Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Special Track on AI4Tech: AI Enabling Critical Technologies

Optimizing the Battery-Swapping Problem in Urban E-Bike Systems with Reinforcement Learning

Wenjing Li[1] , **Zhao Li**[2] , **Xuanwu Liu**[2] , **Ruihao Zhu**[3] , **Zhenzhe Zheng**[1] , **Fan Wu**[1]
[1]Shanghai Jiao Tong University
[2]Hangzhou Yugu Technology Co.,Ltd
[3]Cornell University
wenjingli@sjtu.edu.cn, lzjoey@gmail.com, liuxuanwu@yugu.net.cn, ruihao.zhu@cornell.edu, zhengzhenzhe@sjtu.edu.cn, fwu@cs.sjtu.edu.cn

Abstract

E-bikes (EBs) are a key transportation mode in urban area, especially for couriers of delivery platforms, but underdeveloped EB systems can hinder courier's productivity due to limited battery capacity. Battery-swapping stations address this issue by enabling riders to exchange depleted batteries for fully charged ones. However, managing supply and demand (SnD) imbalances at these stations has become increasingly complex. To address this, we introduce a new approach that formulates the Battery-Swapping Problem (BSP) as a discrete-time Markov Decision Process (MDP) to capture the dynamics of SnD imbalances. Building on it, we propose a Wasserstein-enhanced Proximal Policy Optimization (W-PPO) algorithm, which integrates Wasserstein distance with reinforcement learning to improve the robustness against uncertainty in forecasting SnD. W-PPO provides a BSP-specific, accurate loss function that reflects reward variations between two policies under real-world simulation. The algorithm's effectiveness is assessed using key metrics: Shared Battery Utilization Ratio (SBUR) and Battery Supply Ratio (BSR). Simulations on real-world datasets show that W-PPO achieves a 30.59% improvement in SBUR and a 16.09% increase in BSR ensures practical applicability. By optimizing battery utilization and improving EB delivery systems, this work highlights the potential of AI for creating efficient and sustainable urban transportation solutions.

1 Introduction

E-bikes (EBs) have become the primary mode of transportation for many urban residents due to their affordability and ability to navigate through traffic with ease, significantly impacting service delivery platforms where most couriers rely on EBs. In China, the number of delivery platform-based workers reached 84 million by 2020 [Julie Yujie Chen, Ping Sun2023], with steady growth driven by market expansion. Similarly, in India and South America, the number of couriers exceeds 500,000 [Ezra Fieser2019,Rica Bhattacharyya2022]. Couriers typically earn a base hourly wage supplemented by delivery commissions, making it crucial to optimize EB transportation systems to maximize the number of deliveries within a given time frame. However, underdeveloped EB transportation systems often cause range anxiety [Li *et al.*2024a], which they cannot travel the desired distance due to limited battery capacity.

To address this, battery-swapping stations have been introduced, enabling couriers to exchange depleted batteries for fully charged ones, reducing downtime and improving operational efficiency. With the rapid expansion of the EB demand, managing the imbalance between supply and demand (SnD) at battery-swapping stations has become increasingly complex. As depicted in Figure 1, SnD imbalances can prevent couriers from accessing fully charged batteries at their initial stations, forcing them to travel to alternative locations at a greater distance. This inefficiency not only delays deliveries but also diminishes overall couriers and other EB rider satisfaction of battery needs. The Battery-Swapping Problem (BSP) has thus become a critical challenge in urban transportation, and finding efficient solutions is vital for ensuring fast, cost-effective, and reliable EB transportation. By providing higher earning potential for couriers, as their productivity and route efficiency improve, the BSP further offers significant societal benefits aligned with the United Nations' Sustainable Development Goal 8: Promote sustained, inclusive, and sustainable economic growth, full and productive employment, and decent work for all [United Nations2015].

In the context of urban transportation, the BSP for EBs presents unique challenges that differentiate it from EVs [Sun *et al.*2019, Mao *et al.*2024]. One key issue is the dense distribution of EB battery stations, coupled with the need for rapid, localized battery replacements to support the swift operations of couriers. Additionally, the variation in courier behavior and the operational constraints within the battery station networks introduce a high degree of nonlinearity and unpredictability to the SnD dynamics, , which traditional models struggle to capture. Proximal Policy Optimization (PPO) [Schulman *et al.*2017], the on-policy state-of-the-art reinforcement learning algorithm, seeks to mitigate these issues by constraining policy updates to prevent destabilizing changes. However, its reliance on the clipped surrogate and Kullback-Leibler (KL) divergence is limited in real-time, location-based environments, failing to fully capture the true nature of policy changes and leading to suboptimal perfor-

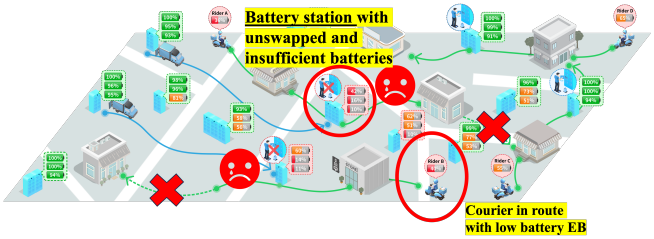

Figure 1: In real-world scenarios without BSP optimization, couriers may be unable to access fully charged batteries, resulting in a decrease in their overall utility.

mance. This creates a critical need for more advanced approaches that can better model the dynamics of BSP for EBs.

