eFlx: Energy Flexibility Provisioning for E-taxi Fleets

Liangkai Zhou
Stony Brook University
Stony Brook, NY, USA
liangkai.zhou@stonybrook.edu

Yue Zhao
Stony Brook University
Stony Brook, NY, USA
yue.zhao.2@stonybrook.edu

Yukun Yuan
University of Tennessee at Chattanooga,
Chattanooga, TN, USA
yukun-yuan@utc.edu

Ce Xu
Stony Brook University
Stony Brook, NY, USA
ce.xu@stonybrook.edu

Shan Lin
Stony Brook University
Stony Brook, NY, USA
shan.x.lin@stonybrook.edu

ABSTRACT

An e-taxi fleet consumes a significant amount of energy daily, making it a substantial electricity consumer. Unlike traditional consumers, such as factories and buildings, a fleet coordinates charging activities across both times and locations, offering considerable flexibility in its energy demand. This allows a fleet to achieve substantial reductions in energy consumption in response to demand response requests while maintaining transportation service quality. To better understand and control this intrinsic energy flexibility, we propose the eFlx framework for managing e-taxi fleets for demand response. In the eFlx framework, we establish a model to characterize the energy flexibility upon receiving a real-time demand response request. We then investigate the energy flexibility provisioning problem, formulated as a bi-level optimal control problem, which aims to optimize and maintain the energy flexibility of the fleet for potential demand response requests that could arise at any time. To achieve real-time flexibility provisioning, we develop an efficient iterative algorithm to solve this problem. Data-driven evaluations with NYC datasets demonstrate that eFlx achieves a 19. 98% greater reduction in energy demand compared to existing solutions, without requiring extra charging or compromising the quality of taxi service.

CCS CONCEPTS

• **Applied computing** → **Transportation**; • **Computing methodologies** → **Planning and scheduling**; • **Hardware** → **Power and energy**;

KEYWORDS

Energy flexibility, e-taxi fleet, real-time provisioning, demand response, grid services

ICCPS '25, May 6–9, 2025, Irvine, CA, USA

ACM ISBN 979-8-4007-1498-6/2025/05…$15.00
https://doi.org/10.1145/3716550.3722026

ACM Reference Format:
Liangkai Zhou, Yue Zhao, Yukun Yuan, Ce Xu, and Shan Lin. 2025. eFlx: Energy Flexibility Provisioning for E-taxi Fleets . In *ACM/IEEE 16th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2025) (ICCPS '25), May 6–9, 2025, Irvine, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3716550.3722026

1 INTRODUCTION

As electric vehicle (EV) technologies mature, electric taxis and buses are rapidly being deployed in numerous cities around the world. Metropolitan areas, in particular, see EV fleets as a popular upgrade to their transportation systems. For example, Shenzhen [1], Amsterdam [2], and London [3] have already implemented e-taxi fleets. Additionally, Tesla-manufactured e-taxis are operating as yellow cabs in New York City. With expanding infrastructure for charging and improvements in battery technology, the global electric vehicle taxi market is expected to grow with a compound annual growth rate of 11.3% from 2024 to 2030 [4].

Notably, a modern e-taxi's daily operations can consume more than 50 kWh [5] energy. As such, the widespread deployment of E-taxi fleets represents a very significant share of the electricity demand. These dynamic loads, together with renewable generation such as wind and solar, introduce substantial uncertainties and variability in power supply and demand, challenging the grid's ability to balance reliably. This issue is further exacerbated by ongoing climate change.

Increased *flexibility* of power system operations has been widely recognized as a key to successfully tackling these challenges. Various Demand Response (DR) programs have been implemented to leverage flexibility on the demand side to maintain the reliability of the power grid in the presence of unpredictable supply-demand imbalance [6]. For example, a building might adjust the HVAC set temperature to reduce the air conditioning load [7], or industrial production may need to slow down [8] in response to an emergency DR request due to extreme weather conditions [9].

An e-taxi fleet, as a significant energy consumer, possesses uniquely desirable characteristics as a DR provider. First, each EV is equipped with batteries that serve as an energy reserve. The stored energy is used for passenger service and replenished through charging at stations. Moreover, in an e-taxi fleet, the workloads and charging activities of numerous e-taxis can be coordinately shifted spatial-temporally without affecting passenger service. This flexibility is distinct from traditional DR resources, such as HVAC or factories,

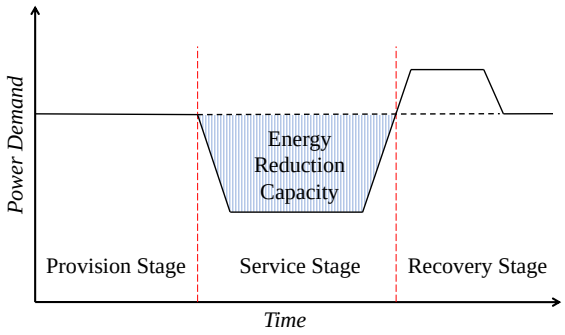


Figure 1: Three stages of a demand response program.

which i) are fixed in locations and ii) require continuous energy supply to provide services. Therefore, an EV fleet offers uniquely valuable flexibility during contingencies, helping the power system manage unpredictable variations in demand and supply.

To address this issue, it is crucial to characterize the intrinsic energy flexibility of an e-taxi fleet. Unlike other flexible demands that can be directly measured, the flexibility of the e-taxi fleet depends on passenger service requirements and e-taxi mobility. In this paper, a model is established to estimate the maximum energy savings that an e-taxi fleet can achieve using a flexible dispatch algorithm for a given DR request, while maintaining the same service quality. This model is valuable for the power grid to analyze and design DR programs and for taxi companies to mitigate any negative impacts on their transportation services. Notably, to quantify the flexibility fairly, we specifically constrain the flexible e-taxi dispatch algorithm to consume no more energy, at any time, than the normal dispatch algorithm in the period leading up to a DR service window. As such, we guarantee that there is no negative impact on the grid from potential preparatory activities by an e-taxi fleet before any DR service window starts. The energy flexibility is thus achieved entirely from the optimized e-taxi workload and mobility via centralized coordination.

To maximize the flexibility of an E-taxi fleet *at all times* — ensuring it can provide flexibility *whenever there is a need* without compromising transportation service quality — we study the *flexibility provisioning* problem. Specifically, how can an e-taxi fleet achieve and maintain a "state" of maximum flexibility so that it can promptly respond to emergency DR requests, even without any advance notice? We note that optimal provisioning of flexibility at all times would provide the highest level of energy system resilience against unforeseeable DR needs. The problem is formulated as a bi-level optimization problem whose computational complexity is exponential. Considering the scale of our problem, existing algorithms cannot solve it directly while meeting the runtime requirements. To reach optimal energy flexibility provisioning in real time, an efficient iterative algorithm is designed.

The contributions of this paper are as follows:

- To the best of our knowledge, this is the first work to investigate energy flexibility of an e-taxi fleet. This work reveals that an e-taxi fleet has intrinsic flexibility to achieve a substantial reduction in energy demand by shifting workloads and charging activities of e-taxis without affecting taxi services.
- An energy flexibility model is constructed to quantify the potential energy demand reduction of an e-taxi fleet in response to a DR request.
- An energy flexibility provisioning problem is formulated, which maximizes the energy flexibility of an e-taxi fleet ready to be released at all times in response to any DR request. This is a bi-level linear programming (LP) problem. To achieve optimal real-time flexibility provisioning, an efficient iterative algorithm is developed to solve the provisioning problem for large-scale e-taxi fleets in practice.
- Using New York City datasets, a comprehensive data-driven evaluation is conducted: Without extra charging or reduced taxi service quality, compared to reference DR solutions, our flexibility provisioning solution increases the energy flexibility for a two-hour emergency DR program by an *additional* 7.11 MWh and reduces the energy demand by 19.98% with minimal overhead.

2 ENERGY FLEXIBILITY

2.1 An e-taxi fleet in a DR program

To manage emergencies such as generator failures, inaccurate renewable forecasts, extreme weather, etc. [9], emergency DR programs alert consumers to reduce their energy demand for a set period due to reserve shortages or reliability needs [10]. Participants are rewarded based on the demand reduction they provide. Furthermore, in energy markets around the world, DR ancillary service markets have been established where demand-side participants need to promptly respond to DR activation signals that can arise at any time of the day. For instance, NYISO's Demand-Side Ancillary Services Program (DSASP) [11] allows DR providers to bid in the day-ahead market for spinning reserve services. In such markets, participants must promptly respond to the Independent System Operator (ISO)'s real-time instructions throughout the day.

To stay ready to respond to any potential DR request at any time, we consider an energy flexibility provisioning problem for an e-taxi fleet that comprises three conceptual stages (cf. Fig. 1): provision, DR service, and recovery [12]. First, the power system sends requests to consumers inviting them to participate in DR services. Upon receiving the request, an e-taxi fleet willing to participate in the DR program can submit a bid. If the fleet wins the bid, it enters the provision stage, standing by for activation instructions from the power system operator and continuously managing its resources (i.e., provision flexibility) to respond to activation signals that may arise at any time. As such, the fleet is optimized to maximize energy flexibility, ensuring that it is ready to be released at any time.

