to integrate the future charging competition information during model training. Moreover, we extend the centralized attentive critic to multi-critics with a dynamic gradient re-weighting strategy to adaptively guide the optimization direction of multiple divergent recommendation objectives. Extensive experiments on two real-world datasets demonstrated the effectiveness of Master compared with nine baselines.

ACKNOWLEDGMENTS

This research is supported in part by grants from the National Natural Science Foundation of China (Grant No.91746301, 71531001, 62072423).

REFERENCES

[1] Yue Cao, Tao Jiang, Omprakash Kaiwartya, Hongjian Sun, Huan Zhou, and Ran Wang. 2019. Toward pre-empted EV charging recommendation through V2V-based reservation system. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* (2019).

[2] Tianshu Chu, Jie Wang, Lara Codecà, and Zhaojian Li. 2019. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems* 21, 3 (2019), 1086–1095.

[3] Bowen Du, Yongxin Tong, Zimu Zhou, Qian Tao, and Wenjun Zhou. 2018. Demand-aware charger planning for electric vehicle sharing. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1330–1338.

[4] Wei Fan, Kunpeng Liu, Hao Liu, Pengyang Wang, Yong Ge, and Yanjie Fu. 2020. AutoFS: Automated Feature Selection via Diversity-aware Interactive Reinforcement Learning. In *IEEE 20th International Conference on Data Mining*.

[5] Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in neural information processing systems*. 2137–2145.

[6] Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual Multi-Agent Policy Gradients. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. 2974–2982.

[7] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. 2020. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[8] Tianci Guo, Pengcheng You, and Zaiyue Yang. 2017. Recommendation of geographic distributed charging stations for electric vehicles: A game theoretical approach. In *IEEE Power & Energy Society General Meeting*. 1–5.

[9] Jayesh K Gupta, Maxim Egorov, and Mykel Kochenderfer. 2017. Cooperative multi-agent control using deep reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems*. 66–83.

[10] Jindong Han, Hao Liu, Hengshu Zhu, Hui Xiong, and Dejing Dou. 2021. Joint Air Quality and Weather Prediction Based on Multi-Adversarial Spatiotemporal Networks. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*.

[11] Shariq Iqbal and Fei Sha. 2019. Actor-attention-critic for multi-agent reinforcement learning. In *International Conference on Machine Learning*. 2961–2970.

[12] Jiechuan Jiang and Zongqing Lu. 2018. Learning attentional communication for multi-agent cooperation. In *Advances in neural information processing systems*. 7254–7264.

[13] Jiarui Jin, Ming Zhou, Weinan Zhang, Minne Li, Zilong Guo, Zhiwei Qin, Yan Jiao, Xiaocheng Tang, Chenxi Wang, Jun Wang, et al. 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*. 1983–1992.

[14] Fanxin Kong, Qiao Xiang, Linghe Kong, and Xue Liu. 2016. On-line event-driven scheduling for electric vehicle charging via park-and-charge. In *IEEE Real-Time Systems Symposium*. 69–78.

[15] Minne Li, Zhiwei Qin, Yan Jiao, Yaodong Yang, Jun Wang, Chenxi Wang, Guobin Wu, and Jieping Ye. 2019. Efficient ridesharing order dispatching with mean field multi-agent reinforcement learning. In *The World Wide Web Conference*. 983–994.

[16] Yanhua Li, Jun Luo, Chi-Yin Chow, Kam-Lam Chan, Ye Ding, and Fan Zhang. 2015. Growing the charging station network for electric vehicles with trajectory data analytics. In *IEEE 31st International Conference on Data Engineering*. 1376–1387.

[17] Yexin Li, Yu Zheng, and Qiang Yang. 2018. Dynamic bike reposition: A spatio-temporal reinforcement learning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1724–1733.

[18] Yexin Li, Yu Zheng, and Qiang Yang. 2019. Efficient and Effective Express via Contextual Cooperative Reinforcement Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 510–519.

[19] 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).

[20] Kaixiang Lin, Renyu Zhao, Zhe Xu, and Jiayu Zhou. 2018. Efficient large-scale fleet management via multi-agent deep reinforcement learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1774–1783.