To tackle these challenges, we formulate the BSP as a Markov Decision Process (MDP) and introduce a new on-policy RL algorithm enhanced by the Wasserstein distance, noting that the problem exhibits Markovian characteristics, where the information received on a given day is primarily influenced by the immediate previous day and a limited set of today's user behaviors, without being influenced by prior days. This algorithm is specifically designed to optimize the battery-swapping process across stations, aiming to maximize battery utilization, improve rider satisfaction, and increase the overall robustness of the BSP system. Grounded in real-world datasets, the algorithm incorporates a carefully designed RL environment to ensure the robust performance guaranteed in highly dynamic environments. Unlike the traditional PPO methods that minimize divergence between old and updated policies, our Wasserstein-enhanced Proximal Policy Optimization (W-PPO) algorithm calculates the actual differences between policies, leveraging real-world simulations for improved precision. This end-to-end solution integrates Wasserstein-improved penalty calculations, delivering high capability and efficiency. To assess the effectiveness of the algorithm, we introduce two key performance metrics: the Shared Battery Utilization Ratio (SBUR) and the Battery Supply Ratio (BSR). The SBUR quantifies the proportion of riders receiving fully charged batteries relative to the total number of batteries at a station, while the BSR assesses the increase in battery availability and the number of riders the system can serve, reflecting the overall performance of EB transportation.

The main contributions can be summarized as follows:

- Formulating the Battery-Swapping Problem (BSP) to address the imbalance between SnD across battery stations within a EB rider transportation network, and modeling the BSP as a discrete-time Markov Decision Process (MDP).
- Introducing a Wasserstein-enhanced Proximal Policy Optimization (W-PPO) reinforcement learning algorithm as a comprehensive solution to the BSP, designed to improve the robustness against SnD dynamic. This algorithm integrates the Wasserstein distance with on-policy reinforcement learning. W-PPO does not minimizing the divergence between probabilities, as allowing policy changes better reflects the dynamics of the real-world simulation. By leveraging the strengths W-PPO, our approach effectively addresses the challenges posed by SnD imbalances in BSP.
- Developing a well-structured algorithm and conducting theoretical evaluations, leading to significant structural insights into the complex BSP transportation network, which demonstrates superior capability and flexibility compared to traditional deep learning methods. The SBUR increased for the average of 30.59% (7 stations per route) and the BSR increased for 16.09%. These two metrics demonstrate our work's improvement of the urban EB transportation system.

2 Battery-Swapping Problem

This section introduces the formal problem statement and its corresponding mathematical models.

From the previous section, we know that customer will swap batteries when the current one they are using is on low battery mode. Depends on the distance they need per day, customers will swap batteries for different number of times. They can return the battery at the station they swapped or at another nearby station, so the number of batteries in each station can be different from one night to the next.

The scenario involving Sharing EB Battery with shared batteries presents a complex landscape that necessitates a detailed representation of the dynamic interactions between riders and EB batteries within a reinforcement learning framework. Delivery companies assign their riders to operate within a designated region, ensuring that battery replacements occur within a confined area. Consequently, the BSP we address is inherently location-based. When users replace batteries, the same battery can appear at multiple exchange stations based on their convenience, allowing it to be reused by a succession of different users. This leads to unbalanced battery resources. To balance the resource, we design the model to learn and match the supply and demand of surplus and shortage battery stations.

We define the problem by building a battery-swapping network: $G = (\mathcal{C}, \mathcal{R})$, to be more specific:

- Each station $C_i \in \mathcal{C}$ allows customers to swap battery and generate corresponding SnD. The number of boxes that can contain battery for each C_i is denoted as X_i. The number of EB battery users for each C_i is denoted as Y_i^t as the amount is increasing in a timely manner.
- The initial number of batteries in the station at C_i as B_i^0, and we use B_i^t, D_i^t, and L_i^t, $\forall t \in (t = 1, 2, ..., T)$ to represent the number of batteries, battery demand, and battery returned loaded at different time, respectively.
- The battery demand D_i^t for each station C_i is the total order number each day.
- The battery returned at each station is denoted as the normal distribution of $L_i : \mathcal{N}(L_i, \sigma^2)$. The distribution follows a normal pattern because the location dynamics remain stable, with consistent daily patterns observed. However, notable variations arise between weekdays and weekends. The formulation of L_i is $B_i^{t+1} - B_i^t + D_i^t - x_i^t$.

- Under the problem formulation, each route $R_i \in \mathcal{R}$ is a cycle in the battery-swapping network, consisting of a sequence of consecutive stations $C_{i_1}, C_{i_2}, ..., C_{i_{|R_i|}}$, where $|R_i|$ is the number of stops on R_i, and the next destination of $C_{i_{|R_i|}}$ is C_{i_1}, which generates a looped path for the vehicle. Each route cannot intersect with others in the network to ensure transportation efficiency. On each route R_i, there is one vehicle cycling around the route, a capacity Cap_j^t (the maximum number of batteries it can convey). When a vehicle arrives at a station, it can load batteries or discharge its batteries to the station.

The objective of the battery-swapping network is to maximize the Sharing Battery Used Rate to optimize utility for individual EB riders. At a specific time t, the station can only use the batteries in the last date, i.e., B_i^{t-1}, to fulfill the current demand D_i^t. Once the battery number is not balanced among stations on one route, customer cannot pick up a fully charged battery, which leads to decreased amount of distance they can travel and delivery order they can finish. Accordingly, re-balancing the surplus and shortage stations can increase the amount of fully charged orders and increase the Shared Battery Used Rate (SBUR).