The service stage begins upon receiving activation instructions, during which the fleet is required to provide load reduction (i.e., to release flexibility) promptly (e.g., within 10 minutes [11]). Throughout the service stage, the fleet minimizes its power demand to meet the DR need, helping balance the overall supply and demand of the power grid. Once the service stage ends, as specified by the activation instructions and contractual terms, the fleet transitions to the recovery stage, during which its load gradually ramps up, and its flexibility is restored. We note that, in the context of DR ancillary markets, a participating e-taxi fleet would in effect perform flexibility provisioning *continuously at all times* before a DR

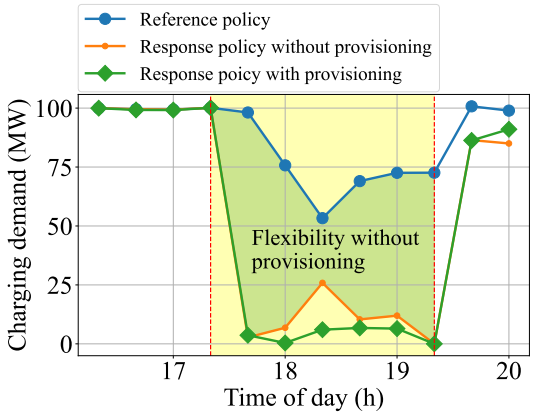


Figure 2: Charging demand of an e-taxi fleet in the presence of a DR service request.

activation request arises, at which time the flexibility can then be released to provide demand reduction.

2.2 A motivational example

This work is motivated by the observation that an e-taxi fleet has significant potential to provide DR services without compromising the quality of passenger service. To demonstrate this, we conduct a case study to explore the power demand flexibility of a fleet participating in a DR program. We evaluate the fleet's charging demand under three distinct coordination algorithms: (i) Reference policy: it focuses solely on optimizing passenger service quality, instead of participating in DR programs; (ii) Response policy without flexibility provisioning: It does not proactively deviate from the above reference policy. Upon the arrival of a DR service stage, it switches from the reference policy to flexible operations to minimize power demand; (iii) Response policy with flexibility provisioning: it is developed in this work with details in Sec. 5. Unlike (i) and (ii), this method proactively schedules e-taxis during the "provisioning stage" to ensure readiness for a potential upcoming DR service stage. Meanwhile, all passengers must be picked up.

Fig. 2 illustrates the charging demand of an e-taxi fleet under the three policies. All three policies maintain the same overall passenger service. The vertical dashed lines divide the timeline into the three stages of the DR program. During the service stage, from 17:20 to 19:20, the response policy significantly reduces energy consumption while maintaining the quality of transportation. The green area between the reference policy and the response policy highlights the reduction in energy demand achieved by the e-taxi fleet without flexibility provisioning. During the two-hour period, the energy demand is reduced by 127.86 MWh when flexibility provisioning is not implemented before the service stage. Comparison of response policies with and without provisioning shows that provisioning increases this reduction to 139.37 MWh. In summary, Fig. 2 demonstrates the viability of leveraging the energy flexibility of the fleet to provide DR services, with provisioning proving essential for enhancing the quality of the DR service.

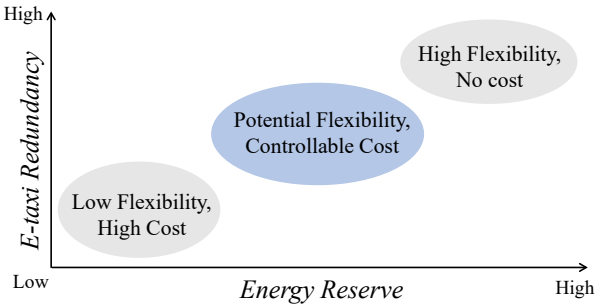


Figure 3: Flexibility vs. cost

2.3 Understanding e-taxis Fleet's flexibility

The energy flexibility of an e-taxi fleet primarily comes from two underlying characteristics: redundancy of e-taxis and energy reserve, as shown in Fig. 3. Here, the redundancy of e-taxis represents the extra e-taxi supply after matching passenger demand, whereas the energy reserve represents the amount of energy stored in the batteries of all e-taxis. Intuitively, the fleet can provide high flexibility when there are a large number of extra e-taxis and many e-taxis have high remaining battery levels, i.e., high e-taxi redundancy and high energy reserve. Conversely, when there are few extra e-taxis and many e-taxis have low battery levels, the fleet can hardly provide any energy demand reduction. To provide energy flexibility in this latter scenario, the fleet has to sacrifice service quality.

Between the two extremes of e-taxi fleet states of very high and low flexibility, transportation service needs may barely be met and energy reduction cannot be easily achieved, since the fleet has some but limited redundancy and energy reserve for energy-flexible operations. This work focuses on such challenging situations and studies the research question of how to optimize e-taxi fleet operations to achieve the maximum flexibility.

2.4 eFlx framework overview

Fig. 4 shows the energy flexibility (eFlx) framework design for the e-taxi service. The core component is the e-taxi coordination algorithm, which operates in three stages corresponding to the stages of the DR service as shown in Fig. 1. In the provision stage, the algorithm optimally dispatches e-taxis to enhance future flexibility for DR services, ensuring that charging demand remains at normal levels. During the service phase, the algorithm reduces charging activities to deliver flexibility and meet the required energy demand reductions. Importantly, transportation service quality should generally remain unaffected during this phase, although the e-taxi fleet can opt to allow a controlled level of service compromise to contribute more effectively to the DR service. In the final recovery stage, the algorithm increases charging activities to rapidly restore energy reserves, ensuring a swift resumption of full transportation service.

The eFlx framework operates with two closed loops. The first loop connects the power system and the e-taxi system. Power system managers transmit various DR-related information, such as service requests, bid selections, and activation instructions, to the e-taxi fleet coordinator. Based on this received information, the

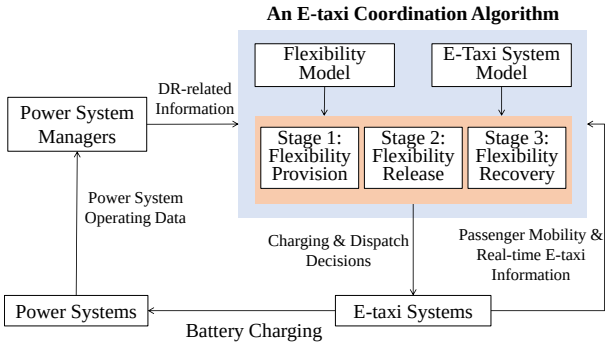


Figure 4: eFlx framework

appropriate phase of the coordination algorithm is triggered to determine the optimal charging and dispatch decisions for the e-taxis. These charging activities impact the overall power demand and this feedback is continuously relayed to the power system managers. The second loop functions within the e-taxi system itself. During daily operations, the e-taxi coordinator generates dispatch commands for passenger pickups. Passenger trip data is collected and processed to monitor real-time e-taxi statuses and predict future passenger demand. This information is then used by the coordinator to optimize dispatch decisions, ensuring efficient fleet management.

3 E-TAXI SYSTEM MODEL

3.1 E-taxi system

3.1.1 ***System state.*** We begin by discretizing the spatial and temporal domains. A day is divided into uniform time slots, indexed by t. The city area served by the e-taxi fleet is partitioned into n regions, aligned with the power grid structure, i.e., each region is powered by a common area substation. Let $\bar{n}$ represent the number of charging stations distributed across the city. The battery capacity of an e-taxi is divided into $\hat{L}$ discrete levels. In our evaluation, we set $\hat{L}$ to 15, which is sufficient to capture the energy dynamics of the e-taxis. Let RE^t denote the remaining energy level of an e-taxi at the beginning of slot t. If an e-taxi charges its battery during slot t, its energy for slot $t+1$ is $RE^{t+1} = RE^t + \hat{L}_2$, where $\hat{L}_2$ represents the energy gained after charging for one time slot. Conversely, if an e-taxi is in operation, either serving passengers or searching for them, its energy level decreases to $RE^{t+1} = RE^t - \hat{L}_1$, where $\hat{L}_1$ represents the energy consumption after working for one time slot. If an e-taxi is idle, waiting at a charging station for an available port, its energy level remains unchanged: $RE^{t+1} = RE^t$.

Based on the location and operational status of an e-taxi, we define four states for the e-taxi in this model: vacant, charging, waiting, and occupied. Vacant: an e-taxi is cruising the streets, searching for or picking up passengers. Charging: an e-taxi is actively being charged at a charging station. Waiting: an e-taxi is idle at a charging station, waiting for an available charging port. Occupied: an e-taxi is in service, transporting passengers to their destination.

We define the state of the e-taxi system as the distribution of e-taxis across different states and energy levels in spatial-temporal dimensions, represented by $V_i^{l,t}$, $O_i^{l,t}$, and $D_j^{l,t}$. Here, $V_i^{l,t}$ and $O_i^{l,t}$ denote the number of vacant and occupied e-taxis, respectively, in region i with energy level l at the beginning of time slot t. Additionally, $D_j^{l,t}$ represents the number of e-taxis at the charging station j, with remaining energy level l, which includes both e-taxis actively charging and those waiting for a charging port.

3.1.2 ***Decision variables.*** In this work, we focus on dispatching unoccupied e-taxis, specifically, vacant, charging, and waiting taxis at the beginning of each time slot, using two types of dispatch decisions: dispatch for charging and dispatch for serving passengers. Let $x_{i,j}^{l,t} \in \mathbb{N}$ represent the number of vacant e-taxis with remaining energy level l that are dispatched from region i to charging station j during time slot t. Similarly, we define $xd_{j,j'}^{l,t} \in \mathbb{N}$ as the number of e-taxis with energy level l dispatched from charging station j to station j' in the same time slot. The total number of e-taxis initially located in region i and dispatched to charging station j is expressed as $X_{i,j}^{l,t} = x_{i,j}^{l,t} + \sum_{j'=1}^{\bar{n}} R_{i,j'}^{rc} xd_{j',j}^{l,t}$, where $R_{i,j'}^{rc} \in \{0,1\}^{n \times \bar{n}}$ indicates the geographical relationship between regions and charging stations. Specifically, $R_{i,j'}^{rc} = 1$ when charging station j' is located in region i, and $R_{i,j'}^{rc} = 0$ otherwise.