[21] Hao Liu, Jindong Han, Yanjie Fu, Jingbo Zhou, Xinjiang Lu, and Hui Xiong. 2020. Multi-modal transportation recommendation with unified route representation learning. *Proceedings of the VLDB Endowment* (2020), 342–350.

[22] Hao Liu, Ting Li, Renjun Hu, Yanjie Fu, Jingjing Gu, and Hui Xiong. 2019. Joint Representation Learning for Multi-Modal Transportation Recommendation. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. 1036–1043.

[23] Hao Liu, Yongxin Tong, Jindong Han, Panpan Zhang, Xinjiang Lu, and Hui Xiong. 2020. Incorporating Multi-Source Urban Data for Personalized and Context-Aware Multi-Modal Transportation Recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[24] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in neural information processing systems*. 6379–6390.

[25] Gary S Lynch. 2009. *Single point of failure*. Wiley Online Library.

[26] Laetitia Matignon, Guillaume J Laurent, and Nadine Le Fort-Piat. 2012. Independent reinforcement learners in cooperative Markov games: a survey regarding

coordination problems. *Knowledge Engineering Review* (2012), 1–31.
[27] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* (2015), 529–533.
[28] Georgios Papoudakis, Filippos Christianos, Arrasy Rahman, and Stefano V Albrecht. 2019. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737* (2019).
[29] Peng Peng, Quan Yuan, Ying Wen, Yaodong Yang, Zhenkun Tang, Haitao Long, and Jun Wang. 2017. Multiagent bidirectionally-coordinated nets for learning to play starcraft combat games. *arXiv preprint arXiv:1703.10069* (2017).
[30] George F Savari, Vijayakumar Krishnasamy, Jagabar Sathik, Ziad M Ali, and Shady HE Abdel Aleem. 2020. Internet of Things based real-time electric vehicle load forecasting and charging station recommendation. *ISA transactions* (2020), 431–447.
[31] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
[32] Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2016. Learning multiagent communication with backpropagation. In *Advances in neural information processing systems*. 2244–2252.
[33] Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente. 2017. Multiagent cooperation and competition with deep reinforcement learning. *PloS one* (2017).
[34] Ming Tan. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the 10th International Conference on Machine Learning*. 330–337.
[35] Zhiyong Tian, Taeho Jung, Yi Wang, Fan Zhang, Lai Tu, Chengzhong Xu, Chen Tian, and Xiang-Yang Li. 2016. Real-time charging station recommendation system for electric-vehicle taxis. *IEEE Transactions on Intelligent Transportation Systems* (2016), 3098–3109.
[36] Enshu Wang, Rong Ding, Zhaoxing Yang, Haiming Jin, Chenglin Miao, Lu Su, Fan Zhang, Chunming Qiao, and Xinbing Wang. 2020. Joint Charging and Relocation Recommendation for E-Taxi Drivers via Multi-Agent Mean Field Hierarchical Reinforcement Learning. *IEEE Transactions on Mobile Computing* (2020).
[37] Guang Wang, Xiaoyang Xie, Fan Zhang, Yunhuai Liu, and Desheng Zhang. 2018. bCharge: Data-driven real-time charging scheduling for large-scale electric bus fleets. In *IEEE Real-Time Systems Symposium*. 45–55.
[38] Guang Wang, Fan Zhang, and Desheng Zhang. 2019. tCharge-A fleet-oriented real-time charging scheduling system for electric taxi fleets. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 440–441.
[39] Guang Wang, Yongfeng Zhang, Zhihan Fang, Shuai Wang, Fan Zhang, and Desheng Zhang. 2020. FairCharge: A data-driven fairness-aware charging recommendation system for large-scale electric taxi fleets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (2020), 1–25.
[40] Yanan Wang, Tong Xu, Xin Niu, Chang Tan, Enhong Chen, and Hui Xiong. 2020. STMARL: A Spatio-Temporal Multi-Agent Reinforcement Learning Approach for Cooperative Traffic Light Control. *IEEE Transactions on Mobile Computing* (2020).
[41] 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. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1913–1922.
[42] Haoran Xin, Xinjiang Lu, Tong Xu, Hao Liu, Jingjing Gu, Dejing Dou, and Hui Xiong. 2021. Out-of-Town Recommendation with Travel Intention Modeling. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*.
[43] Zaiyue Yang, Lihao Sun, Min Ke, Zhiguo Shi, and Jiming Chen. 2014. Optimal charging strategy for plug-in electric taxi with time-varying profits. *IEEE Transactions on Smart Grid* (2014), 2787–2797.
[44] Yukun Yuan, Desheng Zhang, Fei Miao, Jimin Chen, Tian He, and Shan Lin. 2019. pˆ 2Charging: Proactive Partial Charging for Electric Taxi Systems. In *IEEE 39th International Conference on Distributed Computing Systems*. IEEE, 688–699.
[45] Zixuan Yuan, Hao Liu, Yanchi Liu, Denghui Zhang, Fei Yi, Nengjun Zhu, and Hui Xiong. 2020. Spatio-temporal dual graph attention network for query-poi matching. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 629–638.
[46] Weijia Zhang, Hao Liu, Yanchi Liu, Jingbo Zhou, and Hui Xiong. 2020. Semi-Supervised Hierarchical Recurrent Graph Neural Network for City-Wide Parking Availability Prediction. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. 1186–1193.
[47] Jun Zhao, Guang Qiu, Ziyu Guan, Wei Zhao, and Xiaofei He. 2018. Deep reinforcement learning for sponsored search real-time bidding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1021–1030.
[48] Yanchong Zheng, Ziyun Shao, Yumeng Zhang, and Linni Jian. 2020. A systematic methodology for mid-and-long term electric vehicle charging load forecasting: The case study of Shenzhen, China. *Sustainable Cities and Society* (2020), 102084.
[49] Ming Zhou, Jiarui Jin, Weinan Zhang, Zhiwei Qin, Yan Jiao, Chenxi Wang, Guobin Wu, Yong Yu, and Jieping Ye. 2019. Multi-agent reinforcement learning for order-dispatching via order-vehicle distribution matching. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2645–2653.