Definition 1. *Given the total number of batteries in each station, B_i^t, and the fully charged and swapped order number, D_i^t, we define the Shared Battery Used Rate (SBUR) in the Battery-Swapping Problem:*

$$SBUR_i^t = \frac{D_i^t}{B_i^t}. \tag{1}$$

After the current demand is processed, returning battery loaded and those discharged from the vehicle will be added to the station battery number. Thus, we can compute the new battery number as $B_i^t = max(B_i^{t-1} - D_i^t, 0) + L_i^t + x_j^t$, where $x_j^t \in N$ denotes the number of batteries to delivery or pickup at station C_i by vehicle V_j at time t. x_j^t can be negative to denote the discharged amount of resources from the vehicle.

Upon the completion of the battery relocation process, the number of batteries available for swapping increases. This increase in availability can be quantified through the calculation of the Battery Supply Ratio (BSR), based on the variables X_i and Y_i^t. Consequently, the increase in the BSR facilitates the determination of the additional number of EB riders that the system can now accommodate, thereby enhancing the overall efficiency of the EB transportation network.

Definition 2. *Given Shared Battery Utilization Rate $SBUR_i^t$, battery capacity X_i, and the number of EB riders utilizing the shared battery at each cabinet Y_i^t, we define Battery Supply Rate (BSR) within the context of the BSP as:*

$$BSR_i^t = \frac{Y_i^t}{(1 + SBUR_i^t) \times X_i}. \tag{2}$$

3 Designs of W-PPO

Traditional method of matching the demand and supply has limitation as of failures in front of uncertainty of SnD, complex business constraints, and high complexity of transportation networks. To address these issues, we implement the reinforcement learning problem.

3.1 Discrete-time Markov Decision Process

In this section, we will formally define the BSP as a discrete-time Markov Decision Process (MDP) (S, A, P, r, γ, T) where S is the state space, A is the action space, $P(\cdot|s, a)$ is the transition probability distribution function, r is the reward function, γ is the discount factor, and T is the total number of days to training in one episode.

State Space S We validated the inclusion of states in successive steps and evaluated the impact of the choices based on rewards. The states transitions are looping in the order of Night (N), Morning (M), and Next Night (N+1).

$$S = S_N, S_M, S_{N+1}$$

From S_N to S_M, each agent will take action to match the demand and the supply. The next night will be the morning state adding the daily usage. It is also important to note that batteries in the stations without transporting will be fully charged throughout the night.

Action Space Each agent, route R_i defined in the previous section, can choose to pick up $n \in [0, 1, 2, .., k], k \in \mathbb{Z}$ battery and can drop off n battery as well. n is the capacity of each route R_i. Two constraints are set in the action space. First, the sum of action space on one route is defined to be zero: $\sum A_i = 0 \forall i \in R$. Second, the number of battery picked up or dropped off at each station needs to exceed 0: $a_t \geq 0 \; \forall a \in A_i$.

Transition probability function P defined as a mapping S × A × S → [0, 1], which can be specified by the definition of S, A, and R and the distribution behind SnD within particular battery-swapping networks.

Reward Function We define the reward as maximizing the SBUR until it reaches 1:

$$r_t := \min_i SBUR_i^t. \tag{3}$$

The state-value function, the discounted expected reward stating from state s under policy π, is defined as [Sutton and Barto2018]:

$$V^\pi(s) := \mathbb{E}^\pi \left[\sum \gamma^k r_{t+k} | s_t = s \right]. \tag{4}$$

To reduce computational overhead, we use the temporal difference (TD) error approximation as advantage function [Schulman *et al.*2017]:

$$\hat{A}(s, a) = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1}, \tag{5}$$
$$\text{where } \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t).$$

We use the policy of maximizing the reward to take action. Since the battery number of a new day only depends on the amount of the previous day, the Markov property is satisfied.

A policy π is a function from state to a distribution over actions. The goal of the BSP model is to find policies that maximize episodic return, which is maximizing the SBUR. Running a policy π in the MDP generates a state-action trajectory $(s_1, a_1, r_1, s_2, a_2, r_2, ..., r_T) =: \tau$. θ_k denotes the parameter for the udated policy $\pi_{\theta_k}(a|s)$.

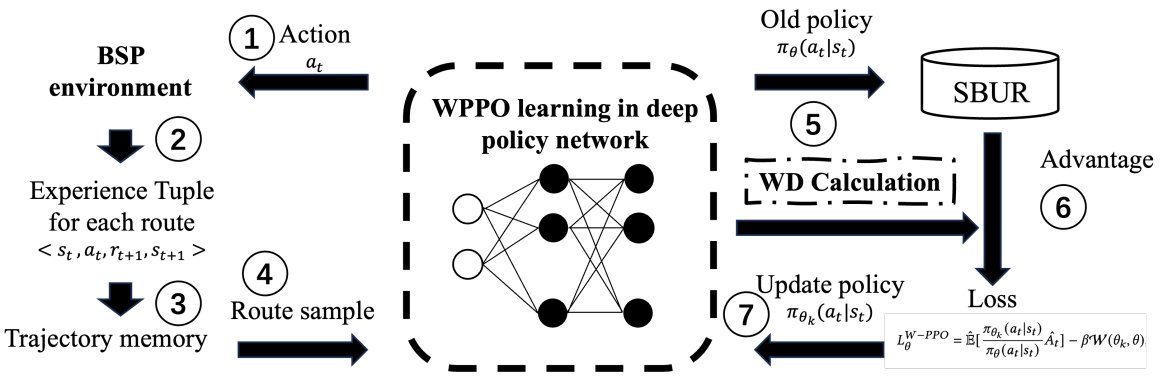

Figure 2: The W-PPO model (WD: Wasserstein distance) within the BSP framework iteratively traverses each step, updating the policy to optimize SBUR as illustrated in the plot, over the specified number of episodes.