For dispatching e-taxis to serve passengers, let $y_{i,i'}^{l,t} \in \mathbb{N}$ denote the number of vacant taxis with energy level l dispatched from region i to region i' in slot t. $yd_{j,i}^{l,t}$ represents the e-taxis with energy level l dispatched from charging station j to region i in time slot t for passenger service. The total number of e-taxis initially in region i dispatched to region i' is given by $Y_{i,i'}^{l,t} = y_{i,i'}^{l,t} + \sum_{j'=1}^{\bar{n}} R_{i',j}^{rc} yd_{j,i'}^{l,t}$. For simplicity, we represent the decision variables for charging dispatches as $X^t = \{x_{i,j}^{l,t}, xd_{j,j'}^{l,t}\}_{l,i,j,j'}$, and those for passenger service dispatches as $Y^t = \{y_{i,i'}^{l,t}, yd_{j,i}^{l,t}\}_{l,i,i',j}$ at the start of time slot t.

Finally, the number of available e-taxis must equal the number of dispatched e-taxis, leading to the following constraints:

$$\sum_{j=1}^{\bar{n}} x_{i,j}^{l,t} + \sum_{i'=1}^{n} y_{i,i'}^{l,t} = V_i^{l,t}, \quad \sum_{j'=1}^{\bar{n}} xd_{j,j'}^{l,t} + \sum_{i=1}^{n} yd_{j,i}^{l,t} = D_j^{l,t}. \quad (1)$$

Due to the inefficiency of mixed-integer programming for handling large-scale taxi networks because of the problem's size, we relax $V_i^{l,t}$, $O_i^{l,t}$, $D_j^{l,t}$ and all decision variables to be within $\mathbb{R}$.

3.2 E-taxi Supply and Passenger Demand

For any time slot of the day, the number of passenger requests with specific departure and destination regions can be estimated using historical records. Let r_i^t denote the number of passenger requests in region i during time slot t. Moreover, let $S_i^{l,t}$ represent the number of e-taxis with remaining energy l that are available for serving passengers in region i at time slot t after the dispatch. The transition of the e-taxi system can be described as follows:

$$S_i^{l,t} = \sum_{i'=1}^{n} Y_{i',i}^{l,t}, \tag{2a}$$

$$V_i^{l,t+1} = \sum_{i'=1}^{n} Pv_{i',i}^{l,t} S_{i'}^{l+\hat{L}_1,t} + \sum_{i'=1}^{n} Qv_{i',i}^{l,t} O_{i'}^{l+\hat{L}_1,t}, \tag{2b}$$

$$O_i^{l,t+1} = \sum_{i'=1}^{n} Po_{i',i}^{l,t} S_{i'}^{l+\hat{L}_1,t} + \sum_{i'=1}^{n} Qo_{i',i}^{l,t} O_{i'}^{l+\hat{L}_1,t}. \tag{2c}$$

Here, $Pv_{i',i}^{l,t}$, $Qv_{i',i}^{l,t}$, $Po_{i',i}^{l,t}$, and $Qo_{i',i}^{l,t} \in [0,1]$ represent the mobility patterns of e-taxis during the time slot t:

- $Pv_{i',i}^{l,t}$ denotes the probability that the unoccupied taxis in region i' travel to region i and remain vacant, either while searching for passengers or after completing a passenger drop-off.
- $Po_{i',i}^{l,t}$ represents the probability that the unoccupied taxis in region i' travel to region i while still delivering a passenger (i.e., the delivery is ongoing).
- $Qv_{i',i}^{l,t}$ indicates the probability that the occupied taxis in region i' travel to region i without completing the passenger delivery.
- $Qo_{i',i}^{l,t}$ denotes the probability that the occupied taxis in region i' travel to region i and complete their passenger delivery.

The mobility probabilities must satisfy the following constraints: $\sum_{i=1}^{n} Pv_{i',i}^{l,t} + Po_{i',i}^{l,t} = 1$, $\sum_{i=1}^{n} Qv_{i',i}^{l,t} + Qo_{i',i}^{l,t} = 1$. The values for these mobility patterns are derived from historical taxi mobility data.

3.3 Charging Supply and Request Model

This section illustrates the interaction between the e-taxi fleet and the power grid. First, we define the charging supply from the power grid and the charging requests from the e-taxi fleet. We then derive the relationship between the charging decision variables and the dynamic behavior of the e-taxi fleet.

The charging supply refers to the maximum number of e-taxis that can be charged simultaneously at a given station. Let e_j^t represent the charging supply, meaning no more than e_j^t e-taxis can be charged at charging station j during time slot t. On the other hand, the charging request refers to the number of e-taxis that are actually charged at station j during time slot t, with an initial remaining energy level of l, denoted as $u_j^{l,t}$. Since the number of charging ports may be insufficient to accommodate all e-taxis dispatched to a station, only a portion of the e-taxis can charge. The others will wait idly. The variable $u_j^{l,t}$ represents the number of e-taxis being charged in region j during time slot t, whose initial energy level is l. We express this with the following constraints:

$$u_j^{l,t} \le \sum_{i=1}^{n} x_{i,j}^{l,t} + \sum_{j'=1}^{\bar{n}} x d_{j',j}^{l,t}, \quad \sum_{l=1}^{\hat{L}} u_j^{l,t} \le e_j^t, \tag{3}$$

which ensures that the number of e-taxis charging at station j is limited by the number of e-taxis dispatched to the station and the available charging ports. And we can derive the number of e-taxis in the charging station j at the beginning of next time slot $t+1$ as:

$$D_j^{l,t+1} = u_j^{l-\hat{L}_2,t} + \sum_{i=1}^{n} X_{i,j}^{l,t} - u_j^{l,t}, \tag{4}$$

where the first term represents the e-taxis that gained energy during charging, while the combination of the second and third terms represents the e-taxis waiting idly for an available charging port. For simplicity, we denote the charging decision variables as $U^t = \{u_j^{l,t}\}_{l,j}$. Additionally, let $\mathcal{S}^t = \{V_i^{l,t}, O_i^{l,t}, D_j^{l,t}, S_i^{l,t}, u_j^{l,t-1}\}$ represent the state of the e-taxi system at time slot t.

Let $\hat{t}$ represent the time when the fleet starts provisioning flexibility. Additionally, let T_b and τ denote the duration of the service stage and the provision stage, respectively, as specified in the DR program. We also introduce an additional temporal parameter, T_c, which defines the optimization horizon. Letting t' denote the current time slot, we denote the decision variables of policy π in the provision stage as $\mathcal{X}_{\text{pro}}^{\pi,t'} = \{X^{t_{ps}:t_{pe}}, Y^{t_{ps}:t_{pe}}, U^{t_{ps}:t_{pe}}\}$, where $t_{ps} = \max\{t, \hat{t}\}$ and $t_{pe} = \max\{t', \hat{t}+\tau-1\}$. Similarly, the decision variables of policy π after the provision stage are denoted as $\mathcal{X}_{\text{ser}}^{\pi,t'} = \{X^{t_{pe}:t_{se}}, Y^{t_{pe}:t_{se}}, U^{t_{pe}:t_{se}}\}$, where $t_{se} = \max\{t', \hat{t}+T_c-1\}$.

4 ENERGY FLEXIBILITY MODEL AND PROVISIONING

4.1 Energy Flexibility Definition

In this work, we investigate the e-taxi fleet coordination policy designed to provision energy flexibility for an upcoming DR request that may arise at any time. The flexibility of an e-taxi fleet is defined based on a specific fleet state $\mathcal{S}^t$ and a coordination (of charging and dispatch) policy π. Given a *known* upcoming DR service window, the flexibility of an e-taxi fleet refers to the amount of *energy consumption for charging that the fleet can reduce* during the DR service stage, without any loss of passenger service. This reduction in charging energy of π captures the difference between two fleet coordination policies: 1) the "baseline" policy π itself, which does not respond to the DR activation signal, and 2) a "response" policy π^{res}, which switches from baseline π during the DR service stage to respond to the DR activation signal and deliver flexibility.

As an example from Fig. 2, the flexibility of the reference policy π^{ref} is highlighted by the green area. It is defined as the gap in charging energy between π^{ref} (i.e., the baseline) and the response policy without provisioning (i.e., switching from π^{ref} to DR response), denoted by π^{res} (with a slight abuse of notation). At a high level, we aim to achieve the optimal flexibility provisioning policy π^*, which maximizes the flexibility of the fleet state at the beginning of a DR service stage. The following sections will provide a detailed explanation of these concepts.