A SIMULATOR DESIGN

We develop a simulator[1] based on historical real-time availability (supplies) records data, historical real-time charging prices data, and charging powers data of charging stations, along with historical electric vehicles (EVs) charging data, and road network data [21], to simulate how the system runs for a day. We take Baidu API[2] to compute ETA [23] between the location of charging request and charging station. Besides, we train a MLP model to predict the future demands based on the historical charging requests data, and leverage the learned MLP model to predict the future demands in evaluation phase.

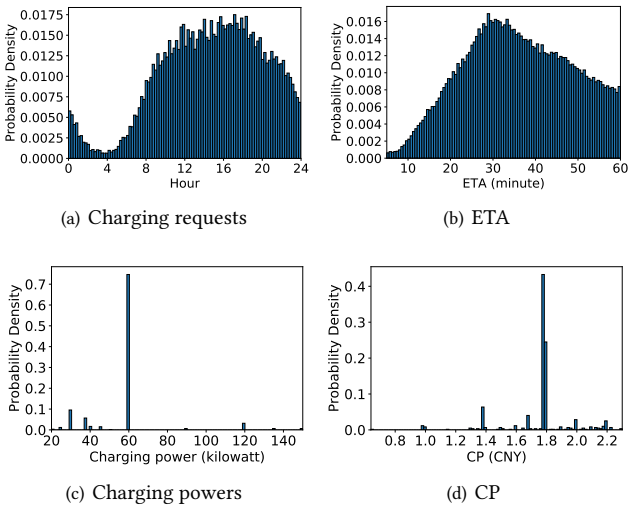
Simulation step. Initially, our simulator loads basic supplies, charging prices and charging powers data of charging stations and charging requests by minutes of a day from the real-world dataset. For each minute T in simulator, there are three events to be disposed:

- **Electric vehicles' departure and arrival**: In this event, we dispose EVs' departure and arrival at T. For EV leaving station, the charging stations will free corresponding number of charging spots. For an arrived EV, the vehicle successfully charges if there still have available spots in the charging station, and the environment will return a negative charging wait time (CWT) and negative charging price (CP) as the rewards if the arrived EV accepted our recommendation. For each successful charging request, the vehicle will block one charging spot in a duration which obeys a Gaussian distribution associated to the charging power of each charging station. If the charging station is full, this EV has to queue until there have an available charging spot, or fail to charge when the CWT exceeds a predefined threshold (45 minutes). The penal rewards ϵ_{cwt} and ϵ_{cp} of failed request are -60 and -2.8, respectively. In the implementation, the number of available charging spots can be a negative number, indicating how many EVs are queuing at station for charging.
- **Charging requests recommendation**: The policy will make recommendations for each charging request at T. The EV of the charging request will drive to the recommended charging station based on the real recommendation acceptance probability, and otherwise will go to the ground truth charging station.
- **Transition collection**: If there have charging requests at T, the transitions $(o^i_{t'}, a^i_{t'}, o^i_t, R_{t':t})$ for charging stations will be collected and stored into replay memory [27] for learning RL algorithms.

B DATASET STATISTICAL ANALYSIS

We take Beijing as an example for more statistical analysis. Figure 10 shows the distribution of the quantities of charging requests in different time, and the distributions of ETA, charging powers,

[1] https://github.com/Vvrep/MASTER-electric_vehicle_charging_recommendation
[2] http://lbsyun.baidu.com/index.php?title=webapi



(a) Charging requests (b) ETA

(c) Charging powers (d) CP

Figure 10: Distributions of charging requests quantity, ETA, charging powers, and CP on Beijing.

and CP. We can observe the charging requests quantity changes notably along with time in a day, which indicates unbalanced charging demands. Besides, we observe it is significantly different by comparing the distributions of ETA with CP, where the latter is much more concentrated. This observation indicates the large differences between the two objectives.

C PROPOSITION

Proposition C.1. *Define r^1 (e.g., r^{cwt}) and r^2 (e.g., r^{cp}) as two different reward functions, then the following two policy gradients to update the policy $\boldsymbol{b}$ are equivalent:*

$$\nabla_{\theta_{\boldsymbol{b}}} J_1(\boldsymbol{b}) = \mathbb{E}_{s_t \sim D}\left[\nabla_{\theta_{\boldsymbol{b}}} \boldsymbol{b}(s|\theta_{\boldsymbol{b}})|_{s=s_t} \nabla_a Q_{\boldsymbol{b}}(s,a)|_{s=s_t, a=\boldsymbol{b}(s_t)}\right] \quad (16)$$

$$\begin{aligned}\nabla_{\theta_{\boldsymbol{b}}} J_2(\boldsymbol{b}) &= \mathbb{E}_{s_t \sim D}\Big[\nabla_{\theta_{\boldsymbol{b}}} \boldsymbol{b}(s|\theta_{\boldsymbol{b}})|_{s=s_t} \nabla_a Q^1_{\boldsymbol{b}}(s,a)|_{s=s_t, a=\boldsymbol{b}(s_t)} \quad (17)\\ &\quad + \nabla_{\theta_{\boldsymbol{b}}} \boldsymbol{b}(s|\theta_{\boldsymbol{b}})|_{s=s_t} \nabla_a Q^2_{\boldsymbol{b}}(s,a)|_{s=s_t, a=\boldsymbol{b}(s_t)}\Big]\\ &= \mathbb{E}_{s_t \sim D}\left[\nabla_{\theta_{\boldsymbol{b}}} \boldsymbol{b}(s|\theta_{\boldsymbol{b}})|_{s=s_t} \nabla_a \left(Q^1_{\boldsymbol{b}}(s,a) + Q^2_{\boldsymbol{b}}(s,a)\right)|_{s=s_t, a=\boldsymbol{b}(s_t)}\right]\end{aligned}$$

where $Q_{\boldsymbol{b}}(s,a) = \mathbb{E}_{\boldsymbol{b}}\left[\sum_{i=t}^{\infty} \gamma^{(i-t)}\left(r^1(s_i,a_i) + r^2(s_i,a_i)\right)|s=s_t, a=a_t\right]$

and $Q^k_{\boldsymbol{b}}(s,a) = \mathbb{E}_{\boldsymbol{b}}\left[\sum_{i=t}^{\infty} \gamma^{(i-t)} r^k(s_i,a_i)|s=s_t, a=a_t\right], k=\{1,2\}.$