Algorithm 1 Wasserstein Enhanced PPO

Input: initialize policy and penalty parameters θ, β
Parameter: Optional list of parameters
Output: Your algorithm's output

for iteration=1, 2, ... **do**
  Compute the SBUR (reward in the BSP) for each route and day under θ
  for route=1, 2, ..., R **do**
    Run policy π_θ in environment for T time steps
    if $\mathcal{W} > \mathcal{W}_{threshold}$ **then**
      $\beta_{k+1} = 2\beta_k$
    else if $\mathcal{W} < \mathcal{W}_{threshold}$ **then**
      $\beta_{k+1} = \beta_k/2$
    end if
    Compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$ using equation 5
    Calculate the loss for each route

$$L_\theta^{W-PPO} = \hat{\mathbb{E}}[\frac{\pi_{\theta_k}(a_t|s_t)}{\pi_\theta(a_t|s_t)}\hat{A}_t] - \beta\mathcal{W}(\theta_k, \theta)$$

  end for
  Compute average loss for all routes under the old policy
  Update the the old policy $\theta \leftarrow \theta_k$
  Updated policy θ_k based on the average loss
end for

3.2 Wasserstein-Enhanced PPO

Building upon the state-of-the-art deep reinforcement learning method, PPO [Schulman *et al.*2017], our analysis reveals that the traditional use of Kullback–Leibler (KL) divergence in PPO fails to ensure robust penalty convergence in the context of BSP. While KL divergence is designed to minimize the divergence between policies, ensuring the new policy θ_k remains closely aligned with the old one θ, this approach falls short of capturing the actual difference between the two policies—a factor that is critical for BSP applications. In the context of BSP, our objective is to evaluate the SBUR, which is derived from a state and action space modeled on real-world data. Therefore, accurately measuring the true difference between policies (how action selections change the battery status in real-world) is essential, as it provides deeper insights into policy performance and adaptability, rather than merely restricting divergence. To address this limitation, we incorporate the Wasserstein distance, derived from optimal transport theory, into the policy optimization algorithm. This modification results in a new distance-regularized policy optimization approach that better evaluates and enforces meaningful distinctions between policies. We use θ_k representing the updated policy and θ representing the old policy.

The Wasserstein distance, a metric for comparing two discrete probability distributions P and Q, is rooted in optimal transport theory. Intuitively, it quantifies the minimum "cost" required to transform one distribution into the other, where the cost is determined by the amount of mass transported and the distance it is moved. For computational efficiency and theoretical analysis, we focus on the dual formulation of the Wasserstein distance, which provides an equivalent but often more tractable representation of the original primal problem. The dual form leverages the properties of duality in optimization, facilitating its application in machine learning and reinforcement learning contexts [Kolouri *et al.*2017, Villani2003]:

$$W_d^p(P, Q) = \sup_{\psi,\phi} \int \psi(y)dQ(y) - \int \phi(x)dP(x). \quad (6)$$

The parameter p specifies the order of the Wasserstein distance, governing how the transport cost is aggregated across the probability distributions. To accurately quantify the difference between two policies in our study of the BSP, we focus on the case $p = 1$, commonly referred to as the 1-Wasserstein distance or Earth Mover's Distance (EMD). This metric represents the minimum cost needed to transform one probability distribution into another, with the cost being directly proportional to the quantity of mass moved and the distance over which it is transported. This interpretation aligns closely with practical scenarios, making it particularly suitable for applications that demand an intuitive and meaningful measure of distributional dissimilarity. The equation for 1-Wasserstein distance:

$$W_1(P,Q) = \sup_{f\in\mathcal{F}} \int f(x)dP(x) - \int f(x)dQ(x), \quad (7)$$

where $\mathcal{F}$ denotes all maps from $\mathbb{R}^d$ to $\mathbb{R}$ such that $|f(y) - f(x)| \leq ||x-y||\forall x, y$. Since the difference between the old policy θ and the updated policy θ_k exists in a one-dimensional flat space, we let $d = 1$, so the functions f in the set $\mathcal{F}$ are 1-Lipschitz, meaning they do not increase faster than the distance between points by the non-expansiveness property of 1-Lipschitz function [Lang1993, Gulrajani *et al.*2017]:

$$W_1(P,Q) = \left(\int_0^1 |F^{-1}(z) - G^{-1}(z)|d(z)\right). \quad (8)$$

F and G are the cumulative distribution function of P and Q. We further expand the $d = 1$ and $p = 1$ Wasserstein distance with the form of probability distribution functions

$$W_1(P,Q) = \sum_{i=1}^{n} ||f(x_i)\Delta x - f(y_i)\Delta y||. \quad (9)$$

With the formulation of the $p = 1, d = 1$ Wasserstein distance, it promises a optimal distance between two policies θ and θ_k. The stochastic policy of reinforcement learning can be written as the probability density function:

$$\pi(a|s) = \mathbb{P}(A = a|S = s). \quad (10)$$

The cumulative distribution function of the policy is:

$$\sum_{a'\leq a} \pi(a'|s)\Delta a. \quad (11)$$

With the cdf of π, we obtain the Wasserstein distance $\mathcal{W}$ of θ and θ_k:

$$\mathcal{W}(\theta,\theta_k) = \sum_{i=1}^{n} ||f_{\theta_k}(a|s)\Delta a - f_\theta(a|s)\Delta a||. \quad (12)$$

With the definition of the Wasserstein distance for policy, we are ready to propose the Wasserstein enhanced PPO:

Definition 3. *Given a MDP $M = (S, A, P, r, \gamma, T)$, the loss function of the Wasserstein-enhanced PPO algorithm is defined with a W-penalty:*

$$L_\theta^{W-PPO} = \hat{\mathbb{E}}[\frac{\pi_{\theta_k}(a_t|s_t)}{\pi_\theta(a_t|s_t)}\hat{A}_t] - \beta\mathcal{W}(\theta_k,\theta), \quad (13)$$

where

$$\mathcal{W}(\theta_k,\theta) = \sum_{i=1}^{n} ||f_{\theta_k}(a|s)\Delta a - f_\theta(a|s)\Delta a||.$$

4 Experiment Results

In this section, we report the experiment results of our algorithm and multiple baselines based on historical data.