4.2 Response Policy

The response policy π^{res} aims to *maximally* reduce the energy demand for charging during the service stage while maintaining the performance of the transportation service. The objective of π^{res} is $f_{\text{res}}(\mathcal{S}^{t'}, \mathcal{X}_{\text{pro}}^{\pi,t'}, \mathcal{X}_{\text{ser}}^{\text{res},t'}) = \sum_{t=t_{\text{pe}}}^{t_{\text{pe}}+T_b} \sum_{j=1}^{n} u_{j,\text{res}}^{l,t}$, where π refers to a baseline policy. This objective aims to minimize the total number of charging e-taxis during the DR service stage. The coordination decisions of the response policy after time slot t' can be derived by solving the following optimization problem:

Problem 1. *Flexibility Delivery Problem:*

$$\min_{\mathcal{X}_{ser}^{res,t'}} f_{res}(\mathcal{S}^{t'}, \mathcal{X}_{pro}^{\pi,t'}, \mathcal{X}_{ser}^{res,t'}), \tag{5a}$$

$$s.t. \quad \sum_{i=1}^{n}\sum_{l=1}^{\hat{L}} u_{i,\pi}^{l,t} \geq \sum_{i=1}^{n}\sum_{l=1}^{\hat{L}} u_{i,res}^{l,t} \qquad t \in [t_{ps}, t_{pe}] \tag{5b}$$

$$J_{trans,res}^{t} \geq J_{trans,\pi}^{t} \quad t \in [t', t'+T_c], \tag{5c}$$

$$\sum_{t=t'}^{t'+T_c} J_{idle,res}^{t} \leq (1+\eta) \sum_{t=t'}^{t'+T_c} J_{idle,\pi}^{t}, \tag{5d}$$

$$X_{i,j}^{l,t} dc_{i,j}^{t} = 0 \quad t \in [t', t'+T_c), \tag{5e}$$

$$Y_{i,i'}^{l,t} ds_{i,i'}^{t} = 0 \quad t \in [t', t'+T_c), \tag{5f}$$

$$S_{i}^{l,t} = 0, \quad 1 \leq l \leq \hat{L}_1, \quad t \in [t', t'+T_c), \tag{5g}$$

$$Eq.\ (1) \sim (4).$$

where $J_{trans}^{t} = \sum_{i=1}^{n} \min\{r_i^t, \sum_{l=1}^{\hat{L}} S_i^{l,t}\}$ *is the number of served passengers in slot* t *and* $J_{idle}^{t} = \sum_{l}(\sum_{i,j}(x_{i,j}^{l,t} + y d_{j,i}^{l,t})\varphi_{i,j} + \sum_{i,i'} y_{i,i'}^{l,t}\mu_{i,i'} + \sum_{j,j'} x d_{j,j'}^{l,t} v_{j,j'})$ *is the idle driving distance caused by the coordination decisions. Here,* $\varphi_{i,j}$ *denotes the distance between region* i *and charging station* j*;* $v_{j,j'}$ *denotes the distance between charging stations* j *and* j'*; and* $\mu_{i,i'}$ *denotes the distance between regions* i *and* i'*.*

In the above problem, the constraint (5b) restricts π^{res} from charging beyond the baseline to avoid negative impacts on the power grid. The constraint (5c) ensures that the response policy does not reduce the taxi service quality compared to the reference policy throughout the service stage. The constraint (5d) ensures that the total idle driving distance under the response policy does not exceed an increase of η compared to the reference policy. All dispatch decisions must be completed within one time slot to ensure practical feasibility. Thus, scheduling e-taxis to regions and charging stations that cannot be reached within one time slot is prohibited, as enforced by constraints (5e) and (5f). $dc_{i,j}^{t} \in \{0,1\}$ and $ds_{i,i'}^{t} \in \{0,1\}$ indicate whether a vehicle can reach charging station j or region i' from region i within one time slot. Specifically, if an e-taxi can reach region i' (or charging station j) within time slot t, then $ds_{i,i'}^{t} = 0$ ($dc_{i,j}^{t} = 0$); otherwise, $ds_{i,i'}^{t} = 1$ ($dc_{i,j}^{t} = 1$). The constraint (5g) ensures that e-taxis with an energy level below $\hat{L}_1$ do not serve passengers, thereby preventing vehicles from running out of energy on the road. Given the relaxation on decision variables, Problem 1 is an LP problem and the computational complexity is polynomial in the problem size V_n, where $V_n = 4n^2\hat{L}T_c + n\hat{L}T_c$. For simplicity, we use $\{g_{\text{res},k}(\mathcal{S}^{t'}, \mathcal{X}_{\text{pro}}^{\pi,t'}, \mathcal{X}_{\text{ser}}^{\pi,t'}, \mathcal{X}_{\text{ser}}^{\text{res},t'})\}$ to denote the constraints in Problem 1.

4.3 Provisioning Flexibility

The flexibility of an e-taxi fleet state $\mathcal{S}^{t'}$ under policy π is quantified as the disparity in the total number of e-taxis that are charged per time slot during the service stage, between π and π^{res}, i.e.,

$$F_{\text{pro}}(\mathcal{S}^{t'}, \mathcal{X}_{\text{pro}}^{\pi,t'}, \mathcal{X}_{\text{ser}}^{\pi,t'}, \mathcal{X}_{\text{ser}}^{\text{res},t'}) = \sum_{t=t_{\text{pe}}}^{t_{\text{pe}}+T_b}\sum_{j=1}^{n} u_{j,\pi}^{l,t} - \sum_{t=t_{\text{pe}}}^{t_{\text{pe}}+T_b}\sum_{j=1}^{n} u_{j,\text{res}}^{l,t}. \tag{6}$$

The first and second terms represent the energy demand under π and π^{res} during the service stage, respectively. Thus, finding the optimal provision policy π^* can be formulated at a high level as:

$$\max_{\pi} F_{\text{pro}}(\mathcal{S}^{t}, \pi, \pi^{\text{res}}). \tag{7}$$

Specifically, we solve the following optimization problem with an outer problem (8a) and an inner problem (8c):

Problem 2. *Flexibility Provisioning Problem:*

$$\max_{\mathcal{X}} F_{pro}(\mathcal{S}^{t'}, \mathcal{X}_{pro}^{\pi,t'}, \mathcal{X}_{ser}^{\pi,t'}, \mathcal{X}_{ser}^{res,t'}), \tag{8a}$$

$$s.t. \quad \mathcal{X}_{ser}^{res,t'} \in \arg\min_{\hat{\mathcal{X}}_{ser}^{res,t'}} \{f_{res}(\mathcal{S}^{t'}, \mathcal{X}_{pro}^{\pi,t'}, \hat{\mathcal{X}}_{ser}^{res,t'}) : \{g_{res,k}(\mathcal{S}^{t'}, \mathcal{X}_{pro}^{\pi,t'}, \mathcal{X}_{ser}^{\pi,t'}, \hat{\mathcal{X}}_{ser}^{res,t'})\}\}, \tag{8b}$$

$$\mathcal{X}_{ser}^{\pi,t'} \in \arg\max_{\hat{\mathcal{X}}_{ser}^{\pi,t'}} \{F_{pro}(\mathcal{S}^{t'}, \hat{\mathcal{X}}_{ser}^{\pi,t'}, \mathcal{X}_{ser'}^{\pi,t'}, \mathcal{X}_{ser'}^{res,t'})\} \tag{8c}$$

$$\sum_{i=1}^{n}\sum_{l=1}^{\hat{L}} u_{i,\pi}^{l,t}, \leq P_{TRC}^{t} \qquad \forall t \in [t_{ps}, t_{pe}] \tag{8d}$$

$$Eq.\ (1) \sim (4),\ and\ (5e) \sim (5g).$$

where (8c) *captures the baseline policy* π *without responding to the DR service request. ser*$'$ *refers to another time period immediately following the service stage.* π *is assumed to deliver flexibility during ser*$'$ *in constraint 8c. In constraint* (8d)*, constant* P_{TRC}^{t} *represents the charging demand of the regular taxi services. This constraint prevents overcharging during provisioning.*

Problem 2 is, however, fundamentally difficult to solve. This is because the provisioning policy π, as the optimization variable, appears both in a) the outer problem to maximize the provisioned flexibility and in b) the inner problem as the non-responding baseline policy from which flexibility is computed. In other words, what we optimize—the provisioning policy π—also determines the *baseline* from which the demand reduction, i.e., flexibility, is computed. This loop results in an essentially infinite-level optimization problem. Specifically, another time period, ser$'$, is introduced immediately following the service stage. This is because we need to extend the provisioning stage to evaluate the non-responding π, and an additional service stage after it is needed. Thus, in order to derive $\mathcal{X}_{\text{ser}'}^{\pi,t'}$, another time period ser$''$ after ser$'$ must be introduced again. As such, Problem 2 becomes an infinitely looped optimization problem, which is difficult to solve directly. To break the infinite loop and compute the flexibility, instead of comparing with the non-responding provisioning policy in the inner problem as the baseline, we employ a reference policy π^{ref} instead, essentially reducing the infinite-level problem to a bi-level problem. Details are provided below.

4.4 Reference Policy

The reference policy π^{ref} aims to provide a consistent and efficient transportation service. Specifically, it maximizes the number of passengers served J^t_{trans} while minimizing idle driving distance J^t_{idle}, which are common objectives in taxi dispatch algorithms [13, 14]. The objective of the reference policy is formulated as $f_{\text{ref}}(\mathcal{S}^{t'}, \mathcal{X}^{\pi,t'}_{\text{pro}}, \mathcal{X}^{\text{ref},t'}_{\text{ser}}) = \sum_{t=t'}^{t'+T_c} J^t_{\text{trans}} - \beta J^t_{\text{idle}}$, where β is a positive parameter that balances the trade-off between optimizing service quality and minimizing idle driving distance. The reference policy π^{ref} can be derived by solving the following problem:

Problem 3. *Taxi Service Quality Optimization Problem:*

$$\max_{\mathcal{X}^{ref,t'}_{ser}} f_{ref}(\mathcal{S}^{t'}, \mathcal{X}^{\pi,t'}_{pro}, \mathcal{X}^{ref,t'}_{ser}), \tag{9a}$$

$$s.t.\ Eq.\ (1) \sim (4), (5e) \sim (5g).$$

In Problem 3, $\mathcal{X}^{\pi,t'}_{\text{pro}}$ represents the coordination decisions that determine the state of the e-taxi fleet at the beginning of the service stage. $\mathcal{X}^{\text{ref},t'}_{ser}$ will serve as a baseline for flexibility calculation during the service stage. Similarly, the computational complexity of Problem 3 is polynomial in the problem size V_n and $V_n = 4n^2\hat{L}T_c + n\hat{L}T_c$. For simplicity, we also use $\{g_{\text{ref},k}(\mathcal{S}^{t'}, X^{\pi,t'}_{\text{pro}}, X^{\text{ref},t'}_{\text{ser}})\}$ to denote the constraints of Problem 3.