4.1 Experiment Settings

The extensive database provided by [Anonymous Company] is crucial for developing and evaluating our models, encompassing detailed information on over 500,000 batteries and 30,000 battery stations distributed across more than 80 cities. The dataset is divided into two key components: time-location data, capturing temporal and spatial relationships, and battery status data, detailing operational metrics such as inventory, capacity, and usage. As most riders tend to swap battery in a certain area, we apply the K-Nearest Neighbor (KNN) algorithm to cluster stations into routes, ensuring the optimization process accounts for the spatial distribution of stations and aligns resource allocation with both operational and geographic constraints. Additionally, the database updates operational data for each station daily, typically by 00:00. In our experiment, we conduct research using the data of one month (31 days).

Based on our dataset, we first configure the environment according to the BSP formulation. To ensure a fair comparison across all approaches, we allocate the same test set for evaluating performance. Initially, we conduct baseline experiments using PPO: without penalty, PPO clipped, and PPO KL penalty [Schulman *et al.*2017]. We further experimented the off-policy state-of-the-art model: Soft Actor-Critic (SAC) [Haarnoja *et al.*2018] to compare the results between on-policy and off-policy models. Given the unique objective in BSP to maximize the SBUR across all routes, the losses and rewards are computed individually for each route. Subsequently, the average loss and reward across all routes are calculated to provide a comprehensive evaluation metric.

As outlined in Algorithm 1, the average loss for all routes within a single episode is used to update the policy. The primary distinction between these models lies in how the loss is calculated and how feedback is incorporated to update policies. This comparative approach ensures insights into the performance impact of different optimization strategies in the BSP context.

In the experiment, the threshold for divergence/distance is determined as the average divergence/distance observed during training. By adopting this average as the threshold, the policy is afforded the flexibility to either update or remain unchanged, depending on the divergence behavior. To ensure reliable evaluation and prevent overfitting, we analyze the loss patterns and select a configuration of running ten episodes for each model. This approach is chosen based on the observation that the loss stabilizes near zero without exhibiting signs of overfitting, which ensures robust performance across the tested models.

Figure 4 illustrates the detailed technical workflow of the W-PPO model. The process begins with the initialization of the policy, where selected actions are applied to the environment for each specified date and route. The environment then generates experience tuples, which are stored in the trajectory memory. Sample routes are drawn from this memory and processed within the W-PPO model, where they are evaluated using several metrics: the SBUR as the reward, the Wasserstein distance as a penalty, and the ratio between the old and new policies. These evaluation variables are utilized to compute the advantage and the loss function, incorporating the

Wasserstein penalty to ensure robust policy learning. Finally, the policy is updated based on these computations, iteratively refining its performance to address the BSP effectively.

4.2 SBUR Results Analysis

Table 1 provides a detailed breakdown of the experimental outcomes. A key performance metric, the percentage improvement in the SBUR, is calculated for each policy update episode to evaluate and compare the models. To benchmark the performance of our proposed Wasserstein-Enhanced PPO (W-PPO) model against baseline models, we compute both the average SBUR improvement across all episodes and the maximum SBUR improvement observed during the experiments. The table reveals that W-PPO demonstrates superior

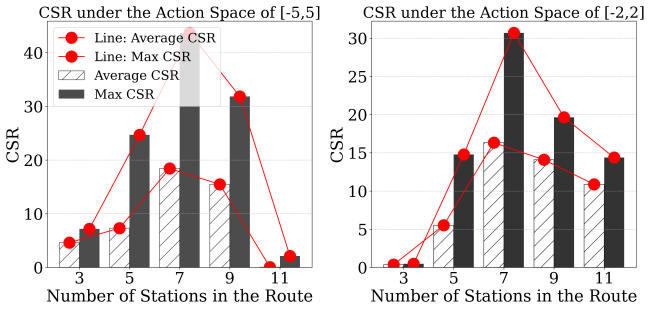

Figure 3: Comparison of SBUR improvements across different route sizes, highlighting optimal performance at 7 stations per route.

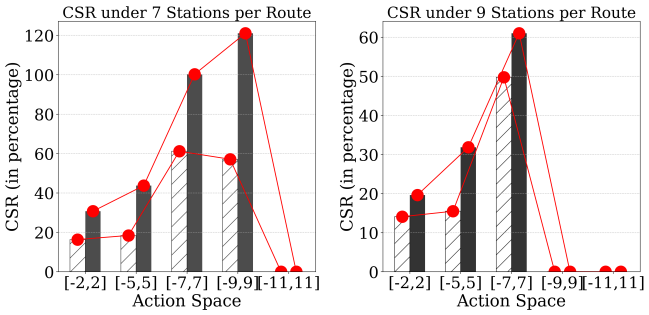

Figure 4: Comparison of SBUR improvements across different action spaces, highlighting optimal performance at the action space: [-9,9].

performance for routing scenarios involving 5 and 7 stations per rout, outperforming the baseline models in both average and maximum SBUR metrics. However, for routes with 3 stations, W-PPO underperforms compared to the baseline models, with the PPO without penalty variant achieving the best results in this scenario. This trend is more pronounced in environments with a smaller action space (e.g., [-2, 2]). The observed instability for smaller route sizes can be attributed to the reduced amount of training data available when fewer stations are included, limiting the model's ability to effectively learn optimal policies. The table also reveals that the off-policy model, SAC, performs less favorably compared to all on-policy methods. From a mathematical perspective, the battery patterns exhibit minimal variation within a given time frame and location cluster. As a result, entropy does not fluctuate significantly between updates, which hinders the SAC's ability to effectively train under the BSP framework.

To further investigate the relationship between route size and performance, we systematically vary the number of stations per route C, conducting experiments with 3, 5, 7, 9, and 11 stations. The baseline configuration assumes 5 stations per route. Figure 4 visualizes the average SBUR improvement for each configuration, highlighting a clear peak at 7 stations per route. Beyond this point, performance begins to decline.

This decline for larger routes can be explained by practical limitations: as the number of stations increases, the geographical area covered by a route expands. This often exceeds an efficient delivery range, leading to greater variability in battery usage patterns and reduced optimization efficiency. On the other hand, routes with too few stations fail to fully exploit the available resources, resulting in suboptimal performance.

We tested various action spaces, ranging from $[-2, 2]$ to $[-11, 11]$, to account for the differing storage capacities of battery stations. Most stations have a storage capacity between 10 and 20 batteries, while some have as few as 7 storage spaces. Experiments were conducted using the top-performing set of stations per route, specifically routes with 7 and 8 stations. As shown in Figure 5, the action space $[-11, 11]$ did not result in any SBUR improvement, as removing too many batteries from a station depleted its reserves, leading to premature experiment termination. Similarly, for routes with 9 stations, the SBUR remained unchanged for an action space of $[-9, 9]$, as the higher variability in battery distribution across stations caused some stations to run out of batteries when taking 9 or more, also terminating the experiment. From the plot, we observe that the average SBUR reaches its peak at an action space of $[-7, 7]$ for routes with 7 stations, achieving an impressive average SBUR increase of 61.14%. The average SBUR improvement for 7 stations per route is **30.59%** for each route on a daily evaluation. Therefore, we conclude that in this experiment, the action space of $[-7, 7]$ for 7 stations per route represents the optimal configuration. Building upon the training results from the W-PPO model for EB battery rebalancing, we further assess the BSR. As presented in Table 2, we observe that under the optimal configuration, the BSR increases by 16.09%. To further investigate the impact of BSR, we evaluate the additional number of EB riders that can be served by this transportation system design. Using the current data from Shanghai, we find that with a BSR increase of 16.09%, the number of EB riders served grows by 45.55%, assuming a constant battery supply. Additionally, to examine the environmental and efficiency benefits, we calculate the number of batteries saved, which, based on the current constant EB rider count in the BSP system, amounts to 3,972. The exploration of the BSR demonstrates the overall BSP robustness.

5 Discussion and Conclusion

This work has addressed the key challenges of optimizing SnD imbalances in battery-swapping stations for EB transportation, which impact courier efficiency, income, and over-

Model	7 Stations per Route									
	SAC		PPO-NC		PPO-KL		PPO-C		W-PPO	
	[-2,2]	[-5, 5]	[-2,2]	[-5, 5]	[-2,2]	[-5, 5]	[-2,2]	[-5, 5]	[-2,2]	[-5, 5]
Average SBUR	-1.32%	0.07%	10.75%	13.09%	20.06%	13.92%	11.02%	9.76%	**16.32%**	**18.41%**
Max SBUR	-0.57%	0.19%	8.09%	12.66%	23.91%	16.72%	28.61%	15.13%	**30.69%**	**43.69%**
Model	5 Stations per Route									
	SAC		PPO-NC		PPO-KL		PPO-C		W-PPO	
	[-2,2]	[-5, 5]	[-2,2]	[-5, 5]	[-2,2]	[-5, 5]	[-2,2]	[-5, 5]	[-2,2]	[-5, 5]
Average SBUR	-1.59%	-0.92%	-2.02%	-0.48%	-1.77%	0.28%	1.56%	2.14%	**5.52%**	**7.31%**
Max SBUR	-0.66%	-0.38%	2.53%	7.73%	12.93%	10.29%	9.86%	17.42%	**14.78%**	**24.64%**
Model	3 Stations per Route									
	SAC		PPO-NC		PPO-KL		PPO-C		W-PPO	
	[-2,2]	[-5, 5]	[-2,2]	[-5, 5]	[-2,2]	[-5, 5]	[-2,2]	[-5, 5]	[-2,2]	[-5, 5]
Average SBUR	-0.79%	-0.11%	**1.29%**	**9.43%**	2.77%	6.91%	1.02%	3.96%	0.36%	4.62%
Max SBUR	-0.49%	0.00%	**2.02%**	**10.29%**	3.12%	9.63%	2.79%	4.56%	0.48%	7.17%

Table 1: Performance of different models (SAC, PPO-NC: PPO no penalty; PPO-KL: PPO KL penalty; PPO-C: PPO clipping; W-PPO: Wassertein-enhanced PPO) on different action spaces ([-2,2] & [-5,5] represent two different action spaces) and routing numbers. All the values are the percentage increase in SBUR.