4.5 Bi-level formulation for the flexibility provisioning problem

As mentioned above, the non-responding π in constraints (8b) and (8c) results in an infinitely looped optimization. Therefore, we replace π with π^{ref} in constraints (8b) and (8c). Since π^{ref} aims to solely maximize passenger service, it introduces tighter constraints on passenger demand, which thus guarantees the feasibility of the optimized provisioning policy π^*. The problem of flexibility provisioning becomes:

Problem 4. *Flexibility Provisioning Problem with reference policy as baseline during the service stage:*

$$\max_{\mathcal{X}} F_{pro}(\mathcal{S}^{t'}, \mathcal{X}^{\pi,t'}_{pro}, \mathcal{X}^{ref,t'}_{ser}, \mathcal{X}^{res,t'}_{ser}), \tag{10a}$$

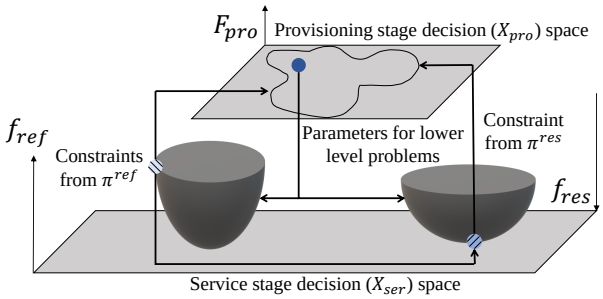
$$s.t.\quad \mathcal{X}^{res,t'}_{ser} \in \arg\min_{\hat{\mathcal{X}}^{res,t'}_{ser}} \{f_{res}(\mathcal{S}^{t'}, \mathcal{X}^{\pi,t'}_{pro}, \hat{\mathcal{X}}^{res,t'}_{ser}) : \{g_{res,k}(\mathcal{S}^{t'}, \mathcal{X}^{\pi,t'}_{pro}, \mathcal{X}^{ref,t'}_{ser}, \hat{\mathcal{X}}^{res,t'}_{ser})\}\}, \tag{10b}$$

$$\mathcal{X}^{ref,t'}_{ser} \in \arg\max_{\hat{\mathcal{X}}^{ref,t'}_{ser}} \{f_{ref}(\mathcal{S}^{t'}, \mathcal{X}^{\pi,t'}_{pro}, \hat{\mathcal{X}}^{ref,t'}_{ser}) : \{g_{ref,k}(\mathcal{S}^{t'}, \mathcal{X}^{\pi,t'}_{pro}, \hat{\mathcal{X}}^{ref,t'}_{ser})\}\}, \tag{10c}$$

$$Eq.\ (1) \sim (4),\ (5e) \sim (5g)\ and\ (8d).$$

Remark 1. *Problem 4 involves two levels, as shown in Fig. 5. Both Problem 3 and Problem 1 constitute the lower-level problem, where* $\mathcal{X}^{ref,t'}_{ser}$ *and* $\mathcal{X}^{res,t'}_{ser}$ *are lower-level decision variables and* $\mathcal{X}^{\pi,t'}_{pro}$ *act as parameters. Equation* (10a) *represents the upper-level optimization problem, with Problem 3 and Problem 1 serving as constraints. Only those* $\mathcal{X}^{ref,t'}_{ser}$ *and* $\mathcal{X}^{res,t'}_{ser}$ *that are optimal in the low-level problems and also satisfy the upper-level constraints are considered feasible. As a result, Problem 4 suffers from non-convexity and disjoint feasible regions due to constraints* (10c) *and* (10b). *Moreover, evaluating the feasibility of any specific combination of decision variables* $\{\mathcal{X}^{\pi,t'}_{pro}, \mathcal{X}^{ref,t'}_{ser}, \mathcal{X}^{res,t'}_{ser}\}$ *involves solving Problem 3 and 1 whose computational complexity is polynomial in* $4n^2\hat{L}T_c + n\hat{L}T_c$.

One classic method to solve such a bi-level programming problem is to convert the lower-level problem to its Karush–Kuhn–Tucker (KKT) conditions[15], so that the problem can be reduced to a single-level one. However, since one Lagrange multiplier is introduced for every constraint, the scale of the converted problem is significantly larger. Additionally, due to the complementary slackness condition, the resulting problem is a quadratic programming (QP) problem. This QP problem can be further transformed into a mixed integer linear programming (MILP) problem[15], where the computational complexity grows exponentially with the problem size. Under our evaluation setting, Problem 4 consists of 1,482,570 *variables before conversion. Given this scale, it is impractical to solve it directly while meeting the runtime requirement.*



(●, ▨, ▨): A feasible bi-level solution for upper-level optimization problem

Figure 5: Bi-level structure of the flexibility provisioning problem.

It is notable that when $\tau = 1$, solving Problem 4 maximizes the flexibility of the fleet in the next time slot. This allows the e-taxi fleet to respond to DR requests arriving at any time and start delivering flexibility immediately. Such a capability is especially valuable for the DR programs designed for emergencies, e.g., DSASP[11] by NYISO, where the start time is not scheduled in advance and a quick response is required.

5 ENERGY FLEXIBILITY PROVISIONING ALGORITHM

Due to the scale of both the upper-level problem and lower-level problems, existing algorithms cannot solve Problem 4 while meeting runtime requirements. To address this challenge, we propose an iterative algorithm that effectively solves Problem 4.

We first convert Problem 4 from a bi-level problem to a single level problem by introducing a prediction of the reference policy. By leveraging historical data from the e-taxi fleet operations, we can predict the performance metrics $\hat{J}^{t,0}_{\text{trans,ref}}$ and $\hat{J}^{t,0}_{\text{idle,ref}}$ of the reference policy for future time slots. Here, the superscript "0" indicates that it is the initial estimation of $J^t_{\text{trans,ref}}$ and $J^t_{\text{idle,ref}}$. By replacing real-time predictions with historical predictions $\hat{J}^{t,0}_{\text{trans,ref}}$ and $\hat{J}^{t,0}_{\text{idle,ref}}$, we can eliminate the constraint (10c) from Problem 4.

Algorithm 1 Flexibility provisioning algorithm

Input: Fleet state $\mathcal{S}^{t'}$; reference policy metrics $\hat{J}^{t,0}_{\text{trans,ref}}$ and $\hat{J}^{t,0}_{\text{idle,ref}}$ learned from historical data; number of charging ports e^t_j; parameters $\hat{L}$, $\hat{L}_1$, $\hat{L}_2$, T_c, T_b, τ, t'.

Output: Provision stage dispatch decisions $\mathcal{X}^{*,t'}_{\text{pro}}$.

1: Initialize $i = 1$.
2: Solve Problem 5 for initial solution $\mathcal{X}^{*,t',0}_{pro}$ using $\hat{J}^{t,0}_{\text{trans,ref}}$ and $\hat{J}^{t,0}_{\text{idle,ref}}$.
3: **while** $\mathcal{X}^{*,t',i}_{\text{pro}}$ not converged and maximum iteration not reached **do**
4: Solve Problem 3 with fixed $\mathcal{X}^{*,t',i}_{\text{pro}}$ to update $\hat{J}^{t,i}_{\text{trans,ref}}$ and $\hat{J}^{t,i}_{\text{idle,ref}}$.
5: Solve Problem 5 with $\hat{J}^{t,i}_{\text{trans,ref}}$ and $\hat{J}^{t,i}_{\text{idle,ref}}$ to update $\mathcal{X}^{*,t',i}_{\text{pro}}$.
6: Update $i = i + 1$.
7: **end while**

Consequently, we can also remove $\mathcal{X}^{\text{ref},t'}_{\text{ser}}$ from Problem 1, so that the constraints $g_{\text{res},k}(\mathcal{S}^{t'}, \mathcal{X}^{\pi,t'}_{\text{pro}}, \mathcal{X}^{\text{ref},t'}_{\text{ser}}, \mathcal{X}^{\text{res},t'}_{\text{ser}})$ can be simplified to $g_{\text{res},k}(\mathcal{S}^{t'}, \mathcal{X}^{\pi,t'}_{\text{pro}}, \mathcal{X}^{\text{res},t'}_{\text{ser}}, \hat{J}^{t,0}_{\text{trans,ref}}, \hat{J}^{t,0}_{\text{idle,ref}})$, where , $\hat{J}^{t,0}_{\text{trans,ref}}$ and $\hat{J}^{t,0}_{\text{idle,ref}}$ are treated as constants. Moreover, the objective of Problem 4 in equation (6) is equivalent to

$$F_{\text{pro}}(\mathcal{S}^{t'}, \mathcal{X}^{\pi,t'}_{\text{pro}}, \mathcal{X}^{\text{res},t'}_{\text{ser}}) = -\sum_{t=\hat{t}}^{\hat{t}+T_c} \sum_{j=1}^{n} u^{l,t}_{j,\text{res}}, \tag{11}$$

which is identical to (5a), the objective of Problem 1. Thus, in Problem 4, the upper-level problem and the lower-level problem share the same decision variables and objective. The problem can be further reduced to a single-level problem:

Problem 5. *Single-level formulation for the provisioning flexibility problem with a prediction of the reference policy:*

$$\min_{\mathcal{X}} \ f_{res}(\mathcal{S}^{t'}, \mathcal{X}^{\pi,t'}_{pro}, \mathcal{X}^{res,t'}_{ser}) \tag{12a}$$

$$s.t. \quad \{g_{res,k}(\mathcal{S}^{t'}, \mathcal{X}^{\pi,t'}_{pro}, \mathcal{X}^{res,t'}_{ser}, \hat{J}^{t,0}_{trans,ref}, \hat{J}^{t,0}_{idle,ref})\}, \tag{12b}$$

$$Eq.\ (1) \sim (4),\ (5e) \sim (5g), (8d).$$

Let the solution policy of Problem 5 be denoted as $\pi^{*,0}$ and the coordination decision variables be denoted as $\mathcal{X}^{*,t',0}_{\text{pro}}$ and $\mathcal{X}^{\text{res},t',0}_{\text{ser}}$. Again, the superscript "0" indicates that it is the initial solution. Similarly, the computational complexity of Problem 5 is polynomial in the size of decision variables $V_n = 4n^2\hat{L}T_c + n\hat{L}T_c$.