SBUR	BSR	EB Rider Growth	Batteries Saved
30.59%	**16.09%**	**45.55%**	**3972**

Table 2: SBUR, BSR, percentage increase of the EB riders for service, and number of batteries saved performance of the optimal configuration of the action space [-7,7] for 7 stations per route.

all EB rider satisfaction. By formulating the BSP as a discrete-time MDP, we provide a robust mathematical foundation to manage these imbalances in complex transportation networks. The proposed W-PPO algorithm, combining PPO with Wasserstein distance, effectively handles the uncertainties in SnD forecasting. Our approach demonstrates notable advantages in both theoretical insights and practical outcomes, achieving a 30.59% average improvement in the SBUR and a 16.09% increase in BSR for 7 stations per route, offering a significant enhancement over traditional methods.

Several limitations and future works is considered in this study. Data protection restrictions hinder the evaluation of cost efficiency, and the applicability of our method to other transportation systems like EVs and AGVs requires further adaptation due to differing operational dynamics. Additionally, the method's performance is sensitive to hyperparameters, affecting its generalizability. Future work will focus on developing advanced hyperparameter optimization techniques, adapting the W-PPO algorithm for broader transportation contexts, and extending the W-PPO algorithm to real-time SnD predictions.

6 Related Work

Electric transportation management has become a key research area, with significant efforts in optimizing various objectives in different types of electric transportation, like electric vehicles (EVs), automated guided vehicles (AGVs), E-Scooters [He and Shin2020], and EBs. Most works studied the EVs discussing its state of charge [Baccari *et al.*2024, Sun *et al.*2019, Cui *et al.*2023], system costs [Widrick *et al.*2016], battery system [Mao *et al.*2024, Zhang *et al.*2018], and company profits [Jin *et al.*2023, Shalaby *et al.*2023, Liang *et al.*2023, Zheng *et al.*2014, Gao *et al.*2020]. The studies related to AGVs used RL and are implemented in warehouses to increase logistics and operational efficiency. Additionally, AI models like transformers and RL techniques have played a vital role in the studies of the EB transportation system. Previous studies focused on monitoring and detecting battery status and predicting riding range in EBs [Li *et al.*2024c, Li *et al.*2024b]. These AI-driven approaches have greatly advanced the efficiency and reliability of electric transportation management systems. Moreover, AI has been extensively applied to optimize logistical systems, such as express delivery networks [Li *et al.*2020], empty container repositioning [Li *et al.*2019], and resource-balancing challenges in various domains [Naderializadeh *et al.*2020, Yi *et al.*2019], further improving operational decision-making and system efficiency across the electric transportation landscape.

In RL, Proximal Policy Optimization (PPO) has proven effective in diverse applications, including robotics and logistics, due to its stability and efficiency [Schulman *et al.*2017]. Optimal transportation theory has also been integrated into RL to address resource allocation and distribution challenges. Studies have combined optimal transportation with RL in multi-agent systems [Gulrajani *et al.*2017, Ali Baheri2024] and curriculum learning [Huang *et al.*2022], showcasing its potential. Our work extends this by incorporating the Wasserstein distance [Kolouri *et al.*2017, Villani2003] with PPO to create a more meaningful penalty structure for policy optimization, especially in environments with limited data. This combination enhances the model's ability to make precise and efficient updates, making a significant contribution to RL applications in complex resource-balancing scenarios.

Acknowledgments

This work was supported, in part by China NSF grant No. U2268204, 62322206, 62025204, 62132018, 62272307,

62372296, in part by Alibaba Group through Alibaba Innovative Research Program. The opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies or the government.

Zhenzhe Zheng is the corresponding author.

References

[Ali Baheri, 2024] Mykel J. Kochenderfer Ali Baheri. The synergy between optimal transport theory and multi-agent reinforcement learning. *arXiv preprint arXiv:2401.10949*, 2024. https://arxiv.org/abs/2401.10949.

[Baccari *et al.*, 2024] Silvio Baccari, Massimo Tipaldi, and Valerio Mariani. Deep reinforcement learning for cell balancing in electric vehicles with dynamic reconfigurable batteries. *IEEE Transactions on Intelligent Vehicles*, pages 1–12, 2024.

[Cui *et al.*, 2023] Dingsong Cui, Zhenpo Wang, Peng Liu, Shuo Wang, David G. Dorrell, Xiaohui Li, and Weipeng Zhan. Operation optimization approaches of electric vehicle battery swapping and charging station: A literature review. *Energy*, 263:126095, 2023.

[Ezra Fieser, 2019] Ezra Fieser. The brains behind rappi, latin america's super app. https://www.bloomberg.com/news/articles/2019-12-04/rappi-is-the-super-app-that-s-transforming-latin-america, 2019.

[Gao *et al.*, 2020] Yuan Gao, Jiajun Yang, Ming Yang, and Zhengshuo Li. Deep reinforcement learning based optimal schedule for a battery swapping station considering uncertainties. *IEEE Transactions on Industry Applications*, 56(5):5775–5784, 2020.

[Gulrajani *et al.*, 2017] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 5767–5777, 2017.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Alborz Zhou, Matti Hartikainen, Sriram Reddy, and Sergey Levine. Soft actor-critic: Algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[He and Shin, 2020] Suining He and Kang G. Shin. Dynamic flow distribution prediction for urban dockless e-scooter sharing reconfiguration. In *Proceedings of The Web Conference 2020*, pages 133–143, 2020.

[Huang *et al.*, 2022] Peide Huang, Mengdi Xu, Jiacheng Zhu, Laixi Shi, Fei Fang, and Ding Zhao. Curriculum reinforcement learning using optimal transport via gradual domain adaptation. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, pages 10656–10670, 2022.

[Jin *et al.*, 2023] Jiangliang Jin, Shuai Mao, and Yunjian Xu. Optimal priority rule-enhanced deep reinforcement learning for charging scheduling in an electric vehicle battery swapping station. *IEEE Transactions on Smart Grid*, 14(6):4581–4593, 2023.

[Julie Yujie Chen, Ping Sun, 2023] Julie Yujie Chen, Ping Sun. Digital labour platforms and national employment policies in china: Studying the case of food delivery platforms. https://webapps.ilo.org/static/english/intserv/working-papers/wp099/index.html, 2023.

[Kolouri *et al.*, 2017] Soheil Kolouri, Gustavo K. Park, Matthew Thorpe, Dejan Slepčev, and Gustavo K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

[Lang, 1993] Serge Lang. *Real and Functional Analysis*, volume 142 of *Graduate Texts in Mathematics*. Springer, 3rd edition, 1993.

[Li *et al.*, 2019] Xihan Li, Jia Zhang, Jiang Bian, Yunhai Tong, and Tie-Yan Liu. A cooperative multi-agent reinforcement learning framework for resource balancing in complex logistics network. In *Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 980–988, 2019.

[Li *et al.*, 2020] Yexin Li, Yu Zheng, and Qiang Yang. Cooperative multi-agent reinforcement learning in express system. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 805–814, 2020.

[Li *et al.*, 2024a] Zhao Li, Yang Liu, Chuan Zhou, Xuanwu Liu, Xuming Pan, Buqing Cao, and Xindong Wu. Transformer-based graph neural networks for battery range prediction in aiot battery-swap services. In *2024 IEEE International Conference on Web Services (ICWS)*, pages 1168–1176, 2024.

[Li *et al.*, 2024b] Zhao Li, Yang Liu, Chuan Zhou, Xuanwu Liu, Xuming Pan, Buqing Cao, and Xindong Wu. Transformer-based graph neural networks for battery range prediction in aiot battery-swap services. In *2024 IEEE International Conference on Web Services (ICWS)*, pages 1168–1176, 2024.

[Li *et al.*, 2024c] Zhao Li, Guoqi Ren, Yongchun Gu, Siwei Zhou, Xuanwu Liu, Jiaming Huang, and Ming Li. Real-time e-bike route planning with battery range prediction. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1070–1073, 2024.

[Liang *et al.*, 2023] Yanchang Liang, Zhaohao Ding, Tianyang Zhao, and Wei-Jen Lee. Real-time operation management for battery swapping-charging system via multi-agent deep reinforcement learning. *IEEE Transactions on Smart Grid*, 14(1):559–571, 2023.

[Mao *et al.*, 2024] Shuai Mao, Jiangliang Jin, and Yunjian Xu. Routing and charging scheduling for ev battery swapping systems: Hypergraph-based heterogeneous multiagent deep reinforcement learning. *IEEE Transactions on Smart Grid*, 15(5):4903–4916, 2024.

[Naderializadeh *et al.*, 2020] Navid Naderializadeh, Jaroslaw Sydir, Meryem Simsek, and Hosein Nikopour. Resource management in wireless networks via multi-agent deep reinforcement learning. In *2020 IEEE 21st*

International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pages 1–5, 2020.

[Rica Bhattacharyya, 2022] Rica Bhattacharyya. More than 400,000 jobs up for grabs fuelled by quick commerce. https://economictimes.indiatimes.com/tech/startups/demand-for-delivery-executives-on-the-rise-\amid-waning-covid-19-infections/articleshow/90523972.cms?from=mdr, 2022.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[Shalaby *et al.*, 2023] Ahmed A. Shalaby, Hussein Abdeltawab, and Yasser Abdel-Rady I. Mohamed. Model-free dynamic operations management for ev battery swapping stations: A deep reinforcement learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 24(8):8371–8385, 2023.

[Sun *et al.*, 2019] Bo Sun, Xu Sun, Danny H.K. Tsang, and Ward Whitt. Optimal battery purchasing and charging strategy at electric vehicle battery swap stations. *European Journal of Operational Research*, 279(2):524–539, 2019.

[Sutton and Barto, 2018] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

[United Nations, 2015] United Nations. Goal 8: Decent Work and Economic Growth — Sustainable Development Goals. https://sdgs.un.org/goals/goal8, 2015.

[Villani, 2003] Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.

[Widrick *et al.*, 2016] Rebecca S. Widrick, Sarah G. Nurre, and Matthew J. Robbins. Optimal policies for the management of an electric vehicle battery swap station. *Transportation Science*, 52(1):59–79, 2016.

[Yi *et al.*, 2019] Xin Yi, Zhuangzhuang Duan, Teng Li, Jun Zhang, Tianrui Li, and Yu Zheng. Citytraffic: Modeling citywide traffic via neural memorization and generalization approach. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1663–1672, 2019.

[Zhang *et al.*, 2018] Tianyang Zhang, Xi Chen, Zhe Yu, Xiaoyan Zhu, and Di Shi. A monte carlo simulation approach to evaluate service capacities of ev charging and battery swapping stations. *IEEE Transactions on Industrial Informatics*, 14(9):3914–3923, 2018.

[Zheng *et al.*, 2014] Yu Zheng, Zhao Yang Dong, Yan Xu, Ke Meng, Jun Hua Zhao, and Jing Qiu. Electric vehicle battery charging/swap stations in distribution systems: Comparison study and optimal planning. *IEEE Transactions on Power Systems*, 29(1):221–229, 2014.