However, a reference policy predicted from historical data may deviate due to uncertainty in passenger demand and e-taxi fleet dynamics in real time. To mitigate the error caused by reference policy deviation, we iteratively estimate the reference policy with real-time e-taxi fleet state and passenger demand, and subsequently update the provisioning policy. Specifically, in the i-th iteration, we first predict the reference policy $\pi^{\text{ref},i}$ by solving Problem 3, with control variables in the provisioning stage fixed as $\mathcal{X}^{*,t',i-1}_{\text{pro}}$. We denote the performance metrics of $\pi^{\text{ref},i}$ as $\hat{J}^{t,i}_{\text{trans,ref}}$ and $\hat{J}^{t,i}_{\text{idle,ref}}$. Next, we solve Problem 5 using $\hat{J}^{t,i}_{\text{trans,ref}}$ and $\hat{J}^{t,i}_{\text{idle,ref}}$ to update $\mathcal{X}^{*,t'}_{\text{pro}}$ and $\mathcal{X}^{\text{res},t'}_{\text{ser}}$. The solution in the i-th iteration is denoted as $\mathcal{X}^{*,t',i}_{\text{pro}}$. The complete algorithm is shown in Algorithm 1.

Algorithm 2 E-taxi coordination algorithm

Input: Time of starting provisioning: $\hat{t}$; DR activation signal arriving at $\hat{t} + \tau$; duration of time slots: t_1 minutes; time horizon T time slots; number of charging ports e^t_j; parameters $\hat{L}$, $\hat{L}_1$, $\hat{L}_2$, T_c, T_b, τ.

Output: Dispatch decisions $x^{l,t}_{i,j}, xd^{l,t}_{i,j}, y^{l,t}_{i,i'}, yd^{l,t}_{i,i'}, u^{l,t}_i$ where $i,\ i' \in [1, n]$, $j \in [1, \bar{n}]$, $l \in [1, \hat{L}]$, $t \in [\hat{t}, \hat{t} + T]$

1: **while** At the beginning of time slot **do**
2: Update the current time slot as t'; collect e-taxi status and update $V^{l,t'}_i$, $D^{l,t'}_i$, $O^{l,t'}_i$; Update driving distance constraint parameters $dc^{t'}_{i,j}$, $ds^t_{i,i'}$; Update passenger demand prediction based on historical data.
3: **if** $t' \in [\hat{t}, \hat{t} + \tau)$ **then**
4: Derive dispatch decisions provisioning flexibility by applying Algorithm 1 assuming $\tau = 1$. ▷ Provision stage
5: **else if** $t' \in [\hat{t} + \tau, \hat{t} + \tau + T_b)$ **then**
6: Solve the Problem (1) for dispatch decisions delivering the flexibility. ▷ Service stage
7: **else**
8: Solve the Problem (3) for dispatch decisions optimizing taxi service. ▷ Recovery stage
9: **end if**
10: Send the coordination decisions of current time slot: $x^{l,t'}_{i,j}$, $xd^{l,t'}_{i,j}$, $y^{l,t'}_{i,i'}$, $yd^{l,t'}_{i,i'}$, $u^{l,t}_i$
11: **end while**

6 E-TAXI COORDINATION ALGORITHM

Due to the uncertainty of passenger demand and deviations between coordination decisions and actual e-taxi behaviors, we have developed a real-time scheduling algorithm based on Model Predictive Control (MPC), as presented in Alg. 2. At the beginning of each time slot, the algorithm updates the e-taxi system state including both spatial and energy distributions of the e-taxis. Passenger demand and mobility patterns are also predicted based on historical taxi trips. Before receiving the activation signal of the DR program, the fleet stays in the provision stage. Alg. 1 is applied to provision flexibility. Since the service stage is not scheduled in advance, we always assume that the activation signal will arrive at the next time slot ($\tau = 1$), ensuring that the fleet is prepared for a rapid response. When the activation signal arrives, the fleet transitions to the service stage. Throughout this stage, the algorithm solves the Problem 1 to deliver flexibility. In the recovery stage, the algorithm solves the Problem 3 to optimize the quality of transportation service. This stage often witnesses a rebound effect as the system is restored.

7 EVALUATION

7.1 Methodology

The dataset used to evaluate the flexibility model is sourced from Manhattan and includes (i) data related to the power distribution network [16], (ii) information on charging stations for EVs [17], and

(iii) records of historical taxi trips [18]. Our data-driven analysis aims to provide a comprehensive understanding of the model's performance in a real-world urban environment. The city is segmented into 38 regions based on the power grid structure in Manhattan [16]. There are approximately 13,000 taxis operating throughout the city. We assume that all the taxis in the dataset are e-taxis. The battery energy of an e-taxi is divided into 15 levels, i.e. $\hat{L} = 15$. We set $\hat{L}_2 = 3$ for charging energy and $\hat{L}_1 = 1$ for energy consumption.

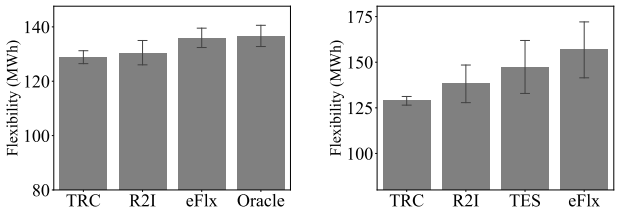
In our evaluation, the length of a time slot is 20 minutes. Multiple experiments are conducted using historical passenger data from various dates. The service stage of the DR program starts at 17:20 and lasts 2 hours, which is consistent with real-world DR programs [11, 19]. Based on historical data, the duration closely aligns with *both the peak power demand and passenger demand* in urban areas. This scenario presents the most challenging situation for the e-taxi fleet to provision its flexibility, highlighting the robustness and adaptability of the proposed system.

During the provision stage, we operate the e-taxi fleet using various strategies to compare their performance:

- Taxi service with regular charging (TRC) [13]: it applies reference policy introduced in Sec. 4.4 to optimize passenger service by maximizing the number of passengers served, ignoring DR.
- Flexibility provisioning with DR window *known in advance* (Oracle): it aims to minimize charging energy consumption during the service stage. The start time of the DR service stage is known at the beginning of the provision stage. With this additional information, it acts as an oracle baseline.
- Taxi service with energy storage (TES): it maximizes overall energy storage for the next time slot while meeting the quality requirement of transportation service.
- Reactive charging to the e-taxis's idling time (R2I): it provisions flexibility by dispatching idle e-taxis to charging stations instead of having them roam the streets in search of passengers, as in the TRC strategy. If all of the charging stations are occupied, the idle e-taxis remain stationary to conserve energy.

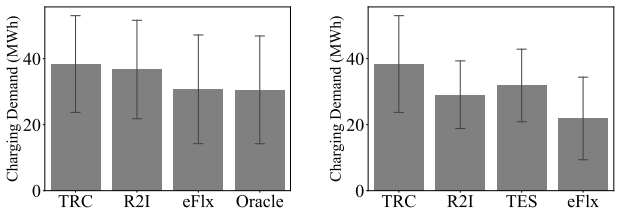
Except for Oracle, all other solutions do not have the information about the DR window during the provisioning stage and always prepare for a possible DR service stage starting in the next time slot. During the service stage, all solutions adopt the response policy (Problem 1) to reduce charging demand. In the recovery stage, they shift to the reference policy (Problem 3) to maximize passenger service. For each trial, we also operate the fleet with the reference policy throughout all stages for comparison.

The length of the control horizon T_c is configured as 200 minutes, covering both the provision and service stages to ensure accuracy while maintaining computational efficiency. Throughout the simulation, all the solutions maintain a 100% transportation service quality compared to TRC at each time slot. Moreover, during the provisioning stage, no solution except TES is permitted to charge more than the reference policy for fairness concerns. We show that *by properly controlling the spatiotemporal and energy distribution, the fleet can increase its flexibility even without additional charging*. To further explore the potential of flexibility provisioning, we also conducted a parallel experiment where the charging limit is removed during the provisioning stage, i.e., constraint (8d) is excluded from Problem 5. This enables us to observe the flexibility gain achieved by allowing more charging before the DR window.



(a) w/ charging limit **(b) w/o charging limit**

Figure 6: Comparison for flexibility



(a) w/ charging limit **(b) w/o charging limit**

Figure 7: Charging demand during service stage

Several measurement metrics are used in this work. (i) Energy flexibility: the flexibility of each solution, except for the Oracle, is estimated by Eq. (6), which calculates the charging gap between one response policy and its non-responding baseline. For the Oracle, which has knowledge of the DR window during provisioning, flexibility is evaluated by directly calculating the charging demand reduction compared to the reference policy, using the same initial fleet state as the Oracle at the start of provisioning. (ii) Charging demand: the total energy consumed during the provision and service stages. We also assess the overhead of flexibility provisioning based on idle driving distance and idle waiting time. (iii) Idle driving distance: the average driving distance per e-taxi due to dispatching. The distance covered by vacant e-taxis searching for passengers is not considered idle driving, as it does not represent overhead from coordination decisions. (iv) Idle waiting time: the average duration an e-taxi spends waiting at charging stations. For R2I, the duration for which the e-taxis remain stationary is also taken into account.

The experiment is carried out on a Windows 10 PC with an Intel 13700K CPU. The average run time to solve Algorithm 1 for each time slot is 254.7 seconds.

7.2 Results

7.2.1 Flexibility and Energy Demand. Fig. 6 illustrates the flexibility provided by each solution. When total charging is limited to remain below the TRC during the provision stage, as shown in Figure 6a, eFlx demonstrates a notable advantage. On average, eFlx provides an additional 7.11 MWh flexibility over TRC (5.52% improvement) and 5.45 MWh over R2I (4.18% improvement). In the best-case scenario, eFlx delivers an extra 8.58% flexibility compared to TRC and 10.25% compared to R2I. This is the additional

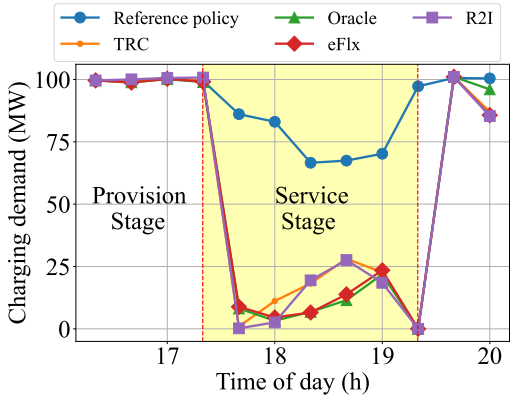
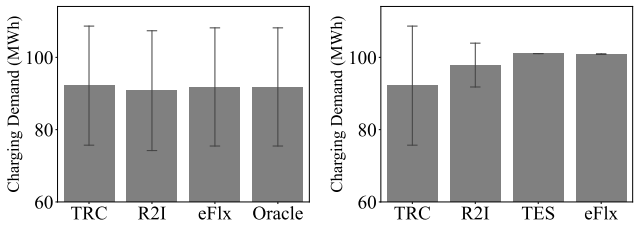


Figure 8: Charging demand during a DR program

flexibility achieved by eFlx from the optimized temporal, spatial, and energy distribution of the e-taxis. Furthermore, with the additional information of the DR service window, Oracle can serve as a performance bound for the flexibility provisioning problem. Notably, eFlx offers only 0.52% less flexibility than Oracle on average, indicating its high efficiency and its extreme closeness to the *optimal* flexibility provisioning. When the charging limit is lifted, eFlx shows yet another substantial improvement, as depicted in Fig. 6b. In this scenario, TES and R2I serve as baseline solutions without charging limits for comparison. On average, eFlx provides an additional 21.67% flexibility over TRC, 13.48% over R2I, and 6.35% over TES. In the best-case scenario, eFlx outperforms TRC by 43.49%, R2I by 21.53% and TES by 9.67%, respectively. R2I, TES and eFlx can provision extra flexibility when the charging limit is removed since more energy is stored during the provisioning stage. Among these, eFlx continues to outperform other solutions significantly.

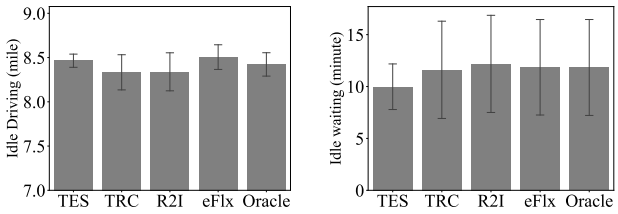
Figure 7 illustrates the charging demand during the service stage for each solution. When total charging is limited during the provision stage, as shown in Fig. 7a, eFlx decreases charging demand by 19.98% relative to TRC and by 16.36% compared to R2I on average. In the best-case scenario, eFlx achieves a remarkable reduction of 59.77% compared to TRC and 63.46% compared to R2I. Again, the reduction in charging demand shows that eFlx optimizes the fleet state more effectively for delivering flexibility. Compared to Oracle, eFlx charges only 0.51% more energy on average, again demonstrating its efficiency. When the charging limit is removed during the provision stage, as depicted in Fig. 7b, eFlx continues to show significant improvements. With the extra charging during provisioning, eFlx can also shift the demand temporally, further enhancing its flexibility during DR service. In this scenario, eFlx reduces charging demand by 42.97% compared to TRC, 24.73% compared to R2I, and 31.37% compared to TES on average. In the best-case scenario, eFlx achieves reductions in charging demand of 89.01% compared to TRC and 68.85% compared to R2I.

In Fig. 8, we also show the charging demand of a single trial with charging limit to illustrate how eFlx behaves. We can see that eFlx significantly reduces charging demand during the service stage compared to TRC and R2I. Also, throughout the entire service stage, eFlx's charging demand is very close to that of Oracle.



(a) w/ charging limit **(b) w/o charging limit**

Figure 9: Charging demand during provision stage



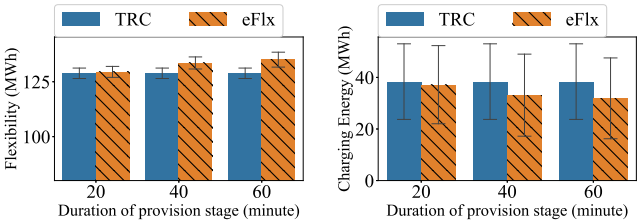
(a) Idle driving distance **(b) Idle waiting time**

Figure 10: Overhead during provision stage.

7.2.2 Overhead of provisioning flexibility. We also analyzed the overhead of provisioning flexibility. Fig. 9 presents the charging demand during the provision stage. When total charging is limited, eFlx has no overhead in charging. On the other hand, when charging is not limited, eFlx charges 9.45% more than TRC on average. However, it also provides increased flexibility in return, as discussed in Section 7.2.1. Fig. 10a and Fig. 10b display the average idle driving distance and idle waiting time during the provision stage. Specifically, idle driving distance increases on average by 2.06%, while idle waiting time rises by 2.12%. The increased idle waiting time suggests that more vacant e-taxis remain stationary rather than roaming on the streets. The slight increase in idle driving distance indicates that eFlx leads to an increase in dispatch decisions. The minimal overhead shows the effectiveness of eFlx.

7.2.3 Duration of provision stage. Fig. 11 illustrates the performance of eFlx during the service stage for various durations of the provisioning stage. Both the flexibility and charging demand during the service stage are shown, with TRC included as a reference. In all trials, the start and end times of the service stage are fixed, while the start time of the provision stage varies based on τ. Provisioning flexibility for only 20 minutes yields an additional 0.64 MWh of flexibility on average. As the duration of the provision stage increases, the capacity to optimize flexibility also improves, highlighting the substantial impact of extended provisioning on enhancing flexibility.

7.2.4 Provisioning under passenger demand prediction error. To assess the robustness of eFlx, we evaluate its performance under passenger demand prediction errors. The state-of-the-art taxi demand prediction methods achieve a mean absolute error (MAE) of less than 10, in terms of the number of passengers per region in Manhattan for each half-hour interval [20, 21]. Considering the difference in region size and time slot duration in our setting, it



(a) Flexibility **(b) Charging Demand**

Figure 11: Performance vs. duration of provision stage.

	Flexibility (MWh)	Charging demand (MWh)
w/o error	135.96	30.70
w/ error	134.16	33.06

Table 1: eFlx flexibility and charging demand during service stage under passenger demand prediction error.

is equivalent to a 10% error of the ground-truth demand. Therefore, we introduce random noise to mimic the prediction error. As presented in Table 1, eFlx's flexibility reduces by 1.33% on average, and the charging demand during the service stage increases by 7.69%. Despite these deviations, eFlx still outperforms TRC and other baselines, even when they have perfect knowledge of future demand. This demonstrates the effectiveness and robustness of the eFlx design in handling uncertainties in the passenger demand predictions.

8 RELATED WORK

Electric Vehicle Coordination: The emergence of electric vehicles has sparked a significant body of research focused on optimizing the coordination of electric vehicle charging activities to enhance the quality of transportation service, as evidenced by a range of studies [13, 14, 22, 23]. [14] addresses electric vehicle rebalancing, focusing on uncertainties in taxi supply and passenger demand to enhance service efficiency. [24] proposes a bi-level spatiotemporal optimization framework to improve the long-term profits of the e-taxi fleets. [22] presents a multi-agent reinforcement learning framework for EV charging stations, using a user incentive scheme to indirectly rebalance EVs and enhance operational efficiency. Furthermore, researchers have been actively addressing the adverse impacts of EV charging behaviors on the stability of power systems [25–27]. Alizadeh et al. [26] propose a collaborative scheme for power and transportation systems to achieve a socially optimal outcome without sharing sensitive private information. [28] proposes a charging system for a ride-hailing fleet which jointly plans charging stations and battery swapping. [27] presents an innovative e-taxi fleet coordination algorithm that enhances the stability of power systems while concurrently maintaining transportation services. In summary, this research distinguishes itself from existing studies by quantifying e-taxi fleet energy flexibility and presenting a novel approach to maximize this flexibility for proactive responses to DR.

Energy Flexibility: The flexibility of power systems refers to the ability to efficiently and reliably handle fluctuations and uncertainties in electricity demand and supply across different time frames [29]. This concept underscores the potential for both the power supply [30–32] and the demand sides to contribute to the adaptability of power systems. Demand-side management enhances energy flexibility by adjusting demand levels through reductions or rescheduling, thus mitigating imbalances between power supply and demand [33–35]. This study distinguishes itself from prior research by: (i) exploring the untapped potential of e-taxi fleets in spatial-temporal flexible power demand management, a novel dimension that has not been sufficiently explored in the existing literature. (ii) investigating proactive strategies to achieve an optimal flexible state, strategically preparing for upcoming demand response requests, rather than responding reactively.

9 CONCLUSION

Due to inherent battery energy reserves and mobility, e-taxi fleets possess great potential to participate in DR programs to support the reliability of power grids. To understand the capacity of an e-taxi fleet for reducing charging demand, we propose a flexibility model and formulate the flexibility provisioning problem for an e-taxi fleet. An efficient algorithm is developed to solve the provisioning problem which optimizes the state of an e-taxi fleet to achieve and maintain the highest energy flexibility at all times, ready to respond to DR requests arriving at any time. A coordination strategy is then presented to guide e-taxi fleets participating in such emergency DR programs. Evaluations using real-world datasets illustrate that, compared to existing solutions, our flexibility provisioning solution achieves an additional 19.98% reduction in energy demand during DR service stage without compromising the quality of transportation service. Further experiments of eFlx operating under additional real-world uncertainties are left for future work.

ACKNOWLEDGMENTS

This publication was supported in part by the FY2025 Center of Excellence for Applied Computational Science competition at the University of Tennessee at Chattanooga, and in part by the National Science Foundation under Grant ECCS-2025152 and Grant CNS-2431552.

REFERENCES

[1] G. Wang, X. Chen, F. Zhang, Y. Wang, and D. Zhang, "Experience: Understanding long-term evolving patterns of shared electric vehicle networks," in *ACM Mobicom'19*, 2019.
[2] (2018) Tesla vehicles have completed 70amsterdam airport this year. [Online]. Available: https://electrek.co/2018/12/20/tesla-taxi-rides-amsterdam-airport/
[3] London gets its first official electric black cab in 120 years. [Online]. Available: https://www.greencarreports.com/news/1125710_london-gets-its-first-official-electric-black-cab-in-120-years
[4] A. M. Manufacturing, "Electric vehicle taxi market report 2024: Global trends, forecast and competitive analysis 2018-2030," 2024. [Online]. Available: https://finance.yahoo.com/news/electric-vehicle-taxi-market-report-092800235.html
[5] M. Moniot, C. L. Rames, and E. Burrell, "Feasibility analysis of taxi fleet electrification using 4.9 million miles of real-world driving data," National Renewable Energy Lab. (NREL), Golden, CO (United States). Warrendale, PA: SAE International, 04 2019. [Online]. Available: https://www.osti.gov/biblio/1526198
[6] J. Aghaei, M.-I. Alizadeh, P. Siano, and A. Heidari, "Contribution of emergency demand response programs in power system reliability," *Energy*, vol. 103, pp. 688–696, 2016.
[7] S. R. Kumar, R. Rigo-Mariani, B. Delinchant, and A. Easwaran, "Towards safe model-free building energy management using masked reinforcement learning," in *2023 IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE)*, 2023, pp. 1–5.

[8] Retail electricity consumer opportunities for demand response in pjm's wholesale markets. [Online]. Available: https://nj.gov/bpu/pdf/publicnotice/stakeholder/20180205/PJM%20Flyer%20on%20DR%20Distributed%20to%20BPU%20Stakeholders%202%205%2018%20%20end-use-customer-fact-sheet.pdf

[9] H. Haes Alhelou, M. E. Hamedani-Golshan, T. C. Njenda, and P. Siano, "A survey on power system blackout and cascading events: Research motivations and challenges," *Energies*, vol. 12, 2019.

[10] H. Aalami, G. R. Yousefi, and M. Parsa Moghadam, "Demand response model considering edrp and tou programs," in *2008 IEEE/PES Transmission and Distribution Conference and Exposition*, 2008.

[11] NYISO. (2024) Demand response providing ancillary services with direct-to-nyiso connectivity. [Online]. Available: https://www.nyserda.ny.gov/-/media/Project/Nyserda/Files/Publications/Research/Electic-Power-Delivery/Demand-Response-Providing-Ancillary-Services.pdf

[12] M. Z. Degefa, I. B. Sperstad, and H. Sæle, "Comprehensive classifications and characterizations of power system flexibility resources," *Electric Power Systems Research*, vol. 194, p. 107022, 2021.

[13] Y. Yuan, D. Zhang, F. Miao, J. Chen, T. He, and S. Lin, "p^{2}charging: Proactive partial charging for electric taxi systems," in *IEEE ICDCS'19*, 2019.

[14] S. He, Z. Zhang, S. Han, L. Pepin, G. Wang, D. Zhang, J. A. Stankovic, and F. Miao, "Data-driven distributionally robust electric vehicle balancing for autonomous mobility-on-demand systems under demand and supply uncertainties," *IEEE TITS*, 2023.

[15] J. F. Bard and J. E. Falk, "An explicit solution to the multi-level programming problem," *Computers & Operations Research*, vol. 9, no. 1, pp. 77–100, 1982. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0305054882900077

[16] C. E. Company. (2020) Con edison - power new york city and westchester. [Online]. Available: https://www.coned.com/en

[17] N. Y. State. (2024) Electric vehicle station locator. [Online]. Available: https://www.nyserda.ny.gov/All-Programs/Drive-Clean-Rebate-For-Electric-Cars-Program/Charging-Options/Electric-Vehicle-Station-Locator#/find/nearest

[18] N. Taxi and L. Commission. (2024) Tlc trip record data. [Online]. Available: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[19] NYISO. (2024) Emergency demand response program manual. [Online]. Available: https://www.nyiso.com/documents/20142/2923301/edrp-mnl.pdf/8f34b039-de10-1726-bce4-0e02d8732a03

[20] C. Zhang, F. Zhu, X. Wang, L. Sun, H. Tang, and Y. Lv, "Taxi demand prediction using parallel multi-task learning model," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 794–803, 2022.

[21] J. Zhao, C. Chen, H. Huang, and C. Xiang, "Unifying uber and taxi data via deep models for taxi passenger demand prediction," *Personal Ubiquitous Comput.*, vol. 27, no. 3, p. 523–535, Jul. 2020. [Online]. Available: https://doi-org.proxy.library.stonybrook.edu/10.1007/s00779-020-01426-y

[22] M. Luo, B. Du, W. Zhang, T. Song, K. Li, H. Zhu, M. Birkin, and H. Wen, "Fleet rebalancing for expanding shared e-mobility systems: A multi-agent deep reinforcement learning approach," *IEEE TITS*, 2023.

[23] H. Tan, Y. Yuan, S. Zhong, and Y. Yang, "Joint rebalancing and charging for shared electric micromobility vehicles with energy-informed demand," in *ACM CIKM*, 2023, pp. 2392–2401.

[24] Lyu, Yelin, Wang, Ning, and Tian, Hangqi, "Coordinated charging and dispatching for large-scale electric taxi fleets based on bi-level spatiotemporal optimization," in *WCX SAE World Congress Experience*. SAE International, apr 2024. [Online]. Available: https://doi.org/10.4271/2024-01-2880

[25] S. Sharma and P. Jain, "Integrated tou price-based demand response and dynamic grid-to-vehicle charge scheduling of electric vehicle aggregator to support grid stability," *International Transactions on Electrical Energy Systems*, 2020.

[26] M. Alizadeh, H.-T. Wai, M. Chowdhury, A. Goldsmith, A. Scaglione, and T. Javidi, "Optimal pricing to manage electric vehicles in coupled power and transportation networks," *IEEE TCNS*, 2016.

[27] Y. Yuan, Y. Zhao, and S. Lin, "Poet: Towards power-system-aware e-taxi coordination under dynamic passenger mobility," in *ACM e-Energy'22*, 2022.

[28] Z. Lai and S. Li, "Towards a multimodal charging network: Joint planning of charging stations and battery swapping stations for electrified ride-hailing fleets," *Transportation Research Part B: Methodological*, vol. 183, p. 102928, 2024.

[29] O. M. Babatunde, J. L. Munda, and Y. Hamam, "Power system flexibility: A review," *Energy Reports*, vol. 6, pp. 101–106, 2020.

[30] J. Kiviluoma, E. Rinne, and N. Helistö, "Comparison of flexibility options to improve the value of variable power generation," *International Journal of Sustainable Energy*, 2018.

[31] E. Lannoye, D. Flynn, and M. O'Malley, "Evaluation of power system flexibility," *IEEE Transactions on Power Systems*, 2012.

[32] Q. Wang and B.-M. Hodge, "Enhancing power system operational flexibility with flexible ramping products: A review," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1652–1664, 2016.

[33] E. Loukarakis, C. J. Dent, and J. W. Bialek, "Decentralized multi-period economic dispatch for real-time flexible demand management," *IEEE Transactions on Power Systems*, 2016.

[34] S. M. Hakimi, A. Hajizadeh, M. Shafie-khah, and J. P. Catalão, "Demand response and flexible management to improve microgrids energy efficiency with a high share of renewable resources," *Sustainable Energy Technologies and Assessments*, 2020.

[35] J. Lizana, D. Friedrich, R. Renaldi, and R. Chacartegui, "Energy flexible building through smart demand-side management and latent heat storage," *Applied energy*, vol. 230, pp. 471–485, 2018.